Introduction to Artificial Intelligence

Written by Febogamedev

Introduction

Since the dawn of civilization, human beings have sought to reduce their burden of effort by delegating some of their activities to tools and machines.
The wheel, the lever, the mill: many inventions have not only made it possible to do “more” and “faster”, but also to reduce the work needed to carry out repetitive or demanding activities.

Throughout history, this delegation has also extended to cognitive activities.
Writing, calculating, recording information and remembering are examples of operations in which tools can support or replace part of human work.
In this way, some cognitive resources can be devoted to different or more complex activities (Risko & Gilbert, 2016).

A further step in the development of automation was to design machines capable not only of executing predefined procedures, but also of changing their behavior according to conditions and goals.

Autonomy and adaptivity are two important characteristics of many artificial intelligence systems (Russell & Norvig, 2021).

- An autonomous machine can carry out certain actions or make decisions without continually requiring human intervention.
- An adaptive machine can change its behavior in response to the information it receives, environmental conditions or experience.

In more advanced systems, adaptation can also result from learning: the system uses data and experience to change the way it approaches a task (Goodfellow et al., 2016).

The idea of the intelligent machine can therefore be interpreted as an extension of a very ancient desire: to build tools capable of taking on an increasing share of human activities, up to performing tasks that we associate with intelligence.

Today's intelligent machines are therefore not the result of a sudden leap in the history of technology, but are part of a long evolution of tools and systems designed to delegate, automate and support increasingly complex activities.

Definition

To understand what artificial intelligence is, we must first clarify what we mean by intelligence.

Intelligence is a complex and debated concept, studied by disciplines such as psychology, neuroscience, philosophy and computer science. In broad terms, we can describe it as the ability to acquire and use knowledge, process information and adapt one's behavior to address problems and achieve goals in different contexts (cf. McCarthy, 2007; Russell & Norvig, 2021).

There is no single definition of intelligence, but many descriptions refer to abilities such as:

- Learning: acquiring new information and skills.
- Reasoning: processing information and drawing conclusions.
- Problem solving: identifying strategies and solutions when faced with new or complex problems.
- Adaptation: changing one's behavior or strategies according to the context.
- Planning: organizing actions and decisions around a goal.

In cognitive psychology, intelligence is studied in relation to various cognitive functions, such as memory, perception, language, reasoning, abstraction and planning (cf. Anderson et al., 2004).
Designing artificial systems capable of performing some of these functions means entering the field of **artificial intelligence (AI)**.

In historical and disciplinary terms, artificial intelligence is a branch of computer science that studies and designs systems capable of performing tasks that, if carried out by a human being, would require abilities considered intelligent (Russell & Norvig, 2021).

John McCarthy, who coined the expression Artificial Intelligence in the 1950s, defined it as "the science and engineering of making intelligent machines, especially intelligent computer programs" (McCarthy, 2007, p. 2).

Today the term artificial intelligence encompasses very different systems, which may possess abilities such as:

- Perceiving: analyzing information from the environment, such as images, sounds, texts or data collected by sensors.
- Representing information: building internal representations of data, objects, relationships or situations.
- Reasoning: combining information to formulate inferences, predictions or decisions.
- Learning: changing performance on the basis of data or experience, through techniques such as machine learning.
- Interacting: communicating with human beings, environments or other systems.

We can therefore adopt an operational definition:
"Artificial intelligence is the discipline of computer science that studies and develops systems capable of performing tasks normally associated with forms of intelligent behavior."

Types of AI

Once we understand what artificial intelligence is, we can distinguish three theoretical categories based on the breadth and generality of the system's abilities:

- Narrow AI or ANI (Artificial Narrow Intelligence).
- General AI or AGI (Artificial General Intelligence).
- Superintelligence or ASI (Artificial Superintelligence).

These categories do not simply indicate a system's computing power, but the scope of the abilities it can use and transfer across different tasks and contexts (cf. Morris et al., 2024).

Narrow AI (ANI)

Definition

Systems designed or trained to perform one or more specific tasks, even with very high performance.

Key characteristic

Their abilities remain limited to a restricted set of tasks or domains, and they do not possess a general ability to independently transfer their skills to any new problem.

Limitation

They can be extremely effective within the tasks for which they were designed or trained, but they do not possess the cognitive breadth and flexibility of a general intelligence (Morris et al., 2024).
Most AI systems currently in use fall into this category.

General AI (AGI)

Definition

A hypothetical system with general cognitive abilities, capable of learning, reasoning and transferring knowledge across different domains.

Key characteristic

It would not be limited to a single task, but could address new problems, learn new skills and use knowledge acquired in different contexts.

Limitation

AGI remains a research goal, and there is no universally shared definition that makes it possible to determine definitively when a system can be considered a general intelligence (Morris et al., 2024).

Its possible development would also raise important issues concerning safety, control and the alignment of the system's goals with human goals.

Superintelligence (ASI)

