Artificial Intelligence

'I visualise a time when we will be to robots what dogs are to humans,… and I am rooting for the machines.' — Claude Shannon

Fred Guth

2022-09-30

Table of contents

This chapter defines artificial intelligence, presents the epistemological differences of intelligent agents in history, and discusses their consequences to machine learning theory.

Artificial Intelligence

AI is the branch of Computer Science that studies general principles of intelligent agents and how to construct them (**russell:2010?**).

This definition uses the terms *intelligence* and *intelligent agents*, so let us start from them.

What is intelligence?

Despite a long history of research, there is still no consensual definition of intelligence.[1] Whatever it is, though, humans are particularly proud of it. We even call our species *homo sapiens*, as intelligence was an intrinsic human characteristic.

In this dissertation:

Intelligence is the ability to predict a course of action to achieve success in specific goals.

[1] For a list with 70 definitions of intelligence, see .

Intelligent Agents

Under our generous definition, intelligence is not limited to humans. It applies to any agent[2]: animal or machine. For example, a bacteria can perceive its environment through chemical signals, process them, and then produce chemicals to signal other bacteria. An air-conditioning can observe temperature changes, know its state, and adapt its functioning, turning off if it is cold or on if it is hot — *intelligence exempts understanding.* The air-conditioning does not comprehend what it is doing. The same way a calculator does not know arithmetics.

A strange inversion of reasoning

This competence without comprehension is what the philosopher Daniel Dennett calls *Turing's strange inversion of reasoning*[3]. The idea of a *strange inversion* comes from one of Darwin's 19th-century critics ((**mackenzie:1868?**) as cited by (**dennett:2009?**)):

> *In the theory with which we have to deal, Absolute Ignorance is the artificer; so that we may enunciate as the fundamental principle of the whole system, that,* ***in order to make a perfect and beautiful machine, it is not requisite to know how to make it****. This proposition will be found, on careful examination, to express, in condensed form, the essential purport of the [Evolution] Theory, and to express in a few words all Mr Darwin's meaning; who, by* ***a strange inversion of reasoning****, seems to think Absolute Ignorance fully qualified to take the place of Absolute Wisdom in all of the achievements of creative skill.* — Robert MacKenzie

Counterintuitively to (**mackenzie:1868?**) and many others to this date, intelligence can emerge from absolute ignorance. Turing's strange inversion of reasoning comes from the realisation that his automata can perform calculations by symbol manipulation, proving that it is possible to build agents that behave

[2] An agent is anything that perceives its environment and acts on it.

[3] In his work, Turing discusses if computers can "think", meaning to examine if they can perform indistinguishably from the way thinkers do.

intelligently, even if they are entirely ignorant of the meaning of what they are doing (**turing:2007?**).

Dreaming of robots

From mythology to Logic

The idea of creating an intelligent agent is perhaps as old as humans. There are accounts of artificial intelligence in almost any ancient mythology: Greek, Etruscan, Egyptian, Hindu, Chinese (**mayor:2018?**). For example, in Greek mythology, the story of the bronze automaton of Talos built by Hephaestus, the god of invention and blacksmithing, first mentioned around 700 BC.

This interest may explain why, since ancient times, philosophers have looked for *mechanical* methods of reasoning. Chinese, Indian and Greek philosophers all developed formal deduction in the first millennium BC.In particular, Aristotelian syllogism, *laws of thought*, provided patterns for argument structures to yield irrefutable conclusions, given correct premises. These ancient developments were the beginning of the field we now call *Logic*.

Rationalism: The Cartesian view of Nature

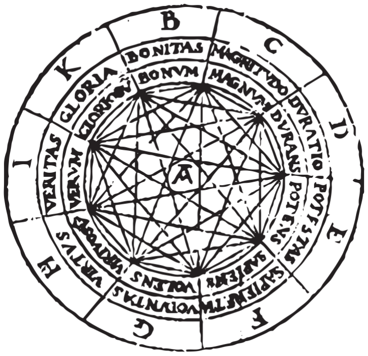


Figure 1: Example of one of Lull's Ars Magna's paper discs.

In the 13th century, the Catalan philosopher Ramon Lull wanted to produce all statements the human mind can think. For this task, he developed *logic paper machines*, discs of paper filled with esoteric coloured diagrams that connected symbols representing statements. Unfortunately, according to (**gardner:1959?**), in a modern reassessment of his work, *"it is impossible, perhaps, to avoid a strong sense of anticlimax"* (**gardner:1959?**). With megalomaniac self-esteem that suggests psychosis, his delusional sense of importance is more characteristic of cult founders. On the bright side, his ideas and books exerted some magic appeal that helped them be rapidly disseminated through all Europe (**gardner:1959?**).

