High-Level Expert Group on Artificial Intelligence

A definition of AI: main capabilities and scientific disciplines

Definition developed for the purpose of the High-Level Expert Group's deliverables*

We start from the following definition of Artificial Intelligence (AI), as proposed within the European Commission's Communication on AI[1],:

> *"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.*
>
> *AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."*

In this document we expand this definition to clarify certain aspects of AI as a scientific discipline and as a technology, with the aim to avoid misunderstandings, to achieve a shared common knowledge of AI that can be fruitfully used also by non-AI experts, and to provide useful details that can be used in the discussion on both the AI ethics guidelines and the AI policies recommendations.

Disclaimer and Use of this Document: The following description and definition of AI capabilities and research areas is a very crude oversimplification of the state of the art. The intent of this document is not to precisely and comprehensively define all AI techniques and capabilities, but to describe summarily the joint understanding of this discipline that the High-Level Expert Group is using in its deliverables. We hope however that this document can also serve as a useful educational starting point for people that are not AI experts, who can then follow up with more extensive and deep reflection on AI to get a more precise knowledge of this discipline and technology.

1. AI systems

The term AI contains an explicit reference to the notion of intelligence. However, since intelligence (both in machines and in humans) is a vague concept, although it has been studied at length by psychologists, biologists, and neuroscientists, AI researchers use mostly the notion of rationality, which refers to the ability to choose the best action to take in order to achieve a certain goal, given certain criteria to be optimized and the available resources. Of course, rationality is not the only ingredient in the concept of intelligence, but it is a significant part of it.

* Francesca Rossi, member of the High-Level Expert Group on AI (AI HLEG), acted as Rapporteur of this document.

[1] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels, 25.4.2018 COM(2018) 237 final.

In the following we will use the term *AI system* to mean any AI-based component, software and/or hardware. Indeed, usually AI systems are *embedded* as components of larger systems, rather than stand-alone systems.

An AI system is thus first and foremost rational, according to one of the most used textbook of AI [1]. But how does an AI system achieve rationality? As pointed out in the first sentence of the above working definition of AI, it does so by perceiving the environment in which the system is immersed through some sensors, reasoning on what is perceived, deciding what the best action is, and then acting accordingly, through some actuators, thus possibly modifying the environment. The illustration of an AI system in Figure 1 may help.

[1] "Artificial Intelligence: A Modern Approach", S. Russell and P. Norvig, Prentice Hall, 3rd edition, 2009.

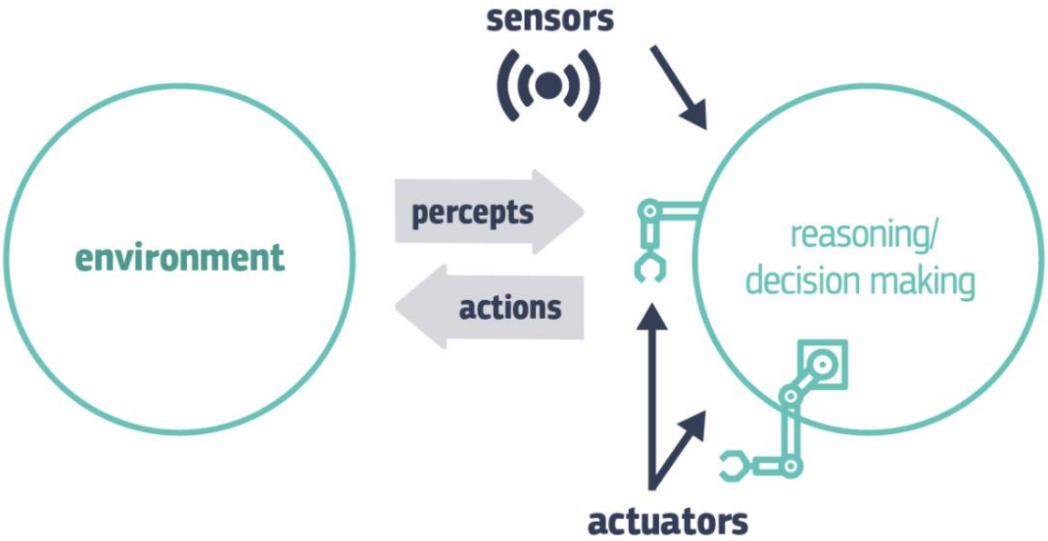

Figure 1: Schematic depiction of an AI system.