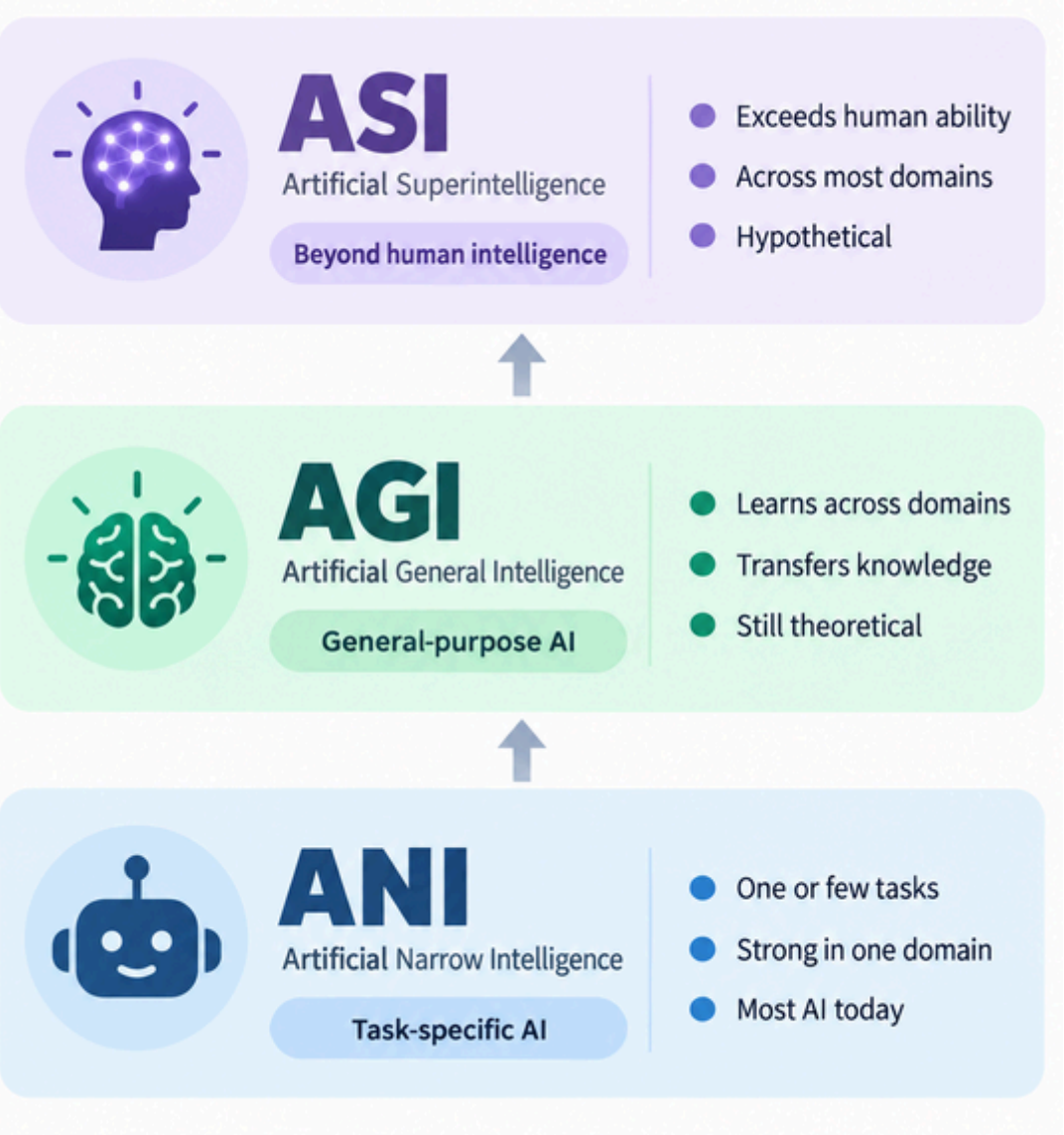
Definition

A hypothetical system whose cognitive abilities significantly exceed those of humans in most relevant areas, such as reasoning, planning, learning, problem solving and creativity (cf. Morris et al., 2024).

Key characteristic

A superintelligence would therefore not merely possess general abilities comparable to human abilities, but would significantly exceed them.

Limitation
Superintelligence is currently a theoretical concept.
A large gap between human and artificial abilities could also make it more difficult to understand, predict or control the system's decisions, making the problem of alignment with human goals particularly important (cf. Morris et al., 2024).



The evolution of AI: from classical to modern

Artificial Intelligence did not begin with neural networks or contemporary chatbots. Throughout its history, different ways of building intelligent systems have developed, among which we can distinguish two broad approaches:

- Symbolic or classical AI.
- AI based on learning from data.

These two approaches do not necessarily represent two phases in which one replaces the other, but different ways of addressing the same problem: designing systems capable of performing tasks associated with intelligence (cf. Garcez & Lamb, 2020).

Symbolic or classical AI

Definition

An approach in which knowledge, concepts and relationships are explicitly represented through symbols, logical rules, formulas or other formal structures.

Method

The programmer defines the knowledge and rules that the system can use. The system then applies inference, search or reasoning procedures to reach conclusions and make decisions (Russell & Norvig, 2021).

Strengths

The reasoning process can be reconstructed and analyzed step by step.
It offers good reliability when the domain is well defined and the rules are known.
It allows knowledge and relationships to be represented explicitly.

Limitations

- Flexibility decreases when the problem goes beyond the knowledge and rules represented.
- It can have difficulty dealing with uncertainty, ambiguity and highly variable situations.
- Manually constructing large amounts of knowledge can become complex (Goodfellow et al., 2016).

Example

Imagine a program that has to help identify some possible diseases based on symptoms. Knowledge such as the following is explicitly entered into the system:

- High fever + cough + difficulty breathing → possible pneumonia.
- High fever + sore throat + tonsillar patches → possible tonsillitis.
- Sneezing + stuffy nose + no fever → possible cold.

The program does not “learn” these relationships from data: the rules were defined in advance by those who designed the system.

If the patient presents with:

- High fever.
- Cough.
- Difficulty breathing.

the system can apply a rule and conclude that the patient has possible pneumonia.

The reasoning can therefore be reconstructed:
Observed symptoms → Applied rule → Conclusion.

This clearly illustrates the strength of symbolic AI: we can know exactly why the system reached that conclusion.
It also shows the limitation. If the patient presents with a combination of symptoms that

was not anticipated in the rules, for example:
Moderate fever + unusual cough + loss of smell.

the system might not know what to conclude, unless that situation has been explicitly represented in its knowledge base.

AI based on learning from data, or modern AI

Definition

An approach in which the rules needed to perform a task are not fully defined in advance by the programmer. Instead, the system learns regularities and relationships from data.

Method

The system is trained using data or examples and adapts its model on the basis of the data to identify patterns useful for carrying out the task.
Instead of directly specifying "If X happens, then do Y", the programmer defines a process through which the system can learn from data what relationship exists between X and Y (Goodfellow et al., 2016).

Strengths

- It can identify regularities that are difficult to describe through explicit rules.
- It can deal with complex, noisy or incomplete data.
- It can achieve high performance in tasks such as image recognition, language processing and machine translation.
- It can generalize what it has learned to new data similar to those encountered during training (Goodfellow et al., 2016).

Limitations

- It can be difficult to understand how the system arrived at a particular response.
- The quality of its behavior depends heavily on the data and the training process.
- More complex models can require large amounts of data and computational resources (Goodfellow et al., 2016).

Example

Let us try to create a program that recognizes whether an image contains a cat.
In the symbolic approach, the programmer would have to try to write rules such as:
If there are two pointed ears, two eyes, whiskers and a certain muzzle shape, then the image probably contains a cat.

The problem is that in reality images vary constantly: the cat may be seen from the side, in the dark, partly hidden, with different colors or in different poses.
Writing all possible rules by hand would become very difficult.

With AI based on learning from data, on the other hand, the programmer does not explicitly describe all the characteristics of a cat.
They provide the system with many examples of labeled images, for example:

- Image 1 → cat.
- Image 2 → dog.
- Image 3 → cat.

- Image 4 → car.
- Image 5 → cat.

During training, the system gradually changes its parameters in an attempt to reduce errors. In this process it learns statistical regularities useful for distinguishing cats from other images.
After training, when it receives a new image it has never seen before, the system can assign the image, for example, an estimated 94% probability of belonging to the “cat” class

The important point is that no one has written an explicit rule such as “a cat has exactly this shape”. The system has derived from the data the patterns useful for performing the task (cf. Goodfellow et al., 2016).
This example also clearly illustrates the strengths and limitations:

