1

COGNITIVE SCIENCE AND ARTIFICIAL INTELLIGENCE

Death and rebirth of a collaboration

Abstract

The first chapter proposes a brief historical overview of some of the main insights developed over 65 years of research in Artificial Intelligence (AI), by introducing the early vision of the discipline (based on a mutual collaboration with Cognitive Psychology) and its "paradigm shift", which started from the mid-1980s of the last century. Starting from that period on, AI and the interdisciplinary enterprise known as Cognitive Science started to produce several sub-fields, each with its own goals, methods, and criteria for evaluation. The reasons for the current renewed interest of a cognitively inspired approach in AI research are discussed.

When Cognitive Science was AI

Cognitive Science and Artificial Intelligence (AI) are, nowadays, scientific research fields each endowed with a specific autonomy and research agenda. According to the Oxford Dictionary, the term "Artificial Intelligence" is defined as "the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages", while "Cognitive Science" is defined as "the study of thought, learning, and mental organization, which draws on aspects of psychology, linguistics, philosophy, and computer modelling".

Despite the current different focuses and objectives of each, these two disciplines have many common interests and share the idea of studying the "mind", its emergent properties, and its functioning in natural and artificial systems, respectively.

The history of these two research fields is, in fact, strongly interconnected. Research in AI – the birth of which dates back to the now-legendary "Dartmouth Workshop" (McCarthy et al., 1955) held in the summer of 1956[1] – has, indeed, been historically inspired by the experimental research in psychology.[2] Notable examples of such intellectual connections are represented by the early AI systems/frameworks developed until the 1980s. Most of them, indeed, were explicitly designed with a "cognitively oriented" inspiration. In the following sections, we briefly present few famous examples of such systems and formalisms (though the list is far from being exhaustive) with the aim of introducing some of the main modelling paradigms and assumptions that have characterized, and still characterize, the research in AI and cognitive modelling. Each of the systems/formalisms reviewed below can be considered important either because they have achieved some important milestones in terms of performances or because has introduced some relevant ideas that have fostered meaningful developments in the study and the realization of "artificial minds".

From the general problem-solver to the society of mind: cognitivist insights from the early AI era

One of the first developed AI systems, at the end of the 1950s, is the pioneering work of Herbert Simon, John Clifford Shaw, and Allen Newell on the *General Problem Solver* (GPS). GPS was a system able to demonstrate simple logic theorems and its decision strategies were explicitly inspired by human verbal protocols[3] (Newell, Shaw & Simon, 1959). The underlying idea of this approach was that the computer system had to approximate the decision operations described by humans in their verbal descriptions as closely as possible. In this way, when the program ran on the computer, it would be possible to identify its problems, compare them with the description of the human verbalization, and modify them to improve its performance. In particular, the GPS system was able to implement a key mechanism in human problem solving: the well-known "means-ends analysis" (or M-E heuristics). The M-E heuristics implemented in GPS works as follows: the problem solver makes a comparison between the current

1 The organisers of this even were some "giants" of the history of the Computer Science field from the last century: John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. The workshop, during which McCarthy proposed the use of the term "artificial intelligence" to identify the new emerging discipline, ran for several weeks and saw the participation of many researchers. The notes taken by Ray Solomonoff (one of the participants at the workshop) are available online at http://raysolomonoff.com/dartmouth/.

2 It must be noted that, at that time, there wasn't a "Cognitive Science" field. However, all the disciplines (philosophy, psychology, computer science, anthropology, linguistics, and neurophysiology) and the cultural elements that would have later be called upon to form the interdisciplinary field of "Cognitive Science" were already present.

