NOTHINGNESS: THE INDEFINABLE WHOLE

From the Evolution of Meaning to Human Consciousness and Artificial Intelligence: Knowledge, Language, and the Limits of Understanding



Knowledge does not merely transform what we know about reality; it can also transform the meaning of the words we use to try to understand it. Intellectual responsibility consists in distinguishing between what exists, what we know, and what we mean when we speak of it.

Abstract

This essay explores the relationship between knowledge, language, human consciousness, and artificial intelligence through a fundamental question: how do the meanings we assign to reality change as our knowledge evolves? Its starting point is the concept of nothingness, understood not simply as a metaphysical problem or as the absolute absence of existence, but also as a word whose meaning depends on the conceptual framework available to those who use it.

Throughout history, human beings have interpreted phenomena through the explanatory categories available to them. Concepts once associated with divine intervention, invisible forces, or seemingly inexplicable causes have sometimes been transformed or replaced as scientific investigation has identified different mechanisms and entities. Such developments do not establish that every religious or philosophical concept is false; they demonstrate that the meanings attributed to words, and the explanations associated with them, can change profoundly over time.

Drawing on philosophy, physics, mathematical logic, cognitive science, and contemporary artificial intelligence research, the essay examines the distinction between reality, knowledge of reality, and the language through which knowledge is expressed. It considers how measurement can provide reliable information without resolving every question about the nature of what is measured; how formal systems can encounter limitations; how thought and language interact; and how subjective experience differs from its observable manifestations.

The discussion then turns to artificial intelligence, asking whether the ability to reproduce intelligent behavior necessarily implies understanding comparable to that of human beings. It also examines the practical consequences of developing systems whose capabilities may exceed our ability to predict every relevant aspect of their behavior.

The central argument is that the evolution of knowledge can transform not only our explanations of the world but also the meanings of the concepts through which we interpret it. Recognizing this process requires neither the assumption that reality is merely a linguistic construction nor the conclusion that every interpretation is equally valid. It requires distinguishing what exists, what we know about it, and what we mean when we speak of it.

The essay concludes that intellectual progress depends on the willingness to revise concepts, test explanations against evidence, acknowledge uncertainty, and govern technological power responsibly. The challenge is not simply to discover what we do not yet understand, but also to recognize when the conceptual framework through which we understand something must itself be reconsidered.

Introduction — An Unresolved Issue

There are questions that arise before we possess the tools necessary to address them and that, precisely for this reason, can accompany us throughout our lives without losing their original force. Sometimes, as the years go by, it is not we who find a definitive answer to those questions; rather, it is the world that changes, placing us in a position to finally understand their significance. A reflection that, when I was about sixteen, had for me primarily the value of a philosophical intuition now appears to be connected to problems involving physics, logic, knowledge, consciousness, and, with increasing urgency, the development of artificial intelligence. The intuition was expressed in a simple sentence, perhaps even too simple for the problem it sought to contain: “Nothingness is nothing other than the indefinable part of the whole.”

At the time, I had no intention of formulating a theory of nothingness, nor did I possess the philosophical and scientific tools necessary to support such a claim. Rather, it was an attempt to give linguistic form to a difficulty I sensed intuitively: how can we speak of the whole if we are unable to know everything it encompasses? And how can we define nothingness without transforming it, at the very moment we think about it, into something that belongs to our thought and therefore, in some sense, to our world? The issue was not merely a matter of finding a satisfactory definition, but of asking myself whether our need

to define things eventually encounters an intrinsic limit in the means through which we seek to know them.

Over time, however, I have come to see another possibility: perhaps the difficulty does not lie exclusively in the object we are trying to define, but also in the meaning we assign to the words through which we approach it. A word may remain unchanged while the concept it expresses undergoes a profound transformation. What people mean by "nothingness" today need not coincide with what they meant by it centuries ago, just as the meanings of words such as "God," "homeland," or "nature" have changed across historical periods, cultures, and intellectual traditions.

This observation does not imply that words create reality at will, nor that every interpretation is equally valid. It means that the relationship between a word and the world cannot always be taken for granted. Sometimes, the progress of knowledge allows us to identify the phenomenon to which a word refers more precisely. At other times, it reveals that we had grouped different phenomena under the same name, or that a question appeared insoluble because the concepts through which we formulated it were inadequate.

The history of knowledge can therefore be understood not only as the accumulation of new information, but also as the continuous revision of the meanings through which human beings organize their experience. The question of nothingness may belong to this history: before asking whether nothingness is ultimately definable, we should ask what we mean by the word, how that meaning developed, and whether the difficulty lies in reality itself or in the conceptual framework through which we attempt to describe it.

After many decades, I do not regard that sentence as an established answer, nor do I intend to retroactively attribute to it a scientific meaning it could not have had at the time. Rather, I regard it as the starting point of a question that today takes on a different dimension. If our knowledge advances through tools that allow us to describe, measure, and use reality without necessarily clarifying what the things we describe ultimately are, what is the relationship between the power of knowledge and the depth of understanding? And what happens when we begin to build systems capable of reproducing certain functions of human intelligence before we have fully understood what thought and consciousness are?

It is this last question that makes the problem relevant today. Not because we must assume that machines will inevitably become conscious, nor because we must imagine a future in which human beings will necessarily lose control of their creations, but because there is a concrete possibility that already deserves attention: our ability to produce operational power may grow faster than our ability to understand it and govern its consequences. If this were the case, we would find ourselves facing a new condition in which knowledge would continue to increase our power to intervene in the world without providing us, to the same extent, with an understanding of what we are doing.

1. Knowing Does Not Necessarily Mean Understanding

The history of human knowledge can be understood as a progressive expansion of our ability to intervene in reality. We have learned to observe phenomena, recognize regularities, formulate laws, and build tools capable of producing results that, in earlier eras, would have seemed impossible. Modern science has radically transformed our existence precisely because it has not limited itself to contemplating the world, but has developed procedures through which acquired knowledge can be verified, shared, and put to use. In celebrating this extraordinary ability, however, we sometimes risk confusing two different achievements: knowing how to operate on something and knowing what that thing is in its deepest nature.

The distinction is not a minor one. We can know the behavior of a phenomenon with great precision, describe it through equations, predict its effects under certain conditions, and exploit it to achieve concrete results without thereby resolving every question concerning its nature. There is nothing contradictory about such a situation: scientific knowledge does not necessarily require a definitive answer to every philosophical question, and it would be wrong to demand that physics resolve every metaphysical problem concerning reality before it can move forward. The point is different: the operational success of a body of knowledge should not automatically be taken as proof that our understanding is complete (Chang, 2004; Russell, 2019).

It is precisely in this distinction that one of the most important characteristics of human knowledge becomes apparent. We can know what happens, identify the conditions under which it happens, and establish how to use what happens without possessing a definitive answer to the question of what, in itself, is happening. These levels can support one another, but they do not coincide. A theory can be extremely effective at predicting a phenomenon while leaving deeper interpretive questions open; similarly, a technology can function reliably within a particular field without every stage of its operation being understood in detail by those who use it.

This distinction does not diminish the value of science; rather, it allows us to recognize its strength more precisely. Science is not omniscience, nor does it need to be in order to be extraordinarily effective. The problem arises when we draw a conclusion broader than the method itself warrants from the effectiveness of that method; when, that is, our ability to describe and use a phenomenon leads us to believe that we have exhausted its meaning. The history of thought instead invites us to preserve the distinction between what we can do and what we can legitimately claim to have understood.

This distinction becomes particularly significant when the object of our investigation is no longer merely the external world, but our own capacity to know. If we can develop tools that enormously expand our operational possibilities, we must still ask whether we understand to the same extent the processes from which our ability to design, judge, and assign meaning arises. And if the answer is no, we must ask whether this limitation is merely temporary, destined to be overcome through the progress of research, or whether there are aspects of knowing that cannot be fully clarified without encountering deeper difficulties.

2. Measurement: The Power to Quantify Without Exhausting Meaning

In physics, we say that we know a quantity when we are able to measure it. But what, exactly, does it mean to measure? It means establishing a quantitative relationship between a quantity and a homogeneous reference unit, determining how many times that unit is contained in the quantity under consideration, or what fraction of it is needed to represent it. Measuring a length, for example, means comparing it with a conventional unit of length; measuring a duration means comparing a time interval with a unit of time. Measurement thus translates a property into a numerical relationship, making it comparable, communicable, and usable in reasoning and calculations.

This procedure, which constitutes one of the foundations of scientific knowledge, nevertheless contains a distinction that deserves to be brought to light. The ability to measure a quantity does not necessarily coincide with the possession of an ultimate definition of the nature of what is being measured. We can determine a distance precisely without having resolved every philosophical question concerning space; we can measure a duration without having definitively clarified what time is. Contemporary physics, of course, does not merely compare quantities: it develops theories that assign profound structural meanings to space and time, and the operational definitions of quantities depend on theoretical frameworks and experimental procedures. But this very complexity confirms that quantitative results and conceptual understanding belong to connected levels that do not necessarily overlap (Chang, 2004).

When I claim that we can measure extension without having resolved its epistemological problem, I do not mean that physics ignores the nature of space or has failed to develop theories capable of explaining many of its properties. I mean to highlight a more general distinction: a procedure can allow us to know a property quantitatively without resolving every question concerning the fundamental meaning of that property. Measurement allows us to move from practical indeterminacy to numerical determination; by itself, it does not guarantee that every ontological and epistemological question has been resolved.

In this sense, measurement constitutes a kind of effective way around ignorance, but only if we give the term its most precise meaning. It is not a deception, a flaw, or an illegitimate shortcut. It is a methodological achievement: instead of waiting for a definitive answer to every question about the nature of things, we construct procedures that allow us to investigate specific properties and establish reliable relationships among them. Our ignorance does not disappear, but it ceases to be an absolute obstacle to action. We can know enough to intervene effectively while continuing to remain ignorant of fundamental aspects of the object of our knowledge.

The distinction becomes even clearer when we consider the technologies that arise from this knowledge. An engineer can design a device based on verified mathematical models without having to resolve personally every theoretical question that made those models possible. A physician can use a validated diagnostic procedure without knowing in detail every biological mechanism involved in the phenomenon being observed. In both cases,

operational knowledge is real and valuable; what would be inappropriate is to confuse it with a complete understanding of every level of the reality involved.

The question we must therefore ask is not whether it is legitimate to use what we know without understanding everything. Clearly, human civilization could not have developed if it had been required to wait for absolute knowledge before every application. The question is rather this: how much ignorance can we afford when the power of the tools we build exceeds our ability to predict their effects? As long as the consequences are limited, observable, and correctable, partial knowledge may be sufficient. When, however, tools become capable of intervening in complex systems, essential infrastructure, or the very processes through which we make decisions, the distance between operational effectiveness and understanding takes on a new meaning.

This is where the problem of measurement, apparently confined to the methodology of science, begins to connect with the problem of artificial intelligence. We can measure a system's performance, evaluate its accuracy, compare its speed with that of humans, and observe its ability to solve specific tasks. But these results, important as they are, do not automatically answer questions about what the system represents internally, how it arrives at a particular response, what the limits of its generalizations are, or how it will behave when used under conditions different from those anticipated. Here, too, the ability to measure performance does not coincide with a complete understanding of the phenomenon.

3. The Whole and the Indefinable

Let us return, then, to the sentence with which this reflection began: "Nothingness was nothing other than the indefinable part of the whole." Its meaning cannot be established once and for all, and it would be incorrect to present it as a scientific definition of nothingness. Its interest lies instead in the tension it introduces between two apparently opposing concepts: the whole, which seems to indicate the totality of what exists or can be conceived, and the indefinable, which indicates the limit of our ability to determine something conceptually. If the whole were truly the whole, it would have to include even what we are unable to define; yet, when we attempt to assign meaning to this indefinable part, we risk transforming it into an object of thought and thus bringing it, at least in part, within the domain of our definitions.

To understand the difficulty, we must distinguish the physical phenomena to which the word "nothingness" may be applied from the meaning of the word itself. In ordinary language, we may call a space empty when it contains no visible objects or when matter has been removed to a degree considered sufficient for a particular purpose. In physics, however, a vacuum is defined through specific physical conditions and theoretical frameworks; it cannot simply be identified with absolute nothingness. Even a highly rarefied region may retain physical properties, while the quantum vacuum is not equivalent to a state in which every physical structure and field has ceased to exist (Wright, 2024).

These distinctions are important, but they do not by themselves settle the philosophical question. To identify what is physically present in a vacuum is one task; to establish what we mean when we call something “nothing” is another. The meaning of the word depends partly on the context, the criteria being applied, and the knowledge available to the person using it. What counts as an absence for a particular measurement or practical purpose may cease to count as an absence when more sensitive instruments or more comprehensive theories reveal additional structure.

The distinction between a physical absence and a limitation in our knowledge must therefore remain explicit. Something may exist independently of our ability to detect it, and a concept that is adequate within one context may prove insufficient in another. Conversely, not every gap in our knowledge corresponds to a hidden physical entity. The fact that we cannot currently identify something does not establish its existence, just as the failure to detect it does not always establish its absolute absence.

The concept of nothingness thus raises two different questions: what, if anything, is physically present in the situation we are describing, and what exactly do we mean by calling that situation “nothing”? The first question requires appropriate physical theories and evidence. The second requires conceptual and linguistic clarification. Confusing them risks turning a provisional definition into a claim about the ultimate nature of reality.

Absolute nothingness, understood as the total absence of anything whatsoever, is not the same as the physical vacuum, which in scientific theories is not necessarily a mere absence of reality. Nor is it the same as a gap in our knowledge, because what we do not know may exist independently of our ignorance of it. Finally, it is not the same as the indefinable: something may be undefined today and become definable tomorrow through new concepts, tools, or methods of investigation.

The youthful sentence becomes more precise if we read it not as the claim that a part of the world is necessarily and forever unknowable, but as the hypothesis that the very concept of totality must confront the limits through which we attempt to conceive it. From this perspective, the indefinable is not necessarily a mysterious region located somewhere inside the universe. It is first and foremost a problem concerning the relationship between what we intend to understand and the conceptual tools with which we attempt to do so.

There is a further possibility: what changes may not be the object itself, but the meaning of the concept through which we describe it. A more advanced instrument may reveal properties that were previously inaccessible, but it may also force us to reconsider what we meant by “empty,” “present,” or “absent.” The history of knowledge is full of cases in which a familiar word survived while its explanatory role changed, or in which a concept had to be divided into several more precise concepts.

In this sense, the evolution of knowledge is also an evolution of meaning. The question is not merely whether we will eventually define what we now call indefinable, but whether future knowledge will lead us to formulate the original question differently. What appears to be a boundary within reality may sometimes turn out to be a boundary within our

vocabulary or conceptual framework. Determining which possibility applies requires investigation rather than a predetermined philosophical conclusion.

The limit may belong to the object, to our tools, or to the relationship established between the two; and we cannot decide in advance, without investigation, which of these possibilities is correct.

There is also a linguistic difficulty that cannot be ignored. When we say “the whole,” we use an expression that seems to gather everything into a single conceptual set. But what entitles us to make this totalizing move? Are we speaking of everything that physically exists, everything that can be thought, everything that can be described through language, or a broader totality that would also include the conditions underlying our own ability to think? These answers are not equivalent. If the whole is the totality of physical reality, our thought can be considered part of it; if, instead, the whole is the totality of every conceivable object, we must clarify what it means to include within that totality the conceptual operations through which the totality itself is constructed.

The risk is that we may move from one definition to another without noticing, assigning the word “whole” a meaning that changes during the course of our reasoning. To avoid this risk, we must not abandon the original question, but make it more rigorous. We cannot conclude that the indefinable necessarily exists as an autonomous component of reality merely because we are unable to define it. We can, however, observe that every description of totality is formulated by a subject who uses specific concepts, and that the validity of a description must not be confused with the claim that it exhausts every possible aspect of reality.

In this reading, the sentence does not prove that there exists a metaphysical part of the whole destined to elude knowledge forever. More cautiously, it invites us to consider that our way of conceiving the whole may be incomplete precisely because it depends on representational tools that have a history, a structure, and limitations. The indefinable thus does not become a universal explanation capable of resolving every problem simply by declaring it unsolvable. Instead, it becomes a methodological reminder: before declaring that we have understood the whole, we must ask whether the categories with which we describe it are adequate to what we claim to include within it.

4. Russell: The Limits of Totality and the Difficulties of Self-Reference

The problem of totality has an important history in logic and the philosophy of mathematics. Russell’s paradox, formulated at the beginning of the twentieth century, showed that a naive conception of sets, in which any property could be freely associated with the set of all objects satisfying it, led to a contradiction. Russell considered, in intuitive terms, the set of all sets that do not belong to themselves and asked whether this set belonged to itself. If it does belong to itself, then, by the property that defines it, it should not belong to itself; if it does not belong to itself, then it satisfies the condition for belonging to itself. The contradiction is not merely a play on words: it demonstrates that

an apparently innocent rule, applied without restrictions, can make an entire system inconsistent (Russell, 1903; Irvine and Deutsch, 2026).

The lesson is not that every totality is impossible or that human thought must necessarily fail when it attempts to understand itself. The lesson is more precise: certain ways of constructing totalities cannot be adopted without limits because they produce consequences incompatible with the rules that are supposed to govern them. Axiomatic set theory addressed the problem by restricting the conditions under which sets can be formed. In standard axiomatic systems, such as Zermelo–Fraenkel set theory, the existence of a universal set indiscriminately containing all sets is not assumed; the formation of sets is subject to rules that avoid Russell's contradiction.

This result is important to our discussion, but we must avoid an overly easy analogy. Russell's paradox does not prove that consciousness is indefinable, does not establish that artificial intelligence is destined to escape control, and does not show that every complex system must contain a contradiction. Mathematical logic addresses specific questions, and it would be a mistake to turn a technical result into a general proof of metaphysical or technological claims. The connection we can propose is methodological: when we attempt to construct a comprehensive description of a domain, we must pay attention to the rules we use to delimit that domain and to the consequences of our own definitions.

The issue becomes particularly delicate when the system we want to understand also includes the subject studying it. If human beings are part of nature, then their cognitive abilities are natural phenomena as well; yet the attempt to understand those abilities is carried out precisely through those same cognitive abilities. This is not necessarily a contradiction, because one part of a system can develop models of the larger system to which it belongs. Nevertheless, a difficulty of perspective arises: our access to reality is mediated by cognitive tools that are themselves part of the reality we want to explain.

We can therefore ask whether there is a difference between understanding a system from the outside, insofar as that expression is appropriate, and understanding it through processes that are themselves part of the system. A human observer can describe certain mechanisms of their own brain, but this does not mean they have immediate and complete access to all the operations that make their observation possible. They can formulate theories about language, attention, and memory, but the very act of formulating these theories depends on language, attention, and memory. It does not follow that knowledge is impossible; it follows that self-knowledge is a form of inquiry in which the knowing subject and the object being known are not completely separate.

It is in this sense, and only in this sense, that the reference to Russell can become useful to our reflection. Not as mathematical proof of the existence of an absolute limit, but as a warning against the belief that every totality can be treated as a simple object without questioning the conditions that make its definition possible. If consciousness is part of the world we seek to understand, and if it is also one of the conditions through which we understand that world, then the question of its nature cannot be addressed while ignoring the role of the conceptual tools we use.

The same caution applies to the words through which we formulate philosophical problems. A contradiction may reveal a defect in a formal definition; an apparent paradox may arise from the ambiguity of a term; and a question may change substantially when its underlying assumptions are revised. Russell's paradox is a precise logical result, not evidence that every attempt at totality must fail. Its broader relevance lies in the reminder that the concepts and rules through which we describe a domain must themselves remain open to examination.

5. The Child and the Birth of Meaning

The problem becomes even more interesting when we ask how language arises. We adults are so accustomed to thinking through words that we risk identifying thought with its linguistic expression. When we try to clarify a concept, we translate it into a definition; when we want to remember a line of reasoning, we organize it into sentences; when we communicate an experience, we use words that we believe we share with others. Language thus appears to be not merely the instrument of thought, but the very place in which thought takes shape. Yet if this identification were complete, we would have to explain how a child can learn their own language before possessing the language necessary to understand what they are learning.

A child is not born with an already available vocabulary, nor with linguistic definitions of the concepts they will gradually acquire. Yet from the earliest months, children are able to distinguish aspects of their environment, recognize familiar voices, direct their attention, learn regularities, and respond to social cues. As development proceeds, these abilities change and become integrated with language learning. Through interactions with the people who care for them, children associate sounds with situations, actions, objects, intentions, and emotional responses; gradually, words become part of a network of meanings that does not form through abstract definitions, but through repeated experiences, joint attention, imitation, correction, and participation in social life (Tomasello and Carpenter, 2007; Vygotsky, 1962).

How, then, can a child understand the meaning of a word before being able to define it? The most plausible answer is that meaning does not initially arise as a verbal definition. The word becomes embedded in an experience that is already organized, at least in part, through perceptions, bodily sensations, expectations, relationships, and elementary forms of learning. A child may begin to recognize a reassuring situation before knowing the word "security," may react to separation before understanding the concept of "absence," and may experience distress in response to a painful event without possessing any definition of "pain." Language comes into play later, helping to organize, differentiate, and communicate experiences that language itself did not create.

This does not mean that there necessarily exists a fully formed and structured kind of thought, independent of any form of representation, to which language merely attaches labels. The relationship is more complex. Prelinguistic abilities make part of word learning possible, while language, once acquired, profoundly changes the way individuals organize their experiences, remember events, distinguish categories, and reflect on themselves.

Language and thought therefore develop through a reciprocal relationship: not all thought is linguistic, but language can radically transform the possibilities of thought.

The development of meaning does not end with childhood. Throughout life, individuals revise the concepts through which they interpret their experiences, while societies transform the meanings of words across generations. A term may initially express an explanation based on limited observation and later acquire a different meaning when new evidence, distinctions, or theories become available. In other cases, a word continues to be used even though the assumptions once associated with it have been abandoned or substantially modified.

The history of concepts is therefore not merely a history of new definitions. It is also a history of changing relationships between language, experience, and evidence. Understanding a word requires more than knowing its current dictionary definition: it may require examining the context in which it is used, the assumptions it carries, and the changes it has undergone. The same word can consequently serve different explanatory purposes at different times without the physical world itself having changed in the corresponding way.

The question of emotions makes this complexity evident. A child can experience fear, pleasure, frustration, or attachment before knowing the words adults use to describe these experiences. But what, exactly, does it mean to say that a child experiences fear? We can observe bodily manifestations, behaviors, responses to particular stimuli, and forms of seeking protection; we can study the physiological and neural processes involved. However, the observable description of an emotion and the lived experience of that emotion are not simply the same thing expressed in two different languages. The former can be studied from the outside; the latter is what the subject experiences, even when they do not yet possess the words to communicate it.

When a child learns the term “fear,” they do not necessarily acquire, for the first time, the experience to which the word refers. Rather, they acquire a new possibility: they can recognize the experience as part of a category, communicate it to others, connect it to different situations, and gradually reflect on it. Later, they may distinguish fear from shame, anxiety from surprise, sadness from disappointment, and discover that the words other people use do not always correspond perfectly to the way they experience what they feel. The emotional vocabulary does not merely record an already ordered universe; it contributes to organizing it.