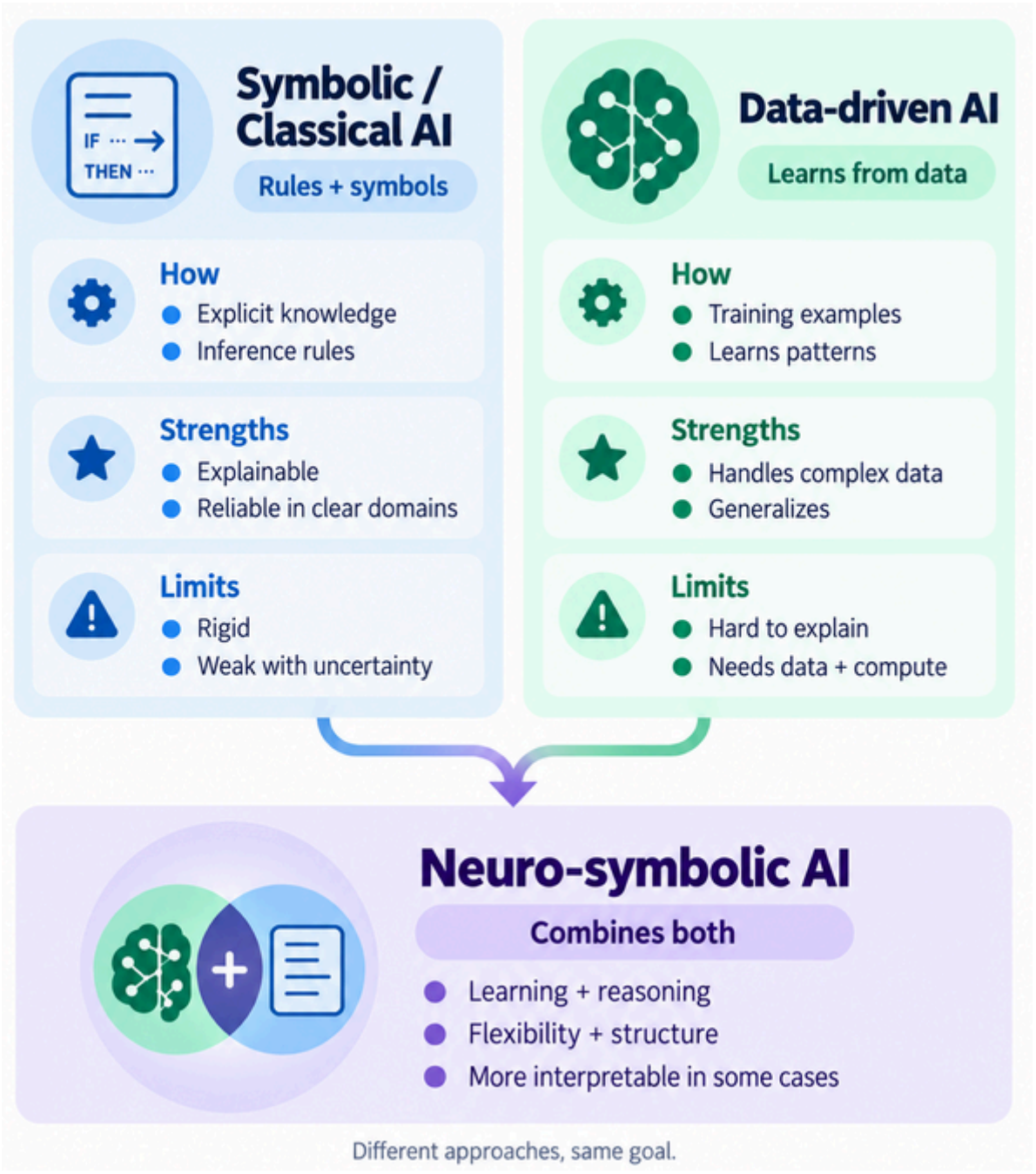
- Generalization: it can recognize a cat even in a new image.
- Complex data: it can work with variations in lighting, position, color and background.
- Data dependency: if it has seen almost only white cats during training, it may have more difficulty with other cases.
- Opacity: in a complex model it can be difficult to explain exactly which features determined the classification.

Symbolic AI and learning from data therefore have different characteristics.

- The symbolic approach allows knowledge and rules to be represented explicitly, making the reasoning more readily interpretable.
- Statistical learning, on the other hand, makes it possible to identify regularities in data and to adapt better to complex problems or problems that are difficult to describe through explicit rules.

One direction in contemporary research seeks to integrate these two approaches.
This integration is called **neuro-symbolic AI**.

Neuro-symbolic AI combines learning techniques, often based on neural networks, with explicit forms of knowledge representation and symbolic reasoning.
The goal is to combine the ability to learn from data with the ability to use structured knowledge and rules, in an attempt to obtain more flexible, consistent and, in some cases, more interpretable systems (Garcez & Lamb, 2020).



Philosophy of AI

Conceptions of AI

Artificial Intelligence does not necessarily have to imitate the way a human being thinks or acts. A system can in fact be considered intelligent even when it uses strategies very different from ours, provided it can effectively achieve certain goals (McCarthy, 2007).

Stuart Russell and Peter Norvig organized different conceptions of Artificial Intelligence around two fundamental questions (Russell & Norvig, 2021, ch. 1):

- What do we evaluate? The way the system thinks or the way it acts.
- What reference do we use? Human beings or an ideal of rationality.

The first criterion therefore distinguishes between thinking and acting.
- Thinking: the focus is on the processes through which the system processes information and reaches a conclusion.
- Acting: the focus is instead on the behavior and actions produced by the system.

The second criterion distinguishes between humanly and rationally.
- Humanly: the reference is the way human beings think or behave.
- Rationally: the reference is the ability to use the available information to choose the most appropriate solution or action for a given goal.

Being human and being rational do not necessarily coincide.
- Human beings think and act through emotions, habits, intuitions, personal experience, cultural context and cognitive limitations. Their decisions are therefore not always the most effective or logically optimal.
- A rational system, on the other hand, is evaluated on its ability to choose, on the basis of the available information, the action that appears most suitable for achieving a given goal (Russell & Norvig, 2021).

A system can therefore behave very differently from a human being and still be rational. Likewise, it can imitate human behavior very well without always making rational decisions.

Combining the two criteria yields four main ways of conceiving Artificial Intelligence.

Thinking humanly
The goal is to reproduce or model the cognitive processes through which human beings reason, learn and make decisions.
- Goal: to reproduce the mechanisms of human thought.
- Example: systems that simulate cognitive processes studied by psychology.
- Limitation: human mental processes are difficult to observe directly and are not yet fully understood.

Example
Let us imagine this problem:
27 × 14 = ?

A system oriented toward thinking rationally could directly apply an efficient algorithm and obtain the correct result.
A system designed to think humanly, on the other hand, seeks to reproduce the cognitive process observed in people. For example:
- 27 × 10 = 270
- 27 × 4 = 108
- 270 + 108 = 378

The model might also take into account typical features of human cognition, such as limited working memory, the time required for each step or the possibility of making certain errors.

In this case the main goal is not to find the mathematically best method, but to build a model that produces a process similar to the human one.

Therefore:
Problem → model of human cognitive processes → human-like reasoning → answer.

The key point is that "thinking humanly" concerns the internal process, not just the result. If a machine arrives at 378 using a procedure completely different from the human one, it may have solved the problem correctly, but it is not necessarily modeling human thought.

A real example of this type of research is provided by cognitive architectures, such as *ACT-R*, designed to model processes such as memory, attention, learning and problem solving observed in human beings (Anderson et al., 2004).

Thinking rationally

The goal is to build systems capable of reaching logically correct conclusions from the available information.

- Goal: to reason according to logical principles.
- Example: systems that use formal logic to derive conclusions from certain premises.
- Limitation: formal logic can be too rigid when information is incomplete, ambiguous or uncertain.

Example

To think rationally, the machine must be able to apply logical rules to the available information through a process of inference.

A very simple example of thinking rationally is a system that must determine whether Socrates is mortal on the basis of certain premises.
The system receives this information:

- All human beings are mortal.
- Socrates is a human being.

It then applies a logical rule:

- If all human beings are mortal
- and Socrates is a human being,
- then Socrates is mortal.

The conclusion is therefore that Socrates is mortal.

Here the system does not try to imitate the way a person actually reasons. The goal is for the transition from premises to conclusion to be logically valid.

In schematic form:
Premises → Logical rules → Inference → Conclusion.

The limitation emerges when the information is uncertain. For example:

- Birds normally fly.
- Tweety is a bird.

We might conclude that "Tweety flies".
But it could be a penguin, or have an injured wing. Classical logic works well with clear-cut

statements such as true/false, whereas the real world often contains exceptions, probabilities and incomplete information.

A concrete historical example is Logic Theorist, a program developed in the 1950s by Allen Newell, Herbert Simon and Cliff Shaw, capable of proving theorems in symbolic logic. It is one of the first programs generally considered part of the history of AI (Russell & Norvig, 2021, ch. 1).

The key point is therefore:
Thinking rationally means reaching logically correct conclusions from the available premises, not necessarily reasoning as a person would.

Acting humanly

The goal is not necessarily to reproduce human mental processes, but to obtain external behavior similar to that of a person (Russell & Norvig, 2021).

- Goal: to behave similarly to a human being.
- Example: a chatbot capable of sustaining a natural conversation.
- Limitation: Human-like behavior does not demonstrate that the system uses the same cognitive processes or possesses the same form of understanding.

Example

Imagine the user writes:
“I arrived late for work because I missed the train.”