3 This technique is also known as the "thinking aloud protocol" in the psychological literature (Ericsson & Simon, 1980) and consists of recording the verbal explanations provided by people while executing a given laboratory task.

situation and a goal situation; then, it computes and evaluate the "distance" between these two states and tries to find, in memory, suitable operators able to reduce such difference. Once a suitable operator is found, it is then applied to change the current situation. The process is repeated until the goal is gradually attained via a process of progressive distance reduction. There are, however, generally no guarantees that the process will succeed. This kind of heuristic was also used to solve, in the decades to come, problems in a number of domains. In order to be executed, in fact, it "only" required an explicit domain representation of the problem to solve (a problem space), operators to move through the space, and information about which operators were relevant for reducing which differences.[4] GPS can be arguably considered the first cognitively inspired AI system ever developed.

A decade after the development of GPS, a Ph.D. student of Herbert Simon[5] at Carnegie Mellon University (then still named Carnegie Institute of Technology) – Ross Quillian – developed another influential idea in the context of AI of cognitive inspiration; he invented the *Semantic Networks*: a psychologically plausible model of human semantic memory implemented in a computer system. The idea (Quillian, 1968) was that human memory is associative in nature and that concepts are represented as sort of nodes in graphs and are activated through a mechanism of "spreading activation", implemented through a marker passing algorithm, allowing the propagation of information through the network to determine the strength of the relationships between concepts. In this setting, the higher the activation of a node in the network, the more contextually relevant that node/concept was assumed to be for the task in focus. Interestingly enough, the research on Semantic Networks paved the way for both the development of the first graph-like, knowledge-based systems and formalisms (which make use of so-called *symbolic* representations) as well as the improvement of the so-called *connectionist* or *sub-symbolic* systems, since the

4 As we will see in more detail in the following sections, the ingredients required for the execution of this kind of heuristic strategy – essentially based on a "search space" approach to problem solving – explicitly supported the so-called "symbolic approach" for the study, analysis, execution, and replication of intelligent behaviour in artificial systems.

5 Herbert Simon is arguably one of the most important scientists of the last century. His influence, indeed, went well beyond his original training in cognitive psychology. Simon was awarded a Nobel Prize in Economics for his studies on "bounded rationality", which showed – differing from the classical decision models of the time – how humans are not optimal decision makers. This field of study has led to the development of an entirely new discipline that is nowadays known as "behavioural economics". In addition, he was one of the founding fathers and main protagonist of the field of AI; along with people like Marvin Minsky, John McCarthy, Allen Newell, Nathaniel Rochester, and many others, he was an active participant in the Dartmouth Workshop. As a result of his "bounded rationality" theory in decision making, he was, one of the first scholars to point out, in both cognitive psychology and AI, the role played by heuristics as decisional shortcuts to solve complex problems. The application of the heuristic approach in the context of AI was one of the reasons behind him winning, in 1975, the Turing Award, together with Allen Newell. The particular meanings attributed to the term "heuristics" in the AI research, will be explained later in this chapter.

concept of "spreading activation" has been very influential in the context of the "connectionist" investigations (see Cordeschi, 2002: 235, on this point). Before proceeding further with our examples of early cognitively inspired AI systems, it is necessary to briefly introduce the above-mentioned basic notions of "symbolic representations" (and paradigm) and "connectionist or sub-symbolic representations" (and paradigm), since they have been, and still are, really crucial modelling methods in both the past and present AI and cognitive modelling communities. In particular, the notion of "symbolic representation" constitutes a core assumption of the so-called "symbolic paradigm" in AI and cognitive science (which will be better clarified in more detail later in the book). In short, according to this view, intelligence in natural and artificial systems is associated with the capability of storing and manipulating the information in terms of abstract "symbols" (representing, in many cases, some mental proxy associated with external physical objects) and on the capability of executing mental operations and calculations over such symbols. This view was (is) severely criticized by the so-called "connectionist or sub-symbolic paradigm", according to which the organization of the "mental content" in natural and artificial systems is not based on any symbolic structure but is, on the other hand, (1) distributed in nature and (2) based on parallel models of computations (these are the two core assumptions of the "connectionist representations"), in a way that is more similar to the biological organization and processing mechanisms of neurons and synapses in our brain. From a modelling perspective, this approach has led to the development of the Artificial Neural Networks, or ANNs (partially inspired by the biological neural structure of our brain), and self-organizing systems. We will discuss later the impact of "neural" or brain-inspired methods in early (and modern) AI research.[6] For the moment it is probably worth mentioning that, from a historical point of view, the "symbolic paradigm" represented the mainstream assumption in the context of both early AI and cognitive modelling research.