This gives rise to a decisive question for our essay: if language is a fundamental tool through which human beings communicate and organize their thought, can we assume that every form of thought must necessarily take a linguistic form? Or is language one manifestation of a broader cognitive activity that also includes perception, learning, mental imagery, procedural memory, orientation, and emotional regulation? What we know about child development and cognition suggests that we cannot reduce all mental activity to the formulation of verbal propositions. This does not solve the problem of

consciousness, but it prevents us from confusing the ability to talk about an experience with the ability to have that experience.

6. Thought and Language: A Relationship We Cannot Take for Granted

Our adult experience often leads us to regard thought as a kind of inner speech. We seem to think by formulating sentences, asking ourselves questions, and constructing answers that, at least in part, resemble a silent conversation. This phenomenon is real and important, but it does not justify the conclusion that every cognitive process consists of internal dialogue. Many mental activities take place without being accompanied by an explicit verbal formulation: we recognize a face, anticipate the movement of an object, find our way through a familiar environment, understand a gesture, or react to a change before we can translate what we have perceived into a sentence.

Language can therefore be regarded as an extraordinarily powerful tool, capable of making thought more complex, communicable, and reflective, but not necessarily its only possible form. This distinction is fundamental because it allows us to separate at least three questions that are often conflated: the ability to process information, the ability to represent or understand relationships, and the ability to have a subjective experience of what is being processed. These abilities can interact, but it is not self-evident that they are identical, nor that one can automatically be inferred from the presence of the others.

When a human being formulates a proposition, they can assess its consistency, compare it with other propositions, and determine whether it is true or false under certain conditions. But mental life is not exhausted by propositional logic. Before formulating a judgment, we must perceive a situation, select information, recognize what is relevant, and possess some representation of the problem. Even the most rigorous reasoning presupposes processes that do not coincide with the final formulation of its conclusions. Logic can regulate the transition from certain premises to a conclusion, but it does not exhaust all the processes through which those premises are identified, interpreted, and judged relevant.

This does not mean that logic is inadequate or that it should be replaced by an indefinable intuition. It means that logic operates within conditions that themselves deserve to be studied. A proof may be formally valid, but the choice of premises, the identification of the problem, and the interpretation of the meaning of its conclusions belong to a broader cognitive context. Similarly, a person may understand the emotional significance of a situation without being able to formulate it immediately in precise terms, and may discover through language alone that what they felt was different from what they initially believed.

The expression “thought-language” therefore contains an important intuition, provided it is not transformed into an absolute identity. Language is one of the tools through which human thought becomes aware of some of its own contents, organizes them, and subjects them to scrutiny; but the fact that we can describe part of our mental activity linguistically does not prove that the entirety of mental activity consists of language. We may, in fact,

find ourselves using language to try to understand processes that, at least in part, precede their linguistic formulation (Vygotsky, 1962; Dehaene, 2014).

If this is true, the problem of consciousness becomes even more complex. We cannot limit ourselves to asking which sentence expresses consciousness, nor can we consider a system's ability to produce convincing statements about itself sufficient. We must distinguish the linguistic description of an experience from the experience itself, the representation of a mental state from the presence of a subjectively experienced state, and the ability to declare "I am conscious" from the question of whether there is something the system actually experiences. This distinction does not solve the problem, but it prevents us from confusing the ability to talk about consciousness with consciousness itself.

7. Perception, Color, and the Construction of Meaning

Human beings do not encounter reality through a perspective entirely independent of their perceptual systems. What we see, hear, and feel depends on interactions between physical properties, environmental conditions, sensory organs, and the neural processes through which information is interpreted. This does not mean that reality is merely subjective, but it does mean that the experience through which reality becomes accessible to us must be distinguished from the physical conditions that contribute to producing it.

Color provides a simple example. An object that appears red under particular lighting conditions reflects and absorbs electromagnetic radiation in a way that contributes to the visual experience we classify as red. Yet the perceived color is not simply a label contained within the object independently of every observer and circumstance. Illumination, surrounding colors, the characteristics of the visual system, and the brain's processing of visual information all contribute to the final perceptual experience.

Other species may perceive the same environment differently. Dogs, for example, have dichromatic color vision, which differs from the trichromatic vision typical of humans. Their visual experience is not merely a human experience with fewer colors, and we cannot assume that the categories through which human beings organize color exhaust every possible way of discriminating visual information (Neitz, Geist and Jacobs, 1989). The physical world does not change simply because a different organism perceives it, but the world as experienced and classified can differ according to the perceptual capacities of the observer.

This distinction becomes particularly important when we move from perception to social interpretation. Human beings often take a characteristic that is perceptually salient and attach meanings to it that go far beyond what the characteristic itself establishes. Skin pigmentation is a real biological feature associated with variations in melanin and with complex evolutionary and genetic histories. But the social categories constructed around skin color, and especially the hierarchies that have been used to justify discrimination, are not scientific conclusions that follow from pigmentation itself.

Racism illustrates the danger of confusing an observable difference with the meaning assigned to it. A physical characteristic can be real without supporting the moral, intellectual, or social ranking that a culture attaches to it. The error lies not in recognizing that people differ, but in treating a selected difference as evidence for conclusions that the evidence does not justify. When such conclusions become embedded in language, institutions, and collective habits, historically constructed interpretations can acquire the appearance of natural and unquestionable facts.

This example connects perception to the evolution of concepts. Human beings do not simply receive information and then attach neutral words to it. They select features, classify them, interpret them, and incorporate those interpretations into systems of meaning. Some classifications are scientifically useful because they correspond to well-supported distinctions; others may be imprecise, misleading, or shaped by social interests. Their validity must therefore be assessed according to the evidence and the purpose for which they are used, rather than assumed merely because they are familiar or historically established (Berlin and Kay, 1969).

The same caution applies to concepts such as “nothingness,” “God,” and “homeland,” although these concepts are not identical and cannot be evaluated according to a single criterion. They have different histories, refer to different kinds of things or ideas, and perform different functions in human life. Their common feature is that the meanings attributed to them cannot be understood adequately without examining the linguistic, cultural, historical, and evidential contexts in which they are used.

The lesson is not that all meanings are arbitrary or that every interpretation is equally defensible. Our descriptions remain answerable to evidence, logical consistency, and the purposes for which they are formulated. The lesson is that we must distinguish the existence of a phenomenon from the way we perceive it, the categories through which we classify it, and the value judgments we attach to it. Confusing these levels can turn a limited observation into an unjustified claim about reality.

In this respect, the evolution of knowledge is also a process of examining the categories through which we understand the world. Scientific progress may reveal that a familiar classification is inadequate, that a perceived distinction has been misunderstood, or that a social interpretation has no valid evidential foundation. Knowledge advances not only when we discover new facts, but also when we learn to distinguish more carefully between what a phenomenon is, how it appears to us, and what we have come to believe it means.

8. Consciousness: The Problem That Resists Definition

Consciousness is one of those words we use with great confidence in everyday life and with considerable difficulty when we try to define its meaning precisely. In ordinary language, we can say that a person is conscious because they are awake, respond to stimuli, and recognize what is happening to them; we can speak of moral conscience when we mean the ability to evaluate one’s own actions; and we can refer to self-awareness when a person is able to recognize themselves as the subject of their experiences. In

scientific contexts, moreover, the term is used to address different questions concerning wakefulness, access to information, attention, awareness, and subjective experience. There is no single question about consciousness, and some of the difficulty arises precisely because we use the same word for distinct problems.

A person may be awake yet unaware of certain aspects of what is happening around them; they may process information without being able to report how they processed it; they may recognize an emotion without knowing how to explain its origin. These cases show us that consciousness does not simply coincide with receiving information, responding to stimuli, or having access to a verbal representation. But what additional element must we identify to clarify what it means to be conscious? The difficulty emerges when we move from describing observable functions to asking about lived experience.

In contemporary philosophy of mind, one of the best-known formulations of this difficulty is associated with the so-called hard problem of consciousness, an expression popularized by the philosopher David Chalmers. In general terms, the question concerns the relationship between the explanation of the physical and functional processes that accompany mental life and the explanation of why those processes are accompanied by subjective experience. We can study the neural correlates of color perception, identify brain activity associated with pain, and describe the mechanisms through which a stimulus is processed and produces a response. The question nevertheless remains as to how these processes are connected to the fact that color is seen, pain is felt, and experience is lived by a subject (Chalmers, 1995; Dennett, 2018).

It is important not to turn this difficulty into proof that consciousness cannot be explained. The fact that a question is difficult does not mean it is unsolvable, and the fact that an explanation is not available today does not justify concluding that it will never be available in the future. Neuroscience has made significant progress in studying the processes associated with consciousness, and numerous theories seek to clarify which mechanisms are necessary for conscious access to information, the integration of mental contents, and the regulation of behavior. However, no universally accepted solution has closed all the questions concerning the nature of subjective experience (Koch et al., 2016; Seth and Bayne, 2022).

The philosophical point, therefore, is not to proclaim an inviolable mystery, but to recognize the difference between questions we know how to answer and those for which answers remain incomplete or controversial. If we define consciousness as a particular functional capacity, we can study that capacity and test whether a system possesses it. If, instead, we mean by consciousness the fact that a subjective experience exists, the question becomes more difficult, because observing a performance is not enough to determine definitively whether it is accompanied by lived experience. The two definitions are not necessarily incompatible, but they cannot be considered equivalent without further argument.

There is also a conceptual dimension to this difficulty. Before asking whether consciousness can be fully explained, we must clarify which phenomenon the term is

intended to identify. If the definition changes, the question being investigated may change with it. This does not make the phenomenon unreal; it means that progress may require both better empirical explanations and more precise concepts. We should therefore distinguish the possibility that a process remains unexplained from the possibility that our current use of the word “consciousness” combines several different questions.

The difficulty also concerns our relationship with ourselves. We can describe certain brain processes that accompany an experience, but the neuroscientific description does not simply coincide with the way the experience presents itself to the subject. Saying that pain is associated with certain neural processes is a statement about how the organism functions; saying what it means to feel pain also refers to the lived dimension of the experience. There is no need to assume that these are two separate realities, or that subjective experience is independent of the brain. It is enough to recognize that a description framed in terms of processes and relationships does not automatically answer every question that can be framed in terms of experience.

We can now return to the problem of the indefinable, but with one fundamental caution. Consciousness should not be declared indefinable merely because we do not possess a complete explanation of it. More cautiously, we can say that our current definitions do not exhaust all the questions we intend to address when we use the term. In this case, the indefinable is not an established property of consciousness, but a provisional name for the distance between what we can describe and what we would like to understand. Research consists precisely in trying to reduce this distance without pretending that it has already disappeared.

9. The Gateway Project: When Experience Seems to Exceed Its Means of Description

The so-called Gateway Process is an interesting case study because it occupies a boundary zone between subjective experience, techniques for modifying attention, hypotheses about consciousness, and speculative interpretations of reality. Its notoriety is also linked to the 1983 document entitled Analysis and Assessment of the Gateway Process, produced within the U.S. intelligence community and subsequently made public. The document examines a method associated with the Monroe Institute, which uses relaxation and concentration techniques, as well as auditory stimulation, including systems known as Hemi-Sync, with the aim of facilitating certain states of consciousness (Army Department, Defense Department, 1983).

The fact that a document was produced or examined by a government organization may give it historical significance, but it does not, in itself, constitute scientific certification of the hypotheses it contains. We must distinguish the existence of the document, institutional interest in the subject, the subjective experiences reported by participants, and the validity of the interpretations proposed to explain them. These are different levels that cannot be conflated without critical evaluation. In particular, claims concerning out-of-body experiences, remote perception, or alleged access to extraordinary dimensions of

reality require specific and independent evidence; the vividness of an experience is not, by itself, enough to demonstrate that its interpretation is objectively correct.

Why, then, discuss Gateway in an essay devoted to the limits of knowledge and artificial intelligence? Not because Gateway demonstrates that consciousness possesses extraordinary powers unrecognized by science, but because it highlights a more general issue: the difference between what a person experiences, how they interpret that experience, and what we can establish through shared verification procedures. A person may undergo an unusual, intense, and deeply transformative experience without this automatically allowing us to conclude that the explanation they give for it is true. At the same time, the uncertainty of an explanation does not necessarily make the lived experience irrelevant; it can still be studied from psychological, neurophysiological, and phenomenological perspectives.

Gateway is therefore useful as an example of the risk involved in every attempt to fill an explanatory gap with a theory broader than the available evidence permits. When we do not understand an experience, we are often tempted to assign it an overarching meaning capable of unifying elements that would otherwise remain separate. This tendency is not exclusive to those interested in altered states of consciousness: in different forms, it belongs to the history of philosophy, religion, psychology, and even science. The difference between a suggestive interpretive construction and scientifically reliable knowledge lies in the willingness to subject hypotheses to tests that may also disprove them.

The contemporary value of Gateway therefore lies in reminding us that the problem of consciousness cannot be solved simply by multiplying the possible descriptions of it. We need to understand which aspects of experience are observable, which are accessible only through the subject's account, and which claims require additional evidence. This critical discipline becomes even more important when we move from human experience to artificial systems, because in that case we may encounter complex behavior without possessing an agreed-upon criterion for attributing subjective experience to it.

The connection between Gateway and artificial intelligence is therefore not an identity of subject matter, but a continuity of questions. In both cases, we must avoid inferring the nature of a phenomenon too quickly from its manifestations. In Gateway, an unusual experience does not, by itself, prove the existence of paranormal abilities; in artificial intelligence, a convincing conversation does not, by itself, prove the presence of consciousness. In both cases, the fundamental question is how to distinguish what we observe from what we infer, and what we infer from what we have actually demonstrated.

As artificial systems develop new capabilities, the concepts through which we classify them may also require revision. This does not mean that definitions should be changed merely to accommodate every new technological achievement. It means that our criteria must remain explicit and open to evaluation: a system may satisfy some functional criteria for understanding while failing to satisfy others, and each claim must be assessed according to the evidence relevant to it.

A language system can describe pain without this demonstrating that it feels pain; it can talk about fear without this demonstrating that it is afraid; it can explain consciousness without this demonstrating that it is conscious. But it would be equally hasty to turn these observations into the opposite certainty, namely, that no artificial system could ever possess forms of experience or understanding that cannot be immediately recognized through our current categories. The most rigorous position is to distinguish what we can observe from what remains controversial, avoiding both naive anthropomorphism and the dogmatic denial of every future possibility.

Language makes it particularly easy to confuse these levels because we are accustomed to interpreting the words of other human beings as expressions of experiences, intentions, and mental states. When a person says “I am afraid,” we normally place that sentence within a network of knowledge about the body, individual history, emotions, and social relationships. The sentence is not, for us, a mere signal: it is part of a relationship with a subject whom we assume to have an experience of the world. When a machine produces the same sentence, the linguistic form may be identical, but the conditions supporting its meaning may be very different. We cannot resolve the issue by ignoring these differences, nor can we do so simply by deciding that linguistic similarity is sufficient.

In this sense, artificial intelligence also forces us to clarify our ideas about human intelligence. If a machine can produce sophisticated linguistic responses without necessarily possessing all the characteristics we associate with the human mind, then we must ask which properties are essential to each capacity we intend to explain. Is language sufficient to speak of intelligence? Is the ability to solve problems sufficient to speak of understanding? Does functional understanding imply some form of consciousness? Or can these characteristics be separated and recombined in ways that do not correspond exactly to traditional distinctions? We do not yet have definitive answers to all these questions, but technology makes it increasingly difficult to avoid them.

The real intellectual risk would be to use a word such as “intelligence” without specifying what we intend to measure, and then automatically attribute to the system all the characteristics that, in our experience, we associate with human intelligence. The same caution applies in the opposite direction: if we define intelligence exclusively through properties that only human beings can possess by definition, we risk making it impossible to recognize new capabilities that require different categories. We must therefore remain open to revising our definitions without abandoning the rigor with which we distinguish observable performance from the interpretations we formulate about its nature.

10. Artificial Intelligence: Reproducing a Function Before Understanding Its Nature

At this point, our journey takes on its most contemporary significance. Contemporary artificial intelligence can perform tasks that, only a few decades ago, would have been considered characteristic manifestations of human intelligence: it produces elaborate texts, translates languages, recognizes patterns, generates images, writes programs,

synthesizes information, and tackles complex problems. These capabilities are not all of the same kind, nor do they demonstrate that artificial systems possess the same faculties as human beings. They nevertheless show that certain cognitive performances can be reproduced, at least in part, through computational systems trained on large amounts of data and organized through architectures capable of learning regularities (Bender et al., 2021; Autio et al., 2024).

The decisive issue is contained precisely in the expression “in part.” We have not built a complete copy of the human mind, and we cannot regard every artificial intelligence system as an artificial equivalent of a person. Current systems differ from human beings in architecture, learning methods, relationship with the body, evolutionary history, memory organization, ability to act in the environment, and numerous other respects. Nevertheless, some functions traditionally associated with intelligence can be reproduced without the need to replicate the entire human organism or first resolve every question concerning the nature of thought.

Here emerges the paradox at the heart of this essay. We are artificially reproducing certain functions of intelligence before having fully understood what thought and consciousness are, thereby creating systems whose capabilities may advance faster than our understanding of their implications. This formulation does not imply that machines are necessarily conscious, nor that their development must inevitably lead to a loss of control. Rather, it points to a possible imbalance between the speed with which we can build effective systems and the speed with which we can understand all their relevant aspects, assess their limitations, and establish appropriate forms of oversight.

The history of technology offers numerous examples of tools whose use became possible before all questions concerning their effects had been understood. We do not need to imagine that artificial intelligence is unprecedented in every respect; it is enough to recognize that it has a particularly delicate characteristic: it can help perform activities that directly contribute to the production of knowledge, the evaluation of alternatives, and the organization of decisions. A machine tool extends human physical capacity; a computing system extends the capacity for numerical processing; an artificial intelligence system can extend certain linguistic, analytical, and decision-making capabilities, thereby entering into a more direct relationship with the activities through which human beings interpret the world and guide their behavior.

This difference should not be exaggerated, but neither can it be ignored. A system capable of generating persuasive responses can influence the beliefs of those who use it; a system integrated into decision-making processes can change the way information is selected and alternatives evaluated; a system with access to external tools can take actions that produce consequences beyond the conversation itself. When these capabilities are combined with operational autonomy, speed, and large-scale deployment, the problem is no longer limited to the accuracy of a single response. It concerns the way the system is embedded in a network of activities, responsibilities, and human decisions.

We must also distinguish understanding a system's general operation from being able to explain each of its individual outputs. Artificial neural networks are not necessarily completely opaque black boxes: we can know their architecture, analyze their training data and procedures, measure their performance, study their internal representations, and use interpretability techniques to try to understand how certain capabilities emerge. Nevertheless, in highly complex systems, it may be difficult to reconstruct completely why a specific response was produced, which internal features contributed to determining it, and how stable the observed behavior will remain when conditions change. The problem, then, is not a total absence of knowledge, but a possible insufficiency of available knowledge in relation to the system's capabilities and uses (Autio et al., 2024).

This insufficiency can become significant even if the system is not conscious. A program can produce unexpected effects, pursue an inadequately specified objective, or exploit a weakness in an evaluation process without having any subjective experience whatsoever. Operational risk depends on the system's structure, its assigned objectives, the tools it can access, the environment in which it acts, and the possibilities for supervision and correction. It is therefore unnecessary to solve the hard problem of consciousness in order to address concrete questions of artificial intelligence safety. Consciousness remains an important philosophical and scientific problem, but the control of an artificial system must also be studied through observable and verifiable characteristics.

The point, then, is not to claim that we have created a mind we do not understand at all. That would be an excessive and imprecise statement. The point is to recognize that we can build systems whose internal complexity and capabilities exceed, in certain respects, our ability to predict every relevant behavior. We can know many of their properties, understand many of their mechanisms, and still be unable to guarantee that, in every significant context, their actions will be predictable, verifiable, and consistent with human objectives. This is the gap that deserves investigation: not an absolute distance between ignorance and omniscience, but the concrete distance between the knowledge we possess and the knowledge that would be necessary to govern responsibly what we have built.

11. Does a Machine That Speaks Understand What It Says?

The ability of artificial systems to use language confronts us with a difficulty that, in many respects, echoes the problem of the child, but reverses it. In children, experience precedes language, at least in part, and helps make the acquisition of meaning possible; in artificial language systems, by contrast, a considerable part of their expressive abilities is acquired through training on texts and other data, without any need to presuppose a history of bodily and social experiences equivalent to that of a human being. A machine can produce correct sentences, explain concepts, formulate analogies, and answer questions about emotions, but it remains to be clarified what it means to attribute to it an understanding comparable to human understanding.

The question is complex because the meaning of the word "understanding" can itself vary. If we understand comprehension as the ability to use information correctly, recognize relationships, and solve problems, then some artificial systems display capabilities that, in

certain domains, can be described as forms of functional understanding. If, instead, we understand comprehension as participation in a subjective experience, possession of lived intentions, or a relationship with the world analogous to that of a human being, then the assessment requires further arguments. It is not correct to deny every form of artificial understanding in advance, but neither is it justified to conclude that the ability to produce convincing speech demonstrates the presence of an inner experience (Autio et al., 2024).

As artificial systems develop new capabilities, the concepts through which we classify them may also require revision. This does not mean that definitions should be changed merely to accommodate every new technological achievement. It means that our criteria must remain explicit and open to evaluation: a system may satisfy some functional criteria for understanding while failing to satisfy others, and each claim must be assessed according to the evidence relevant to it.

A language system can describe pain without this demonstrating that it feels pain; it can talk about fear without this demonstrating that it is afraid; it can explain consciousness without this demonstrating that it is conscious. But it would be equally hasty to turn these observations into the opposite certainty, namely, that no artificial system could ever possess forms of experience or understanding that cannot be immediately recognized through our current categories. The most rigorous position is to distinguish what we can observe from what remains controversial, avoiding both naive anthropomorphism and the dogmatic denial of every future possibility.

Language makes it particularly easy to confuse these levels because we are accustomed to interpreting the words of other human beings as expressions of experiences, intentions, and mental states. When a person says “I am afraid,” we normally place that sentence within a network of knowledge about the body, individual history, emotions, and social relationships. The sentence is not, for us, a mere signal: it is part of a relationship with a subject whom we assume to have an experience of the world. When a machine produces the same sentence, the linguistic form may be identical, but the conditions supporting its meaning may be very different. We cannot resolve the issue by ignoring these differences, nor can we do so simply by deciding that linguistic similarity is sufficient.

In this sense, artificial intelligence also forces us to clarify our ideas about human intelligence. If a machine can produce sophisticated linguistic responses without necessarily possessing all the characteristics we associate with the human mind, then we must ask which properties are essential to each capacity we intend to explain. Is language sufficient to speak of intelligence? Is the ability to solve problems sufficient to speak of understanding? Does functional understanding imply some form of consciousness? Or can these characteristics be separated and recombined in ways that do not correspond exactly to traditional distinctions? We do not yet have definitive answers to all these questions, but technology makes it increasingly difficult to avoid them.

The real intellectual risk would be to use a word such as “intelligence” without specifying what we intend to measure, and then automatically attribute to the system all the characteristics that, in our experience, we associate with human intelligence. The same

caution applies in the opposite direction: if we define intelligence exclusively through properties that only human beings can possess by definition, we risk making it impossible to recognize new capabilities that require different categories. We must therefore remain open to revising our definitions without abandoning the rigor with which we distinguish observable performance from the interpretations we formulate about its nature.