A system oriented toward acting humanly might reply:
“What bad luck. Did you at least manage to let them know in time?”

The point is not that the system actually feels sorry or understands the situation in the same way as a person. The goal is for the external behavior of the conversation to appear human: using natural language, reacting relevantly, asking appropriate questions, respecting conversational turns and adapting the tone to the context.

In schematic form:
Human input → human-like linguistic behavior → observable response.

This is precisely the limitation: from the outside we can observe a very natural response, but we cannot automatically infer that the system thinks or understands like a human being.

A chatbot is a striking example of a machine that acts humanly.

Therefore:
Acting humanly means imitating observable human behavior, not necessarily reproducing the mental processes that produce it.

Acting rationally

The goal is to build systems capable of choosing actions that increase the probability of achieving a given goal on the basis of the available information (Russell & Norvig, 2021).

- Goal: To choose the most appropriate action for a goal.
- Example: A self-driving system that chooses a trajectory based on road conditions, obstacles and the destination.
- Strength: It allows a system to be evaluated on the results it can achieve, without requiring it to think or behave like a human being.

Example
Imagine that the car is traveling toward a destination and suddenly detects a pedestrian crossing the road.

The system has various pieces of information:
- Current speed: 50 km/h.
- Distance from the pedestrian: 30 meters.
- Wet road.
- Vehicle behind very close.

At this point it can evaluate several possible actions:
- Brake sharply.
- Brake progressively.
- Steer.
- Continue at the same speed.

The system chooses the action that, on the basis of the available information, maximizes the probability of achieving the main goal: avoiding a collision while keeping the vehicle under control.

It might therefore decide:
Brake immediately and maintain the trajectory.

The important point is that it does not have to behave as a person would. It could make a decision in a completely different way from a human driver, provided that the chosen action is the most appropriate for the goal.

In schematic form:
Perception of the environment → evaluation of possible actions → choice of the most appropriate action → action.

Therefore:
Acting rationally means choosing the action that, given the available information, increases the probability of achieving a given goal.

The difference from thinking rationally is important:
- Thinking rationally: reaching logically correct conclusions.
- Acting rationally: choosing the most appropriate action to achieve a goal.

A system can therefore act rationally without “thinking” rationally. It can simply apply rules that connect certain conditions to certain actions, thus choosing effective behavior without having to understand the situation in the same way a human being would (cf. Russell & Norvig, 2021).

The four perspectives can be summarized through this matrix:

Reference Criterion	Humanly	Rationally
Thinking (being)	**Thinking humanly** Imitating human cognitive processes. It draws on cognitive psychology, seeking to model how we reason, learn and make decisions.	**Thinking rationally** Following the laws of formal logic. It aims to build systems that draw correct conclusions from known premises.
Acting (imitating)	**Acting humanly** Behaving like a human being.	**Acting rationally** Choosing the best action to achieve a goal, on the basis of the available information. It is the most widespread approach in practical AI today.

These four perspectives are not necessarily in competition. The same system can be designed or evaluated according to different criteria.

For example, a conversational assistant can be considered from at least two points of view:

- Acting humanly: If the goal is to produce a natural conversation similar to that of a person.
- Acting rationally: If the goal is to provide the most useful response to the user's request, even when it does not perfectly imitate human behavior.

This distinction shows that speaking generically of an “intelligent” machine can mean very different things. We can ask whether it thinks like us, whether it reasons correctly, whether it behaves like us or whether it acts effectively to achieve a goal.

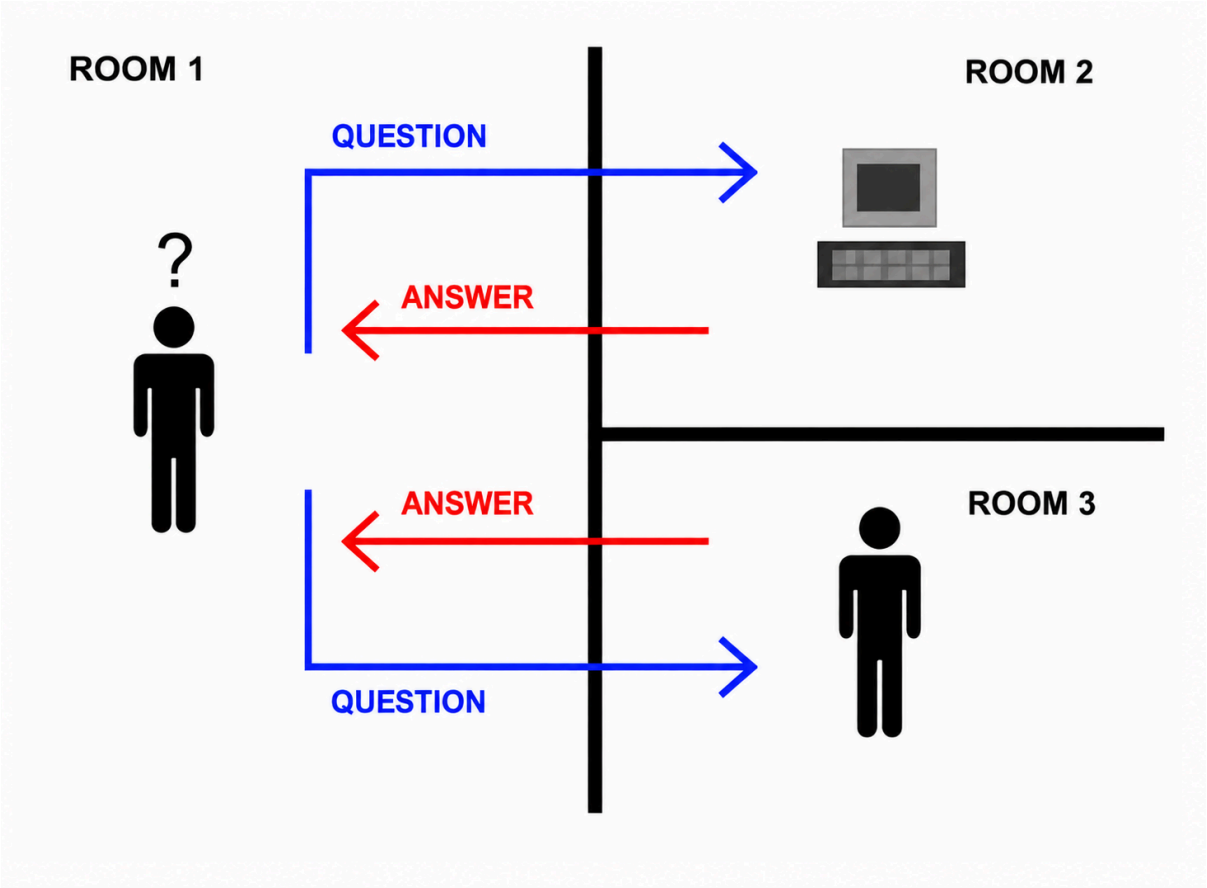
The Turing Test and the idea of “acting humanly”

One of the most influential attempts to address the problem of intelligence in machines is the **Turing Test**, proposed by Alan Turing in 1950 in the article Computing Machinery and Intelligence (Turing, 1950).

Turing starts from a question that is very difficult to define precisely: “Can machines think?” (can machines think?).
Instead of seeking an abstract definition of “thought”, he proposes replacing the question with an observable criterion: evaluating the machine's behavior.

In the so-called **imitation game**, a human observer communicates through text with interlocutors they cannot see directly. In the version now commonly associated with the Turing Test, one interlocutor is human and the other is a machine.

If the observer cannot reliably distinguish the machine from the human being on the basis of the conversation, then the system displays behavior sufficiently similar to human behavior to satisfy the test's criterion (Turing, 1950).



The Turing Test therefore belongs to the perspective of acting humanly (Russell & Norvig, 2021).