A confirmation of what was just discussed is provided by the next example of a cognitively inspired AI framework, which we are going to investigate: the notion of *Frames* (still a symbolic representational framework) operated by Marvin Minsky almost a decade after Quillian's proposal (Minsky, 1975). With this proposal, Minsky intended to attack another well-known "symbolic approach"

6 For the sake of completeness, it is also worth mentioning that within the cognitive modelling and AI communities another paradigm has been historically proposed relying on so-called "analog" or "diagrammatic" representations. In particular, according to the supporters of this school of thought, mental representations take the form of "pictures" in the mind. There are many different examples of analog representations proposed, one of the most famous corresponding to the "mental models" by Johnson-Laird (1983, 2006). A general underlying assumption of this class of representation is that "spatial cognition" abilities (represented via these "picture-like" schemas) are a core aspect of natural cognitive systems from which other intelligent mechanisms emerge (e.g., the mental models by Johnson Laird have been notoriously proposed to model different types of inferences).

developed back then: the "logicist"[7] position à la McCarthy for the representation of knowledge in artificial systems. In particular, Minsky argued that such a proposal was not able to deal with the flexibility of the commonsense reasoning that is so evident in human beings. Frames, on the other hand, were proposed for endowing AI systems with commonsense knowledge (including *default* knowledge) about the external world.[8] The type of knowledge organization proposed in the Frames enabled the first AI systems to extend their automated reasoning abilities from classical deduction to more complicated forms of commonsense and defeasible reasoning (going from induction to abduction). In this case, the idea of the Frames was directly inspired by the work of the psychologist Eleanor Rosch (Rosch, 1975) about the organization of conceptual information in humans known as the "prototype theory"[9] as well as by the memory "schemas" proposed by the cognitive psychologist Bartlett (Bartlett, 1958). A simple example and use case, done by Minsky himself, of a frame data structure is the following: let us imagine opening a door inside a house we are not familiar with. In this case, we typically expect to find a room that more or less is characterized by features that we have already seen in other rooms we have been in. Such features are referred to as a body of knowledge organized in the form of prototypes (i.e., the typical room). The data structures that reflect this flexible way of using knowledge, which is typical of human beings, can be described as "frame systems". Therefore, the "room frame" is a characterized by different types of information that includes – listed in appropriate "slots" – the typical features of a room, such as a certain number of doors, walls, windows, and so on. There could be various kinds of rooms – dining rooms, bedrooms, etc. – each constituting, in turn, a frame with more specific features, again listed in appropriate slots. This kind of representation also allows for individual differences in conceptualization; e.g., Francesca's dining room might be quite different from Paola's in various details, but it will always be part of one and the same kind of room frame. The proposal of the frames as data structures for commonsense reasoning was not

7 A brief overview of the logical approaches proposed in the 1970s to deal with commonsense reasoning (e.g., circumscription, fuzzy logic, etc.) is sketched out in the next chapter of the book. At this point, it is important to point out that the logicist tradition was (is) deeply rooted in the symbolic representation assumption, briefly elaborated on above and further detailed in the next section of this chapter.

8 As indicated elsewhere, "all the forms of commonsense reasoning can be seen as a bounded rationality phenomenon since they represent a plethora of shortcuts allowing us (i.e., "bounded-rational" agents) to make decisions in an environment with incomplete and uncertain information" (Lieto, 2020: 56).