Lull's work greatly influenced Leibniz and Descartes, who, in the 17thcentury, believed that all rational thought could be mechanised. This belief was the basis of **rationalism**, the epistemic view of the *Enlightenment* that regarded reason as the sole source of knowledge. In other words, they believed that reality has a logical structure and that certain truths are *self-evident*, and all truths can be derived from them.

There was considerable interest in developing artificial languages during this period. Nowadays, they are called formal languages.

> *If controversies were to arise, there would be no more need for disputation between two philosophers than between two accountants. For it would suffice to take their pencils in their hands, to sit down to their slates, and to say to each other:* ***Let us calculate.*** — Gottfried Leibniz

The rationalist view of the world has had an enduring impact on society until today. In the 19thcentury, George Boole and others developed a precise notation for statements about all kinds of objects in Nature and their relations. Before them, Logic was philosophical rather than mathematical. The name of Boole's masterpiece, *"The Laws of Thought"*, is an excellent indicator of his Cartesian worldview.

At the beginning of the 20th century, some of the most famous mathematicians, David Hilbert, Bertrand Russel, Alfred Whitehead, were still interested in formalism: they wanted mathematics to be formulated on a solid and complete logical foundation. In particular, Hilbert's *Entscheidungs Problem* (decision problem) asked if there were limits to mechanical Logic proofs (**chaitin:2006?**).

Kurt Gödel's incompleteness theorem (1931) proved that any language expressive enough to describe arithmetics of the natural numbers is either incomplete or inconsistent. This theorem imposes a limit on logic systems. There will always be truths that will not be provable from within such languages: there are "true" statements that are undecidable.

Alan Turing brought a new perspective to the *Entscheidungs Problem*: a function on natural numbers that an algorithm in a formal language cannot represent cannot be computable (**chaitin:2006?**). Gödel's limit appears in this context as functions that are not computable, no algorithm can decide whether another algorithm will stop or not (the halting problem). To prove that, Turing developed a whole new general theory of computation: what is computable and how to compute it, laying out a blueprint to build computers, and making possible Artificial Intelligence research as we know it. An area in which Turing himself was very much invested.

Empiricism: The sceptical view of Nature

The response to **rationalism** was **empiricism**, the epistemological view that knowledge comes from sensory experience, our perceptions of the world. Locke explains this with the peripatetic axiom[4]: *"there is nothing in the intellect that was not previously in the senses"* (**williams:2020?**). Bacon, Locke and Hume were great exponents of this movement, which established the grounds of the scientific method.

Figure 2: David Hume, Scottish Enlightenment philosopher, historian, economist, librarian and essayist.

David Hume, in particular, presented in the 18th century a radical empiricist view: reason only does not lead to knowledge. In (**hume:2009?**), Hume distinguishes *relations of ideas*, propositions that derive from deduction and *matters of facts*, which rely on the connection of cause and effect through experience (induction). Hume's critiques, known as *the Problem of Induction*, added a new slant on the debate of the emerging scientific method.

From Hume's own words:

> *The bread, which I formerly eat, nourished me; that is, a body of such sensible qualities was, at that time, endued with such secret powers: but does it follow, that other bread must also nourish me at another time, and that like sensible qualities must always be attended with like secret powers? The consequence seems nowise necessary.* — David Hume

[4] This citation is the principle from the Peripatetic school of Greek philosophy and is found in Thomas Aquinas' work cited by Locke.

There is no logic to deduce that the future will resemble the past. Still, we expect uniformity in Nature. As we see more examples of something happening, it is *wise* to expect that it will happen in the future just as it did in the past. There is, however, no *rationality*[5] in this expectation.

Hume explains that we see conjunction repeatedly, "bread" and "nourish", and we expect *uniformity in Nature*; we hope that "nourish" will always follow "eating bread"; When we fulfil this expectancy, we misinterpret it as causation. In other words, we *project* causation into phenomena. Hume explained that this connection does not exist in Nature. We do not "see causation"; we create it.

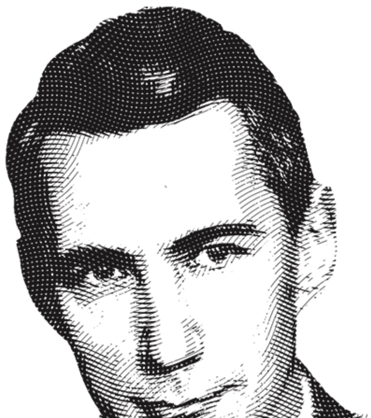
This projection is *Hume's strange inversion of reasoning* (**huebner:2017?**): We do not like sugar because it is sweet; sweetness exists because we like (or need) it. There is no sweetness in honey. We wire our brain so that glucose triggers a labelled desire we call sweetness. As we will see later, sweetness is *information.* This insight shows the pattern matching nature of humans. Musicians have relied on this for centuries. Music is a sequence of sounds in which we expect a pattern. The expectancy is the tension we feel while the chords progress. When the progression finally *resolves*, forming a pattern, we release the tension. We feel pattern matching in our core. It is very human, it can be beneficial and wise, but it is, *stricto sensu*, *irrational.*

The epistemology of the sceptical view of Nature is science: to weigh one's beliefs to the evidence. Knowledge is not absolute truth but justified belief. It is a Babylonian epistemology.

