# Chapter 4 - The Philosophy of AI

## On artificiality and intelligence

### 4.1 - Introduction

Artificial intelligence is still a controversial topic. Many people believe that it poses inherent dangers to our society and our survival as a species. Similarly, not everyone believes we can achieve the same intelligence, consciousness, and sentience in artificial beings. Even defining intelligence and artificiality can be tricky. Let us start with the 'intelligence' part of Artificial Intelligence.

### 4.2 - Intelligence and the Turing test

When quizzed on what makes people intelligent, they will say it means being good at math or a good problem solver. More broadly, they will argue that intelligent people are better at connecting disparate concepts and ideas, are more creative, and can think "outside the box." These are not especially useful definitions to a scientist because they are vague and ambiguous. Furthermore, we know that intelligence exists outside the context of human beings. Animal studies have shown us uniquely intelligent behaviors in other lifeforms, such as monkeys, birds, whales, butterflies, bees, and octopods. However, the logical mind was still the pinnacle of intelligence in the time of Alan Turing when he proposed his unique intelligence test. The Turing test has a human ask question, through some interface, to an agent to determine whether that agent is human or artificial. The interface is there to hide the physical embodiment of the AI because whether the AI appears human is irrelevant to whether it is intelligent, argued Turing. The human interrogator can ask any question to find evidence that the intelligence answering is human or artificial. Should the human mistake the AI for a human, the AI is said to have passed the Turing test. Since the questions do not necessarily involve academic skills, this test is less strict than our above definitions. However, it still relies on a human interaction style: dialogue in natural language. If this is not the natural way of interacting with humans, our AI will fail miserably, even if it has some rudimentary form of intelligence. Like asking your Nest thermostat what her views are on the inevitability of impermanence. Yet, we feel our IoT (Internet of Things) devices at least embody some intelligence. So, where do we draw the line?

|  |  |  |
| --- | --- | --- |
|  | Diagram  Description automatically generated with medium confidence |  |
|  | **Figure 4.1** - Credits to XCDF |  |

### 4.3 - Biological Intelligence

Intelligence as an adaptive strategy cannot be strictly a cognitive process, operating in total isolation on abstract symbols, applying the rules of logic and inference to them. Biological systems are embedded in their physical environments. Our thoughts and ideas ultimately depend on how well our sensory systems are adapted to extracting meaningful signals from our environment and how well we can actuate and influence that same environment. Sensing our environment is far from passive. It is a highly creative process in which what we ultimately learn about the environment through our senses is modified by previous experiences, context, and internal states. Together they combine into our perception of the world: sensory information augmented to include cognitive and behavioral relevance.

|  |  |  |
| --- | --- | --- |
|  | A monkey looking at a computer screen  Description automatically generated with low confidence |  |
|  | **Figure 4.2** - Interestingly, other apes are sometimes far superior at a human perceptual skill. Chimpanzees are far better at a game of memory where one must memorize the location of many objects. Research shows that chimpanzees can remember far longer sequences than humans. For example, the chimp shown above is very briefly shown digits placed randomly on the screen. When a white overlay hides them, the chimp needs to remember the sequence 1 to 9 and click the white overlays in the correct order. Surprisingly, on average, chimpanzees can remember much longer sequences than their human counterparts. |  |

We refer to our ability to reason with such perceptions as cognition. However, there is no clear boundary between perception and cognition. Imagine noticing that your friend looks sad today. You perceive your friend to be sad. The sensory inputs provide evidence for this perception: your friend is hunched over, frowning, sullen, and withdrawn. But it also requires cognition: factoring in context, cultural norms, and expectations will further increase our understanding of our friend's state of mind. The bottom line is that our friend's sadness is both a perceptual and cognitive construct.