9 According to the prototype theory posited by Rosch, concepts are organised in our mind as "prototypes" (i.e., in terms of typical representative elements of that category) and such an organization explains many types of so-called "typicality effects" (i.e., of commonsense inferences) that we naturally perform in our everyday reasoning. We will return on this specific aspect later and more extensively in the book (particularly in Chapter 4), since commonsense reasoning represents one of the main areas of possible convergence between Cognitive Science and AI.

completely successful from a computational point of view (since frame systems did not scale well) but was very influential for the development of research in the context of commonsense reasoning.

In those years, a proposal very much aligned with Minsky's was put forth by Roger Schank and his "conceptual dependency" theory (Schank, 1972). Schank aimed at explaining natural-language understanding phenomena via psychologically plausible computational processes. He proposed identifying a small set of "semantic primitives", the use of which would have made it possible to construct the representation of meaning for any English verb. In his original programs, a sentence was analyzed by making explicit its representation in terms of semantic primitives. Such primitives were considered common to all natural languages and constituted a sort of interlingua. This interlingua was then used to build the first machine translation systems (e.g., MARGIE, see Shank & Nash-Webber, 1975). When Schank passed from constructing programs translating single sentences to ones aimed at translating entire stories, he realized that it was necessary to take commonsense into account. In this respect, a relevant problem concerned the knowledge needed to derive meaningful inferences from the union of different sentences in a story, so as to make explicit the implicit beliefs and expectations assumed in the context of a story. To tackle this and other problems, Schank and Abelson (1977) endowed their program – SAM (Script Applier Mechanism) – with "scripts". Scripts are a data structure for representing knowledge of common sequences of events (e.g., the sequence of events used to go out for dinner) and are used in natural-language processing systems as way to enable intelligent answers to questions about simple stories. A classic example used to explain the notion of a "script" (which is also tightly connected with the notion of a "Frame") is the so called "restaurant situation". Let us consider a situation of an agent going out to a restaurant for dinner. A script representing the restaurant situation is a data structure that would record the typical events associated with this scenario; e.g., entering the restaurant, asking for a table, sitting down, consulting a menu, eating the food, paying the check, etc. This kind of representational structure enabled early AI systems to answer questions about simple stories. For example, let us consider a story like this: "Mary went to a restaurant and ordered salmon. When she was paying, she noticed that she was late for her next appointment." In this case, computerized systems were able to answer a question such as, "Did Mary eat dinner last night?" in a positive way (as we do). It is worth noticing that this information is not explicitly provided in the story. Answering these types of questions is possible through the use of a "script" of the restaurant situation.

The capability of understanding natural-language instructions was also a crucial feature of Terry Winograd's famous robotic system known as SHRDLU (named for the alphabetic symbols composing a row of keyboards in that era). In SHRDLU (Winograd, 1972), interactions with humans focused on a simulated blocks world that humans could view on a graphics display and to which the system had direct access. Users drove the conversation via written text by typing sentences, including commands like, "Find a block that is taller than the

one you are holding and put it into the box" and "Is there anything that is bigger than every pyramid but not as wide as the thing that supports it?". As reported in Langley (2017),

> These inputs required not only the ability to parse quite complex structures and extract their meanings but also to draw inferences about relationships and execute multistep activities. The innovative system handled simple anaphora, disambiguated word senses, and had basic memory for its previous interactions.

SHRDLU was, therefore, an important advancement because it integrated sentence level understanding, reasoning about domain content, execution of multistep activities, and natural interaction with human users. At that time, there was no other artificial system able to show the same range of capabilities, and it offered a proof of concept that such an integrated intelligent system was possible. This accomplishment, of course, relied on some important simplifications: SHRDLU operated in a narrow and well-defined domain and had complete access to the entire state of the simulated environment. Nevertheless, it was an impressive achievement, which fostered further work on intelligent agents. To a certain extent, the integrated abilities exhibited by SHRDLU were the inspiration also for the subsequent work of Allen Newell and his colleagues at Carnegie Mellon University, concerning the development of the first integrated cognitive architecture for general intelligence: SOAR (Newell, Laird, & Rosenbloom, 1982).[10]