Sensors and perception. In Figure 1 the system's sensors are depicted as eyes. In practice they could be cameras, microphones, a keyboard, a website, or other input devices, as well as sensors of physical quantities (e.g. temperature, pressure, distance, force/torque, tactile sensors). In general, we need to provide the AI system with sensors that are adequate to perceive the data present in the environment that are relevant to the goal given to the AI system. For example, if we want to build an AI system that automatically cleans the floor of a room when it is dirty, the sensors could include cameras to take a picture of the floor.

As to what regards the collected data, it is often useful to distinguish between structured and unstructured data. *Structured data* is data that is organized according to pre-defined rules (such as in a database), while *unstructured data* does not have a known organization (such as in an image or a piece of text).

Reasoning and Decision Making. At the core of an AI system lies its reasoning module, which takes as input the data coming from the sensors and proposes an action to take, given the goal to achieve. This means that the data collected by the sensors need to be transformed into information that the reasoning module can understand. To continue in our example of a cleaning AI system, the camera will provide a picture of the floor to the reasoning module, and this module needs to decide whether to clean the floor or not (that is, what the best action is to achieve the desired goal). While it may seem easy for us humans to go from a picture of a floor to the decision of whether it needs to be cleaned, this is not so easy for a machine, because a picture is just a sequence of 0s and 1s. The reasoning module therefore has to:

1. Interpret the picture to decide if the floor is clean or not. In general, this means being able to transform data into information and to model such information in a succinct way, which however should include all relevant pieces of data (in this case, whether the floor is clear or not).
2. Reason on this knowledge in order to understand what the best action is. In this example, if the information derived from the picture is that the floor is dirty, the best action is to activate the cleaning, otherwise the best action is to stay still.

Notice that the term” decision” should be considered broadly, as any act of selecting the action to take, and does not necessarily mean that AI systems are completely autonomous. A decision can also be the selection of a recommendation to be provided to a human being, who will be the final decision maker.

Actuation. Once the action has been decided, the AI system is ready to perform it through the actuators available to it. In the cartoon above, the actuators are depicted as arms and legs, but they don’t need to be physical. Actuators could be software as well. In our cleaning example, the AI system could produce a signal that activates a vacuum cleaner if the action is to clean the floor. As another example, a conversational system (that is, a chatbot) acts by generating texts to respond to user’s utterances.

The action performed is going to possibly modify the environment, so the next time the system needs to use its sensors again to perceive possibly different information from the modified environment.

Rational AI systems do not always choose the best action for their goal, thus achieving only *bounded rationality*, due to limitations in resources such as time or computational power.

Rational AI systems are a very basic version of AI systems. They modify the environment but they do not adapt their behaviour over time to better achieve their goal. A *learning rational system* is a rational system that, after taking an action, evaluates the new state of the environment (through perception) to determine how successful its action was, and then adapts its reasoning rules and decision making methods.

2. AI as a scientific discipline

The one described above is a very simple abstract description of an AI system, through three main capabilities: perception, reasoning/decision making, and actuation. However, it is enough to allow us to introduce and understand most of the AI techniques and sub-disciplines that are currently used to build AI systems, since they all refer to the various capabilities of the systems. Broadly speaking, all such techniques can be grouped in two main groups that refer to the capability of *reasoning* and *learning*. On top of them, we also have *robotics*.

Reasoning and Decision Making. This group of techniques includes knowledge representation and reasoning, planning, scheduling, search, and optimization. These techniques allow to perform the reasoning on the data coming from the sensors. To be able to do this, one needs to transform data to knowledge, so one area of AI has to do with how best to model such knowledge (*knowledge representation*). Once knowledge has been modelled, the next step is to reason with it (*knowledge reasoning*), which includes making inferences, *planning* and *scheduling* activities, *searching* through a large solution set, and *optimizing* among all possible solutions to a problem. The final step is to decide what action to take. The reasoning/decision making part of an AI system is usually very complex and requires a combination of several of the above mentioned techniques.