Regardless of where we place the dividing line, eventually, the incoming signals drive us to engage and respond. Interaction with our environment requires effort since our actions can be highly complex and operate on many temporal and spatial scales. Like the count of Monte Christo, we can plan a set of actions long in advance, carefully planned and executed, to achieve our goal (in the count's case: revenge). Or we can just start throwing punches in response to the same sensory experience of being treated unjustly. In neuroscience, all this falls under the header of executive planning and action. Again, perception, cognition, and executive function/action form a continuous space without clear boundaries. By abstracting further and further away from the original senses, later stages of perception blend into what we more closely define as cognition. Similarly, cognition will slowly crystallize in executive functions and actions. However, there are some things that we can uniquely associate with the core three pillars of biological intelligence.

### 4.4 - Perception – detection, segmentation, localization, and recognition

All perceptual systems aim to isolate and segment signals that belong to a single observable entity from other entities and localize it in space relative to the observer. For example, in vision, we need to combine some signals as they all belong to the same individual and segment them from the background and another individual in our visual field. In addition, we need to keep track of the position of that individual relative to our position in space. Once all signals belonging to a particular observable entity have been combined, the input giving rise to all these signals needs to be identified and categorized. Similarly, we need to group all the soundwaves at a wide range of frequencies into a single voice produced by a single entity from many other voices simultaneously speaking in the background. The neural systems underlying perception are jointly responsible for encoding or representing our sensed environment: the numerous attributes that make up our environment are mapped onto a pattern of neural activity. The entire world around us is encoded by the joint activity of millions of neurons. Our conscious minds do not have access to the ground truth of physical reality, only the brain's interpretation of it. Perceptual systems are usually hierarchical. As the sensory information is processed by a series of cortical stages (i.e., different areas in the brain chained together through neural connections), the information represented in those areas becomes ever more abstract and complex. For example, the visual system starts by recognizing single spots of light on a dark background, combining them into line segments and edges. Edges combine into textures and textures into shapes with colors and patterns. This increase in complexity continues up to where areas placed further down the processing stream are uniquely sensitive to visual images of human faces and objects.

|  |  |  |
| --- | --- | --- |
|  | A picture containing icon  Description automatically generated |  |
|  | **Figure 4.3** - The squares A and B are an identical shade of gray. Our perception is fooled, thinking the shadow needs to be discounted from square B to get to its true gray value. Perception is a creative process where the brain needs to integrate many diverse sources of noise information to the best of its ability. Here, the presence of a shadow tricks the perceptual system into overestimating the brightness of B, reasoning that it is currently in a shadow; its actual appearance outside the shadow must be brighter. |  |

### 4.5 - Cognition - reasoning with concepts, unknowns, and probabilities

The signals biological systems need to operate on are far from perfect. They rely on noisy measurement systems (our senses) and are often ambiguous. As such, cognition has a probabilistic nature as it attempts to get as close to ground truth, by combining information from various sources, either within or between sensory modalities. But the natural dividing line between perception and cognition is that cognition can operate on abstract concepts with no sensory counterpart. We can derive meaningful conclusions from such abstract things as time, love, mathematics, and psychological constructs. Of course, one such abstraction humans show extremely high aptitude in is the understanding and production (through speech and writing) of language. Language tokenizes our perceptual experiences and abstract concepts in a set of symbols that we can combine in infinite ways to form new relationships. Take, for example, the simple utterance of one person to another: 'John will be back tomorrow.' A sentence of modest complexity, one could say. But the amount of information communicated is exceptionally high, even though it only uses a handful of words to express it. 'John' refers to an individual known to both the speaker and the listener. An almost endless number of different sensory inputs map to another person's single abstract conceptual placeholder, depending on how we might encounter John in the here and now. The same is true for the word 'back.' Again, both speakers have a mutual and shared understanding of what we refer to as 'back.' It refers to John's presence at a particular physical location, like an office. Finally, tomorrow suggests that we must consider the passage of time and understand that even though our senses currently do not provide us with any information about John, we can expect this to change within some time window. Finally, this mutual understanding or agreement in exchanging information relies on another ability that humans (and other social mammals) excel. With normal development, humans can imagine a world as if it were seen through someone else's senses. We can simulate this experience and understand what information is available to that individual, even if this information is not available to us, or vice versa. It bestows humans (and some other animals) with empathy and a theory of mind, traits most useful for cooperation among many individuals performing complex tasks that require them to operate as a unit. Many of the philosophical debates on AI are concerned with whether these higher-level abilities (language, theory of mind, and reasoning with abstract concepts) require a set of unique neural mechanisms and processes or whether they emerge automatically from perceptual systems increasing in complexity through evolution. Historically, language was usually the focal point of such discussions. Is the human ability and propensity for language something unique to humans and humans alone? Or can we find systems in other animals that would also support language if we would give it some more evolutionary time and necessity?