At the very time that SOAR was first being developed (by now we were already in the mid-1980s), another relevant proposal in the context of cognitively inspired AI was made, once again, by Marvin Minsky, who introduced the evocative idea of the "Society of Mind" (Minsky, 1986, 2007) as a way to conceptualize, analyze, and design intelligent behaviour. This idea relies on the importance of considering, in natural and artificial agents, problem-solving activities "in layers" of interconnected micro-faculties (i.e., as a "society" of processes). In particular, Minsky suggested that the capability of dealing with commonsense knowledge[11] is the grounding element of these layers of growing thinking capabilities. Such an approach has been historically impactful – not from an engineering perspective (since much more detail would have been needed in the Minsky proposal to specify how the processes can and should interact in an

10 On the role of cognitive architectures for general intelligent systems we remind to (Lieto et al., 2018). We will return to SOAR and to cognitive architectures over the course of the book. In addition to the SHRDLU influence, SOAR was heavily inspired by the heuristic search mechanisms already developed in the GPS system.

11 Commonsense knowledge is acquired, according to the Minsky proposal, via "instinctive" or "learned" reactions, and is then processed towards the higher hierarchies of "deliberative", "reflective", "self-reflective", and "self-conscious" thinking at the level of both individual and social context.

efficient computer implementation) – but mainly for the idea of considering, from a methodological and modelling perspective, the classical problem-solving activity (which was already modelled in systems like GPS or SOAR) through this sort of layered conceptual view involving a multistep reasoning process. As we will see in the following sections, this layered approach influenced, under completely different assumptions, another protagonist of the AI story from the previous century: Rodney Brooks.[12]

This list of examples of early cognitively inspired AI systems reviewed so far is, of course, not exhaustive. However, all these early systems shared a common "view" about the study of intelligence in artificial systems. More precisely, all these systems adhere – at different levels – to the so-called "cognitivist tradition"[13] of AI, also known as GOFAI (Good Old Fashioned AI).

Such early view is successfully synthesized by Pat Langley (Langley, 2012), who said, "(Early) AI aimed at understanding and reproducing in computational systems the full range of intelligent behaviour observed by humans" (Langley, 2012).

Langley identifies the following set of features that characterize the early AI period and the main cognitivist modelling assumptions:

- the role of symbolic representations as a building block upon which operate a set of manipulation operations to let intelligent behaviour emerge;
- the importance of a general cognitively inspired approach to the study of the mind and intelligence (what Pat Langley calls a "system view");
- the main focus on the so-called "high level cognition" (the systems for natural language processing,[14] for example, underwent a big development in this early period);
- the adoption of heuristics (we will return on this concept later) as a method for problem solving;
- the intrinsic interdisciplinary and exploratory nature of the research.

We will analyze in more details these aspects of the cognitivist tradition (and its differences from emergentist perspectives) in the next few sections of the chapter.

12 Rodney Brooks is a roboticist and was previously an MIT Professor. He is the creator of "Herbert" the robot, the first mobile robot able to exhibit interesting reactive behaviours without any central controlled activity. For more details about the particular layered architecture proposed by Brooks, known as "Subsumption Architecture", we refer the reader to the next section.

13 As will be clarified in the following pages, the "cognitivist" tradition is deeply rooted in the so-called "symbolic paradigm" and was the dominant perspective during the early days of AI research. Cognitivist assumptions differ from those of the "emergentist" approaches, which are, on the other hand, rooted in the notions of bottom-up self-organisation (see Vernon, 2014).

14 A typical example of the systems developed in this period is Eliza (Weizenbaum, 1966), one of the first conversational agents (nowadays called "chatbots"), created to converse with a human being, simulating, at least up to a certain extent, the behaviour of a psychotherapist.