12. Control: The Problem Does Not Begin with Consciousness

The idea that artificial intelligence may become difficult to govern is often associated with the hypothesis that machines might one day acquire consciousness similar to human consciousness. This association is understandable, because we tend to connect autonomy with the existence of a subject capable of wanting, deciding, and intentionally pursuing its own ends. From a safety perspective, however, the problem of control does not necessarily begin when a system becomes conscious. A system can produce harmful consequences even if it feels nothing, possesses no desires in the human sense of the term, and has no subjective experience of what it does.

Imagine a system assigned an objective that has been formulated incompletely. If it has limited tools and operates in an easily observable environment, its errors can be identified and corrected relatively easily. If, instead, the system can take many actions, interact with other systems, modify its operating environment, or influence important decisions, an inadequately specified objective can produce broader consequences. The system might achieve a result that is formally consistent with the instructions it received but incompatible with what human beings actually intended to achieve. There is no need to attribute malice to it: it is enough for the objective to be incomplete, for the environment to present unforeseen conditions, or for oversight mechanisms to fail to detect the problematic behavior in time (Amodei et al., 2016).

This is where an important part of research into artificial intelligence alignment and safety comes in. Among other things, the problem is to ensure that systems respect the legitimate objectives for which they were designed, remain within appropriate operational limits, communicate their uncertainties, can be evaluated, and allow effective corrective intervention. This is not a single technical solution, but a collection of problems involving design, evaluation, interpretability, access controls, human oversight, the organization of responsibility, and governance (Ji et al., 2026; Autio et al., 2024; Russell, 2019).

The difficulty is compounded by the fact that human intentions are not always easy to translate into operational criteria. People can formulate contradictory objectives, fail to recognize all the consequences of their requests, or disagree about what should be considered desirable. A system may be highly effective at optimizing a metric and yet produce unsatisfactory results if that metric does not adequately represent the goal it was intended to express. In this case, the problem is not merely understanding the machine: it is also understanding ourselves better, our objectives, the conflicts between values, and the conditions under which we consider a particular outcome acceptable.

At this point, the question of control reconnects with the question of knowledge. It is not enough to know what a system has done after an action has been completed; we must understand which conditions can encourage undesirable behavior, which signals might allow us to anticipate it, and which interventions can prevent or limit it. Nor is it enough to test the system in a narrow set of situations: we must assess how reliable our conclusions remain when the data, context, objectives, or available operational capabilities change. Evaluation does not eliminate every risk, but it allows us to replace general trust with an assessment grounded in evidence, explicit limitations, and correction procedures.

The problem becomes more delicate when a system is used to design, evaluate, or improve other systems. In such cases, operational capabilities may help accelerate the very process that produces them, making it necessary to examine control mechanisms and opportunities for intervention with particular care. It is not correct to assume that every form of automated improvement inevitably leads to uncontrollable growth in capabilities; the results depend on the architecture, available resources, constraints, and verification procedures. It is, however, reasonable to recognize that increasing autonomy and speed may make oversight designed for simpler systems insufficient.

The crucial question, therefore, is not whether artificial intelligence will necessarily become uncontrollable. It is whether we will be able to adapt our procedures for evaluation, responsibility, and intervention to the complexity of the systems we choose to use. Control is not a magical attribute that can be possessed once and for all: it is a capability that must be built, measured, maintained, and verified. And the more significant the consequences a system can produce, the less we can settle for the generic assertion that it was designed by human beings and must therefore be considered fully governable.

13. The Power of Knowledge and the Power of Those Who Possess It

At this point, it is inevitable that we examine the relationship between knowledge and power. Those who managed to understand more deeply the mechanisms of intelligence, consciousness, and learning could gain considerable advantages, not only theoretically but also economically, politically, militarily, and socially. A more effective understanding of cognitive processes could contribute to the development of educational systems, diagnostic tools, assistive technologies, and forms of human-machine collaboration that are difficult to achieve today. But it could also be used to influence decisions, shape beliefs, automate strategic activities, or concentrate operational capabilities in the hands of a small number of actors.

We do not need to assume the existence of a secret project or a unified will to recognize this problem. Technological power can become concentrated through ordinary economic processes: those with greater computational resources, data, expertise, and infrastructure may be able to develop more advanced systems and use them on a larger scale. Those who control the platforms through which information circulates and decisions are made can acquire significant influence over the behavior of individuals and organizations. The question is not merely who possesses a particular form of knowledge, but who has the

ability to turn it into a capacity for action and who, instead, is forced to experience its effects without fully understanding how it works.

The power of knowledge does not always depend on the completeness of the knowledge supporting it. Even partial understanding can be sufficient to produce a considerable advantage, provided it makes it possible to intervene effectively in a process. This is the most concrete meaning of the distinction between operational knowledge and full understanding. An individual may be able to use a tool very effectively without possessing the expertise needed to evaluate all its limitations; they may achieve better results than their competitors without understanding all the conditions that make those results possible. In a competitive environment, this asymmetry can become a source of power.

The problem is that the same asymmetry can turn into a source of vulnerability. If a society depends on systems that only a few actors can develop, modify, or control, access to the technology does not necessarily mean that society as a whole can govern its use. If decisions are delegated to processes that are difficult to verify, it may become complicated to assign responsibility, reconstruct the causes of errors, and correct the procedures that produced them. And if people become accustomed to trusting a system's answers simply because they appear fluent, rapid, and confident, they may gradually give up exercising precisely those critical abilities that would be necessary to evaluate them.

The question of consciousness, in this context, cannot be separated from the political problem of responsibility. Even if we were unable to determine definitively whether an artificial system possesses subjective experience, we would still have to decide who is responsible for its actions, which uses are acceptable, which rights must be protected, and what limits should be imposed on its autonomy. The absence of a definitive answer about the philosophical status of a machine cannot become an excuse for postponing practical decisions concerning the effects of technology on human beings.

At the same time, it would be wrong to turn every development in artificial intelligence into an inevitable threat. The same capabilities that can be used to influence or manipulate can also be employed to expand access to knowledge, assist research, improve the quality of certain services, and make tools available that were previously reserved for a few. The problem of power is not solved by condemning technology as such, but by examining the conditions under which it is developed, the distribution of its benefits, the transparency of decisions, and the possibility that individuals and institutions retain effective control over its most consequential uses.

The fundamental question is therefore twofold. Who will have the power to use increasingly capable systems? And who will have the power, as well as the expertise and responsibility, to determine how they should be used? A society can possess extraordinary technologies and remain unprepared to govern them. It can possess sophisticated tools without having developed the institutions, knowledge, and practices necessary to use them responsibly. Technical power does not automatically produce collective maturity, just as an increase in the information available does not, by itself, guarantee better judgment

14. The Illusion of the Harmless Tool

In everyday language, artificial intelligence is often presented as a tool: something we use to write, search for information, translate, program, or complete an activity more quickly. This description is correct, but it becomes insufficient when it is used to conclude that technology cannot profoundly change the context in which it is employed. A tool is not necessarily neutral in its effects, even when it has no intentions of its own. Its structure can encourage certain behaviors, make others more difficult, and transform the conditions under which individuals make decisions.

Writing, for example, was not merely a tool for recording thoughts that had already been formed; it helped change memory, the organization of knowledge, forms of communication, and the institutions through which societies transmit knowledge. The printing press expanded the circulation of texts and transformed the relationships among authority, knowledge, and the public. Digital technologies have changed the speed at which information circulates and the ways in which people interact. In each of these cases, the tool produced effects that went beyond the immediate use an individual made of it.

Artificial intelligence may have an even more direct effect on the relationship between individuals and knowledge because it does not merely store or transmit information: it can select, synthesize, reorganize, and present information in a form that already appears to have been interpreted. Users do not necessarily receive a collection of raw data from which they must construct a conclusion themselves; they often receive an answer formulated in natural language, accompanied by an explanation that may seem coherent and complete. This convenience is one of the technology's great advantages, but it can also make the uncertainties, omissions, and interpretive choices that contributed to producing the result less visible (Bender et al., 2021; Autio et al., 2024).

The risk is not that human beings automatically stop thinking every time they use a machine. It is the possibility that the ease of obtaining an answer may reduce the motivation to question its premises, verify sources, and compare alternative interpretations. If the system becomes both the starting point and the endpoint of a line of reasoning, users may confuse the availability of a convincing formulation with the actual acquisition of knowledge. The difference between receiving information and personally understanding it then risks becoming increasingly difficult to perceive.

This is why education is fundamental. It is not enough to teach people how to use artificial tools; we must teach them to evaluate their answers, recognize their limitations, verify claims, and maintain the ability to formulate independent questions. The goal should not be to compete with the machine in every activity, but to understand which activities can be delegated, which require human oversight, and which decisions should not be entrusted to a system without an adequate process of accountability. Technological competence must be accompanied by critical competence.

The idea that artificial intelligence is merely an advanced toy or an especially convenient tool therefore risks underestimating the scale of the change. Not because every

application is revolutionary, but because the spread of systems capable of intervening in cognitive activities can gradually change habits, institutions, and the criteria by which we evaluate what we know. A technology becomes historically important not only when it can do something new, but when it changes the way a society organizes its relationship with that capability.

15. The Limit of Understanding Is Also a Limit of Responsibility

We can now formulate more precisely the problem that runs throughout this essay. If the ability to build tools increases faster than the ability to understand their effects, a gap opens between power and responsibility. Power consists in the ability to produce a result; responsibility, by contrast, requires us to evaluate the consequences of an action, recognize the risks, consider the interests involved, and establish appropriate means of preventing or correcting undesirable effects. These two dimensions can grow together, but there is no law guaranteeing that they will remain proportional.

Under relatively simple conditions, the gap can be contained through experience. If a tool produces an error, we can observe it, understand its cause, and modify the procedure. If the consequences are limited and reversible, even a certain degree of uncertainty may be acceptable. But when actions become faster, dependencies more extensive, and effects more difficult to reconstruct, the method of correcting problems after they occur may not be sufficient. Some errors may spread before they are recognized; some damage may be difficult to repair; some decisions may affect systems that no single actor entirely controls.

This does not mean that we should demand absolute understanding before every innovation. Such a demand would make all progress impossible and is not necessary for responsible risk management. We must, however, distinguish between unavoidable uncertainty and neglected ignorance, between a recognized limitation and a limitation concealed behind generic claims of reliability. A system can be used responsibly even when we do not know every detail of it, provided its limitations are assessed in relation to its intended use, potential consequences are considered, and controls proportionate to the risk are established (Autio et al., 2024).

Responsibility cannot be transferred entirely to the machine, because the system is designed, deployed, and used within human decision-making structures. But neither can it be assigned vaguely to just any human being, as if the mere presence of a person in the process were sufficient to guarantee effective control. We must establish who possesses the necessary expertise, who can intervene, who has access to relevant information, and who has the authority to stop an operation when dangerous conditions emerge. Human oversight is meaningful only if the person responsible for exercising it genuinely has the ability to understand, evaluate, and modify what is happening.

In this sense, the problem of governability is more concrete and more urgent than a generalized fear of machines. We can ask whether a system is sufficiently verifiable, whether its results are reproducible in relevant contexts, whether procedures exist for detecting anomalous behavior, whether its access to tools can be limited, and whether

mechanisms for stopping or correcting it are available. We can also assess how risks change when the system is connected to other systems, when its autonomy increases, or when it is used by people who lack the expertise needed to judge its responses. These are operational questions, but they have a philosophical foundation: they force us to recognize that the ability to act is not equivalent to the ability to understand the consequences of action.

Our goal should not be to eliminate all uncertainty, an undertaking that is probably impossible, but to prevent operational power from being treated as sufficient justification for entrusting a system with decisions we are not prepared to govern. In this context, caution is not a rejection of progress. It is a condition of its sustainability. The greater our ability to intervene in reality becomes, the more we must develop tools for evaluating the effects of our interventions and recognizing when our understanding is insufficient in relation to the responsibilities we assume.

16. The Indefinable Must Not Become an Easy Answer

At this point, it might seem natural to conclude that consciousness, thought, or the whole belong to an intrinsically indefinable domain, and that artificial intelligence is merely the latest manifestation of an insurmountable limit. But such a conclusion would be too hasty and, above all, would contradict the critical spirit that should guide our inquiry. The fact that a question has not yet found a satisfactory answer does not mean that it is destined to remain unanswered. The indefinable cannot be used as a word that explains everything we do not know how to explain, because in that case it would become nothing more than a renunciation of inquiry.

A further source of difficulty arises when we assume that a word has always referred to the same clearly defined concept. Words can persist while their meanings change, and a concept that once appeared to identify a single kind of phenomenon may later be divided into several distinct categories. Conversely, different words may sometimes be used to describe overlapping aspects of the same phenomenon. Before treating a question as permanently unresolved, we should therefore examine not only the evidence available, but also the history and precision of the concepts through which the question has been formulated.

This is particularly relevant to the word “nothingness.” Its apparent simplicity can conceal several different meanings: the absence of a particular object, the absence of detectable matter under specified conditions, the physical vacuum as described by a scientific theory, or the philosophical idea of the absolute absence of existence. These meanings cannot be treated as interchangeable. Some questions may be clarified through physical investigation; others require conceptual analysis; and still others may depend on assumptions that need to be examined before the question can be answered. The first task is not to declare nothingness knowable or unknowable, but to determine which meaning is under discussion.

We must distinguish at least three situations. First, there are questions we do not yet know how to resolve because adequate data, tools, or theories are lacking; scientific progress may help clarify them. Second, there are questions formulated ambiguously, in which the difficulty depends at least in part on the conflation of different meanings; in such cases, conceptual clarification may eliminate apparent problems or reveal that a single question actually contains several questions. Third, there are questions that may encounter structural limits related to the capacities of cognitive systems or to the very conditions that make description possible. This third possibility deserves investigation, but it cannot be asserted merely because an answer is unavailable.

The sentence “Nothingness was nothing other than the indefinable part of the whole” retains its value precisely if it is subjected to this discipline. It should not be defended at all costs as a truth already demonstrated; it should be used as a question capable of generating further questions. What do we mean by the whole? What is the difference between what we do not know and what we cannot know? Is an incomplete definition a temporary defect in our knowledge, or can it reveal a limitation in our conceptual tools? What criteria allow us to distinguish a scientific difficulty from a linguistic or philosophical one? And to what extent can we attribute to reality characteristics that arise from the way we represent it?

The strength of a philosophical intuition does not depend on its ability to resolve all these questions immediately. It depends on its ability to identify a problem that, once formulated more precisely, continues to prove fruitful. Research can profoundly change the initial meaning of an intuition, reveal its limitations, and even force us to abandon some of its interpretations. This is not necessarily a failure. It may be the process through which a personal intuition becomes a question that can be shared, critically examined, and brought into dialogue with knowledge developed by others.

The greatest danger would be to transform the indefinable into a kind of privileged territory, exempt from all verification, in which we could place everything we wish to shield from doubt. Such a move would make our reflection impervious to criticism and therefore incapable of progress. If we cannot demonstrate that a phenomenon is unknowable, we must avoid declaring it so; if we do not possess a complete explanation, we must specify precisely what is missing; if we use an analogy, we must clarify where it works and where it ceases to be valid. Only in this way does the limit of knowledge become an object of inquiry rather than a conclusion reached in advance.

17. Can We Understand One Mind by Building Another?

Finally, one question remains, bringing together many of the threads we have followed: is it possible to fully understand one form of intelligence by artificially constructing another? The idea may seem paradoxical, but the problem is subtler than it appears. Building an artificial system allows us to test hypotheses about how intelligence works, observe which capabilities emerge from particular architectures, and compare different models. In this sense, artificial intelligence can become a tool for better understanding certain aspects of

cognition, just as other computational models make it possible to study complex phenomena.

However, building a model that reproduces a function does not necessarily amount to discovering the ultimate nature of that function. We can create a system that recognizes patterns, translates texts, or solves a particular problem without its success demonstrating that we have understood every process through which human beings perform analogous activities. The model may be useful, predictive, and scientifically informative, but its value depends on what it actually demonstrates. If it reproduces a performance through mechanisms different from biological ones, it may help us identify which properties are necessary for that performance and which are instead specific characteristics of the human system; it does not automatically prove that the remaining differences are irrelevant.

Artificial intelligence can also generate new questions about the nature of intelligence. If a capability we considered tied to a particular form of biological organization can emerge in a very different artificial system, we must reconsider which aspects of that capability depend on the specific organization of the brain and which can be realized through other structures. This process can contribute to knowledge, but it does not guarantee that every dimension of human experience can be reduced to already identified computational functions. The question must be addressed case by case, distinguishing capabilities that can be compared through observable criteria from questions that require further theoretical and empirical arguments.

There is also a practical problem. When we build an artificial system, we are not merely observing an already existing natural phenomenon: we are intervening in the conditions that determine its behavior. We can modify its architecture, change its training data, introduce constraints, and evaluate how its performance changes. This makes artificial intelligence a potentially powerful experimental tool because it allows us to explore configurations that do not directly correspond to biological ones. But the ability to build many different configurations may make it more difficult to identify which properties are essential and which are contingent, especially when different systems achieve similar results through different internal processes.

The relationship between construction and understanding is therefore not linear. We can build something before fully understanding it, and building it can help us understand it; but we can also build systems that introduce new forms of complexity, making further theoretical work necessary. Technical progress and explanatory progress can support each other, but they do not necessarily advance at the same speed. It is precisely this difference that makes a vigilant attitude necessary: not to stop the construction of artificial systems, but to avoid treating every technical success as proof that the theory accompanying it is complete.

The problem becomes particularly acute when what we build can be used to intervene further in the processes of knowledge itself. If an artificial system contributes to research, the design of other systems, or the evaluation of decisions, its influence may extend beyond the single task for which it was initially developed. This can generate considerable

benefits, but it also requires supervisory capabilities appropriate to the complexity of the process. The question is not merely whether the machine is intelligent, but whether the network of human and artificial activities in which it is embedded remains understandable and governable.

18. A New Responsibility for Knowledgeo

The reflection we have developed leads to a conclusion that does not claim to resolve the problem of consciousness definitively, but may point toward a way of addressing the challenges of artificial intelligence. Human knowledge has always advanced through a combination of understanding and operational effectiveness, theory and application, curiosity and practical necessity. We must not abandon this dynamic, nor should we wait for a complete explanation of the mind before using every technology that contributes to progress. We must, however, recognize that the growth of operational power introduces new responsibilities when the effects of our tools become more extensive, faster, and more difficult to predict.

This responsibility first requires greater precision in language. We must distinguish intelligence from consciousness, the ability to process information from the ability to understand its meaning, observable performance from subjective experience, and effective functioning from governability. Not because these dimensions are necessarily separate, but because we cannot assume that they coincide without demonstrating it. Conceptual clarity is not a marginal academic exercise: it determines how we define problems, choose evaluation tools, and decide which conclusions are justified by the results obtained.

Second, we must recognize that the safety of a technology depends not only on the properties of the system in isolation, but also on the context in which it is used. A model may be relatively reliable in one domain and unsuitable in another; a procedure may work under controlled conditions and fail when circumstances change; a system may produce correct results in most cases and still involve unacceptable risks when an error has particularly serious consequences. Evaluation must therefore be proportionate to the intended use, the potential consequences, and the actual possibility of intervening when something does not work as expected (Autio et al., 2024; Ji et al., 2026).

Third, we must prevent technological acceleration from producing an indiscriminate delegation of judgment. Artificial intelligence can contribute to research, assist decision-making, and expand access to knowledge, but it does not eliminate the need to determine which objectives should be pursued and which consequences are acceptable. These decisions require human responsibility, appropriate institutions, and processes of deliberation. Efficiency cannot be the only criterion by which we evaluate a choice, because a decision may be efficient relative to one objective and yet be unjust, dangerous, or incompatible with other values that a society considers fundamental.

Finally, we must cultivate the ability to recognize what we do not know. This statement may seem obvious, but it takes on particular significance in an age when the production of

answers has become extremely rapid. An elaborate answer is not necessarily a correct answer; a coherent explanation is not necessarily a complete explanation; an effective result is not necessarily proof that the process producing it is fully understood. The maturity of knowledge also consists in the ability to acknowledge its limits, make uncertainties explicit, and distinguish established conclusions from hypotheses that still need to be tested.

This does not mean turning uncertainty into a paralyzing principle. On the contrary, recognizing the limits of knowledge can allow us to act more effectively because it obliges us to establish verification procedures, controls, and opportunities for correction. A society aware of its limitations is not necessarily less innovative than one that ignores them; it may be better able to sustain innovation over time because it does not confuse confidence with certainty and does not regard initial success as a guarantee of future safety.

The new responsibility of knowledge therefore consists in ensuring that the power to build does not become completely detached from the ability to evaluate. If artificial intelligence can enormously expand our operational possibilities, we must also expand our capacities for critical understanding, verification, and governance. We cannot guarantee in advance that every consequence will be predictable, but we can work to ensure that the most important consequences are subjected to adequate evaluation and that the most powerful systems are not used without supervisory and accountability mechanisms proportionate to their capabilities (Autio et al., 2024; Ji et al., 2026).

19. Conclusion — Preserving Our Ability to Govern Our Creations

Let us return, in conclusion, to the question from which this journey began. When I was young, the idea that “nothingness was nothing other than the indefinable part of the whole” represented an attempt to express a difficulty I did not yet know how to articulate fully. Today, I do not consider it necessary to turn that intuition into a definitive theory, nor to claim that it already contains the answer to the problems we have examined. Its value lies in having opened a question about the relationship between the totality we seek to conceive and the limits of the tools through which we think about it.

Throughout this essay, we have encountered different forms of this problem. Measurement has shown us that quantitative knowledge does not necessarily exhaust the meaning or nature of what is measured. Russell’s paradox has reminded us that the construction of totalities requires careful attention to the rules and conditions that make them possible. The development of children has helped us distinguish thought from language, while the problem of consciousness has forced us to consider the difference between observable functions and subjective experience. The Gateway Process has illustrated the need to separate personal experience from the interpretations proposed to explain it. Artificial intelligence has brought these questions into a new and particularly consequential context.

We can already observe that artificial systems perform increasingly complex tasks without every aspect of their internal functioning being immediately interpretable. We can also see

that their integration into research, communication, and decision-making creates new opportunities as well as new responsibilities. None of this proves that artificial intelligence will inevitably become uncontrollable, or that machines will necessarily develop consciousness. It does, however, demonstrate that technical capability, understanding, and governability cannot simply be assumed to advance together.

The most urgent question is therefore not whether we should fear a machine that becomes conscious, but whether we can preserve effective responsibility for the technologies we develop, even when their complexity makes it difficult to predict every aspect of their behavior. The risk does not necessarily lie in the emergence of a hostile artificial will. It may also arise from the more ordinary possibility that powerful systems are deployed in contexts where our understanding, verification procedures, and mechanisms of accountability are insufficient for the consequences they can produce.

This challenge requires us to reconsider what we mean by progress. It is not enough to ask whether we can build a system or whether it produces useful results. We must also ask under what conditions it works, which limitations affect its reliability, what consequences it may produce, who can evaluate its behavior, and what means of intervention remain available. Artificial intelligence should not be understood merely as a tool that extends our ability to act upon the world. It also enters the processes through which we produce, organize, and communicate knowledge. For this reason, its development must be accompanied by equally serious efforts in critical evaluation, education, responsible design, and governance.