|  |  |  |
| --- | --- | --- |
|  | A close-up of a statue  Description automatically generated with medium confidence |  |
|  | **Figure 4.4** – Rodin's the Thinker. |  |

### 4.6 - Executive function - deciding between actions

Executive function refers to the mechanisms that combine our perceptual experience or cognitive understanding of it with the needs and goals of the system itself and our previous experiences with similar situations into action. It has some abilities to override more automatic, reflexive processes the brain would operate on without it, which seems to translate, to some, into the idea of humans having at least some free will over the actions they take. But regardless of 'free will' being a reality, the executive functions allow us to interact with our environment by moving us through it. This interaction can be taken quite literally, like walking to the subway to go to work. It can also be taken more figuratively, like vocalizing a thought, and therefore move and change our environment because it changes how individuals capable of understanding our vocalization behave.

|  |  |
| --- | --- |
|  |  |

### 4.7 - Complexity and Invariance

In the above example, there are several processes we could describe as embedding some form of intelligence. Which processes are easy for a computer to take on, and which are not? Some intelligent behavior is easy to port to modern computers that excel at rule-based cognitive tasks. We can quickly turn these rules into algorithms. Other intelligence behavior is not easy for a computer. Specifically, sensory and action systems have large state spaces and must cope with invariance, noise, uncertainty, and parallel processes. And these processes need to be learned through repeated experience. And learning through experience is not how we originally designed our computing systems.

|  |  |  |
| --- | --- | --- |
|  | A picture containing text, mammal, primate, posing  Description automatically generated |  |
|  | **Figure 4.5** - Phineas Gage is a fascinating case from the early days of neuroscience. While working blasting rock away to level the land for a train track, he accidentally had a large metal rod blown through his eye and out the top of his skull. Miraculously, he survived, but his whole personality had changed. He was described before the accident as a kind person. But after the accident, he became "gross, profane, coarse, and vulgar, to such a degree that his society was intolerable to decent people" according to one of the doctors studying him. We now know that Gage's frontal lobes were damaged to a point where the executive control and function were severely disrupted, making Gage extremely uninhibited in his actions. |  |

### 4.8 - Chess versus buying cookies

So, what is more challenging for a computer to do? Beating Magnus Carlsen in a game of chess? Buying your favorite cookies from the supermarket? To beat Magnus Carlsen in a chess game, we can use a brute force approach: compute every move. A brute force method is not complicated, provided you have sufficient memory. The average branching factor of the game tree of chess is ~31. This factor means that, on average, a board configuration during a chess game has about 31 allowable moves for the player. We can also compute the total number of possible and permissible board configurations. This 'state space' is about 4E36. So, a 4 with 36 zeros behind it. But even though the state space is vast, it is entirely deterministic: we know the rules and transitions between board configurations beforehand. In other words, from the computer's point of view, there is no uncertainty or probability to the game of chess.

Now, let us imagine buying our favorite cookies. You will have to get up and plan to go to the nearest supermarket. Do not forget to bring your wallet and wait for the lights to change before crossing the street. Find your way to the bakery aisle and find your favorite cookie. Go to the register, pay for the cookies, and eat them. Buying cookies is not a super-intelligent endeavor for an adult human being. But from the computer's point of view, it is utter chaos. There are so many unpredictable variables. What if the elevator is out of order? What if you do not have any money? What if the traffic light does not work? What if the store is closed? Even just finding your favorite cookie amongst all the other products if you are already in the supermarket is a complex problem for a computer algorithm to take on.