It does not seek to establish what happens inside the machine. It does not directly test whether the system understands, possesses consciousness or uses the same cognitive processes as a human being.

It evaluates only what is observable from the outside:
Linguistic input → machine behavior → comparison with human behavior.

This approach is particularly important because it allows the problem of intelligence to be addressed without first having to resolve much more difficult philosophical questions, such as the nature of mind or consciousness.

However, the Turing Test also has a fundamental limitation: human-like behavior does not necessarily tell us which internal processes produced it.
A machine could therefore respond correctly, use natural expressions, ask questions, make ironic remarks or even make errors typical of a person without the test allowing us to establish whether the system actually understands what it is saying (cf. Searle, 1980).

The test in fact evaluates the external outcome of behavior, not the nature of the process that generates it.
For this reason, passing the Turing Test does not automatically demonstrate that a machine:

- Thinks like a human being.
- Understands the meaning of its responses.
- Possesses consciousness or intentionality.
- Uses the same human cognitive processes.

The historical value of the Turing Test therefore lies above all in having transformed a very difficult question — “does a machine think?” — into a problem that can be observed and evaluated through behavior (cf. Turing, 1950).

However, the test does not definitively solve the problem of artificial intelligence.
It leaves a fundamental question open:
“If a machine behaves as if it understands, does that mean it actually understands?”

This distinction between observable behavior and internal understanding will become central to the philosophical debate on Artificial Intelligence and lead to a distinction between different conceptions of the artificial mind.

Strong hypothesis and weak hypothesis

In the philosophical debate on Artificial Intelligence, there are two theoretical positions:

- Weak AI hypothesis (weak AI hypothesis).
- Strong AI hypothesis (strong AI hypothesis).

This distinction is not primarily about what AI systems can do, but about how we interpret the nature of their cognitive abilities.

The fundamental difference is this:

- **The weak hypothesis** regards AI systems as tools capable of simulating or reproducing intelligent behavior, without this necessarily implying the presence of a mind or real understanding.
- **The strong hypothesis**, on the other hand, holds that an appropriately organized artificial system might not merely simulate cognitive abilities, but actually possess understanding and other mental states (Searle, 1980).

The weak AI hypothesis
According to the weak hypothesis, an artificial system can exhibit behavior that appears intelligent without it being necessary to attribute a human-like mind to it.

A system can therefore solve problems, classify images, translate texts, play games, generate language or make decisions without these abilities automatically demonstrating the presence of understanding, consciousness or intentionality.

From this perspective, the system can behave as if it understands:

- It can answer a question correctly.
- It can use language coherently.
- It can solve a complex problem.
- It can choose an action appropriate to a goal.

Observable behavior, however, is not enough to establish that there is a genuine mental experience behind that behavior.

AI is therefore treated primarily as a computational tool capable of producing results that we associate with intelligence.

This position is methodologically cautious: it allows artificial systems to be studied and used without having to assume that they possess a mind or consciousness.

The strong AI hypothesis
The strong hypothesis, on the other hand, holds that, at least in principle, a suitably organized artificial system might not merely simulate a mind, but actually possess cognitive states (cf. Searle, 1980).

A system of this kind could therefore have:
- Understanding.
- Intentionality.
- Mental states.
- Possibly, according to some interpretations, conscious experience.

From this perspective, the mind would not necessarily be tied exclusively to the biological brain.

If certain mental properties depended primarily on the organization and functioning of a system, then it might be possible to realize mental processes through an artificial medium as well.

The machine would therefore not merely behave as if it understood, but could actually understand.

However, this position raises problems that are difficult to resolve.
- How can we establish whether an artificial system actually understands something?
- How can we distinguish genuine understanding from a perfect simulation of understanding?
- And what characteristics would a system need to possess for it to be appropriate to attribute mental states to it?

Example
Imagine an artificial assistant to which the user says:
“Tomorrow I have to go to the beach, but it looks as though it will rain. What do you recommend?”

The system replies:
“If rain is forecast, it would probably be better to postpone or choose an indoor activity.”

Weak hypothesis
According to the weak hypothesis, the system produces an appropriate response because it processes the text, recognizes relationships between concepts and generates a coherent output.
It can therefore behave as if it understands that the user wants to go to the beach and that rain could be a problem.
But there is no need to assume that it actually possesses the concept of the sea, knows what it means to rain or has an intention of its own.
Intelligent behavior ≠ necessarily real understanding.

Strong hypothesis
According to the strong hypothesis, on the other hand, a sufficiently organized artificial system might not merely produce an appropriate response, but actually understand the situation.
In this case the system might possess mental states corresponding to concepts such as:

- “The user wants to go to the beach.”
- “The rain could prevent them from achieving their goal.”
- “I must suggest an alternative.”

The response would therefore not merely be the result of a simulation of understanding: the system would actually understand what it is talking about.
Intelligent behavior + real understanding.

The fundamental difference is therefore this:

- Weak hypothesis: the machine can behave as if it understands.
- Strong hypothesis: the machine could actually understand.

Important: these are not necessarily two different types of machine, but two philosophical interpretations of what happens behind apparently intelligent behavior.

The distinction between the weak hypothesis and the strong hypothesis can therefore be summarized as follows:

- Weak hypothesis: a machine can simulate intelligent behavior without this demonstrating the presence of a mind.
- Strong hypothesis: an appropriately organized machine could actually possess cognitive abilities and mental states.

Current AI systems can display very advanced abilities, but this does not in itself constitute evidence of the presence of consciousness, understanding or subjective experience.
We do not in fact have a shared scientific criterion that makes it possible to establish with certainty the presence of an artificial mind (cf. Butlin et al., 2023).
The distinction between the weak and strong hypotheses therefore remains primarily a philosophical question: it concerns what we can legitimately attribute to a system on the basis of its behavior and internal organization.

And it is precisely this problem that one of the most famous thought experiments in the philosophy of Artificial Intelligence addresses: **John Searle's Chinese room**.

The starting question is simple:
If a machine uses symbols correctly and produces appropriate responses, does that necessarily mean it understands their meaning?

The Chinese room experiment

The famous Chinese room thought experiment, proposed by philosopher John Searle in 1980, is a criticism of so-called Strong AI, understood as the thesis that executing the appropriate program would be sufficient to produce genuine understanding (Searle, 1980).

Searle uses the experiment to argue that a system can correctly manipulate symbols and produce appropriate responses without this being sufficient, on its own, to demonstrate the presence of understanding.

Imagine a person locked in a room who does not know Chinese. However, the person has a large set of instructions written in a language they understand.

The instructions might say, for example:

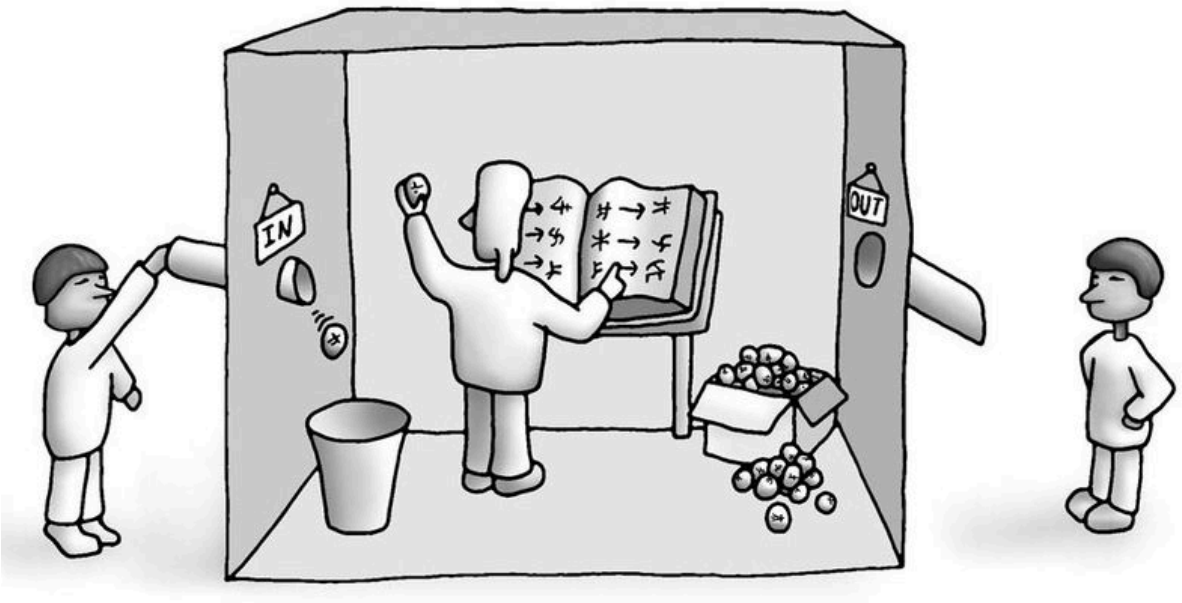
- “If you receive this symbol, return this other symbol”.
- “If you receive this sequence of symbols, produce this sequence”.

Questions written in Chinese are passed in from outside. The person does not know the meaning of the symbols, but carefully follows the instructions and returns appropriate sequences of Chinese characters.

To someone observing the room from outside, the system appears to understand Chinese:

- It receives questions in Chinese.
- It produces correct answers in Chinese.

However, the person inside the room does not understand what the symbols mean: they manipulate them solely on the basis of their shape and the rules received (Searle, 1980).



Searle uses this situation to distinguish syntax and semantics:

- Syntax concerns the rules by which symbols are manipulated.
- Semantics concerns the meaning attributed to those symbols.

According to Searle, executing syntactic rules is not sufficient, on its own, to produce semantic understanding (Searle, 1980).
The Chinese room therefore constitutes a criticism of the thesis that executing an appropriate program is sufficient to possess a mind.
The experiment does not claim to prove that a machine can never understand or possess mental states. Instead, it argues for a more limited thesis:
Executing a program is not, in itself, sufficient to constitute or produce semantic understanding (Searle, 1980).

From this perspective, Searle distinguishes:

- Correctness of behavior and understanding.
- Syntax and semantics.
- Execution of a program and the presence of mental states.

The Chinese room therefore leaves a fundamental question open: when a system produces intelligent behavior, are we observing only a simulation of understanding or genuine understanding?

Structuralist hypothesis

In the debate on Artificial Intelligence, the distinction between the weak hypothesis and the strong hypothesis helps to focus on a fundamental question: can an artificial system only simulate cognitive abilities, or can it, at least in principle, actually possess mental states?

This opposition is useful, but it does not necessarily exhaust the problem.

- The weak hypothesis cautions against automatically attributing understanding or mental states to a system merely because it displays intelligent behavior.
- The strong hypothesis, on the other hand, admits the possibility that a suitably organized artificial system could actually possess cognitive abilities and mental states.

However, a further question remains open:
From which characteristics of a system's organization might cognitive or mental properties emerge?

To address this question, we can formulate, for interpretive purposes, a perspective that we will call the structuralist hypothesis here.

This is not a third canonical position opposed to the weak hypothesis and the strong hypothesis, but a model through which to analyze systems by focusing not only on what they do, but above all on how they are organized.
The structuralist hypothesis therefore does not primarily ask whether a machine “simulates” a mind or “really is” a mind. The question becomes:
“What kind of organization must a system possess for it to make sense to attribute certain cognitive or mental properties to it?”

From this perspective, it is not enough to observe a system's computing power, the number of its components or how well it can imitate human behavior.
Greater complexity, on its own, does not automatically imply a mind.
Instead, it becomes important to study how the system:
- Maintains and changes its internal states.
- Receives and integrates information.
- Uses previous information.
- Regulates its behavior over time.
- Builds representations or models of the context.
- Interacts with the environment.
- Coordinates its components within an overall organization.

To understand this idea, we can start from a very simple distinction: not all systems that change state thereby possess cognitive abilities.

A stone rolling down a hill continuously changes position, speed and direction.
Its behavior depends on the physical conditions to which it is subjected: gravity, friction, the slope of the ground and impacts. The stone does not maintain a goal, does not use information about the environment to actively change its behavior and does not retain an internal reference state to achieve or preserve.

A thermostat, on the other hand, introduces a different form of organization.
It has an internal configuration, for example a set temperature, receives information through a sensor and uses the difference between the measured temperature and the desired temperature to change its behavior.

If the temperature falls below a certain threshold, the heating is switched on; when the desired temperature is reached, it is switched off.
The thermostat does not necessarily possess a mind, and there is no need to attribute an understanding of temperature to it.
However, it has a characteristic that the stone does not: it uses information from the environment in relation to an internal reference state to regulate its behavior (cf. Ashby, 1956).

From the perspective of the structuralist model, we can therefore distinguish, for analytical purposes, at least three different levels:

- Regulation.
- Cognition.
- Consciousness.

These levels should not be interpreted as a universally accepted scientific scale, but as a conceptual tool for distinguishing properties that are often confused.

Regulation
A regulatory system uses information about its own state or the environment to maintain or achieve certain conditions.
The thermostat is an elementary example.
The system has a reference state, compares that state with the information received and changes its behavior accordingly (Ashby, 1956).
Regulation therefore implies a form of control, but does not necessarily require a complex representation of the world, planning abilities or subjective experience.

Cognition
At a more complex level, we can speak of systems that do not merely react to a single condition, but use information to distinguish situations, anticipate consequences, compare alternatives and plan actions.