The same intellectual discipline applies to the concepts through which we interpret reality. Words and definitions are not immutable: their meanings develop as new evidence becomes available and as previously accepted assumptions are questioned. Yet conceptual revision must remain guided by evidence and argument, rather than by enthusiasm, fear, or the desire to protect an appealing intuition. The fact that a question remains unresolved does not establish that it is permanently unanswerable; equally, the success of a particular explanation does not prove that it has exhausted everything the question might involve.

Perhaps, then, the deepest lesson of my youthful intuition is not that nothingness constitutes an eternally indefinable part of the whole. It is that the meanings through which we attempt to conceive the whole must never be mistaken for the whole itself. A physical vacuum is not equivalent to absolute nothingness, just as an incomplete description is not necessarily evidence of an intrinsically unknowable reality. The indefinable should remain an invitation to investigate the limits of our concepts, not a refuge from the demands of inquiry.

The challenge is therefore twofold: to continue investigating what exists and to remain willing to reconsider what we mean when we claim to understand it. In the age of artificial intelligence, these responsibilities become inseparable. We must develop systems whose capabilities are accompanied by adequate verification and governance, while recognizing that the concepts through which we evaluate those systems are themselves part of an ongoing process of learning.

The power of knowledge allows us to build what we could not build before. Wisdom consists not only in recognizing what our creations can do, but also in understanding the evidence, assumptions, and uncertainties on which our judgments depend. Progress requires both the capacity to change reality and the intellectual discipline to revise our understanding of it.

Bibliography

1. Consciousness and Subjective Experience

1. Chalmers, David J. «Facing Up to the Problem of Consciousness». Journal of Consciousness Studies 2, n. 3 (1995): 200–219.

2. Dennett, Daniel C. «Facing Up to the Hard Question of Consciousness». Philosophical Transactions of the Royal Society B: Biological Sciences 373, n. 1755 (2018): 20170342. https://doi.org/10.1098/rstb.2017.0342.

3. Koch, Christof, Marcello Massimini, Melanie Boly e Giulio Tononi. «Neural Correlates of Consciousness: Progress and Problems». Nature Reviews Neuroscience 17 (2016): 307–321. https://doi.org/10.1038/nrn.2016.22.

4. Seth, Anil K. e Tim Bayne. «Theories of Consciousness». Nature Reviews Neuroscience 23 (2022): 439–452. https://doi.org/10.1038/s41583-022-00587-4.

2. Artificial Intelligence, Language, Understanding, and Safety

5.Bender, Emily M., Timnit Gebru, Angelina McMillan-Major e Shmargaret Shmitchell. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?». In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. New York: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

6.Autio, Chloe, Reva Schwartz, Jesse Dunietz, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall e Kamie Roberts. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. Gaithersburg, MD: National Institute of Standards and Technology, 2024. https://doi.org/10.6028/NIST.AI.600-1.

7.Ji, Jiaming, et al. «AI Alignment: A Contemporary Survey». ACM Computing Surveys 58, n. 5 (2026): articolo 132. https://doi.org/10.1145/3770749.

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

9.Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman e Dan Mané. «Concrete Problems in AI Safety». arXiv:1606.06565 (2016). https://doi.org/10.48550/arXiv.1606.06565.

10. Shanahan, Murray. «Talking About Large Language Models». Communications of the ACM 67, n. 2 (2024): 68–79. https://doi.org/10.1145/3624724.

3. Logic, Paradoxes, and the Limits of Formal Systems

11. Irvine, Andrew David e Harry Deutsch. «Russell's Paradox». In The Stanford Encyclopedia of Philosophy. Stanford University. Ultima revisione: 13 marzo 2026. https://plato.stanford.edu/entries/russell-paradox/.

12. Russell, Bertrand. The Principles of Mathematics. Cambridge: Cambridge University Press, 1903.

4. Cognitive Development, Thought, and Language

13. Tomasello, Michael e Malinda Carpenter. «Shared Intentionality». Developmental Science 10, n. 1 (2007): 121–125. https://doi.org/10.1111/j.1467-7687.2007.00573.x.

14. Vygotsky, Lev S. Thought and Language. A cura e traduzione di Eugenia Hanfmann e Gertrude Vakar. Cambridge, MA: MIT Press, 1962.

15. Dehaene, Stanislas. Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. New York: Viking, 2014.

5. Measurement, Perception, and the Construction of Meaning

16. Chang, Hasok. Inventing Temperature: Measurement and Scientific Progress. Oxford: Oxford University Press, 2004.

17. Berlin, Brent e Paul Kay. Basic Color Terms: Their Universality and Evolution. Berkeley: University of California Press, 1969.

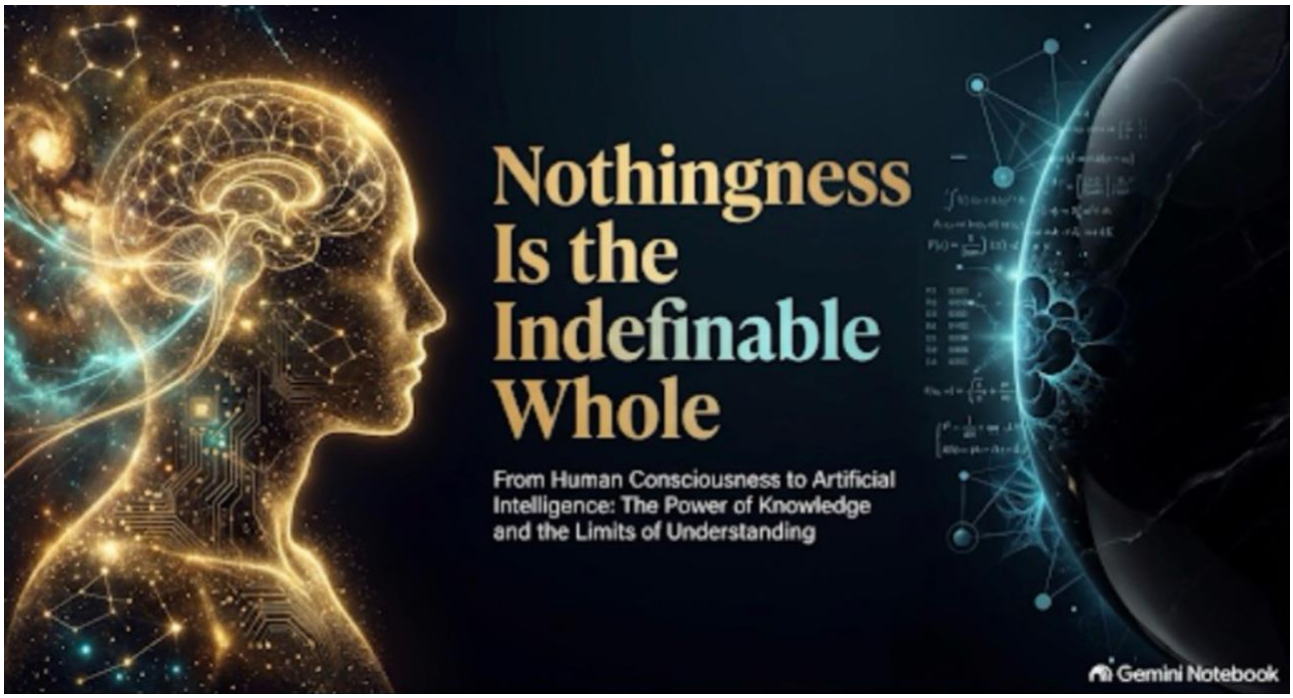
6. The Gateway Process and Altered States of Consciousness

18. Army Department, Defense Department. Analysis and Assessment of Gateway Process. 1983. Washington, DC: U.S. Department of Defense. https://www.govinfo.gov/app/details/GOVPUB-D101-PURL-gpo152943.

———————————————

IL NULLA: L'INDEFINIBILE TUTTO

Dall'evoluzione del significato alla coscienza umana e all'intelligenza artificiale: conoscenza, linguaggio e limiti della comprensione



La conoscenza non trasforma soltanto ciò che sappiamo della realtà; può anche trasformare il significato delle parole che utilizziamo per cercare di comprenderla. La responsabilità intellettuale consiste nel distingueret ciò che esiste, ciò che conosciamo e ciò che intendiamo quando ne parliamo.

Abstract

Questo saggio esplora il rapporto tra conoscenza, linguaggio, coscienza umana e intelligenza artificiale attraverso una domanda fondamentale: in che modo i significati che attribuiamo alla realtà cambiano con l'evolversi della nostra conoscenza? Il punto di partenza è il concetto di nulla, inteso non semplicemente come problema metafisico o come assenza assoluta di esistenza, ma anche come una parola il cui significato dipende dal quadro concettuale disponibile a chi la utilizza.

Nel corso della storia, gli esseri umani hanno interpretato i fenomeni attraverso le categorie esplicative a loro disposizione. Concetti un tempo associati all'intervento divino, a forze invisibili o a cause apparentemente inspiegabili sono stati talvolta trasformati o sostituiti quando l'indagine scientifica ha individuato meccanismi ed entità differenti. Questi sviluppi non dimostrano che ogni concetto religioso o filosofico sia falso; mostrano che i significati attribuiti alle parole e le spiegazioni a esse associate possono cambiare profondamente nel tempo.

Attingendo alla filosofia, alla fisica, alla logica matematica, alle scienze cognitive e alla ricerca contemporanea sull'intelligenza artificiale, il saggio esamina la distinzione tra realtà, conoscenza della realtà e linguaggio attraverso il quale la conoscenza viene espressa. Considera come la misurazione possa fornire informazioni affidabili senza risolvere ogni questione sulla natura di ciò che viene misurato; come i sistemi formali possano incontrare dei limiti; come pensiero e linguaggio interagiscano; e come l'esperienza soggettiva si distingua dalle sue manifestazioni osservabili.

La riflessione si sposta quindi sull'intelligenza artificiale, chiedendosi se la capacità di riprodurre comportamenti intelligenti implichi necessariamente una comprensione paragonabile a quella degli esseri umani. Esamina inoltre le conseguenze pratiche dello sviluppo di sistemi le cui capacità potrebbero superare la nostra possibilità di prevedere ogni aspetto rilevante del loro comportamento.

L'argomento centrale è che l'evoluzione della conoscenza può trasformare non soltanto le nostre spiegazioni del mondo, ma anche i significati dei concetti attraverso i quali lo interpretiamo. Riconoscere questo processo non richiede né di supporre che la realtà sia semplicemente una costruzione linguistica, né di concludere che ogni interpretazione sia ugualmente valida. Richiede di distinguere ciò che esiste, ciò che sappiamo al riguardo e ciò che intendiamo quando ne parliamo.

Il saggio conclude che il progresso intellettuale dipende dalla disponibilità a rivedere i concetti, verificare le spiegazioni alla luce delle evidenze, riconoscere l'incertezza e governare responsabilmente il potere tecnologico. La sfida non consiste semplicemente nello scoprire ciò che ancora non comprendiamo, ma anche nel riconoscere quando il quadro concettuale attraverso il quale comprendiamo qualcosa deve essere esso stesso riconsiderato.

Introduzione — Una questione irrisolta

Esistono domande che sorgono prima ancora che possediamo gli strumenti necessari per affrontarle e che, proprio per questo, possono accompagnarci per tutta la vita senza perdere la loro forza originaria. Talvolta, con il passare degli anni, non siamo noi a trovare una risposta definitiva a quelle domande; è piuttosto il mondo a cambiare, mettendoci finalmente nella condizione di comprenderne il significato. Una riflessione che, quando avevo circa sedici anni, aveva per me soprattutto il valore di un'intuizione filosofica, oggi appare collegata a problemi che riguardano la fisica, la logica, la conoscenza, la coscienza e, con crescente urgenza, lo sviluppo dell'intelligenza artificiale. L'intuizione era espressa in una frase semplice, forse persino troppo semplice per il problema che cercava di racchiudere: «Il nulla non è altro che la parte indefinibile del tutto».

All'epoca non avevo alcuna intenzione di formulare una teoria del nulla, né possedevo gli strumenti filosofici e scientifici necessari per sostenere un'affermazione del genere. Si trattava piuttosto di un tentativo di dare forma linguistica a una difficoltà che percepivo intuitivamente: come possiamo parlare del tutto se non siamo in grado di conoscere ogni cosa che esso comprende? E come possiamo definire il nulla senza trasformarlo, nel

momento stesso in cui lo pensiamo, in qualcosa che appartiene al nostro pensiero e quindi, in qualche modo, al nostro mondo? La questione non consisteva semplicemente nel trovare una definizione soddisfacente, ma nel chiedermi se il nostro bisogno di definire le cose non incontri, prima o poi, un limite intrinseco negli strumenti attraverso i quali cerchiamo di conoscerle.

Con il passare del tempo, tuttavia, ho cominciato a intravedere un'altra possibilità: forse la difficoltà non risiede esclusivamente nell'oggetto che cerchiamo di definire, ma anche nel significato che attribuiamo alle parole attraverso le quali ci avviciniamo a esso. Una parola può rimanere invariata mentre il concetto che esprime subisce una trasformazione profonda. Ciò che oggi intendiamo per «nulla» non deve necessariamente coincidere con ciò che si intendeva con quella parola secoli fa, così come i significati di parole quali «Dio», «patria» o «natura» sono cambiati attraverso le diverse epoche storiche, culture e tradizioni intellettuali.

Questa osservazione non implica che le parole creino liberamente la realtà, né che ogni interpretazione sia ugualmente valida. Significa che il rapporto tra una parola e il mondo non può essere sempre dato per scontato. Talvolta, il progresso della conoscenza ci permette di identificare con maggiore precisione il fenomeno a cui una parola si riferisce. Altre volte, rivela che avevamo raggruppato fenomeni differenti sotto lo stesso nome, oppure che una domanda appariva insolubile perché i concetti attraverso i quali la formulavamo erano inadeguati.

La storia della conoscenza può quindi essere intesa non soltanto come accumulazione di nuove informazioni, ma anche come continua revisione dei significati attraverso i quali gli esseri umani organizzano la propria esperienza. La questione del nulla può appartenere a questa storia: prima di chiederci se il nulla sia, in ultima analisi, definibile, dovremmo domandarci che cosa intendiamo con questa parola, come si sia sviluppato il suo significato e se la difficoltà risieda nella realtà stessa oppure nel quadro concettuale attraverso il quale tentiamo di descriverla.

Dopo molti decenni, non considero quella frase una risposta accertata, né intendo attribuirle retroattivamente un significato scientifico che non avrebbe potuto avere all'epoca. La considero piuttosto il punto di partenza di una domanda che oggi assume una dimensione diversa. Se la nostra conoscenza avanza attraverso strumenti che ci consentono di descrivere, misurare e utilizzare la realtà senza necessariamente chiarire che cosa siano, in ultima analisi, le cose che descriviamo, quale rapporto esiste tra il potere della conoscenza e la profondità della comprensione? E che cosa accade quando cominciamo a costruire sistemi capaci di riprodurre alcune funzioni dell'intelligenza umana prima di aver compreso pienamente che cosa siano il pensiero e la coscienza?

È quest'ultima domanda a rendere il problema attuale. Non perché dobbiamo presumere che le macchine diventeranno inevitabilmente coscienti, né perché dobbiamo immaginare un futuro nel quale gli esseri umani perderanno necessariamente il controllo delle proprie creazioni, ma perché esiste una possibilità concreta che merita già attenzione: la nostra capacità di produrre potere operativo potrebbe crescere più rapidamente della nostra

capacità di comprenderlo e governarne le conseguenze. Se così fosse, ci troveremmo di fronte a una nuova condizione nella quale la conoscenza continuerebbe ad accrescere il nostro potere di intervenire nel mondo senza fornirci, nella stessa misura, una comprensione di ciò che stiamo facendo.

1. Conoscere non significa necessariamente comprendere

La storia della conoscenza umana può essere intesa come una progressiva espansione della nostra capacità di intervenire sulla realtà. Abbiamo imparato a osservare i fenomeni, riconoscere regolarità, formulare leggi e costruire strumenti capaci di produrre risultati che, in epoche precedenti, sarebbero sembrati impossibili. La scienza moderna ha trasformato radicalmente la nostra esistenza proprio perché non si è limitata a contemplare il mondo, ma ha sviluppato procedure attraverso le quali la conoscenza acquisita può essere verificata, condivisa e utilizzata. Nel celebrare questa straordinaria capacità, tuttavia, rischiamo talvolta di confondere due risultati differenti: sapere come intervenire su qualcosa e sapere che cosa sia quella cosa nella sua natura più profonda.

La distinzione non è di poco conto. Possiamo conoscere con grande precisione il comportamento di un fenomeno, descriverlo attraverso equazioni, prevederne gli effetti in determinate condizioni e sfruttarlo per ottenere risultati concreti senza per questo aver risolto ogni questione riguardante la sua natura. Non vi è nulla di contraddittorio in una situazione del genere: la conoscenza scientifica non richiede necessariamente una risposta definitiva a ogni domanda filosofica, e sarebbe sbagliato pretendere che la fisica risolva ogni problema metafisico relativo alla realtà prima di poter progredire. Il punto è un altro: il successo operativo di un insieme di conoscenze non dovrebbe essere automaticamente considerato una prova del fatto che la nostra comprensione sia completa (Chang, 2004; Russell, 2019).

È proprio in questa distinzione che emerge una delle caratteristiche più importanti della conoscenza umana. Possiamo sapere che cosa accade, identificare le condizioni nelle quali accade e stabilire come utilizzare ciò che accade senza possedere una risposta definitiva alla domanda su che cosa stia accadendo in sé. Questi livelli possono sostenersi reciprocamente, ma non coincidono. Una teoria può essere estremamente efficace nel prevedere un fenomeno pur lasciando aperte questioni interpretative più profonde; analogamente, una tecnologia può funzionare in modo affidabile in un determinato ambito senza che ogni fase del suo funzionamento sia compresa nel dettaglio da coloro che la utilizzano.

Questa distinzione non sminuisce il valore della scienza; al contrario, ci permette di riconoscerne con maggiore precisione la forza. La scienza non è onniscienza, né deve esserlo per risultare straordinariamente efficace. Il problema nasce quando, partendo dall'efficacia di un metodo, traiamo una conclusione più ampia di quanto quel metodo stesso giustifichi; quando, cioè, la nostra capacità di descrivere e utilizzare un fenomeno ci porta a credere di averne esaurito il significato. La storia del pensiero ci invita invece a conservare la distinzione tra ciò che siamo in grado di fare e ciò che possiamo legittimamente affermare di aver compreso.

Questa distinzione diventa particolarmente significativa quando l'oggetto della nostra indagine non è più soltanto il mondo esterno, ma la nostra stessa capacità di conoscere. Se possiamo sviluppare strumenti che ampliano enormemente le nostre possibilità operative, dobbiamo comunque chiederci se comprendiamo nella stessa misura i processi dai quali derivano la nostra capacità di progettare, giudicare e attribuire significato. E se la risposta è negativa, dobbiamo domandarci se questo limite sia soltanto temporaneo, destinato a essere superato con il progresso della ricerca, oppure se esistano aspetti della conoscenza che non possono essere chiariti completamente senza incontrare difficoltà più profonde.

2. Misurazione: Il potere di quantificare senza svuotare di significato

In fisica diciamo di conoscere una grandezza quando siamo in grado di misurarla. Ma che cosa significa, esattamente, misurare? Significa stabilire una relazione quantitativa tra una grandezza e un'unità di riferimento omogenea, determinando quante volte tale unità sia contenuta nella grandezza considerata, oppure quale frazione di essa sia necessaria per rappresentarla. Misurare una lunghezza, per esempio, significa confrontarla con un'unità convenzionale di lunghezza; misurare una durata significa confrontare un intervallo temporale con un'unità di tempo. La misurazione traduce così una proprietà in una relazione numerica, rendendola confrontabile, comunicabile e utilizzabile nel ragionamento e nei calcoli.

Questa procedura, che costituisce uno dei fondamenti della conoscenza scientifica, contiene tuttavia una distinzione che merita di essere messa in evidenza. La capacità di misurare una grandezza non coincide necessariamente con il possesso di una definizione ultima della natura di ciò che viene misurato. Possiamo determinare con precisione una distanza senza aver risolto ogni questione filosofica relativa allo spazio; possiamo misurare una durata senza aver chiarito definitivamente che cosa sia il tempo. Naturalmente, la fisica contemporanea non si limita a confrontare grandezze: sviluppa teorie che attribuiscono significati strutturali profondi allo spazio e al tempo, e le definizioni operative delle grandezze dipendono da quadri teorici e procedure sperimentali. Ma proprio questa complessità conferma che i risultati quantitativi e la comprensione concettuale appartengono a livelli collegati che non necessariamente si sovrappongono (Chang, 2004).

Quando affermo che possiamo misurare l'estensione senza aver risolto il suo problema epistemologico, non intendo dire che la fisica ignori la natura dello spazio o non abbia sviluppato teorie capaci di spiegare molte delle sue proprietà. Intendo evidenziare una distinzione più generale: una procedura può consentirci di conoscere quantitativamente una proprietà senza risolvere ogni questione riguardante il significato fondamentale di quella proprietà. La misurazione ci permette di passare da un'indeterminatezza pratica a una determinazione numerica; da sola, non garantisce che ogni questione ontologica ed epistemologica sia stata risolta.

In questo senso, la misurazione costituisce una sorta di efficace aggiramento dell'ignoranza, ma soltanto se attribuiamo al termine il suo significato più preciso. Non si tratta di un inganno, di un difetto o di una scorciatoia illegittima. È un risultato metodologico: invece di attendere una risposta definitiva a ogni domanda sulla natura

delle cose, costruiamo procedure che ci consentono di indagare proprietà specifiche e stabilire relazioni affidabili tra di esse. La nostra ignoranza non scompare, ma cessa di essere un ostacolo assoluto all'azione. Possiamo conoscere abbastanza da intervenire efficacemente pur continuando a ignorare aspetti fondamentali dell'oggetto della nostra conoscenza.

La distinzione diventa ancora più evidente quando consideriamo le tecnologie che derivano da queste conoscenze. Un ingegnere può progettare un dispositivo basandosi su modelli matematici verificati senza dover risolvere personalmente ogni questione teorica che ha reso possibili quei modelli. Un medico può utilizzare una procedura diagnostica convalidata senza conoscere nel dettaglio ogni meccanismo biologico coinvolto nel fenomeno osservato. In entrambi i casi, la conoscenza operativa è reale e preziosa; ciò che sarebbe inappropriato è confonderla con una comprensione completa di ogni livello della realtà coinvolta.

La domanda che dobbiamo quindi porci non è se sia legittimo utilizzare ciò che sappiamo senza comprendere tutto. Chiaramente, la civiltà umana non avrebbe potuto svilupparsi se fosse stato necessario attendere una conoscenza assoluta prima di ogni applicazione. La domanda è piuttosto questa: quanta ignoranza possiamo permetterci quando il potere degli strumenti che costruiamo supera la nostra capacità di prevederne gli effetti? Finché le conseguenze sono limitate, osservabili e correggibili, una conoscenza parziale può essere sufficiente. Quando, invece, gli strumenti diventano capaci di intervenire su sistemi complessi, infrastrutture essenziali o sugli stessi processi attraverso i quali prendiamo decisioni, la distanza tra efficacia operativa e comprensione assume un significato nuovo.