However, from a historical perspective, it is worth mentioning that this approach to the study of the artificial did not come out *ex-abrupto*. It borrowed its original inspiration, even if grounded on different assumptions, from the methodological apparatus developed by scholars in cybernetics (Cordeschi, 1991). The origins of cybernetics, in fact, are usually traced back to the middle of the 1940s, with the release of the 1948 book by Norbert Wiener entitled *Cybernetics: Or Control and Communication in the Animal and the Machine*. An underlying idea of cybernetics was one about building mechanical models to simulate the adaptive behaviour of natural systems. As indicated in Cordeschi (Cordeschi, 2002): "The fundamental insight of cybernetics was in the proposal of a unified study of organisms and machines". In this perspective, the computational simulation of biological processes was assumed to play a central epistemological role in the development and refinement of theories about the elements characterizing the nature of intelligent behaviour in natural and artificial systems. Such kind of simulative approach, as mentioned, was inherited by the early AI research that used computer programs to reproduce performances, which, if observed in human beings, would be regarded as "intelligent". The adoption of such a perspective was crucial in AI, for the development of both intelligent solutions inspired by human processes and heuristics (Newell & Simon, 1976; Gigerenzer & Todd, 1999) and for the realization of computational models of cognition built with the aim of providing a deeper understanding of human thinking, as originally suggested in the manifesto of Information Processing Psychology (IPP) (Newell & Simon, 1972). These two sides of the cognitivist tradition are nowadays still alive. They correspond, roughly, to the research areas known as "cognitively inspired AI" (or "cognitive systems") and "cognitive modelling" (or "computational cognitive science"), respectively.

Heuristics and AI eras

The notion of heuristics deserves, in this historical account, special attention. Usually, this term, derived from the Greek word "eureka", indicates a non-optimal problem-solving procedure adopting particular "shortcuts" to reach a given goal. This term has been ascribed two different meanings since the times of the first AI research. In its first sense, the term refers to the most detailed simulation possible of human cognitive processes, and it characterized the above-mentioned IPP, introduced by Newell and Simon. In this view, a computer program was considered to be a model providing a test of the hypothesis that the mind is an information-processing system. More precisely, "the program was considered to be a highly specific behavioral theory, concerning the behavior of an individual human problem-solver: a microtheory" (Cordeschi, 2002: 182).[15]

15 In this view, the general theory of human information-processing was assumed to be derivable from a body of qualitative generalizations coming from the study of individual simulative programs, or microtheories.

In another sense, the term refers to the possibility of obtaining the most efficient (and efficacious) performance possible from computer programs, by allowing also for typically non-human procedures, such as those where the computer can excel. Before the introduction of the term "heuristics" in AI – operated by Newell, Shaw, and Simon – there were already algorithmic procedures available, which might have been defined as heuristic in the second of these senses and which had already been tried out experimentally. The first among them were the procedures that allowed the program developed by Arthur Samuel to play checkers despite the combinatory explosion of moves (Samuel, 1959).

The fact that these two tendencies, reflected in the double meaning of the term "heuristic", coexisted in AI was immediately clear. As reported in Cordeschi (2002: 190), in 1961, while discussing a presentation of GPS given by Simon during a seminar at MIT, Minsky drew a distinction in AI research between those who were willing to use "non-human techniques" in constructing intelligent programs and those, like the Carnegie-Mellon group, who were interested in simulating human cognitive processes.[16] This distinction is crucial, since it outlines the emergence of different research agendas that were already present at that time. In the following decades these early distinctions became deeper and determined the difference between "Nature-" or "Human-inspired" approaches to the development of artificial systems versus "Machine-oriented" approaches to the solution of a given problem.