|  |  |  |
| --- | --- | --- |
|  | A person in a suit and tie  Description automatically generated with low confidence |  |
|  | **Figure 4.6** - On the left: chess grandmaster Magnus Carlsen. On the right: an Oreo cookie |  |

### 4.9 - Invariance and context

The distinction between the two tasks described above is that one is performed in a controlled environment with minimal inputs and outputs. Of course, there are an extremely high number of chess boards, but any board can be described in a very concise way: 64 numbers that represent the type of piece (if any) on any 64 positions of a chess board. Our AI playing chess does not need to know any beyond that, like their shape, the eye color of your opponent, etc., to play well and win. On the other hand, buying cookies is performed and embedded in an extraordinarily complex environment that is impossible to describe in any detail and is ever-changing. It is even unknown what information is necessary to buy cookies successfully. Therefore, our sensory systems deal with a lot of ambiguity in their inputs. There is nothing that is not subject to minor variations, and there are no invariant properties. Looking at the packaging of

|  |  |  |
| --- | --- | --- |
|  | A picture containing table, dining table  Description automatically generated |  |
|  | **Figure 4.7** - Top: there is an infinite number of ways one can design a chair. Yet our brains see all or perhaps most of them as a chair, even though their appearances vary greatly. Yet the brain has no trouble seeing them as the same. The visual image can change on viewing angle, lighting conditions, and distance. Even for the same object. |  |

Two cookies we can see very quickly that they are indeed the same type of cookie. But at the pixel level, they are nothing alike except for a general sense of hue. In our sensory world, two relationships make defining strict rules hard, if not outright impossible. Invariance is a many-to-one relationship. There are countless examples of chairs, and we can even see the same chair from an infinite number of light conditions, angles, and positions. Yet they still are all chairs: an object we can sit in. Context also relates to the ambiguity of sensory systems. The same bit of visual information can change its meaning by its larger context. Degraded sensory signals force us to rely on context to fill in the missing pieces. Another exceedingly complicated process to capture in terms of fixed rules and relationships. How our brain deals with these problems is fascinating. The current generation of neural networks to deal with these problems is beginning to offer insights into how both artificial and biological systems achieve such complex tasks.

### 4.10 - Chess Revisited

Let us return briefly to computers and humans playing chess. Computers do not have the same memory limitations humans have, allowing them to brute force solutions like Deep Blue. But why, then, are humans then still a match? Because even in something as highly rule-based as chess, grandmasters do not compute sequences of moves. Instead, they assign an intuitive value due to experience. The skill has been internalized, like tying your shoelaces.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | A picture containing text, crossword puzzle  Description automatically generated |  |
|  | **Figure 4.8** - I will admit to not being good enough at chess to see who is at an advantage. But to experienced players, especially grandmasters, the board's configuration provides them with an immediate sense of value and possibility, like how most of us open our eyes and have an immediate sense of where we are and what is around us. | |  |

### 4.11 - Artificiality

Besides the question of what constitutes intelligence, it is still heavily debated what properties are necessary and sufficient for AI. Discussing AI typically goes beyond just defining intelligence. It will generally touch on perception, awareness, consciousness, free will, morality, and understanding. But the question for all these properties is the same: what level of explanation is necessary (at a lower level, we lack the explanatory power), and what level of description is sufficient? In other words, do we need hardware with physical properties that match those found for biological systems? Or is the abstract concept of computing alone enough to generate a sentient agent? Or how much does it matter how much we embed our artificial intelligence into the environment? An intelligence disconnected from a physical reality cannot embody real intelligence or sentience. But there is no consensus on how much embodiment is necessary.

|  |  |  |
| --- | --- | --- |
|  | A picture containing container, glass, linedrawing  Description automatically generated |  |
|  | **Figure 4.9** - The Chinese Room thought experiment as proposed by John Searle. Locked in a room, you receive messages in a language or script you do not understand. You can, however, look up the message by matching it to a pattern in an extensive lexicon. For any message coming in, there is a matching message that you output to the outside of the room. This mapping has meaning to the people sending and receiving these messages, but you are merely matching patterns without understanding. Are you an intelligent being? Or are you just a pattern-matching machine functioning without understanding what these messages mean? |  |