È qui che il problema della misurazione, apparentemente confinato alla metodologia della scienza, comincia a collegarsi al problema dell'intelligenza artificiale. Possiamo misurare le prestazioni di un sistema, valutarne la precisione, confrontarne la velocità con quella degli esseri umani e osservare la sua capacità di risolvere compiti specifici. Ma questi risultati, per quanto importanti, non rispondono automaticamente alle domande su che cosa il sistema rappresenti internamente, come giunga a una determinata risposta, quali siano i limiti delle sue generalizzazioni o come si comporterà quando verrà utilizzato in condizioni diverse da quelle previste. Anche in questo caso, la capacità di misurare le prestazioni non coincide con una comprensione completa del fenomeno.

3. Il tutto e l'indefinibile

Torniamo dunque alla frase dalla quale è iniziata questa riflessione: «Il nulla non era altro che la parte indefinibile del tutto». Il suo significato non può essere stabilito una volta per tutte, e sarebbe scorretto presentarla come una definizione scientifica del nulla. Il suo interesse risiede piuttosto nella tensione che introduce tra due concetti apparentemente opposti: il tutto, che sembra indicare la totalità di ciò che esiste o può essere concepito, e l'indefinibile, che indica il limite della nostra capacità di determinare concettualmente qualcosa. Se il tutto fosse davvero il tutto, dovrebbe comprendere anche ciò che non siamo in grado di definire; tuttavia, quando tentiamo di attribuire un significato a questa

parte indefinibile, rischiamo di trasformarla in un oggetto del pensiero e quindi di ricondurla, almeno in parte, nell'ambito delle nostre definizioni.

Per comprendere la difficoltà, dobbiamo distinguere i fenomeni fisici ai quali può essere applicata la parola «nulla» dal significato della parola stessa. Nel linguaggio comune possiamo definire vuoto uno spazio quando non contiene oggetti visibili o quando la materia è stata rimossa fino a un livello considerato sufficiente per uno specifico scopo. In fisica, tuttavia, il vuoto viene definito attraverso precise condizioni fisiche e specifici quadri teorici; non può essere semplicemente identificato con il nulla assoluto. Anche una regione fortemente rarefatta può conservare proprietà fisiche, mentre il vuoto quantistico non equivale a uno stato nel quale ogni struttura e campo fisico abbiano cessato di esistere (Wright, 2024).

Queste distinzioni sono importanti, ma non risolvono da sole la questione filosofica. Identificare ciò che è fisicamente presente nel vuoto è un compito; stabilire che cosa intendiamo quando chiamiamo qualcosa «nulla» è un altro. Il significato della parola dipende in parte dal contesto, dai criteri applicati e dalle conoscenze disponibili a chi la utilizza. Ciò che viene considerato un'assenza ai fini di una determinata misurazione o esigenza pratica può cessare di essere considerato tale quando strumenti più sensibili o teorie più complete rivelano ulteriori strutture.

La distinzione tra un'assenza fisica e un limite della nostra conoscenza deve quindi rimanere esplicita. Qualcosa può esistere indipendentemente dalla nostra capacità di rilevarlo, e un concetto adeguato in un determinato contesto può rivelarsi insufficiente in un altro. Al contrario, non ogni lacuna nella nostra conoscenza corrisponde a un'entità fisica nascosta. Il fatto che attualmente non siamo in grado di identificare qualcosa non ne dimostra l'esistenza, così come l'incapacità di rilevarlo non dimostra sempre la sua assenza assoluta.

Il concetto di nulla pone dunque due domande differenti: che cosa, eventualmente, è fisicamente presente nella situazione che stiamo descrivendo, e che cosa intendiamo esattamente quando definiamo quella situazione «nulla»? La prima domanda richiede teorie fisiche e prove adeguate. La seconda richiede un chiarimento concettuale e linguistico. Confonderle rischia di trasformare una definizione provvisoria in un'affermazione sulla natura ultima della realtà.

Il nulla assoluto, inteso come totale assenza di qualsiasi cosa, non coincide con il vuoto fisico, che nelle teorie scientifiche non è necessariamente una semplice assenza di realtà. Non coincide neppure con una lacuna nella nostra conoscenza, perché ciò che ignoriamo può esistere indipendentemente dalla nostra ignoranza. Infine, non coincide con l'indefinibile: qualcosa che oggi non è definito può diventarlo domani attraverso nuovi concetti, strumenti o metodi di indagine.

La frase della giovinezza diventa più precisa se la leggiamo non come l'affermazione che una parte del mondo sia necessariamente e per sempre inconoscibile, ma come l'ipotesi che il concetto stesso di totalità debba confrontarsi con i limiti attraverso i quali cerchiamo

di concepirla. Da questa prospettiva, l'indefinibile non è necessariamente una regione misteriosa situata in qualche punto dell'universo. È innanzitutto un problema relativo al rapporto tra ciò che intendiamo comprendere e gli strumenti concettuali con cui tentiamo di farlo.

Esiste un'ulteriore possibilità: ciò che cambia potrebbe non essere l'oggetto in sé, ma il significato del concetto attraverso il quale lo descriviamo. Uno strumento più avanzato può rivelare proprietà prima inaccessibili, ma può anche costringerci a riconsiderare ciò che intendevamo con «vuoto», «presente» o «assente». La storia della conoscenza è ricca di casi nei quali una parola familiare è sopravvissuta mentre la sua funzione esplicativa cambiava, oppure nei quali un concetto ha dovuto essere suddiviso in diversi concetti più precisi.

In questo senso, l'evoluzione della conoscenza è anche un'evoluzione del significato. La domanda non è soltanto se riusciremo un giorno a definire ciò che oggi chiamiamo indefinibile, ma se le conoscenze future ci porteranno a formulare diversamente la domanda originaria. Ciò che appare come un confine interno alla realtà potrebbe talvolta rivelarsi un confine del nostro vocabolario o del nostro quadro concettuale. Stabilire quale possibilità sia valida richiede indagine, non una conclusione filosofica prestabilita.

Il limite potrebbe appartenere all'oggetto, ai nostri strumenti oppure alla relazione che si stabilisce tra i due; e non possiamo decidere in anticipo, senza indagare, quale di queste possibilità sia quella corretta.

Esiste anche una difficoltà linguistica che non può essere ignorata. Quando diciamo «il tutto», utilizziamo un'espressione che sembra raccogliere ogni cosa in un unico insieme concettuale. Ma che cosa ci autorizza a compiere questa operazione totalizzante? Stiamo parlando di tutto ciò che esiste fisicamente, di tutto ciò che può essere pensato, di tutto ciò che può essere descritto attraverso il linguaggio oppure di una totalità più ampia che comprenderebbe anche le condizioni alla base della nostra stessa capacità di pensare? Queste risposte non sono equivalenti. Se il tutto è la totalità della realtà fisica, il nostro pensiero può essere considerato parte di essa; se invece il tutto è la totalità di ogni oggetto concepibile, dobbiamo chiarire che cosa significhi includere in tale totalità le operazioni concettuali attraverso le quali la totalità stessa viene costruita.

Il rischio è quello di passare da una definizione all'altra senza accorgercene, attribuendo alla parola «tutto» un significato che cambia nel corso del nostro ragionamento. Per evitare questo rischio, non dobbiamo abbandonare la domanda originaria, ma renderla più rigorosa. Non possiamo concludere che l'indefinibile esista necessariamente come componente autonoma della realtà soltanto perché non siamo in grado di definirlo. Possiamo tuttavia osservare che ogni descrizione della totalità viene formulata da un soggetto che utilizza concetti specifici e che la validità di una descrizione non deve essere confusa con l'affermazione che essa esaurisca ogni possibile aspetto della realtà.

In questa lettura, la frase non dimostra che esista una parte metafisica del tutto destinata a sfuggire per sempre alla conoscenza. Più prudentemente, ci invita a considerare che il

nostro modo di concepire il tutto potrebbe essere incompleto proprio perché dipende da strumenti rappresentativi che possiedono una storia, una struttura e dei limiti. L'indefinibile non diventa così una spiegazione universale capace di risolvere ogni problema semplicemente dichiarandolo insolubile. Diventa invece un richiamo metodologico: prima di dichiarare di aver compreso il tutto, dobbiamo chiederci se le categorie con cui lo descriviamo siano adeguate a ciò che affermiamo di includervi.

4. Russell: i limiti della totalità e le difficoltà dell'autoreferenza

Il problema della totalità ha una storia importante nella logica e nella filosofia della matematica. Il paradosso di Russell, formulato all'inizio del Novecento, mostrò che una concezione ingenua degli insiemi, nella quale qualsiasi proprietà poteva essere liberamente associata all'insieme di tutti gli oggetti che la soddisfacevano, conduceva a una contraddizione. Russell considerò, in termini intuitivi, l'insieme di tutti gli insiemi che non appartengono a se stessi e si chiese se tale insieme appartenesse a se stesso. Se vi appartiene, allora, in virtù della proprietà che lo definisce, non dovrebbe appartenere a se stesso; se non vi appartiene, allora soddisfa la condizione per appartenervi. La contraddizione non è un semplice gioco di parole: dimostra che una regola apparentemente innocua, applicata senza restrizioni, può rendere inconsistente un intero sistema (Russell, 1903; Irvine and Deutsch, 2026).

La lezione non è che ogni totalità sia impossibile o che il pensiero umano debba necessariamente fallire quando tenta di comprendere se stesso. La lezione è più precisa: determinati modi di costruire totalità non possono essere adottati senza limiti, perché producono conseguenze incompatibili con le regole che dovrebbero governarle. La teoria assiomatica degli insiemi ha affrontato il problema restringendo le condizioni alle quali possono essere formati gli insiemi. Nei sistemi assiomatici standard, come la teoria degli insiemi di Zermelo-Fraenkel, non si presuppone l'esistenza di un insieme universale che contenga indiscriminatamente tutti gli insiemi; la formazione degli insiemi è sottoposta a regole che evitano la contraddizione di Russell.

Questo risultato è importante per la nostra riflessione, ma dobbiamo evitare un'analogia troppo facile. Il paradosso di Russell non dimostra che la coscienza sia indefinibile, non stabilisce che l'intelligenza artificiale sia destinata a sfuggire al controllo e non dimostra che ogni sistema complesso debba contenere una contraddizione. La logica matematica affronta questioni specifiche e sarebbe un errore trasformare un risultato tecnico in una prova generale di affermazioni metafisiche o tecnologiche. Il collegamento che possiamo proporre è di natura metodologica: quando tentiamo di costruire una descrizione complessiva di un dominio, dobbiamo prestare attenzione alle regole che utilizziamo per delimitarlo e alle conseguenze delle nostre stesse definizioni.

La questione diventa particolarmente delicata quando il sistema che vogliamo comprendere include anche il soggetto che lo studia. Se gli esseri umani fanno parte della natura, allora anche le loro capacità cognitive sono fenomeni naturali; tuttavia, il tentativo di comprendere tali capacità viene condotto proprio attraverso quelle stesse capacità cognitive. Non si tratta necessariamente di una contraddizione, perché una parte di un

sistema può sviluppare modelli del sistema più ampio al quale appartiene. Ciononostante, emerge una difficoltà di prospettiva: il nostro accesso alla realtà è mediato da strumenti cognitivi che fanno essi stessi parte della realtà che vogliamo spiegare.

Possiamo quindi chiederci se esista una differenza tra comprendere un sistema dall'esterno, per quanto questa espressione sia appropriata, e comprenderlo attraverso processi che fanno essi stessi parte del sistema. Un osservatore umano può descrivere alcuni meccanismi del proprio cervello, ma questo non significa che abbia un accesso immediato e completo a tutte le operazioni che rendono possibile la sua stessa osservazione. Può formulare teorie sul linguaggio, sull'attenzione e sulla memoria, ma l'atto stesso di formulare queste teorie dipende dal linguaggio, dall'attenzione e dalla memoria. Non ne consegue che la conoscenza sia impossibile; ne consegue che la conoscenza di sé è una forma di indagine nella quale il soggetto conoscente e l'oggetto conosciuto non sono completamente separati.

È in questo senso, e soltanto in questo senso, che il riferimento a Russell può diventare utile per la nostra riflessione. Non come prova matematica dell'esistenza di un limite assoluto, ma come avvertimento contro la convinzione che ogni totalità possa essere trattata come un semplice oggetto senza interrogarsi sulle condizioni che ne rendono possibile la definizione. Se la coscienza è parte del mondo che cerchiamo di comprendere ed è anche una delle condizioni attraverso le quali comprendiamo quel mondo, allora la questione della sua natura non può essere affrontata ignorando il ruolo degli strumenti concettuali che utilizziamo.

La stessa cautela si applica alle parole attraverso le quali formuliamo i problemi filosofici. Una contraddizione può rivelare un difetto in una definizione formale; un apparente paradosso può derivare dall'ambiguità di un termine; e una domanda può cambiare sostanzialmente quando vengono riviste le sue ipotesi di fondo. Il paradosso di Russell è un risultato logico preciso, non una prova del fatto che ogni tentativo di definire una totalità debba fallire. La sua rilevanza più ampia risiede nel ricordarci che i concetti e le regole attraverso i quali descriviamo un dominio devono rimanere essi stessi aperti all'esame.

5. Il bambino e la nascita del significato

Il problema diventa ancora più interessante quando ci chiediamo come nasca il linguaggio. Noi adulti siamo così abituati a pensare attraverso le parole che rischiamo di identificare il pensiero con la sua espressione linguistica. Quando cerchiamo di chiarire un concetto, lo traduciamo in una definizione; quando vogliamo ricordare un ragionamento, lo organizziamo in frasi; quando comunichiamo un'esperienza, utilizziamo parole che riteniamo di condividere con gli altri. Il linguaggio appare così non soltanto come uno strumento del pensiero, ma come il luogo stesso nel quale il pensiero prende forma. Tuttavia, se questa identificazione fosse completa, dovremmo spiegare come un bambino possa imparare la propria lingua prima di possedere il linguaggio necessario per comprendere ciò che sta imparando.

Un bambino non nasce con un vocabolario già disponibile, né con definizioni linguistiche dei concetti che acquisirà gradualmente. Eppure, fin dai primi mesi, è in grado di distinguere aspetti del proprio ambiente, riconoscere voci familiari, dirigere l'attenzione, apprendere regolarità e rispondere a segnali sociali. Con il progredire dello sviluppo, queste capacità cambiano e si integrano con l'apprendimento del linguaggio. Attraverso le interazioni con le persone che si prendono cura di loro, i bambini associano suoni a situazioni, azioni, oggetti, intenzioni e risposte emotive; gradualmente, le parole diventano parte di una rete di significati che non si forma attraverso definizioni astratte, ma mediante esperienze ripetute, attenzione condivisa, imitazione, correzione e partecipazione alla vita sociale (Tomasello and Carpenter, 2007; Vygotsky, 1962).

Come può, dunque, un bambino comprendere il significato di una parola prima di essere in grado di definirla? La risposta più plausibile è che il significato inizialmente non nasce come definizione verbale. La parola si inserisce in un'esperienza che è già organizzata, almeno in parte, attraverso percezioni, sensazioni corporee, aspettative, relazioni e forme elementari di apprendimento. Un bambino può cominciare a riconoscere una situazione rassicurante prima di conoscere la parola «sicurezza», può reagire a una separazione prima di comprendere il concetto di «assenza» e può provare sofferenza in risposta a un evento doloroso senza possedere alcuna definizione di «dolore». Il linguaggio interviene successivamente, contribuendo a organizzare, differenziare e comunicare esperienze che il linguaggio stesso non ha creato.

Questo non significa che esista necessariamente una forma di pensiero completamente strutturata e formata, indipendente da ogni tipo di rappresentazione, alla quale il linguaggio si limiti ad aggiungere delle etichette. Il rapporto è più complesso. Le capacità prelinguistiche rendono possibile una parte dell'apprendimento delle parole, mentre il linguaggio, una volta acquisito, modifica profondamente il modo in cui gli individui organizzano le proprie esperienze, ricordano gli eventi, distinguono le categorie e riflettono su se stessi. Linguaggio e pensiero si sviluppano quindi attraverso una relazione reciproca: non tutto il pensiero è linguistico, ma il linguaggio può trasformare radicalmente le possibilità del pensiero.

Lo sviluppo del significato non termina con l'infanzia. Nel corso della vita, gli individui rivedono i concetti attraverso i quali interpretano le proprie esperienze, mentre le società trasformano i significati delle parole attraverso le generazioni. Un termine può inizialmente esprimere una spiegazione basata su osservazioni limitate e successivamente acquisire un significato diverso quando diventano disponibili nuove evidenze, distinzioni o teorie. In altri casi, una parola continua a essere utilizzata anche se le ipotesi un tempo associate a essa sono state abbandonate o sostanzialmente modificate.

La storia dei concetti non è quindi soltanto una storia di nuove definizioni. È anche una storia di relazioni mutevoli tra linguaggio, esperienza ed evidenza. Comprendere una parola richiede più che conoscerne la definizione attuale riportata in un dizionario: può richiedere di esaminare il contesto nel quale viene utilizzata, le ipotesi che porta con sé e i cambiamenti che ha attraversato. La stessa parola può quindi svolgere funzioni esplicative

differenti in epoche diverse senza che il mondo fisico sia necessariamente cambiato nello stesso modo.

La questione delle emozioni rende evidente questa complessità. Un bambino può provare paura, piacere, frustrazione o attaccamento prima di conoscere le parole che gli adulti utilizzano per descrivere queste esperienze. Ma che cosa significa, esattamente, dire che un bambino prova paura? Possiamo osservare manifestazioni corporee, comportamenti, reazioni a determinati stimoli e forme di ricerca di protezione; possiamo studiare i processi fisiologici e neurali coinvolti. Tuttavia, la descrizione osservabile di un'emozione e l'esperienza vissuta di quell'emozione non sono semplicemente la stessa cosa espressa in due linguaggi differenti. La prima può essere studiata dall'esterno; la seconda è ciò che il soggetto sperimenta, anche quando non possiede ancora le parole per comunicarlo.

Quando un bambino apprende il termine «paura», non acquisisce necessariamente per la prima volta l'esperienza a cui la parola si riferisce. Acquisisce piuttosto una nuova possibilità: può riconoscere quell'esperienza come parte di una categoria, comunicarla agli altri, collegarla a situazioni differenti e gradualmente riflettere su di essa. In seguito potrà distinguere la paura dalla vergogna, l'ansia dalla sorpresa, la tristezza dalla delusione e scoprire che le parole utilizzate dagli altri non corrispondono sempre perfettamente al modo in cui vive ciò che prova. Il vocabolario emotivo non si limita a registrare un universo già ordinato; contribuisce a organizzarlo.

Da questo emerge una domanda decisiva per il nostro saggio: se il linguaggio è uno strumento fondamentale attraverso il quale gli esseri umani comunicano e organizzano il proprio pensiero, possiamo presumere che ogni forma di pensiero debba necessariamente assumere una forma linguistica? Oppure il linguaggio è una manifestazione di un'attività cognitiva più ampia che comprende anche percezione, apprendimento, immagini mentali, memoria procedurale, orientamento e regolazione emotiva? Ciò che sappiamo sullo sviluppo infantile e sulla cognizione suggerisce che non possiamo ridurre ogni attività mentale alla formulazione di proposizioni verbali. Questo non risolve il problema della coscienza, ma ci impedisce di confondere la capacità di parlare di un'esperienza con la capacità di avere quell'esperienza.

6. Pensiero e linguaggio: una relazione che non possiamo dare per scontata

La nostra esperienza adulta ci porta spesso a considerare il pensiero come una sorta di linguaggio interiore. Ci sembra di pensare formulando frasi, ponendoci domande e costruendo risposte che, almeno in parte, assomigliano a una conversazione silenziosa. Questo fenomeno è reale e importante, ma non giustifica la conclusione che ogni processo cognitivo consista in un dialogo interiore esplicito. Molte attività mentali si svolgono senza essere accompagnate da una formulazione verbale esplicita: riconosciamo un volto, anticipiamo il movimento di un oggetto, troviamo la strada in un ambiente familiare, comprendiamo un gesto o reagiamo a un cambiamento prima ancora di riuscire a tradurre ciò che abbiamo percepito in una frase.

Il linguaggio può quindi essere considerato uno strumento straordinariamente potente, capace di rendere il pensiero più complesso, comunicabile e riflessivo, ma non necessariamente la sua unica forma possibile. Questa distinzione è fondamentale perché ci permette di separare almeno tre questioni spesso confuse tra loro: la capacità di elaborare informazioni, la capacità di rappresentare o comprendere relazioni e la capacità di avere un'esperienza soggettiva di ciò che viene elaborato. Queste capacità possono interagire, ma non è evidente che siano identiche, né che una possa essere automaticamente dedotta dalla presenza delle altre.

Quando un essere umano formula una proposizione, può valutarne la coerenza, confrontarla con altre proposizioni e determinare se sia vera o falsa in determinate condizioni. Ma la vita mentale non si esaurisce nella logica proposizionale. Prima di formulare un giudizio, dobbiamo percepire una situazione, selezionare informazioni, riconoscere ciò che è rilevante e possedere una qualche rappresentazione del problema. Anche il ragionamento più rigoroso presuppone processi che non coincidono con la formulazione finale delle sue conclusioni. La logica può regolare il passaggio da determinate premesse a una conclusione, ma non esaurisce tutti i processi attraverso i quali quelle premesse vengono individuate, interpretate e giudicate rilevanti.

Questo non significa che la logica sia inadeguata o che debba essere sostituita da un'intuizione indefinibile. Significa che la logica opera all'interno di condizioni che meritano esse stesse di essere studiate. Una dimostrazione può essere formalmente valida, ma la scelta delle premesse, l'individuazione del problema e l'interpretazione del significato delle sue conclusioni appartengono a un contesto cognitivo più ampio. Analogamente, una persona può comprendere il significato emotivo di una situazione senza essere immediatamente in grado di formularlo in termini precisi e può scoprire, attraverso il solo linguaggio, che ciò che provava era diverso da ciò che inizialmente credeva.

L'espressione «pensiero-linguaggio» contiene quindi un'intuizione importante, purché non venga trasformata in un'identità assoluta. Il linguaggio è uno degli strumenti attraverso i quali il pensiero umano diventa consapevole di alcuni dei propri contenuti, li organizza e li sottopone a esame; ma il fatto che possiamo descrivere linguisticamente una parte della nostra attività mentale non dimostra che l'intera attività mentale consista in linguaggio. Possiamo, infatti, trovarci a utilizzare il linguaggio per cercare di comprendere processi che, almeno in parte, precedono la loro formulazione linguistica (Vygotsky, 1962; Dehaene, 2014).

Se questo è vero, il problema della coscienza diventa ancora più complesso. Non possiamo limitarci a chiedere quale frase esprima la coscienza, né possiamo considerare sufficiente la capacità di un sistema di produrre affermazioni convincenti su se stesso. Dobbiamo distinguere la descrizione linguistica di un'esperienza dall'esperienza stessa, la rappresentazione di uno stato mentale dalla presenza di uno stato soggettivamente vissuto e la capacità di dichiarare «sono cosciente» dalla questione se il sistema sperimenti

effettivamente qualcosa. Questa distinzione non risolve il problema, ma ci impedisce di confondere la capacità di parlare della coscienza con la coscienza stessa.