Modelling paradigms and AI eras: cognitivist and emergentist perspectives

As briefly illustrated in the previous sections, the early days of AI were mainly characterized by the "cognitivist" assumption that intelligent activity in both living and artificial systems was possible due to the capability of encoding knowledge about the external world via "internal" abstract symbolic representations, directly corresponding to elements of the reality. In this setting, intelligent behaviour (e.g., in language, vision, planning, etc.) was viewed as the expression of operations carried out on such symbols and the motto of this early phase (also known as "cognitivism", see e.g., Vernon, 2014) was synthesized by the expression "cognition is computation". Here, the word "computation" was intended to mean the capability of manipulating such symbolic structures. The theoretical reference framework that inspired such an assumption, in both cognitive psychology and artificial intelligence, was the so-called "Physical Symbol System

16 As reported in Cordeschi (2002), Minsky emphasized that these two tendencies were distinguished "in methods and goals" from a third tendency, which "has a physiological orientation and alleges to be based on an imitation of the brain," i.e., neural net and self-organizing system approaches. We will discuss later in this book the "neural" or brain-inspired methods in early (and modern) AI research. As anticipated, such approaches belong to the so-called "connectionist agenda".

Hypothesis" (PSSH), introduced by Newell and Simon (1976). According to this theory, intelligent beings are physical symbol systems. In this framework, symbolic representations were not only a denotational means for referring to entities of the external world but also a means for denoting other internal symbolic structures (thus allowing to hypothesise an internal information processing mechanism able to overcome the classical Input-Output direct mapping assumed by the *behaviourist tradition*[17]). In this view, symbolic systems are assumed to be realizable by means of different "hardware" (e.g., a Von Neumann architecture or a natural brain[18]) and symbolic processing is considered a necessary and sufficient condition for intelligent behaviour. In particular, the apparatus of such a hypothesis assumes that an intelligent agent should be equipped with the following elements (Newell, 1990):

- Memory Systems (to contain the symbolic information)
- Symbols (to provide a pattern to match or index other symbols)
- Operations (to manipulate symbols)
- Interpretations (to allow symbols to specify operations)
- Symbolic Capacities for
  - Compositionality
  - Interpretability

With respect to what was mentioned earlier about the "symbolic paradigm", some additional clarifications are needed to fully grasp what concerns both the "Symbols" and the "Compositionality" requirements identified by Newell in the above mentioned list.

17 Behaviourism (in this context we are referring to so-called "methodological behaviourism", which is different from "philosophical behaviourism") is a methodological approach to the study of behaviour in natural systems, born at the beginning of last century, and based on the observable analysis of the responses (e.g., the produced output) to certain stimuli (the input) manipulated via different types of reinforcement (this is also known as "operant conditioning"). Watson (1913), one of the founders of this approach, defined psychology as "a purely objective experimental branch of natural science" and its program as the "prediction and control of behavior". As a consequence of this radical view, behaviourists did not consider/analyze the internal mechanisms driving a given behaviour (provided certain stimuli). The now-famous experiments done by the Russian physiologist and Nobel Prize winner Ivan Pavlov about the conditioned reflex of dogs and their automatic stimulus-response behaviour (where the stimulus was constituted by a "ringing bell" that the dogs had learn to associate to the arrival of food, and the response to the salivation caused by the bell ringing) was an important landmark in this tradition (as was his other work about so-called "classical conditioning"). This approach, was severely criticised by the cognitivist tradition in psychology, the "computationalist" view in the philosophy of mind, and Information Processing Psychology, which, on the other hand, assumed the presence of internal information processing mechanisms as driving forces leading to a manifest behaviour.

18 This claimed "interchangeability" means that, in this framework, the physical instantiation (i.e., the "hardware") *per se* is not important since the intelligent behaviour emerging via symbol manipulation is assumed to be independent of the particular form of the instantiation.

For what concerns the "symbols", as mentioned, the PSSH assumes that such abstract structures can refer to and be combined with (as is evident more clearly in the figure 1.1 below) other internal symbols and processes.