In rationalism, Logic connects knowledge and good actions. In empiricism, the connection between knowledge and justifiable actions is determined by probability. More specifically, Bayes' theorem. As Jaynes puts it, probability theory is the "Logic of Science" . [6]

The birth of AI as a research field

In 1943, McCulloch and Pitts, a neurophysiologist and a logician, demonstrated that neuron-like electronic units



[5] In the philosophical sense.

[6] The Bayes' theorem is attributed to the Reverend Thomas Bayes after the posthumous publication of his work. By the publication time, it was an already known theorem, derived by Laplace.

could be wired together, act and interact by physiologically plausible principles and perform complex logical calculations (**russell:2010?**). Moreover, they showed that any computable function could be computed by some network of connected neurons (**mcculloch:1943?**). Their work marks the birth of ANNs, even before the field of AI had this name. It was also the birth of **Connectionism**, using artificial neural networks, loosely inspired by biology, to explain mental phenomena and imitate intelligence.

Their work inspired John von Neumann's demonstration of how to create a universal Turing machine out of electronic components, which lead to the advent of computers and programming languages. Ironically, these advents hastened the ascent of the formal logicist approach called **Symbolism**, disregarding Connectionism.

In 1956, John McCarthy, Claude Shannon, Marvin Minsky and Nathaniel Rochester organised a 2-month summer workshop in Dartmouth College to bring researchers of different fields concerned with *"thinking machines"* (cybernetics, information theory, automata theory). The workshop attendees became a community of researchers and chose the term *"artificial intelligence"* for the field.

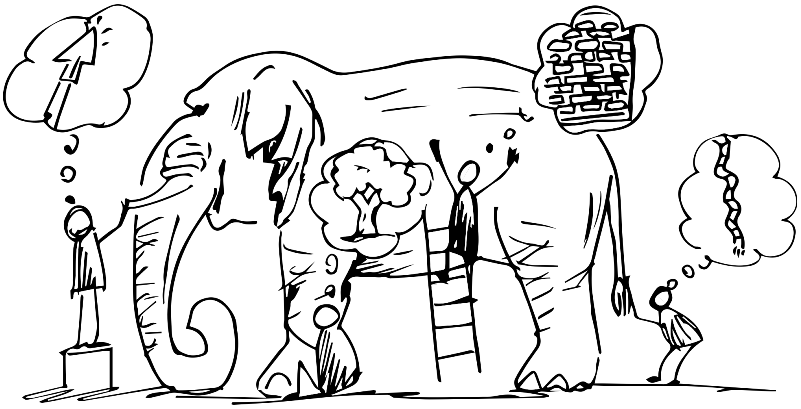


Figure 4: The Blind Men and the Elephant.

It was six men of Indostan
To learning much inclined,

Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind
—*John Godfrey Saxe,*

The Blind Men and the Elephant

Building Intelligent Agents

Anatomy of intelligent agents

Like the blind men in the parable, an intelligent agent shall model her understanding of Nature from limited sensory data.

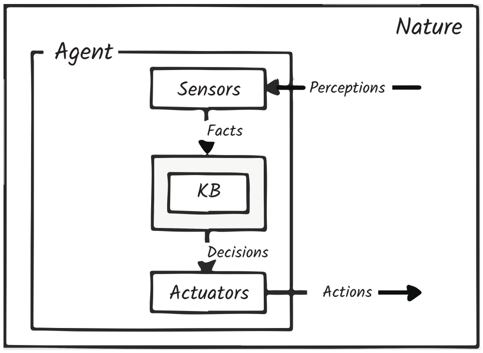


Figure 5: Anatomy of an Intelligent Agent. Inspired by art in ((**russell:2010?**))

Thus, an agent perceives her environment with sensors, treat sensory data as facts and use these facts to possibly update her model of Nature, use the model to decide her actions, and acts via her actuators. In a way, agents continually communicate with Nature in a perception/action conversation ([fig-anatomy]).