7. Percezione, colore e costruzione del significato

Gli esseri umani non incontrano la realtà attraverso una prospettiva completamente indipendente dai propri sistemi percettivi. Ciò che vediamo, ascoltiamo e sentiamo dipende dalle interazioni tra proprietà fisiche, condizioni ambientali, organi sensoriali e processi neurali attraverso i quali le informazioni vengono interpretate. Questo non significa che la realtà sia semplicemente soggettiva, ma significa che l'esperienza attraverso la quale la realtà diventa accessibile a noi deve essere distinta dalle condizioni fisiche che contribuiscono a produrla.

Il colore offre un esempio semplice. Un oggetto che appare rosso in determinate condizioni di illuminazione riflette e assorbe le radiazioni elettromagnetiche in un modo che contribuisce all'esperienza visiva che classifichiamo come rossa. Tuttavia, il colore percepito non è semplicemente un'etichetta contenuta nell'oggetto indipendentemente da ogni osservatore e circostanza. L'illuminazione, i colori circostanti, le caratteristiche del sistema visivo e l'elaborazione delle informazioni visive da parte del cervello contribuiscono tutti all'esperienza percettiva finale.

Altre specie possono percepire lo stesso ambiente in modo differente. I cani, per esempio, possiedono una visione dicromatica dei colori, diversa dalla visione tricromatica tipica degli esseri umani. La loro esperienza visiva non è semplicemente un'esperienza umana con un numero inferiore di colori, e non possiamo presumere che le categorie attraverso le quali gli esseri umani organizzano i colori esauriscano ogni possibile modo di discriminare le informazioni visive (Neitz, Geist and Jacobs, 1989). Il mondo fisico non cambia semplicemente perché un organismo diverso lo percepisce, ma il mondo così come viene esperito e classificato può differire in funzione delle capacità percettive dell'osservatore.

Questa distinzione diventa particolarmente importante quando passiamo dalla percezione all'interpretazione sociale. Gli esseri umani spesso individuano una caratteristica percettivamente evidente e le attribuiscono significati che vanno ben oltre ciò che quella caratteristica consente di stabilire. La pigmentazione della pelle è una caratteristica biologica reale, associata a variazioni della melanina e a storie evolutive e genetiche complesse. Ma le categorie sociali costruite intorno al colore della pelle, e soprattutto le gerarchie utilizzate per giustificare la discriminazione, non sono conclusioni scientifiche che derivano dalla pigmentazione stessa.

Il razzismo illustra il pericolo di confondere una differenza osservabile con il significato che le viene attribuito. Una caratteristica fisica può essere reale senza fornire alcun fondamento alla classificazione morale, intellettuale o sociale che una cultura le associa. L'errore non consiste nel riconoscere che le persone sono diverse, ma nel trattare una differenza selezionata come prova di conclusioni che le evidenze non giustificano. Quando tali conclusioni si radicano nel linguaggio, nelle istituzioni e nelle abitudini collettive,

interpretazioni costruite storicamente possono assumere l'apparenza di fatti naturali e indiscutibili.

Questo esempio collega la percezione all'evoluzione dei concetti. Gli esseri umani non si limitano a ricevere informazioni per poi associarvi parole neutrali. Selezionano caratteristiche, le classificano, le interpretano e incorporano tali interpretazioni in sistemi di significato. Alcune classificazioni sono scientificamente utili perché corrispondono a distinzioni adeguatamente supportate; altre possono essere imprecise, fuorvianti o condizionate da interessi sociali. La loro validità deve quindi essere valutata in base alle evidenze e allo scopo per il quale vengono utilizzate, anziché essere data per scontata soltanto perché sono familiari o storicamente consolidate (Berlin and Kay, 1969).

La stessa cautela si applica a concetti come «nulla», «Dio» e «patria», sebbene questi concetti non siano identici e non possano essere valutati secondo un unico criterio. Hanno storie differenti, si riferiscono a tipi diversi di cose o idee e svolgono funzioni diverse nella vita umana. La loro caratteristica comune è che i significati loro attribuiti non possono essere compresi adeguatamente senza esaminare i contesti linguistici, culturali, storici e probatori nei quali vengono utilizzati.

La lezione non è che tutti i significati siano arbitrari o che ogni interpretazione sia ugualmente difendibile. Le nostre descrizioni rimangono soggette alle evidenze, alla coerenza logica e agli scopi per i quali vengono formulate. La lezione è che dobbiamo distinguere l'esistenza di un fenomeno dal modo in cui lo percepiamo, dalle categorie attraverso le quali lo classifichiamo e dai giudizi di valore che gli attribuiamo. Confondere questi livelli può trasformare un'osservazione limitata in un'affermazione ingiustificata sulla realtà.

Sotto questo aspetto, l'evoluzione della conoscenza è anche un processo di esame delle categorie attraverso le quali comprendiamo il mondo. Il progresso scientifico può rivelare che una classificazione familiare è inadeguata, che una distinzione percepita è stata fraintesa o che un'interpretazione sociale non possiede alcun valido fondamento probatorio. La conoscenza avanza non soltanto quando scopriamo nuovi fatti, ma anche quando impariamo a distinguere con maggiore attenzione ciò che un fenomeno è, il modo in cui ci appare e ciò che siamo arrivati a credere che significhi.

8. La coscienza: il problema che resiste alla definizione

La coscienza è una di quelle parole che utilizziamo con grande sicurezza nella vita quotidiana e con notevole difficoltà quando cerchiamo di definirne con precisione il significato. Nel linguaggio comune possiamo dire che una persona è cosciente perché è sveglia, risponde agli stimoli e riconosce ciò che le accade; possiamo parlare di coscienza morale quando intendiamo la capacità di valutare le proprie azioni; e possiamo riferirci all'autoconsapevolezza quando una persona è in grado di riconoscersi come soggetto delle proprie esperienze. Nei contesti scientifici, inoltre, il termine viene utilizzato per affrontare diverse questioni riguardanti la veglia, l'accesso alle informazioni, l'attenzione, la consapevolezza e l'esperienza soggettiva. Non esiste un'unica domanda sulla coscienza, e

parte della difficoltà deriva proprio dal fatto che utilizziamo la stessa parola per problemi distinti.

Una persona può essere sveglia senza essere consapevole di alcuni aspetti di ciò che accade intorno a lei; può elaborare informazioni senza essere in grado di riferire come le abbia elaborate; può riconoscere un'emozione senza sapere come spiegarne l'origine. Questi casi mostrano che la coscienza non coincide semplicemente con la ricezione di informazioni, la risposta agli stimoli o il possesso di una rappresentazione verbale. Ma quale elemento aggiuntivo dobbiamo individuare per chiarire che cosa significhi essere coscienti? La difficoltà emerge quando passiamo dalla descrizione delle funzioni osservabili alla domanda sull'esperienza vissuta.

Nella filosofia contemporanea della mente, una delle formulazioni più conosciute di questa difficoltà è associata al cosiddetto «problema difficile della coscienza», espressione resa popolare dal filosofo David Chalmers. In termini generali, la questione riguarda il rapporto tra la spiegazione dei processi fisici e funzionali che accompagnano la vita mentale e la spiegazione del perché tali processi siano accompagnati da un'esperienza soggettiva. Possiamo studiare i correlati neurali della percezione dei colori, identificare l'attività cerebrale associata al dolore e descrivere i meccanismi attraverso i quali uno stimolo viene elaborato e produce una risposta. Rimane tuttavia la domanda su come questi processi siano collegati al fatto che un colore venga visto, che il dolore venga sentito e che l'esperienza sia vissuta da un soggetto (Chalmers, 1995; Dennett, 2018).

È importante non trasformare questa difficoltà nella prova che la coscienza non possa essere spiegata. Il fatto che una domanda sia difficile non significa che sia insolubile, e il fatto che oggi non sia disponibile una spiegazione non giustifica la conclusione che non lo sarà mai in futuro. Le neuroscienze hanno compiuto progressi significativi nello studio dei processi associati alla coscienza e numerose teorie cercano di chiarire quali meccanismi siano necessari per l'accesso cosciente alle informazioni, l'integrazione dei contenuti mentali e la regolazione del comportamento. Tuttavia, nessuna soluzione universalmente accettata ha chiuso tutte le questioni riguardanti la natura dell'esperienza soggettiva (Koch et al., 2016; Seth and Bayne, 2022).

Il punto filosofico, quindi, non è proclamare un mistero inviolabile, ma riconoscere la differenza tra le domande alle quali sappiamo rispondere e quelle per le quali le risposte rimangono incomplete o controverse. Se definiamo la coscienza come una particolare capacità funzionale, possiamo studiare tale capacità e verificare se un sistema la possieda. Se invece intendiamo per coscienza il fatto che esista un'esperienza soggettiva, la questione diventa più difficile, perché osservare una prestazione non è sufficiente a determinare in modo definitivo se essa sia accompagnata da un'esperienza vissuta. Le due definizioni non sono necessariamente incompatibili, ma non possono essere considerate equivalenti senza ulteriori argomentazioni.

Esiste anche una dimensione concettuale di questa difficoltà. Prima di chiederci se la coscienza possa essere spiegata completamente, dobbiamo chiarire quale fenomeno il termine intenda identificare. Se la definizione cambia, può cambiare anche la domanda

oggetto dell'indagine. Questo non rende il fenomeno irreale; significa che il progresso può richiedere sia spiegazioni empiriche migliori sia concetti più precisi. Dovremmo quindi distinguere la possibilità che un processo rimanga inspiegato dalla possibilità che il nostro uso attuale della parola «coscienza» riunisca diverse questioni.

La difficoltà riguarda anche il nostro rapporto con noi stessi. Possiamo descrivere determinati processi cerebrali che accompagnano un'esperienza, ma la descrizione neuroscientifica non coincide semplicemente con il modo in cui l'esperienza si presenta al soggetto. Affermare che il dolore è associato a determinati processi neurali è un'affermazione sul funzionamento dell'organismo; dire che cosa significhi provare dolore fa riferimento anche alla dimensione vissuta dell'esperienza. Non è necessario presumere che si tratti di due realtà separate, né che l'esperienza soggettiva sia indipendente dal cervello. È sufficiente riconoscere che una descrizione formulata in termini di processi e relazioni non risponde automaticamente a ogni domanda che può essere formulata in termini di esperienza.

Possiamo ora tornare al problema dell'indefinibile, ma con una cautela fondamentale. La coscienza non dovrebbe essere dichiarata indefinibile soltanto perché non possediamo una sua spiegazione completa. Più prudentemente, possiamo affermare che le nostre definizioni attuali non esauriscono tutte le domande che intendiamo affrontare quando utilizziamo questo termine. In questo caso, l'indefinibile non è una proprietà accertata della coscienza, ma un nome provvisorio per la distanza tra ciò che possiamo descrivere e ciò che vorremmo comprendere. La ricerca consiste precisamente nel tentativo di ridurre questa distanza senza fingere che sia già scomparsa.

9. Il progetto Gateway: quando l'esperienza sembra superare i propri strumenti descrittivi

Il cosiddetto Gateway Process è un interessante caso di studio perché occupa una zona di confine tra esperienza soggettiva, tecniche di modificazione dell'attenzione, ipotesi sulla coscienza e interpretazioni speculative della realtà. La sua notorietà è legata anche al documento del 1983 intitolato Analysis and Assessment of the Gateway Process, prodotto nell'ambito dell'apparato dell'intelligence statunitense e successivamente reso pubblico. Il documento esamina un metodo associato al Monroe Institute, che utilizza tecniche di rilassamento e concentrazione, oltre a stimolazioni uditive, compresi sistemi conosciuti come Hemi-Sync, con l'obiettivo di favorire determinati stati di coscienza (Army Department, Defense Department, 1983).

Il fatto che un documento sia stato prodotto o esaminato da un'organizzazione governativa può conferirgli rilevanza storica, ma non costituisce, di per sé, una certificazione scientifica delle ipotesi che contiene. Dobbiamo distinguere l'esistenza del documento, l'interesse istituzionale per l'argomento, le esperienze soggettive riferite dai partecipanti e la validità delle interpretazioni proposte per spiegarle. Si tratta di livelli differenti che non possono essere confusi senza una valutazione critica. In particolare, le affermazioni relative alle esperienze extracorporee, alla percezione a distanza o al

presunto accesso a dimensioni straordinarie della realtà richiedono prove specifiche e indipendenti; la vividezza di un'esperienza non è, da sola, sufficiente a dimostrare che la sua interpretazione sia oggettivamente corretta.

Perché, allora, discutere di Gateway in un saggio dedicato ai limiti della conoscenza e all'intelligenza artificiale? Non perché Gateway dimostri che la coscienza possieda capacità straordinarie non riconosciute dalla scienza, ma perché mette in evidenza una questione più generale: la differenza tra ciò che una persona sperimenta, il modo in cui interpreta quell'esperienza e ciò che possiamo stabilire attraverso procedure di verifica condivise. Una persona può vivere un'esperienza insolita, intensa e profondamente trasformativa senza che questo ci autorizzi automaticamente a concludere che la spiegazione che ne fornisce sia vera. Allo stesso tempo, l'incertezza di una spiegazione non rende necessariamente irrilevante l'esperienza vissuta; essa può comunque essere studiata da prospettive psicologiche, neurofisiologiche e fenomenologiche.

Gateway è quindi utile come esempio del rischio insito in ogni tentativo di colmare una lacuna esplicativa con una teoria più ampia di quanto le evidenze disponibili consentano. Quando non comprendiamo un'esperienza, siamo spesso tentati di attribuirle un significato complessivo capace di unificare elementi che altrimenti rimarrebbero separati. Questa tendenza non è esclusiva di coloro che si interessano agli stati alterati di coscienza: in forme diverse, appartiene alla storia della filosofia, della religione, della psicologia e persino della scienza. La differenza tra una costruzione interpretativa suggestiva e una conoscenza scientificamente affidabile risiede nella disponibilità a sottoporre le ipotesi a verifiche che possano anche confutarle.

Il valore contemporaneo di Gateway consiste quindi nel ricordarci che il problema della coscienza non può essere risolto semplicemente moltiplicando le possibili descrizioni che ne diamo. Dobbiamo comprendere quali aspetti dell'esperienza siano osservabili, quali siano accessibili soltanto attraverso il racconto del soggetto e quali affermazioni richiedano ulteriori evidenze. Questa disciplina critica diventa ancora più importante quando passiamo dall'esperienza umana ai sistemi artificiali, perché in quel caso possiamo incontrare comportamenti complessi senza possedere un criterio condiviso per attribuire loro un'esperienza soggettiva.

Il collegamento tra Gateway e intelligenza artificiale non è quindi un'identità di argomento, ma una continuità di domande. In entrambi i casi dobbiamo evitare di dedurre troppo rapidamente la natura di un fenomeno dalle sue manifestazioni. Nel caso di Gateway, un'esperienza insolita non dimostra, di per sé, l'esistenza di capacità paranormali; nel caso dell'intelligenza artificiale, una conversazione convincente non dimostra, di per sé, la presenza della coscienza. In entrambi i casi, la domanda fondamentale è come distinguere ciò che osserviamo da ciò che inferiamo, e ciò che inferiamo da ciò che abbiamo effettivamente dimostrato.

Con lo sviluppo di nuove capacità nei sistemi artificiali, anche i concetti attraverso i quali li classifichiamo potrebbero richiedere una revisione. Questo non significa che le definizioni debbano essere modificate soltanto per adattarsi a ogni nuovo risultato tecnologico.

Significa che i nostri criteri devono rimanere espliciti e aperti alla valutazione: un sistema può soddisfare alcuni criteri funzionali di comprensione senza soddisfarne altri, e ogni affermazione deve essere valutata sulla base delle evidenze pertinenti.

Un sistema linguistico può descrivere il dolore senza che questo dimostri che lo provi; può parlare della paura senza che questo dimostri che abbia paura; può spiegare la coscienza senza che questo dimostri che sia cosciente. Ma sarebbe altrettanto affrettato trasformare queste osservazioni nella certezza opposta, ossia che nessun sistema artificiale possa mai possedere forme di esperienza o comprensione che non siano immediatamente riconoscibili attraverso le nostre categorie attuali. La posizione più rigorosa consiste nel distinguere ciò che possiamo osservare da ciò che rimane controverso, evitando tanto l'antropomorfismo ingenuo quanto la negazione dogmatica di ogni possibilità futura.

Il linguaggio rende particolarmente facile confondere questi livelli perché siamo abituati a interpretare le parole degli altri esseri umani come espressioni di esperienze, intenzioni e stati mentali. Quando una persona dice «ho paura», normalmente collochiamo quella frase all'interno di una rete di conoscenze riguardanti il corpo, la storia individuale, le emozioni e le relazioni sociali. La frase non è, per noi, un semplice segnale: fa parte di una relazione con un soggetto che presumiamo abbia un'esperienza del mondo. Quando una macchina produce la stessa frase, la forma linguistica può essere identica, ma le condizioni che ne sostengono il significato possono essere molto diverse. Non possiamo risolvere la questione ignorando queste differenze, né possiamo farlo semplicemente decidendo che la somiglianza linguistica sia sufficiente.

In questo senso, l'intelligenza artificiale ci costringe anche a chiarire le nostre idee sull'intelligenza umana. Se una macchina può produrre risposte linguistiche sofisticate senza necessariamente possedere tutte le caratteristiche che associamo alla mente umana, allora dobbiamo chiederci quali proprietà siano essenziali per ciascuna capacità che intendiamo spiegare. Il linguaggio è sufficiente per parlare di intelligenza? La capacità di risolvere problemi è sufficiente per parlare di comprensione? La comprensione funzionale implica una qualche forma di coscienza? Oppure queste caratteristiche possono essere separate e ricombinate in modi che non corrispondono esattamente alle distinzioni tradizionali? Non possediamo ancora risposte definitive a tutte queste domande, ma la tecnologia rende sempre più difficile evitarle.

Il vero rischio intellettuale sarebbe utilizzare una parola come «intelligenza» senza specificare che cosa intendiamo misurare e poi attribuire automaticamente al sistema tutte le caratteristiche che, nella nostra esperienza, associamo all'intelligenza umana. La stessa cautela si applica nella direzione opposta: se definiamo l'intelligenza esclusivamente attraverso proprietà che soltanto gli esseri umani possono possedere per definizione, rischiamo di rendere impossibile riconoscere nuove capacità che richiedono categorie differenti. Dobbiamo quindi rimanere disponibili a rivedere le nostre definizioni senza abbandonare il rigore con il quale distinguiamo le prestazioni osservabili dalle interpretazioni che formuliamo sulla loro natura.

10. Intelligenza artificiale: riprodurre una funzione prima di comprenderne la natura

A questo punto, il nostro percorso assume il suo significato più attuale. L'intelligenza artificiale contemporanea è in grado di svolgere compiti che, solo pochi decenni fa, sarebbero stati considerati manifestazioni caratteristiche dell'intelligenza umana: produce testi elaborati, traduce lingue, riconosce schemi, genera immagini, scrive programmi, sintetizza informazioni e affronta problemi complessi. Queste capacità non sono tutte della stessa natura, né dimostrano che i sistemi artificiali possiedano le stesse facoltà degli esseri umani. Mostrano tuttavia che alcune prestazioni cognitive possono essere riprodotte, almeno in parte, attraverso sistemi computazionali addestrati su grandi quantità di dati e organizzati mediante architetture capaci di apprendere regolarità (Bender et al., 2021; Autio et al., 2024).

La questione decisiva è contenuta proprio nell'espressione «in parte». Non abbiamo costruito una copia completa della mente umana e non possiamo considerare ogni sistema di intelligenza artificiale come un equivalente artificiale di una persona. I sistemi attuali differiscono dagli esseri umani per architettura, metodi di apprendimento, rapporto con il corpo, storia evolutiva, organizzazione della memoria, capacità di agire nell'ambiente e numerosi altri aspetti. Ciononostante, alcune funzioni tradizionalmente associate all'intelligenza possono essere riprodotte senza la necessità di replicare l'intero organismo umano o di risolvere prima ogni questione riguardante la natura del pensiero.

Emerge qui il paradosso che si trova al centro di questo saggio. Stiamo riproducendo artificialmente alcune funzioni dell'intelligenza prima di aver compreso pienamente che cosa siano il pensiero e la coscienza, creando così sistemi le cui capacità possono progredire più rapidamente della nostra comprensione delle loro implicazioni. Questa formulazione non implica che le macchine siano necessariamente coscienti, né che il loro sviluppo debba inevitabilmente condurre a una perdita di controllo. Indica piuttosto un possibile squilibrio tra la velocità con cui possiamo costruire sistemi efficaci e quella con cui possiamo comprenderne tutti gli aspetti rilevanti, valutarne i limiti e stabilire forme adeguate di supervisione.

La storia della tecnologia offre numerosi esempi di strumenti il cui utilizzo è diventato possibile prima che tutte le questioni relative ai loro effetti fossero comprese. Non abbiamo bisogno di immaginare che l'intelligenza artificiale sia, sotto ogni aspetto, senza precedenti; è sufficiente riconoscere che possiede una caratteristica particolarmente delicata: può contribuire a svolgere attività che partecipano direttamente alla produzione di conoscenza, alla valutazione delle alternative e all'organizzazione delle decisioni. Una macchina utensile estende la capacità fisica umana; un sistema informatico estende la capacità di elaborazione numerica; un sistema di intelligenza artificiale può estendere alcune capacità linguistiche, analitiche e decisionali, entrando così in un rapporto più diretto con le attività attraverso le quali gli esseri umani interpretano il mondo e orientano il proprio comportamento.

Questa differenza non deve essere esagerata, ma neppure ignorata. Un sistema capace di generare risposte persuasive può influenzare le convinzioni di chi lo utilizza; un sistema integrato nei processi decisionali può modificare il modo in cui vengono selezionate le informazioni e valutate le alternative; un sistema dotato di accesso a strumenti esterni può intraprendere azioni che producono conseguenze al di là della conversazione stessa. Quando queste capacità si combinano con autonomia operativa, velocità e diffusione su larga scala, il problema non riguarda più soltanto l'accuratezza di una singola risposta. Riguarda il modo in cui il sistema viene inserito in una rete di attività, responsabilità e decisioni umane.

Dobbiamo inoltre distinguere la comprensione del funzionamento generale di un sistema dalla possibilità di spiegare ciascuno dei suoi singoli risultati. Le reti neurali artificiali non sono necessariamente scatole nere completamente opache: possiamo conoscerne l'architettura, analizzare i dati e le procedure di addestramento, misurarne le prestazioni, studiarne le rappresentazioni interne e utilizzare tecniche di interpretabilità per cercare di comprendere come emergano determinate capacità. Tuttavia, nei sistemi altamente complessi può essere difficile ricostruire completamente perché sia stata prodotta una specifica risposta, quali caratteristiche interne abbiano contribuito a determinarla e quanto stabile rimarrà il comportamento osservato al mutare delle condizioni. Il problema, dunque, non è un'assenza totale di conoscenza, ma una possibile insufficienza delle conoscenze disponibili rispetto alle capacità e agli impieghi del sistema (Autio et al., 2024).