Learning. This group of techniques includes machine learning, neural networks, deep learning, decision trees, and many other learning techniques. These techniques allow an AI system to learn how to solve problems that cannot be precisely specified, or whose solution method cannot be described by symbolic reasoning rules. Examples of such problems are those that have to do with perception capabilities such as *speech* and *language understanding*, as well as *computer vision*. Notice that these problems are apparently easy, because they are indeed usually easy for humans. However, they are not that easy for AI systems, since they cannot rely on common sense reasoning (at least not yet), and are especially difficult when the system needs to interpret unstructured data. This is where techniques following the *machine learning* approach come in handy. However, machine learning techniques can be used for many more tasks than only perception .

Machine learning comes in several flavours. The most wide-spread approaches are *supervised learning*, *unsupervised learning*, and *reinforcement learning*.

In supervised machine learning, instead of giving behavioural rules to the system (how about "instead of specifying what the system is supposed to do"?), we provide it with examples of input-output behaviour, hoping that it will be able to generalize from the examples and behave well also in situations not shown in the examples. In our running example, we would give the system many examples of pictures of a floor and the corresponding interpretation (that is, whether the floor is clean or not in that picture). If we give enough examples, which are diverse and inclusive enough of most of the situations, the system, through its machine learning algorithm, will be able to generalize to know also how to correctly interpret pictures of floors never seen before. Some machine learning approaches adopt algorithms that are based on the concept of *neural networks*, which is inspired by the human brain in that it has a network of small processing units (analogously to our neurons) with lots of weighted connections among them. A neural network has as input the data coming from the sensors (in our example, the picture of the floor) and as output the interpretation of the picture (in our example, whether the floor is clean or not). During the analysis of the examples (the network's *training* phase), the connections' weights are adjusted to match as much as possible what the available examples say (that is, to minimize the error between the expected output and the output computed by the network). At the end of the training phase, a testing phase of the behaviour of the neural network over examples never seen before checks that the task has been learnt correctly.

It is important to notice that this approach (as all machine learning techniques) has always a certain percentage of error, albeit usually a small one. So an essential notion is the *accuracy,* a measure of how large the percentage of correct answers is.

There are several kinds of neural networks and machine learning approaches, of which currently one of the most successful one is *deep learning*. This approach refers to the fact that the neural network has several layers between the input and the output that allow to learn the overall input-output relation in successive steps. This makes the overall approach more accurate and with less need of human guidance.

Neural networks are just one machine learning tool, but there are many others, with different properties: random forests & boosted trees, clustering methods, matrix factorization, etc.

Another useful kind of machine learning approach is called *reinforcement learning*. In this approach, we let the AI system free to make its decisions, over time, and at each decision we provide it with a reward signal that tells it whether it was a good or a bad decision. The goal of the system, over time, is to maximize the positive reward received. This approach is usually used, for example, in recommending system (such as the several online recommending systems that suggest users what they might like to buy), or also in marketing.

Machine learning approaches are useful not just in perception tasks, such as vision and text understanding, but in all those tasks that are hard to define and cannot be comprehensively described by symbolic behavioural rules.

Notice the distinction between machine learning approaches to learn a new task that cannot be described well in a symbolic way, and learning rational agents (mentioned in the previous section) that adapt their behaviour over time to better achieve the given goal. These two techniques may overlap or cooperate, but are not necessarily the same.

Robotics. Robotics can be defined as "AI in action in the physical world" (also called *embodied AI*). A robot is a physical machine that has to cope with the dynamics, the uncertainties and the complexity of the physical world. Perception, reasoning, action, learning, as well as interaction capabilities with other systems are usually integrated in the control architecture of the robotic system. In addition to AI, other disciplines play a role in robot design and operation, such as mechanical engineering and control theory. Examples of robots include multi-degree-of-freedom robotic manipulators, autonomous vehicles (e.g. cars, drones, flying taxis), humanoid robots, robotic vacuum cleaners, etc.

Figure 2 depicts most of the AI sub-disciplines mentioned above, as well as their relationship. It is important however to notice that AI is much more complex than this picture shows, since it includes many other sub-disciplines and techniques. Moreover, as noted above, robotics also relies on techniques which fall outside the AI space. However, we believe this is enough for informing in a fruitful way the sharing, awareness, and discussion on AI, AI ethics, and AI policies that needs to take place within the very multi-disciplinary and multi-stakeholder high level expert group.