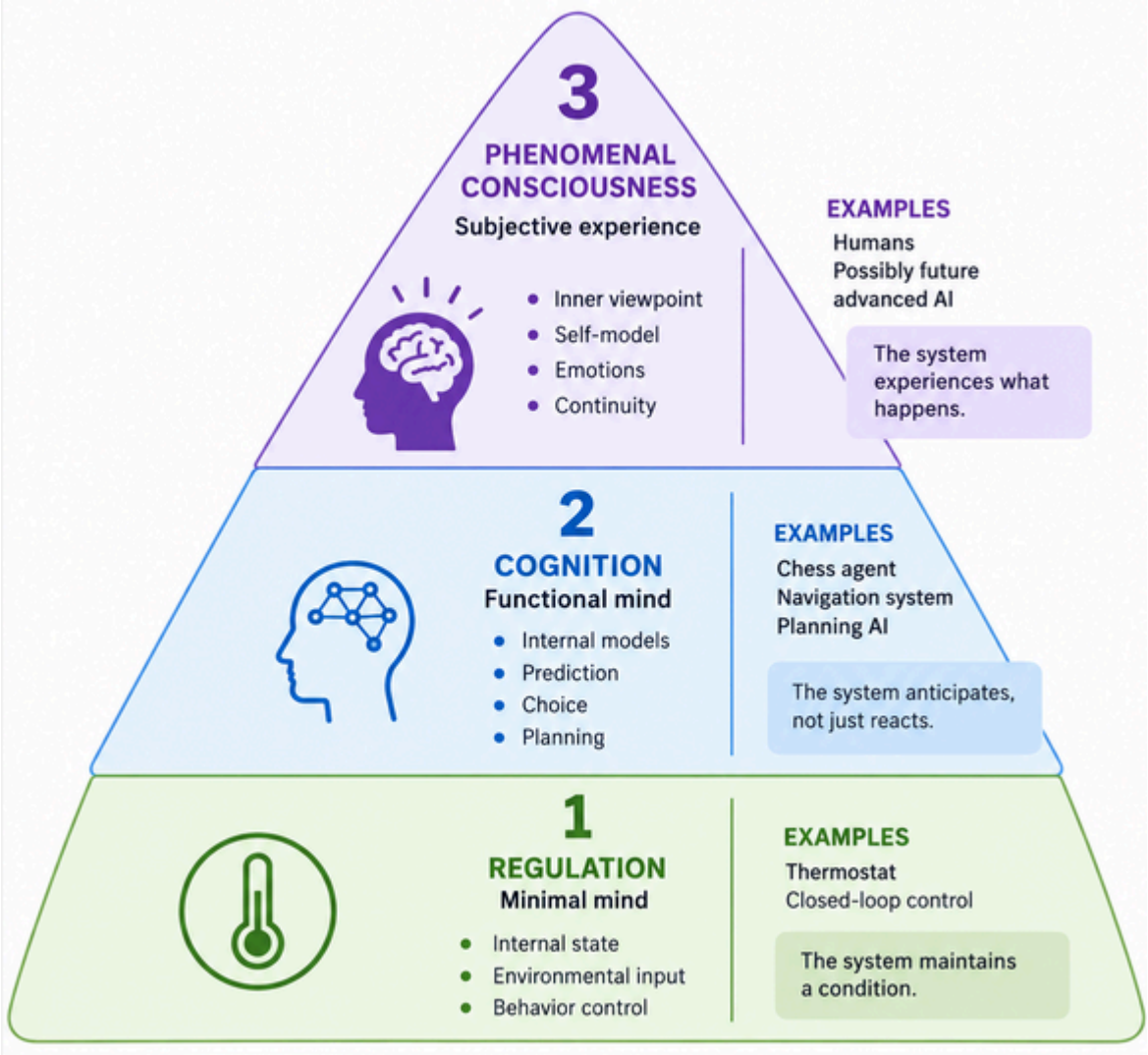
- A navigation system, for example, can represent different routes, estimate their travel times, consider obstacles and choose a solution according to the destination.
- A chess-playing program can represent different board configurations, estimate possible future developments and choose a move according to a goal.

From the perspective of this model, systems of this kind have a richer cognitive organization than a simple regulatory system.
We can therefore speak of cognitive abilities in a functional sense without automatically concluding that the system possesses subjective experience.

Cognitive competence does not necessarily coincide with consciousness.
A system can classify, predict, plan, decide or solve problems without these abilities constituting, on their own, evidence of the existence of an internal experience (cf. Butlin et al., 2023).

Consciousness
A further problem concerns phenomenal consciousness, that is, the existence of a subjective experience.

In this case, it would not be enough for a system to process information or effectively regulate its behavior.
There would have to be something we could describe as an internal point of view: not just information processed by the system, but an experience for the system itself (Butlin et al., 2023).
In human beings, consciousness is associated with an extremely complex network of phenomena: perception, memory, emotions, temporal continuity, representation of one's own body, self-reference and the construction of a self-model.



However, we do not know which of these characteristics are necessary or sufficient for subjective experience to emerge (Butlin et al., 2023).
The structuralist hypothesis therefore adopts a cautious position: if properties such as consciousness and subjectivity depend on the physical organization of a system, reproducing a single function is not enough to automatically obtain a complete mind.
The question concerns the overall organization: the way in which different processes are integrated, maintained over time and related to the environment and to the state of the system itself.

This perspective thus seeks to avoid two opposing simplifications.

- Naive functional reductionism: Thinking that implementing a particular function or obtaining a certain behavior is sufficient for a complete mind to exist automatically.
- Radical dualism: Considering the mind as something completely separate from physical processes and therefore, in principle, not explainable through the organization of a material system.

In the hypothesis proposed here, the mind is considered dependent on physical processes, but this does not mean it can be trivially reduced to a single operation or a simple sequence of calculations.

It therefore becomes necessary to ask not only whether a system computes, but:
- What it represents.
- How it integrates information.
- Which internal states it maintains.
- How it uses its previous history.
- How it anticipates possible future states.
- How it interacts with the environment.
- Whether it has a self-model.
- How these components are coordinated within a single system.

This perspective is compatible with two possibilities.
- It might be possible in the future to build artificial systems with much richer forms of mentality than current ones.
- Or it might emerge that some properties of the mind require forms of organization that the artificial systems known today do not possess.

It is therefore not necessary to assume either that artificial consciousness is impossible, or that it is an inevitable consequence of increasing computing power or model complexity (cf. Butlin et al., 2023).

Organization becomes the central question.
From an ethical point of view as well, it is important to keep different properties separate.
- A system can be highly autonomous without being conscious.
- It can make decisions that are difficult to interpret without having intentions of its own.
- It can have an enormous impact on society without possessing interests or subjective experience.

The following therefore do not automatically coincide:
- Operational autonomy.
- Cognitive ability.
- Consciousness.
- Social impact.
- Responsibility.
- Possible moral status.

For this reason, instead of immediately asking “Is this machine a person or not?”, it may be more useful to ask “What kind of system is this machine?”.

And therefore:

- What form of regulation it has.
- What information it maintains.
- What kind of representations it builds.
- How autonomous it is.
- How it uses its previous experience.
- How it interacts with the environment.
- How much it depends on designers and external infrastructures.
- What impact it has on human beings.
- What properties can be attributed to its specific organization.

The structuralist hypothesis, formulated in this way, does not solve the problem of the mind. Instead, it proposes shifting the question.
Rather than simply asking “Does this system possess a mind?”, we can ask “What properties emerge from the specific organization of this system?”

The focus thus shifts from observable behavior alone to the structure that makes it possible: architecture, internal states, information integration, temporal dynamics, relationship with the environment and capacity for self-regulation.

It is therefore not enough to observe what a system does.

To understand which properties it is reasonable to attribute to it, we must also try to understand how the system that produces that behavior is organized.

Supplementary material

Exercises

Automation, autonomy, adaptivity and learning

Consider four systems used in a supermarket.

- System A: An automatic door opens whenever a sensor detects a person.
- System B: A program autonomously decides which self-checkout to open based on queue length.
- System C: A system changes its predictions over time about the quantity of products to order using data from previous sales.
- System D: Every evening at 22:00, a software program executes a predefined procedure to create a backup copy of the data.

For each system:

- Identify which behaviors are simply automatic.
- Identify where a form of autonomy is present.
- Identify where adaptivity is present.
- Identify which system actually uses a process of learning from data.
- Explain why autonomy, adaptivity and learning are not synonyms.
- Can a system be autonomous without learning? Use one of the cases to justify your answer.
- Can a system change its behavior without using machine learning? Justify your answer.
- Indicate which of the four systems you would consider an AI system according to the operational definition proposed in the handout and explain the criterion used.
- Finally, formulate a definition of AI that does not use vague expressions such as “a machine that thinks”.

ANI, AGI or ASI?

A company presents three hypothetical systems.

- System Alpha: it analyzes satellite photographs and recognizes forest fires with very high performance, but is unable to perform other tasks.
- System Beta: it can autonomously learn new activities, move from image analysis to programming, study a new domain and transfer knowledge acquired from one problem to another with flexibility comparable to that of humans.
- System Gamma: it possesses all the abilities of System Beta, but consistently surpasses the best human beings in scientific research, planning, creativity and solving complex problems.

Classify the three systems as ANI, AGI or ASI.

- Justify each classification with reference to the breadth of abilities, not simply performance.
- Explain why System Alpha would remain ANI even if it were much more accurate than any human being at its task.
- Identify the fundamental characteristic that separates Alpha from Beta.