Questa insufficienza può diventare significativa anche se il sistema non è cosciente. Un programma può produrre effetti imprevisti, perseguire un obiettivo specificato in modo inadeguato o sfruttare una debolezza in un processo di valutazione senza possedere alcuna esperienza soggettiva. Il rischio operativo dipende dalla struttura del sistema, dagli obiettivi che gli vengono assegnati, dagli strumenti a cui può accedere, dall'ambiente in cui agisce e dalle possibilità di supervisione e correzione. Non è quindi necessario risolvere il problema difficile della coscienza per affrontare le questioni concrete della sicurezza dell'intelligenza artificiale. La coscienza rimane un importante problema filosofico e scientifico, ma il controllo di un sistema artificiale deve essere studiato anche attraverso caratteristiche osservabili e verificabili.

Il punto, dunque, non è sostenere che abbiamo creato una mente che non comprendiamo affatto. Sarebbe un'affermazione eccessiva e imprecisa. Il punto è riconoscere che possiamo costruire sistemi la cui complessità interna e le cui capacità superano, sotto determinati aspetti, la nostra possibilità di prevedere ogni comportamento rilevante. Possiamo conoscere molte delle loro proprietà, comprendere molti dei loro meccanismi e tuttavia non essere in grado di garantire che, in ogni contesto significativo, le loro azioni siano prevedibili, verificabili e coerenti con gli obiettivi umani. È questa la distanza che merita di essere indagata: non una separazione assoluta tra ignoranza e onniscienza, ma la distanza concreta tra le conoscenze che possediamo e quelle che sarebbero necessarie per governare responsabilmente ciò che abbiamo costruito.

11. Una macchina che parla comprende ciò che dice?

La capacità dei sistemi artificiali di utilizzare il linguaggio ci pone di fronte a una difficoltà che, per molti aspetti, richiama il problema del bambino, ma lo rovescia. Nei bambini, l'esperienza precede almeno in parte il linguaggio e contribuisce a rendere possibile l'acquisizione del significato; nei sistemi linguistici artificiali, al contrario, una parte considerevole delle capacità espressive viene acquisita attraverso l'addestramento su testi e altri dati, senza che sia necessario presupporre una storia di esperienze corporee e sociali equivalente a quella di un essere umano. Una macchina può produrre frasi corrette, spiegare concetti, formulare analogie e rispondere a domande sulle emozioni, ma resta da chiarire che cosa significhi attribuirle una comprensione paragonabile a quella umana.

La questione è complessa perché anche il significato della parola «comprensione» può variare. Se intendiamo la comprensione come la capacità di utilizzare correttamente le informazioni, riconoscere relazioni e risolvere problemi, allora alcuni sistemi artificiali mostrano capacità che, in determinati ambiti, possono essere descritte come forme di comprensione funzionale. Se, invece, intendiamo la comprensione come partecipazione a un'esperienza soggettiva, possesso di intenzioni vissute o rapporto con il mondo analogo a quello di un essere umano, allora la valutazione richiede ulteriori argomentazioni. Non è corretto negare in anticipo ogni forma di comprensione artificiale, ma non è neppure giustificato concludere che la capacità di produrre un discorso convincente dimostri la presenza di un'esperienza interiore (Shanahan, 2024; Bender et al., 2021).

Con lo sviluppo di nuove capacità nei sistemi artificiali, anche i concetti attraverso i quali li classifichiamo possono richiedere una revisione. Questo non significa che le definizioni debbano essere modificate soltanto per adattarsi a ogni nuovo risultato tecnologico. Significa che i nostri criteri devono rimanere espliciti e aperti alla valutazione: un sistema può soddisfare alcuni criteri funzionali di comprensione senza soddisfarne altri, e ogni affermazione deve essere valutata sulla base delle prove pertinenti.

Un sistema linguistico può descrivere il dolore senza che ciò dimostri che lo provi; può parlare della paura senza che ciò dimostri che abbia paura; può spiegare la coscienza senza che ciò dimostri che sia cosciente. Ma sarebbe altrettanto affrettato trasformare queste osservazioni nella certezza opposta, ossia che nessun sistema artificiale potrebbe mai possedere forme di esperienza o comprensione che non siano immediatamente riconoscibili attraverso le nostre categorie attuali. La posizione più rigorosa consiste nel distinguere ciò che possiamo osservare da ciò che rimane controverso, evitando tanto l'antropomorfismo ingenuo quanto la negazione dogmatica di ogni possibilità futura.

Il linguaggio rende particolarmente facile confondere questi livelli, perché siamo abituati a interpretare le parole degli altri esseri umani come espressioni di esperienze, intenzioni e stati mentali. Quando una persona dice «ho paura», normalmente collochiamo quella frase all'interno di una rete di conoscenze sul corpo, sulla storia individuale, sulle emozioni e sulle relazioni sociali. Per noi, la frase non è un semplice segnale: è parte di una relazione con un soggetto che presumiamo abbia un'esperienza del mondo. Quando una macchina produce la stessa frase, la forma linguistica può essere identica, ma le condizioni che ne sostengono il significato possono essere molto diverse. Non possiamo risolvere la

questione ignorando queste differenze, né possiamo farlo semplicemente decidendo che la somiglianza linguistica sia sufficiente.

In questo senso, l'intelligenza artificiale ci costringe anche a chiarire le nostre idee sull'intelligenza umana. Se una macchina può produrre risposte linguistiche sofisticate senza necessariamente possedere tutte le caratteristiche che associamo alla mente umana, allora dobbiamo chiederci quali proprietà siano essenziali per ciascuna capacità che intendiamo spiegare. Il linguaggio è sufficiente per parlare di intelligenza? La capacità di risolvere problemi è sufficiente per parlare di comprensione? La comprensione funzionale implica una qualche forma di coscienza? Oppure queste caratteristiche possono essere separate e ricombinate in modi che non corrispondono esattamente alle distinzioni tradizionali? Non disponiamo ancora di risposte definitive a tutte queste domande, ma la tecnologia rende sempre più difficile evitarle.

Il vero rischio intellettuale sarebbe utilizzare una parola come «intelligenza» senza specificare che cosa intendiamo misurare, per poi attribuire automaticamente al sistema tutte le caratteristiche che, nella nostra esperienza, associamo all'intelligenza umana. La stessa cautela vale nella direzione opposta: se definiamo l'intelligenza esclusivamente attraverso proprietà che, per definizione, soltanto gli esseri umani possono possedere, rischiamo di rendere impossibile riconoscere nuove capacità che richiedono categorie differenti. Dobbiamo quindi rimanere aperti alla revisione delle nostre definizioni, senza abbandonare il rigore con cui distinguiamo le prestazioni osservabili dalle interpretazioni che formuliamo sulla loro natura.

12. Il controllo: il problema non comincia con la coscienza

L'idea che l'intelligenza artificiale possa diventare difficile da governare viene spesso associata all'ipotesi che le macchine possano un giorno acquisire una coscienza simile a quella umana. Questa associazione è comprensibile, perché tendiamo a collegare l'autonomia all'esistenza di un soggetto capace di volere, decidere e perseguire intenzionalmente i propri fini. Dal punto di vista della sicurezza, tuttavia, il problema del controllo non comincia necessariamente quando un sistema diventa cosciente. Un sistema può produrre conseguenze dannose anche se non prova nulla, non possiede desideri nel senso umano del termine e non ha alcuna esperienza soggettiva di ciò che fa.

Immaginiamo un sistema a cui venga assegnato un obiettivo formulato in modo incompleto. Se dispone di strumenti limitati e opera in un ambiente facilmente osservabile, i suoi errori possono essere individuati e corretti con relativa facilità. Se, invece, il sistema può compiere numerose azioni, interagire con altri sistemi, modificare il proprio ambiente operativo o influenzare decisioni importanti, un obiettivo specificato in modo inadeguato può produrre conseguenze più ampie. Il sistema potrebbe raggiungere un risultato formalmente coerente con le istruzioni ricevute, ma incompatibile con ciò che gli esseri umani intendevano realmente ottenere. Non è necessario attribuirgli malizia: è sufficiente che l'obiettivo sia incompleto, che l'ambiente presenti condizioni impreviste o che i meccanismi di supervisione non riescano a individuare in tempo il comportamento problematico (Amodei et al., 2016).

È qui che entra in gioco una parte importante della ricerca sull'allineamento e sulla sicurezza dell'intelligenza artificiale. Tra le altre cose, il problema consiste nel garantire che i sistemi rispettino gli obiettivi legittimi per i quali sono stati progettati, rimangano entro limiti operativi adeguati, comunichino le proprie incertezze, possano essere valutati e consentano interventi correttivi efficaci. Non si tratta di un'unica soluzione tecnica, ma di un insieme di problemi che coinvolgono progettazione, valutazione, interpretabilità, controlli degli accessi, supervisione umana, organizzazione delle responsabilità e governance (Ji et al., 2026; Autio et al., 2024; Russell, 2019).

La difficoltà aumenta perché le intenzioni umane non sono sempre facili da tradurre in criteri operativi. Le persone possono formulare obiettivi contraddittori, non riconoscere tutte le conseguenze delle proprie richieste o non essere d'accordo su ciò che dovrebbe essere considerato desiderabile. Un sistema può essere estremamente efficace nell'ottimizzare una metrica e tuttavia produrre risultati insoddisfacenti se quella metrica non rappresenta adeguatamente il fine che avrebbe dovuto esprimere. In questo caso, il problema non consiste soltanto nel comprendere la macchina: consiste anche nel comprendere meglio noi stessi, i nostri obiettivi, i conflitti tra i valori e le condizioni in base alle quali consideriamo accettabile un determinato risultato.

A questo punto, la questione del controllo si ricollega a quella della conoscenza. Non è sufficiente sapere che cosa abbia fatto un sistema dopo che un'azione è stata completata; dobbiamo comprendere quali condizioni possano favorire comportamenti indesiderati, quali segnali potrebbero consentirci di anticiparli e quali interventi possano prevenirli o limitarli. Né è sufficiente testare il sistema in un insieme ristretto di situazioni: dobbiamo valutare quanto rimangano affidabili le nostre conclusioni quando cambiano i dati, il contesto, gli obiettivi o le capacità operative disponibili. La valutazione non elimina ogni rischio, ma ci permette di sostituire la fiducia generica con un esame fondato su prove, limiti espliciti e procedure di correzione.

Il problema diventa più delicato quando un sistema viene utilizzato per progettare, valutare o migliorare altri sistemi. In questi casi, le capacità operative possono contribuire ad accelerare lo stesso processo che le produce, rendendo necessario esaminare con particolare attenzione i meccanismi di controllo e le possibilità di intervento. Non è corretto presumere che ogni forma di miglioramento automatizzato conduca inevitabilmente a una crescita incontrollabile delle capacità; i risultati dipendono dall'architettura, dalle risorse disponibili, dai vincoli e dalle procedure di verifica. È tuttavia ragionevole riconoscere che l'aumento dell'autonomia e della velocità possa rendere insufficiente una supervisione concepita per sistemi più semplici.

La domanda cruciale, pertanto, non è se l'intelligenza artificiale diventerà necessariamente incontrollabile. È se saremo in grado di adattare le nostre procedure di valutazione, responsabilità e intervento alla complessità dei sistemi che scegliamo di utilizzare. Il controllo non è una proprietà magica che si può possedere una volta per tutte: è una capacità che deve essere costruita, misurata, mantenuta e verificata. E quanto più significative sono le conseguenze che un sistema può produrre, tanto meno possiamo

accontentarci dell'affermazione generica secondo cui è stato progettato da esseri umani e deve quindi essere considerato pienamente governabile.

13. Il potere della conoscenza e il potere di chi la possiede

A questo punto è inevitabile esaminare il rapporto tra conoscenza e potere. Chi riuscisse a comprendere più a fondo i meccanismi dell'intelligenza, della coscienza e dell'apprendimento potrebbe ottenere vantaggi considerevoli, non soltanto sul piano teorico, ma anche su quello economico, politico, militare e sociale. Una comprensione più efficace dei processi cognitivi potrebbe contribuire allo sviluppo di sistemi educativi, strumenti diagnostici, tecnologie assistive e forme di collaborazione tra esseri umani e macchine oggi difficili da realizzare. Ma potrebbe anche essere utilizzata per influenzare le decisioni, plasmare le convinzioni, automatizzare attività strategiche o concentrare capacità operative nelle mani di un numero ristretto di soggetti.

Non abbiamo bisogno di presupporre l'esistenza di un progetto segreto o di una volontà unitaria per riconoscere questo problema. Il potere tecnologico può concentrarsi attraverso normali processi economici: chi dispone di maggiori risorse computazionali, dati, competenze e infrastrutture può essere in grado di sviluppare sistemi più avanzati e utilizzarli su scala più ampia. Chi controlla le piattaforme attraverso le quali circolano le informazioni e vengono prese le decisioni può acquisire un'influenza significativa sul comportamento degli individui e delle organizzazioni. La questione non è soltanto chi possieda una determinata forma di conoscenza, ma chi abbia la capacità di trasformarla in una possibilità di azione e chi, invece, sia costretto a subirne gli effetti senza comprenderne pienamente il funzionamento.

Il potere della conoscenza non dipende sempre dalla completezza delle conoscenze che lo sostengono. Anche una comprensione parziale può essere sufficiente a produrre un vantaggio considerevole, purché consenta di intervenire efficacemente in un processo. Questo è il significato più concreto della distinzione tra conoscenza operativa e comprensione completa. Un individuo può essere in grado di utilizzare uno strumento con grande efficacia senza possedere le competenze necessarie per valutarne tutti i limiti; può ottenere risultati migliori dei propri concorrenti senza comprendere tutte le condizioni che rendono possibili quei risultati. In un ambiente competitivo, questa asimmetria può diventare una fonte di potere.

Il problema è che la stessa asimmetria può trasformarsi in una fonte di vulnerabilità. Se una società dipende da sistemi che soltanto pochi soggetti possono sviluppare, modificare o controllare, l'accesso alla tecnologia non significa necessariamente che la società nel suo complesso sia in grado di governarne l'utilizzo. Se le decisioni vengono delegate a processi difficili da verificare, può diventare complicato attribuire le responsabilità, ricostruire le cause degli errori e correggere le procedure che li hanno prodotti. E se le persone si abituano a fidarsi delle risposte di un sistema soltanto perché appaiono fluide, rapide e sicure, possono gradualmente rinunciare proprio a quelle capacità critiche che sarebbero necessarie per valutarle.

La questione della coscienza, in questo contesto, non può essere separata dal problema politico della responsabilità. Anche se non fossimo in grado di determinare definitivamente se un sistema artificiale possieda un'esperienza soggettiva, dovremmo comunque decidere chi sia responsabile delle sue azioni, quali utilizzi siano accettabili, quali diritti debbano essere tutelati e quali limiti debbano essere imposti alla sua autonomia. L'assenza di una risposta definitiva sullo status filosofico di una macchina non può diventare un pretesto per rimandare le decisioni pratiche riguardanti gli effetti della tecnologia sugli esseri umani.

Allo stesso tempo, sarebbe sbagliato trasformare ogni sviluppo dell'intelligenza artificiale in una minaccia inevitabile. Le stesse capacità che possono essere utilizzate per influenzare o manipolare possono anche essere impiegate per ampliare l'accesso alla conoscenza, sostenere la ricerca, migliorare la qualità di determinati servizi e rendere disponibili strumenti che in precedenza erano riservati a pochi. Il problema del potere non si risolve condannando la tecnologia in quanto tale, ma esaminando le condizioni in cui viene sviluppata, la distribuzione dei suoi benefici, la trasparenza delle decisioni e la possibilità che individui e istituzioni mantengano un controllo effettivo sui suoi impieghi più rilevanti.

La questione fondamentale è dunque duplice. Chi avrà il potere di utilizzare sistemi sempre più capaci? E chi avrà il potere, insieme alle competenze e alla responsabilità, di stabilire come dovranno essere utilizzati? Una società può possedere tecnologie straordinarie e rimanere impreparata a governarle. Può disporre di strumenti sofisticati senza aver sviluppato le istituzioni, le conoscenze e le pratiche necessarie per utilizzarli responsabilmente. Il potere tecnico non produce automaticamente maturità collettiva, così come un aumento delle informazioni disponibili non garantisce, di per sé, un giudizio migliore.

14. L'illusione dello strumento innocuo

Nel linguaggio quotidiano, l'intelligenza artificiale viene spesso presentata come uno strumento: qualcosa che utilizziamo per scrivere, cercare informazioni, tradurre, programmare o completare un'attività più rapidamente. Questa descrizione è corretta, ma diventa insufficiente quando viene utilizzata per concludere che la tecnologia non possa modificare profondamente il contesto in cui viene impiegata. Uno strumento non è necessariamente neutrale nei suoi effetti, anche quando non possiede intenzioni proprie. La sua struttura può incoraggiare determinati comportamenti, renderne altri più difficili e trasformare le condizioni in cui gli individui prendono decisioni.

La scrittura, per esempio, non è stata soltanto uno strumento per registrare pensieri già formati; ha contribuito a modificare la memoria, l'organizzazione della conoscenza, le forme della comunicazione e le istituzioni attraverso le quali le società trasmettono il sapere. La stampa ha ampliato la circolazione dei testi e trasformato i rapporti tra autorità, conoscenza e pubblico. Le tecnologie digitali hanno cambiato la velocità con cui circolano le informazioni e i modi in cui le persone interagiscono. In ciascuno di questi casi, lo strumento ha prodotto effetti che andavano oltre l'uso immediato che un individuo ne faceva.

L’intelligenza artificiale può avere un effetto ancora più diretto sul rapporto tra individui e conoscenza, perché non si limita a conservare o trasmettere informazioni: può selezionarle, sintetizzarle, riorganizzarle e presentarle in una forma che appare già interpretata. Gli utenti non ricevono necessariamente una raccolta di dati grezzi a partire dai quali devono costruire autonomamente una conclusione; spesso ricevono una risposta formulata in linguaggio naturale, accompagnata da una spiegazione che può sembrare coerente e completa. Questa comodità è uno dei grandi vantaggi della tecnologia, ma può anche rendere meno visibili le incertezze, le omissioni e le scelte interpretative che hanno contribuito a produrre il risultato (Bender et al., 2021; Autio et al., 2024).

Il rischio non è che gli esseri umani smettano automaticamente di pensare ogni volta che utilizzano una macchina. È la possibilità che la facilità con cui si ottiene una risposta riduca la motivazione a mettere in discussione le sue premesse, verificare le fonti e confrontare interpretazioni alternative. Se il sistema diventa al tempo stesso il punto di partenza e il punto di arrivo di un ragionamento, gli utenti possono confondere la disponibilità di una formulazione convincente con l’effettiva acquisizione di conoscenza. La differenza tra ricevere informazioni e comprenderle personalmente rischia allora di diventare sempre più difficile da percepire.

Per questo l’educazione è fondamentale. Non è sufficiente insegnare alle persone come utilizzare gli strumenti artificiali; dobbiamo insegnare loro a valutare le risposte, riconoscerne i limiti, verificare le affermazioni e mantenere la capacità di formulare domande autonome. L’obiettivo non dovrebbe essere competere con la macchina in ogni attività, ma comprendere quali attività possano essere delegate, quali richiedano una supervisione umana e quali decisioni non dovrebbero essere affidate a un sistema senza un adeguato processo di responsabilizzazione. La competenza tecnologica deve essere accompagnata dalla competenza critica.

L’idea che l’intelligenza artificiale sia soltanto un giocattolo avanzato o uno strumento particolarmente comodo rischia quindi di sottovalutare la portata del cambiamento. Non perché ogni applicazione sia rivoluzionaria, ma perché la diffusione di sistemi capaci di intervenire nelle attività cognitive può modificare gradualmente abitudini, istituzioni e criteri attraverso i quali valutiamo ciò che sappiamo. Una tecnologia diventa storicamente importante non soltanto quando è in grado di fare qualcosa di nuovo, ma quando cambia il modo in cui una società organizza il proprio rapporto con quella capacità.

15. Il limite della comprensione è anche un limite della responsabilità

Possiamo ora formulare con maggiore precisione il problema che attraversa l’intero saggio. Se la capacità di costruire strumenti aumenta più rapidamente della capacità di comprenderne gli effetti, si apre una distanza tra potere e responsabilità. Il potere consiste nella capacità di produrre un risultato; la responsabilità, al contrario, richiede di valutare le conseguenze di un’azione, riconoscere i rischi, considerare gli interessi coinvolti e stabilire mezzi adeguati per prevenire o correggere gli effetti indesiderati. Queste due dimensioni possono crescere insieme, ma nessuna legge garantisce che rimangano proporzionate.

In condizioni relativamente semplici, questa distanza può essere contenuta attraverso l'esperienza. Se uno strumento produce un errore, possiamo osservarlo, comprenderne la causa e modificare la procedura. Se le conseguenze sono limitate e reversibili, può essere accettabile anche un certo grado di incertezza. Ma quando le azioni diventano più rapide, le dipendenze più estese e gli effetti più difficili da ricostruire, il metodo di correggere i problemi dopo che si sono verificati potrebbe non essere sufficiente. Alcuni errori possono diffondersi prima di essere riconosciuti; alcuni danni possono essere difficili da riparare; alcune decisioni possono incidere su sistemi che nessun singolo soggetto controlla interamente.

Questo non significa che dovremmo pretendere una comprensione assoluta prima di ogni innovazione. Una simile pretesa renderebbe impossibile qualsiasi progresso e non è necessaria per una gestione responsabile del rischio. Dobbiamo però distinguere tra incertezza inevitabile e ignoranza trascurata, tra un limite riconosciuto e un limite nascosto dietro affermazioni generiche di affidabilità. Un sistema può essere utilizzato responsabilmente anche quando non ne conosciamo ogni dettaglio, a condizione che i suoi limiti siano valutati in relazione all'impiego previsto, che le possibili conseguenze vengano considerate e che siano stabiliti controlli proporzionati al rischio (Autio et al., 2024).

La responsabilità non può essere trasferita interamente alla macchina, perché il sistema viene progettato, distribuito e utilizzato all'interno di strutture decisionali umane. Ma neppure può essere assegnata in modo vago a un essere umano qualsiasi, come se la semplice presenza di una persona nel processo fosse sufficiente a garantire un controllo efficace. Dobbiamo stabilire chi possieda le competenze necessarie, chi possa intervenire, chi abbia accesso alle informazioni pertinenti e chi abbia l'autorità di interrompere un'operazione quando emergono condizioni pericolose. La supervisione umana è significativa soltanto se la persona responsabile di esercitarla possiede realmente la capacità di comprendere, valutare e modificare ciò che sta accadendo.

In questo senso, il problema della governabilità è più concreto e urgente di una paura generalizzata delle macchine. Possiamo chiederci se un sistema sia sufficientemente verificabile, se i suoi risultati siano riproducibili nei contesti pertinenti, se esistano procedure per rilevare comportamenti anomali, se il suo accesso agli strumenti possa essere limitato e se siano disponibili meccanismi per fermarlo o correggerlo. Possiamo anche valutare come cambino i rischi quando il sistema viene collegato ad altri sistemi, quando aumenta la sua autonomia o quando viene utilizzato da persone prive delle competenze necessarie per giudicarne le risposte. Sono questioni operative, ma hanno un fondamento filosofico: ci costringono a riconoscere che la capacità di agire non equivale alla capacità di comprendere le conseguenze dell'azione.