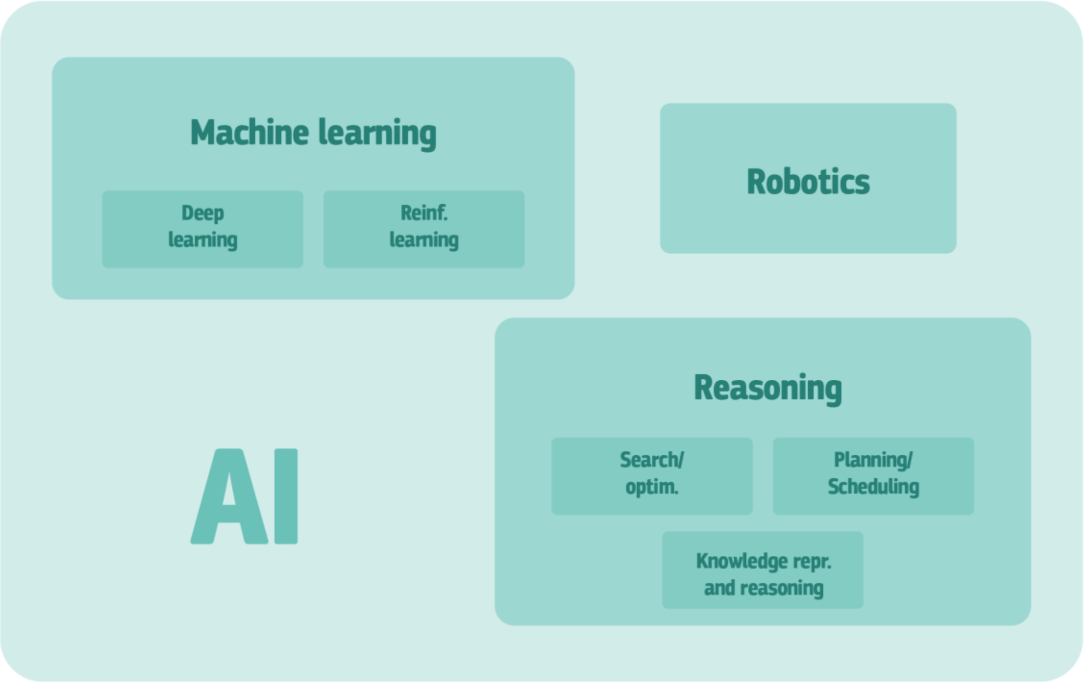

Figure 2: AI's sub-disciplines and their relationship.

3. Other important AI notions and issues

Narrow (or weak) and general (or strong) AI. A general AI system is intended to be a system that can perform most activities that humans can do. Narrow AI systems are instead systems that can perform one or few specific tasks. Currently deployed AI systems are examples of narrow AI. In the early days of AI, researchers used a different terminology (weak and strong AI). There are many open scientific and technological challenges to build the capabilities that are needed to achieve general AI, such as common sense reasoning, self-awareness, and the ability of the machine to define its own purpose.

Data issues and bias. Since many AI systems, such as those including supervised machine learning components, rely on huge amounts of data to perform well, it is important to understand how data are influencing the behaviour of the AI system. For example, if the training data is biased, that is, it is not balanced or inclusive enough, the AI system trained on such data will not be able to generalize well and will possibly make unfair decisions that can favour some groups over others. Recently the AI community has been working on methods to detect and mitigate bias in training datasets and also in other parts of an AI system.

Black-box AI and explainability. Some machine learning techniques, although very successful from the accuracy point of view, are very opaque in terms of understanding how they make decisions. The notion of *black-box AI* refers to such scenarios, where it is not possible to trace back to the reason for certain decisions. Explainability is a property of those AI systems that instead can provide a form of explanation for their actions.

Goal-directed AI. Current AI systems are goal-directed, meaning that they receive the specification of a goal to achieve from a human being and use some techniques to achieve such goal. They do not define their own goals. However, some AI systems (such as those based on certain machine learning techniques) can have more freedom to decide which path to take to achieve the given goal.

4. Updated definition of AI

We propose to use the following updated definition of AI:

> "Artificial intelligence (AI) systems are software (and possibly also hardware) systems that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions.
>
> As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

and to refer to this document as a source of additional information to support this definition.