The expected result of this conversation is a change in the agent's KB, therefore in her model and, more importantly, her future decisions. The model is an abstraction of how the agent "thinks" the world is (her "mental picture" of the environment). Therefore, it should be consistent with it: if something is true in Nature, it is equally valid, *mutatis mutandis*, in the model. A Model should also be as simple as possible so that the agent can make decisions that maximise a chosen performance measure, but not simpler. As the agent knows more about Nature, less it gets surprised by it.

This rudimentary anatomy is flexible enough to entail different epistemic views, like the rationalist (mathematical) and the empiricist (scientific); different approaches to how to implement the knowledge base (it can be learned, therefore updatable, or it can be set in stone from an expert prior knowledge); and also from how to implement it (a robot or software).

Noteworthy, though, is that the model that transforms input data into decisions should be the target of our focus.

Symbolism

Symbolism is the pinnacle of rationalism. In the words of Thomas Hobbes, one of the forerunners of rationalism, *"thinking is the manipulation of symbols and reasoning is computation".* Symbolism is the approach to building intelligent agents that does just that. It attempts to represent knowledge with a formal language and explicitly connects the knowledge with actions. It is *competence from comprehension.* In other words, it is *programmed.*

Even though McCulloch and Pitts work on artificial neural networks predates Von Neumann's computers, Symbolism dominated AI until the 1980s. It was so ubiquitous that symbolic AI is even called "good old fashioned AI" (**russell:2010?**).

The symbolic approach can be traced back to Nichomachean Ethics (**aristotle:2000?**):

> *We deliberate not about ends but means. For a doctor does not deliberate whether he shall heal, nor an orator whether he shall persuade, nor a statesman whether he shall produce law and order, nor does anyone else deliberate about his end. They assume*

> *the end and consider how and by what means it is to be attained; and if it seems to be produced by several means, they consider by which it is most easily and best produced, while if it is achieved by one only they consider how it will be achieved by this and by what means this will be achieved, till they come to the first cause, which in the order of discovery is last.*
>
> — Aristotle

This perspective is so entrenched that (**russell:2010?**) still says: *"(…) Only by understanding how actions can be justified can we understand how to build an agent whose actions are justifiable"*; even though, in the same book, they cover machine learning (which we will address later in this chapter) without noticing it is proof that there are other ways to build intelligent agents. Moreover, it is also a negation of competence without comprehension. It seems that even for AI researchers, the strange inversion of reasoning is uncomfortable ([ch:introduction]).

All humans, even those in prisons and under mental health care, think their actions are justifiable. Is that not an indication that we rationalise our actions *ex post facto*? We humans tend to think our rational assessments lead to actions, but it is also likely possible that we act and then rationalise afterwards to justify what we have done, fullheartedly believing that the rationalisation came first.

Claude Shannon's Theseus

After writing what is probably the most important master's dissertation of the 20th century and "inventing" IT, what made possible the Information Age we live in today, Claude Shannon enjoyed the freedom to pursue any interest to which his curious mind led him (**soni:2017?**). In the 1950s, his interest shifted to building artificial intelligence. He was not a typical academic, in any case. A lifelong tinkerer, he liked to "think" with his hand as much as with his mind. Besides developing an algorithm to play chess (when he even did not have a computer to run it), one of his most outstanding achievements in AI was Theseus, a robotic maze-solving mouse.[7]

To be more accurate, Theseus was just a bar magnet covered with a sculpted wooden mouse with copper whiskers; the maze was the "brain" that solved itself (**klein:2018?**).

> *"Under the maze, an electromagnet mounted on a motor-powered carriage can move north, south, east, and west; as it moves, so does Theseus. Each time its copper whiskers touch one of the metal walls and complete the electric circuit, two things happen. First, the corresponding relay circuit's switch flips from"on" to "off," recording that space as having a wall on that side. Then Theseus rotates* 90° *clockwise and moves forward. In this way, it systematically moves through the maze until it reaches the target, recording the exits and walls for each square it passes through." — (**klein:2018?**).*

Symbolic AI problems

Several symbolic AI projects sought to hard-code knowledge about domains in formal languages, but it has always been a costly, slow process that could not scale.

Anyhow, by 1965, there were already programs that could solve any solvable problem described in logical notation (**russell:2010?**). However, hubris and lack of philosophical perspective made computer scientists believe that "intelligence was a problem about to be solved[8]."

Those inflated expectations lead to disillusionment and funding cuts[9] (**russell:2010?**). They failed to estimate the inherent difficulty in slating informal knowledge in formal terms: the world has many shades of grey. Besides, complexity theory had yet to be developed: they did not count on the exponential explosion of their problems.

[7] Many AI students will recognise in Theseus the inspiration to Russel and Norvig's Wumpus World .

[8] Marvin Minsky, head of the artificial intelligence laboratory at MIT (1967)

[9] Sometimes called *winters*.