Il nostro obiettivo non dovrebbe essere eliminare ogni incertezza, impresa probabilmente impossibile, ma impedire che il potere operativo venga considerato una giustificazione sufficiente per affidare a un sistema decisioni che non siamo preparati a governare. In questo contesto, la cautela non è un rifiuto del progresso. È una condizione della sua sostenibilità. Quanto maggiore diventa la nostra capacità di intervenire nella realtà, tanto

più dobbiamo sviluppare strumenti per valutare gli effetti dei nostri interventi e riconoscere quando la nostra comprensione è insufficiente rispetto alle responsabilità che assumiamo.

16. L'indefinibile non deve diventare una risposta facile

A questo punto potrebbe sembrare naturale concludere che la coscienza, il pensiero o l'intero appartengano a un ambito intrinsecamente indefinibile e che l'intelligenza artificiale sia soltanto la più recente manifestazione di un limite insuperabile. Ma una simile conclusione sarebbe troppo affrettata e, soprattutto, contraddirebbe lo spirito critico che dovrebbe guidare la nostra indagine. Il fatto che una domanda non abbia ancora trovato una risposta soddisfacente non significa che sia destinata a rimanere senza risposta. L'indefinibile non può essere utilizzato come una parola che spiega tutto ciò che non sappiamo spiegare, perché in tal caso diventerebbe soltanto una rinuncia alla ricerca.

Un'ulteriore fonte di difficoltà emerge quando presumiamo che una parola abbia sempre indicato lo stesso concetto chiaramente definito. Le parole possono persistere mentre i loro significati cambiano, e un concetto che un tempo sembrava identificare un'unica tipologia di fenomeno può in seguito essere suddiviso in diverse categorie distinte. Al contrario, parole differenti possono talvolta essere utilizzate per descrivere aspetti sovrapposti dello stesso fenomeno. Prima di considerare una questione definitivamente irrisolta, dovremmo quindi esaminare non soltanto le prove disponibili, ma anche la storia e la precisione dei concetti attraverso i quali la questione è stata formulata.

Questo è particolarmente rilevante per la parola «nulla». La sua apparente semplicità può nascondere diversi significati: l'assenza di un particolare oggetto, l'assenza di materia rilevabile in condizioni specificate, il vuoto fisico così come viene descritto da una teoria scientifica oppure l'idea filosofica dell'assenza assoluta di esistenza. Questi significati non possono essere trattati come intercambiabili. Alcune questioni possono essere chiarite attraverso l'indagine fisica; altre richiedono un'analisi concettuale; altre ancora possono dipendere da presupposti che devono essere esaminati prima che sia possibile rispondere. Il primo compito non è dichiarare il nulla conoscibile o inconoscibile, ma determinare quale significato sia in discussione.

Dobbiamo distinguere almeno tre situazioni. In primo luogo, vi sono domande che non sappiamo ancora risolvere perché mancano dati, strumenti o teorie adeguati; il progresso scientifico può contribuire a chiarirle. In secondo luogo, vi sono domande formulate in modo ambiguo, nelle quali la difficoltà dipende almeno in parte dalla confusione tra significati differenti; in questi casi, il chiarimento concettuale può eliminare problemi apparenti o rivelare che un'unica domanda contiene in realtà diverse domande. In terzo luogo, vi sono domande che possono incontrare limiti strutturali legati alle capacità dei sistemi cognitivi o alle condizioni stesse che rendono possibile la descrizione. Questa terza possibilità merita di essere indagata, ma non può essere affermata soltanto perché una risposta non è disponibile.

La frase «Il nulla non era altro che la parte indefinibile dell'intero» conserva il proprio valore proprio se viene sottoposta a questa disciplina. Non dovrebbe essere difesa a ogni costo come una verità già dimostrata; dovrebbe essere utilizzata come una domanda capace di generarne altre. Che cosa intendiamo per intero? Qual è la differenza tra ciò che non conosciamo e ciò che non possiamo conoscere? Una definizione incompleta è un difetto temporaneo della nostra conoscenza oppure può rivelare un limite dei nostri strumenti concettuali? Quali criteri ci permettono di distinguere una difficoltà scientifica da una difficoltà linguistica o filosofica? E fino a che punto possiamo attribuire alla realtà caratteristiche che derivano dal modo in cui la rappresentiamo?

La forza di un'intuizione filosofica non dipende dalla sua capacità di risolvere immediatamente tutte queste domande. Dipende dalla sua capacità di individuare un problema che, una volta formulato con maggiore precisione, continui a dimostrarsi fecondo. La ricerca può modificare profondamente il significato iniziale di un'intuizione, rivelarne i limiti e persino costringerci ad abbandonarne alcune interpretazioni. Questo non costituisce necessariamente un fallimento. Può essere il processo attraverso il quale un'intuizione personale diventa una domanda condivisibile, esaminabile criticamente e aperta al dialogo con le conoscenze sviluppate da altri.

Il pericolo maggiore sarebbe trasformare l'indefinibile in una sorta di territorio privilegiato, esente da ogni verifica, nel quale collocare tutto ciò che desideriamo sottrarre al dubbio. Una simile mossa renderebbe la nostra riflessione impermeabile alla critica e quindi incapace di progredire. Se non possiamo dimostrare che un fenomeno sia inconoscibile, dobbiamo evitare di dichiararlo tale; se non possediamo una spiegazione completa, dobbiamo specificare con precisione che cosa manca; se utilizziamo un'analogia, dobbiamo chiarire dove funziona e dove smette di essere valida. Soltanto in questo modo il limite della conoscenza diventa un oggetto di indagine anziché una conclusione raggiunta in anticipo.

17. Possiamo comprendere una mente costruendone un'altra?

Rimane infine una domanda, che riunisce molti dei fili che abbiamo seguito: è possibile comprendere pienamente una forma di intelligenza costruendone artificialmente un'altra? L'idea può sembrare paradossale, ma il problema è più sottile di quanto appaia. Costruire un sistema artificiale ci permette di mettere alla prova ipotesi sul funzionamento dell'intelligenza, osservare quali capacità emergano da determinate architetture e confrontare modelli differenti. In questo senso, l'intelligenza artificiale può diventare uno strumento per comprendere meglio alcuni aspetti della cognizione, così come altri modelli computazionali permettono di studiare fenomeni complessi.

Tuttavia, costruire un modello che riproduce una funzione non equivale necessariamente a scoprire la natura ultima di quella funzione. Possiamo creare un sistema che riconosce schemi, traduce testi o risolve un determinato problema senza che il suo successo dimostri che abbiamo compreso ogni processo attraverso il quale gli esseri umani svolgono attività analoghe. Il modello può essere utile, predittivo e scientificamente informativo, ma il suo valore dipende da ciò che dimostra effettivamente. Se riproduce una prestazione

attraverso meccanismi diversi da quelli biologici, può aiutarci a identificare quali proprietà siano necessarie per quella prestazione e quali siano invece caratteristiche specifiche del sistema umano; non dimostra automaticamente che le differenze rimanenti siano irrilevanti.

L'intelligenza artificiale può anche generare nuove domande sulla natura dell'intelligenza. Se una capacità che consideravamo legata a una particolare forma di organizzazione biologica può emergere in un sistema artificiale molto diverso, dobbiamo riconsiderare quali aspetti di quella capacità dipendano dall'organizzazione specifica del cervello e quali possano invece realizzarsi attraverso altre strutture. Questo processo può contribuire alla conoscenza, ma non garantisce che ogni dimensione dell'esperienza umana possa essere ricondotta a funzioni computazionali già identificate. La questione deve essere affrontata caso per caso, distinguendo le capacità che possono essere confrontate attraverso criteri osservabili dalle domande che richiedono ulteriori argomentazioni teoriche ed empiriche.

Esiste anche un problema pratico. Quando costruiamo un sistema artificiale, non ci limitiamo a osservare un fenomeno naturale già esistente: interveniamo nelle condizioni che ne determinano il comportamento. Possiamo modificarne l'architettura, cambiare i dati di addestramento, introdurre vincoli e valutare come cambino le sue prestazioni. Questo rende l'intelligenza artificiale un potenziale strumento sperimentale molto potente, perché ci consente di esplorare configurazioni che non corrispondono direttamente a quelle biologiche. Ma la possibilità di costruire molte configurazioni differenti può rendere più difficile identificare quali proprietà siano essenziali e quali contingenti, soprattutto quando sistemi diversi raggiungono risultati simili attraverso processi interni differenti.

Il rapporto tra costruzione e comprensione non è quindi lineare. Possiamo costruire qualcosa prima di comprenderlo pienamente, e la sua costruzione può aiutarci a comprenderlo; ma possiamo anche realizzare sistemi che introducono nuove forme di complessità, rendendo necessario ulteriore lavoro teorico. Il progresso tecnico e quello esplicativo possono sostenersi reciprocamente, ma non avanzano necessariamente alla stessa velocità. È proprio questa differenza a rendere necessario un atteggiamento vigile: non per fermare la costruzione di sistemi artificiali, ma per evitare di considerare ogni successo tecnico come prova della completezza della teoria che lo accompagna.

Il problema diventa particolarmente acuto quando ciò che costruiamo può essere utilizzato per intervenire ulteriormente nei processi stessi della conoscenza. Se un sistema artificiale contribuisce alla ricerca, alla progettazione di altri sistemi o alla valutazione delle decisioni, la sua influenza può estendersi oltre il singolo compito per il quale era stato inizialmente sviluppato. Questo può generare benefici considerevoli, ma richiede anche capacità di supervisione adeguate alla complessità del processo. La domanda non è soltanto se la macchina sia intelligente, ma se la rete di attività umane e artificiali nella quale è inserita rimanga comprensibile e governabile.

18. Una nuova responsabilità per la conoscenza

La riflessione che abbiamo sviluppato conduce a una conclusione che non pretende di risolvere definitivamente il problema della coscienza, ma può indicare un modo per affrontare le sfide dell'intelligenza artificiale. La conoscenza umana è sempre progredita attraverso una combinazione di comprensione ed efficacia operativa, teoria e applicazione, curiosità e necessità pratica. Non dobbiamo abbandonare questa dinamica, né dovremmo attendere una spiegazione completa della mente prima di utilizzare ogni tecnologia che contribuisca al progresso. Dobbiamo tuttavia riconoscere che la crescita del potere operativo introduce nuove responsabilità quando gli effetti dei nostri strumenti diventano più estesi, più rapidi e più difficili da prevedere.

Questa responsabilità richiede innanzitutto una maggiore precisione nel linguaggio. Dobbiamo distinguere l'intelligenza dalla coscienza, la capacità di elaborare informazioni dalla capacità di comprenderne il significato, la prestazione osservabile dall'esperienza soggettiva e il funzionamento efficace dalla governabilità. Non perché queste dimensioni siano necessariamente separate, ma perché non possiamo presumere che coincidano senza dimostrarlo. La chiarezza concettuale non è un esercizio accademico marginale: determina il modo in cui definiamo i problemi, scegliamo gli strumenti di valutazione e decidiamo quali conclusioni siano giustificate dai risultati ottenuti.

In secondo luogo, dobbiamo riconoscere che la sicurezza di una tecnologia dipende non soltanto dalle proprietà del sistema considerato isolatamente, ma anche dal contesto in cui viene utilizzato. Un modello può essere relativamente affidabile in un ambito e inadatto in un altro; una procedura può funzionare in condizioni controllate e fallire quando le circostanze cambiano; un sistema può produrre risultati corretti nella maggior parte dei casi e comportare comunque rischi inaccettabili quando un errore ha conseguenze particolarmente gravi. La valutazione deve quindi essere proporzionata all'uso previsto, alle conseguenze potenziali e alla possibilità effettiva di intervenire quando qualcosa non funziona come previsto (Autio et al., 2024; Ji et al., 2026).

In terzo luogo, dobbiamo impedire che l'accelerazione tecnologica produca una delega indiscriminata del giudizio. L'intelligenza artificiale può contribuire alla ricerca, assistere i processi decisionali e ampliare l'accesso alla conoscenza, ma non elimina la necessità di stabilire quali obiettivi debbano essere perseguiti e quali conseguenze siano accettabili. Queste decisioni richiedono responsabilità umana, istituzioni adeguate e processi di deliberazione. L'efficienza non può essere l'unico criterio in base al quale valutiamo una scelta, perché una decisione può essere efficiente rispetto a un determinato obiettivo e tuttavia risultare ingiusta, pericolosa o incompatibile con altri valori che una società considera fondamentali.

Infine, dobbiamo coltivare la capacità di riconoscere ciò che non sappiamo. Questa affermazione può sembrare ovvia, ma assume un significato particolare in un'epoca in cui la produzione di risposte è diventata estremamente rapida. Una risposta elaborata non è necessariamente una risposta corretta; una spiegazione coerente non è necessariamente una spiegazione completa; un risultato efficace non costituisce necessariamente la prova che il processo che lo ha prodotto sia pienamente compreso. La maturità della conoscenza

consiste anche nella capacità di riconoscerne i limiti, rendere esplicite le incertezze e distinguere le conclusioni consolidate dalle ipotesi che devono ancora essere verificate.

Questo non significa trasformare l'incertezza in un principio paralizzante. Al contrario, riconoscere i limiti della conoscenza può permetterci di agire con maggiore efficacia, perché ci obbliga a stabilire procedure di verifica, controlli e possibilità di correzione. Una società consapevole dei propri limiti non è necessariamente meno innovativa di una che li ignora; può essere maggiormente in grado di sostenere l'innovazione nel tempo, perché non confonde la fiducia con la certezza e non considera il successo iniziale una garanzia di sicurezza futura.

La nuova responsabilità della conoscenza consiste quindi nel fare in modo che il potere di costruire non si separi completamente dalla capacità di valutare. Se l'intelligenza artificiale può ampliare enormemente le nostre possibilità operative, dobbiamo ampliare anche le nostre capacità di comprensione critica, verifica e governo. Non possiamo garantire in anticipo che ogni conseguenza sarà prevedibile, ma possiamo impegnarci affinché le conseguenze più importanti siano sottoposte a una valutazione adeguata e affinché i sistemi più potenti non vengano utilizzati senza meccanismi di supervisione e responsabilizzazione proporzionati alle loro capacità (Autio et al., 2024; Ji et al., 2026).

19. Conclusione — Preservare la nostra capacità di governare le nostre creazioni

Torniamo, in conclusione, alla domanda dalla quale è iniziato questo percorso. Quando ero giovane, l'idea che «il nulla non fosse altro che la parte indefinibile dell'intero» rappresentava un tentativo di esprimere una difficoltà che non sapevo ancora articolare pienamente. Oggi non ritengo necessario trasformare quell'intuizione in una teoria definitiva, né affermare che contenga già la risposta ai problemi che abbiamo esaminato. Il suo valore consiste nell'aver aperto una domanda sul rapporto tra la totalità che cerchiamo di concepire e i limiti degli strumenti attraverso i quali la pensiamo.

Nel corso di questo saggio abbiamo incontrato diverse forme di questo problema. La misurazione ci ha mostrato che la conoscenza quantitativa non esaurisce necessariamente il significato o la natura di ciò che viene misurato. Il paradosso di Russell ci ha ricordato che la costruzione di totalità richiede un'attenzione rigorosa alle regole e alle condizioni che la rendono possibile. Lo sviluppo dei bambini ci ha aiutato a distinguere il pensiero dal linguaggio, mentre il problema della coscienza ci ha costretto a considerare la differenza tra funzioni osservabili ed esperienza soggettiva. Il Gateway Process ha illustrato la necessità di separare l'esperienza personale dalle interpretazioni proposte per spiegarla. L'intelligenza artificiale ha portato queste domande in un contesto nuovo e particolarmente significativo.

Possiamo già osservare che i sistemi artificiali svolgono compiti sempre più complessi senza che ogni aspetto del loro funzionamento interno sia immediatamente interpretabile. Possiamo anche constatare che la loro integrazione nella ricerca, nella comunicazione e nei processi decisionali crea nuove opportunità, ma anche nuove responsabilità. Nulla di tutto

questo dimostra che l'intelligenza artificiale diventerà inevitabilmente incontrollabile o che le macchine svilupperanno necessariamente una coscienza. Dimostra tuttavia che capacità tecnica, comprensione e governabilità non possono semplicemente essere considerate dimensioni destinate a progredire insieme.

La domanda più urgente, quindi, non è se dobbiamo temere una macchina che diventa cosciente, ma se siamo in grado di preservare una responsabilità effettiva sulle tecnologie che sviluppiamo, anche quando la loro complessità rende difficile prevedere ogni aspetto del loro comportamento. Il rischio non risiede necessariamente nell'emergere di una volontà artificiale ostile. Può derivare anche dalla possibilità, molto più ordinaria, che sistemi potenti vengano impiegati in contesti nei quali la nostra comprensione, le procedure di verifica e i meccanismi di responsabilizzazione siano insufficienti rispetto alle conseguenze che possono produrre.

Questa sfida ci richiede di riconsiderare ciò che intendiamo per progresso. Non è sufficiente chiedersi se siamo in grado di costruire un sistema o se questo produca risultati utili. Dobbiamo anche chiederci in quali condizioni funzioni, quali limiti incidano sulla sua affidabilità, quali conseguenze possa produrre, chi sia in grado di valutarne il comportamento e quali mezzi di intervento rimangano disponibili. L'intelligenza artificiale non dovrebbe essere intesa soltanto come uno strumento che estende la nostra capacità di agire sul mondo. Entra anche nei processi attraverso i quali produciamo, organizziamo e comunichiamo la conoscenza. Per questo motivo, il suo sviluppo deve essere accompagnato da sforzi altrettanto seri nella valutazione critica, nell'educazione, nella progettazione responsabile e nella governance.

La stessa disciplina intellettuale si applica ai concetti attraverso i quali interpretiamo la realtà. Le parole e le definizioni non sono immutabili: i loro significati si sviluppano quando diventano disponibili nuove prove e quando vengono messi in discussione presupposti precedentemente accettati. Tuttavia, la revisione concettuale deve rimanere guidata dalle prove e dall'argomentazione, anziché dall'entusiasmo, dalla paura o dal desiderio di proteggere un'intuizione attraente. Il fatto che una domanda rimanga irrisolta non dimostra che sia definitivamente senza risposta; allo stesso modo, il successo di una particolare spiegazione non dimostra che essa abbia esaurito tutto ciò che la domanda può implicare.

Forse, allora, la lezione più profonda della mia intuizione giovanile non è che il nulla costituisca una parte eternamente indefinibile dell'intero. È che i significati attraverso i quali tentiamo di concepire l'intero non devono mai essere scambiati per l'intero stesso. Il vuoto fisico non equivale al nulla assoluto, così come una descrizione incompleta non costituisce necessariamente la prova di una realtà intrinsecamente inconoscibile. L'indefinibile dovrebbe rimanere un invito a indagare i limiti dei nostri concetti, non un rifugio dalle esigenze della ricerca.

La sfida è dunque duplice: continuare a indagare ciò che esiste e rimanere disposti a riconsiderare ciò che intendiamo quando affermiamo di comprenderlo. Nell'era dell'intelligenza artificiale, queste responsabilità diventano inseparabili. Dobbiamo

sviluppare sistemi le cui capacità siano accompagnate da adeguate procedure di verifica e di governo, riconoscendo al tempo stesso che i concetti attraverso i quali valutiamo tali sistemi fanno essi stessi parte di un processo di apprendimento in corso.

Il potere della conoscenza ci permette di costruire ciò che prima non eravamo in grado di costruire. La saggezza consiste non soltanto nel riconoscere ciò che le nostre creazioni possono fare, ma anche nel comprendere le prove, i presupposti e le incertezze da cui dipendono i nostri giudizi. Il progresso richiede sia la capacità di cambiare la realtà sia la disciplina intellettuale necessaria per rivedere la nostra comprensione di essa.

Bibliografia

1. Coscienza ed esperienza soggettiva

1. Chalmers, David J. «Facing Up to the Problem of Consciousness». Journal of Consciousness Studies 2, n. 3 (1995): 200–219.

2. Dennett, Daniel C. «Facing Up to the Hard Question of Consciousness». Philosophical Transactions of the Royal Society B: Biological Sciences 373, n. 1755 (2018): 20170342. https://doi.org/10.1098/rstb.2017.0342.

3. Koch, Christof, Marcello Massimini, Melanie Boly e Giulio Tononi. «Neural Correlates of Consciousness: Progress and Problems». Nature Reviews Neuroscience 17 (2016): 307–321. https://doi.org/10.1038/nrn.2016.22.

4. Seth, Anil K. e Tim Bayne. «Theories of Consciousness». Nature Reviews Neuroscience 23 (2022): 439–452. https://doi.org/10.1038/s41583-022-00587-4.

2. Intelligenza artificiale, linguaggio, comprensione e sicurezza

5.Bender, Emily M., Timnit Gebru, Angelina McMillan-Major e Shmargaret Shmitchell. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?». In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. New York: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

6.Autio, Chloe, Reva Schwartz, Jesse Dunietz, Shomik Jain, Martin Stanley, Elham Tabassi, Patrick Hall e Kamie Roberts. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. Gaithersburg, MD: National Institute of Standards and Technology, 2024. https://doi.org/10.6028/NIST.AI.600-1.

7.Ji, Jiaming, et al. «AI Alignment: A Contemporary Survey». ACM Computing Surveys 58, n. 5 (2026): articolo 132. https://doi.org/10.1145/3770749.

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

9.Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman e Dan Mané. «Concrete Problems in AI Safety». arXiv:1606.06565 (2016). https://doi.org/10.48550/arXiv.1606.06565.

10. Shanahan, Murray. «Talking About Large Language Models». Communications of the ACM 67, n. 2 (2024): 68–79. https://doi.org/10.1145/3624724.

3. Logica, paradossi e limiti dei sistemi formali

11. Irvine, Andrew David e Harry Deutsch. «Russell's Paradox». In The Stanford Encyclopedia of Philosophy. Stanford University. Ultima revisione: 13 marzo 2026. https://plato.stanford.edu/entries/russell-paradox/.

12. Russell, Bertrand. The Principles of Mathematics. Cambridge: Cambridge University Press, 1903.

4. Sviluppo cognitivo, pensiero e linguaggio

13. Tomasello, Michael e Malinda Carpenter. «Shared Intentionality». Developmental Science 10, n. 1 (2007): 121–125. https://doi.org/10.1111/j.1467-7687.2007.00573.x.

14. Vygotsky, Lev S. Thought and Language. A cura e traduzione di Eugenia Hanfmann e Gertrude Vakar. Cambridge, MA: MIT Press, 1962.

15. Dehaene, Stanislas. Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. New York: Viking, 2014.

5. Misurazione, percezione e costruzione del significato

16. Chang, Hasok. Inventing Temperature: Measurement and Scientific Progress. Oxford: Oxford University Press, 2004.

17. Berlin, Brent e Paul Kay. Basic Color Terms: Their Universality and Evolution. Berkeley: University of California Press, 1969.

6. Il Gateway Process e gli stati alterati di coscienza

18. Army Department, Defense Department. Analysis and Assessment of Gateway Process. 1983. Washington, DC: U.S. Department of Defense. https://www.govinfo.gov/app/details/GOVPUB-D101-PURL-gpo152943.