### 4.12 - The Chinese Room

An often-used metaphor in this debate is the Chinese room thought experiment envisioned by John Searle in his 1980 article Minds, Brains, and Machines. In it, he imagines a person placed in a room. Occasionally, somebody will slide a piece of paper under the door. On this piece of paper is a set of Chinese characters, a language our person does not comprehend. However, they have a book to look up a new set of characters based on the characters he received. He copies those characters onto a piece of paper and slides that back under the door. Outside the room, the people sending and receiving the pieces of paper speak Chinese. They write down a question and slide them under the door. When they receive a new piece of paper, it answers their question. So, from the outsider's perspective, whoever is in the room is intelligent. But on the inside, our person has no clue what is going on. They simply take the piece of paper, find a set of new symbols that go with the ones given, and give it back. So, is this system intelligent? Or not? Searle's paper divided the AI and general scientific community into two interpretations: the weak and strong definitions of AI. In weak AI, non-biological systems may exhibit intelligence behavior. But at their core, they still simply mimic these behaviors. In contrast, in strong AI, A non-biological system can be intelligent, conscious, sentient, and aware. We should not expect the hardware and exact implementation to pose a limitation on intelligence.

### 4.13 - On ethics, morality, and the singularity

Regardless of our stance on weak/strong AI, we entrust critical decisions into the hands of AI and continue to empower them with more decision-making authority driven by real-world choices. As such, human designers need to be aware of biases in their algorithms and actively guard against them. Something we will discuss later in chapter 29. Similarly, we might have to think about the rules of robotics, as suggested by Isaac Asimov. Asimov argued that if we eventually create AI, we need to endow it with agreed-upon rules to remove ambiguity and uncertainty in the set of available actions. But hardwiring such laws seems an outdated concept. Instead, ethics and morality must develop naturally to be meaningful and useful. In the same way we teach our children to express gratitude, we should teach our AI that certain actions are impermissible under any circumstances. All of this is in anticipation of the likelihood of the singularity. This point in time, introduced by futurist Ray Kurzweil, is when AI will match and overtake humankind in the most significant areas humans currently still are superior. It may already have happened, but not as most of us expected or imagined. We will return to this topic at the end of the course.

|  |  |  |
| --- | --- | --- |
|  | A group of people running  Description automatically generated with low confidence |  |
|  | **Figure 4.10** - Is the Singularity far away? Or already here? |  |

### 4.14 - References

* Hauser, L. S. (1993). Searle's Chinese Box: The Chinese Room Argument and Artificial Intelligence. Michigan State University. Department of Philosophy.
* Searle, J. R. (1984). Minds, Brains, and Science. United Kingdom: Harvard University Press.
* Godfrey-Smith, P. (2018). Other Minds: The Octopus and the Evolution of Intelligent Life. United Kingdom: William Collins.
* Bridle, J. (2022). Ways of Being: Animals, Plants, Machines: The Search for a Planetary Intelligence. United States: Farrar, Straus, and Giroux.
* Kurzweil, R. (2005). The Singularity Is Near: When Humans Transcend Biology. United States: Penguin Publishing Group.
* de Waal, F. (2016). Are We Smart Enough to Know How Smart Animals Are? United States: W. W. Norton.
* Buss, D. (2015). Evolutionary Psychology: The New Science of the Mind. United Kingdom: Taylor & Francis.
* Butler, B. (2019). The Grandmaster: Magnus Carlsen and the Match That Made Chess Great Again. United States: Simon & Schuster.
* Siegelbaum, S. A., Mack, S. H., Koester, J. D., Kandel, E. R. (2021). Principles of Neural Science, Sixth Edition. United States: McGraw-Hill Education.
* Biologically Inspired Cognitive Architectures 2018: Proceedings of the Ninth Annual Meeting of the BICA Society. (2018). Germany: Springer International Publishing.
* Marr, D. (2010). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. United Kingdom: MIT Press.