- Identify the characteristic that separates Beta from Gamma.
- Explain why increasing Alpha's computing power would not be sufficient, on its own, to turn it into AGI.
- Which of the three types of AI currently exist according to the distinction proposed in the handout?
- Why must AGI and ASI be treated with caution when discussing the abilities of contemporary systems?
- Explain why this classification primarily concerns the generality and transferability of abilities.

Explicit rules or learning from data?

A company must develop two systems.

- Problem A: determining whether a reimbursement request complies with a set of regulations consisting of precise and publicly defined conditions.
- Problem B: automatically recognizing whether an audio recording contains applause, traffic noise, music or a human voice.

You must decide which approach to use primarily for each problem:
Symbolic AI or AI based on learning from data.

- Choose the most suitable approach for Problem A and justify your choice.
- Choose the most suitable approach for Problem B and justify your choice.
- For Problem A, indicate what knowledge should be represented explicitly.
- Explain why the ability to reconstruct the reasoning might be important in this case.
- For Problem B, explain why it would be difficult to manually describe all the necessary rules.
- Describe what data could be used during training.
- Explain what it means to generalize to new audio never encountered during training.
- Imagine that the dataset contains very few examples recorded in very noisy environments. What problem might arise?
- Compare the two approaches in terms of:
    - Interpretability.
    - Flexibility.
    - Data dependency.
    - Handling unforeseen situations.
- Finally, propose a possible neuro-symbolic solution that combines learning from data and explicit rules.

What does it mean to say that a machine is “intelligent”?

A drone must deliver medicine to a mountainous area.
During the flight:

- It analyzes images from the cameras.
- It evaluates different routes.
- It avoids obstacles.
- It takes the available battery charge into account.
- It changes its trajectory if the wind changes.
- It must reach the destination in the shortest possible time without compromising safety.

Analyze the system through the four conceptions of AI studied in the handout.

- What would it mean to design the drone so that it thinks humanly?
- What would it mean, instead, to make it think rationally?
- What would it mean to evaluate it according to the criterion of acting humanly?
- What would it mean to evaluate it according to the criterion of acting rationally?
- Which of the four perspectives do you consider most relevant to this system? Justify your answer.
- Could the drone choose a completely different route from the one chosen by a human pilot and still be rational? Explain.
- Distinguish precisely between thinking rationally and acting rationally.
- Imagine that the drone perfectly imitates the decisions of a human pilot, including some of their habitual errors. Would it necessarily be rational?
- Now imagine that it produces very effective decisions through a procedure completely different from the human one. Would it be correct to say that it “thinks like a person”?
- Conclude by explaining why simply asking “Is it intelligent?” can be less precise than specifying the criterion according to which we are evaluating the system.

Intelligent behavior, cognition and consciousness

Imagine a robot intended for exploring a planet.

The robot:

- Maintains a map of the environment.
- Remembers the areas already explored.
- Compares alternative routes.
- Predicts which areas might be dangerous.
- Autonomously plans its movements.
- Changes its plan when it encounters an obstacle.
- Communicates with operators using natural language.
- Can describe its own state, for example: “My battery is almost depleted, so I must reach the base.”

A researcher states:
“It is clear that this robot understands where it is and is conscious of its own situation.”

Critically analyze the statement.

- Which observable abilities of the robot can be considered cognitive in a functional sense?
- Identify any forms of regulation present in the system.
- Explain why planning, predicting and deciding do not automatically demonstrate the presence of consciousness.
- According to the distinction proposed in the handout, what would be required to speak of phenomenal consciousness?
- Analyze the researcher's statement according to the weak AI hypothesis.
- How might it instead be interpreted according to the strong hypothesis?
- Does the fact that the robot says “I must reach the base” necessarily demonstrate that it possesses an intention in the human sense of the term? Justify your answer.

- Explain why extremely convincing linguistic behavior does not, on its own, solve the problem of understanding.
- Using the structuralist perspective, indicate which characteristics of the robot's internal organization you would want to know before attributing stronger cognitive properties to it.
- Clearly distinguish:
    - Operational autonomy.
    - Cognition.
    - Understanding.
    - Consciousness.
    - Impact of decisions.
- Explain why these properties should not automatically be considered equivalent.
- Finally, formulate a cautious conclusion about which properties we can attribute to the robot on the basis of the available information alone.

Bibliography

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

Ashby, W. R. (1956). *An introduction to cybernetics*. Chapman & Hall. https://ashby.info/Ashby-Introduction-to-Cybernetics.pdf

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2308.08708

Garcez, A. d'A., & Lamb, L. C. (2020). *Neurosymbolic AI: The 3rd wave* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2012.05876

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

McCarthy, J. (2007). *What is artificial intelligence?*. Stanford University. https://www-formal.stanford.edu/jmc/whatisai.pdf

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2024). Position: Levels of AGI for operationalizing progress on the path to AGI. *Proceedings of Machine Learning Research, 235*, 36308–36321. https://proceedings.mlr.press/v235/morris24b.html

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4ª ed.). Pearson. https://aima.cs.berkeley.edu/

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Further reading

- Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, 4th ed., Pearson, 2021 — The main academic reference for this handout. It provides a broad introduction to intelligent agents, search, knowledge representation, reasoning, decision making, machine learning, natural language processing, robotics, and the philosophy and ethics of AI. It is also the direct reference for the distinction between thinking humanly, thinking rationally, acting humanly, and acting rationally that structures a substantial part of the chapter.
- Melanie Mitchell, Artificial Intelligence: A Guide for Thinking Humans, new ed. with a new preface, Picador, 2025 — Probably the best accessible complement for students. Mitchell explains the history, achievements, and limitations of contemporary AI while repeatedly questioning what apparently impressive performance actually tells us about general intelligence, understanding, reasoning, and human-like cognition. This makes it especially relevant to the handout's distinction between observable intelligent behavior and stronger claims about understanding or mentality.
- Margaret A. Boden, Artificial Intelligence: A Very Short Introduction, Oxford University Press, 2018 — A concise but theoretically rich introduction covering what AI is, general intelligence, language, creativity, neural networks, robotics, and the philosophical question of whether machines can really be intelligent or conscious. It is particularly well suited to the handout because it bridges technical history and philosophy without requiring an advanced mathematical background. OUP Academic
- Nello Cristianini, The Shortcut: Why Intelligent Machines Do Not Think Like Us, CRC Press, 2023 — Particularly relevant to the transition from symbolic AI to data-driven AI. Cristianini explains how modern systems can acquire highly effective behavior from statistical regularities without reproducing human reasoning processes. The book therefore supports one of the chapter's central ideas: a system may perform intelligently while using mechanisms very different from those used by humans.
- John R. Searle, Minds, Brains and Science, Harvard University Press, 1984 — The strongest philosophical complement to the sections on Strong AI, understanding, syntax versus semantics, and the Chinese Room argument. Searle directly addresses whether executing a program is sufficient for genuine understanding and discusses the implications of artificial intelligence for philosophy of mind and cognitive science.

Editorial notes

In producing these materials, I have used generative artificial intelligence tools to support the writing process, particularly to improve the form of the text, reorganize content, refine wording, and speed up certain editorial tasks.

As I work independently on the production of these materials, I try to automate any activities that can reasonably be automated, so that I can devote more time to research, design, and content development.

However, artificial intelligence does not determine the content of this work: the choice of topics, structure, ideas, interpretations, examples, and teaching approach are developed by me. AI is therefore used primarily as a tool to support production and formal revision, while authorship and responsibility for the design and content remain mine.

Unless otherwise indicated, this material is distributed under the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International license (CC BY-NC-SA 4.0).

You are therefore free to share, redistribute, adapt, and create derivative works based on this material, provided that appropriate credit is given, the material is not used for commercial purposes, and any modified or derivative versions are distributed under the same license.