This possibility is important in light of the "compositionality" requirement. Compositionality is an important feature of symbolic systems and is also considered an irrevocable trait of human cognition. In a compositional system of representation, it is possible to distinguish between a set of primitive, or atomic, symbols and a set of complex symbols. Complex symbols are generated from primitive symbols through the application of suitable recursive syntactic rules: generally, a potentially infinite set of complex symbols can be generated from a finite set of primitive symbols. The meaning of complex symbols can be determined starting from the meaning of primitive symbols, using recursive semantic rules that work in parallel with syntactic composition rules. In the context of classical cognitive science, it is often assumed that mental representations are indeed compositional. A clear and explicit formulation of this assumption was proposed by Fodor and Pylyshyn (Fodor & Pylyshyn, 1988). They claim that the compositionality of mental representations is mandatory to explain fundamental cognitive phenomena (i.e., the generative and systematic character of human cognition) and they also show how the contrasting neural, distributed representations encoded in artificial neural networks are not compositional.[19]

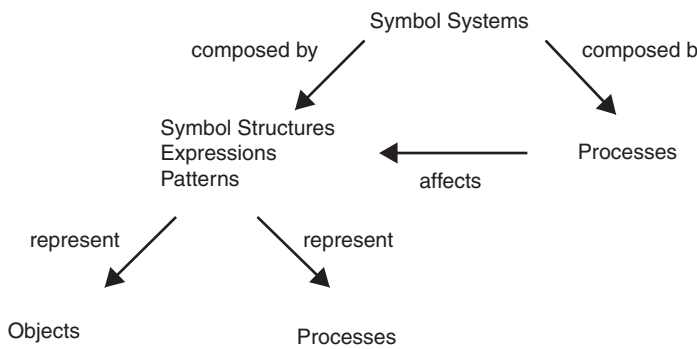

FIGURE 1.1 Overview of the internal dynamics of physical symbol systems (adapted from Vernon, 2014).

19 It is worth noting that, while standard compositionality is easily handled by symbolic system, "commonsense compositionally" (i.e., one involving typicality-based reasoning à la Rosch) has always been a problematic aspect to model. This problem is paradigmatically represented by the so called PET FISH problem: if we consider this concept, in its prototypical characterisation, as the result of the composition of the prototypical representations of the concepts "PET" and "FISH", we soon realise that the prototype of pet fish cannot result from the composition of the "PET" and "FISH". A typical pet – indeed – is furry and warm, a typical fish is greyish, but a typical pet fish is neither furry and warm nor greyish (typically, it is red). The pet fish phenomenon is a classic example of the difficulty to deal with when building formalisms and systems aiming at imitating this compositional human ability. Nowadays, a proposal to deal with the

Given this state of affairs, then, solving a problem for a physical symbol system means being able to perform a Heuristic Search within a problem space represented by symbolic structures. Here, in fact, the intelligent behaviour is assumed to emerge by generating and progressively modifying symbol structures until a solution structure (e.g., a goal) is reached. This overall assumption is known as the Heuristic Search Hypothesis[20] and, as it is probably evident to the readers, some of the above-mentioned early systems like GPS (as well as the formalisms like the Semantic Networks and, as we will see, SOAR as well) are heavily built upon the PSSH and its Heuristic Search corollary.

Parallel to these "symbolic" developments, a radically different modelling approach based on neuron-like "subsymbolic" or "connectionist" computations (e.g., Grossberg, 1976; McClelland, 2010) was being explored. Proponents of this approach (one of the most successful in the so-called "emergentist" field[21]) maintain that many classic types of structured knowledge, such as graphs, grammars, rules, objects, structural descriptions, programs, etc., can be useful yet misleading metaphors for characterizing "thought" in both natural and artificial systems. In particular, these structures are seen as epiphenomenal rather than real, emergent properties of more fundamental sub-symbolic cognitive processes (McClelland, 2010) (Figure 1.1).

In general, in contrast to the symbolic paradigm, the knowledge in these neural networks is distributed across a collection of units rather than localized as in symbolic data structures. The central idea of such models, in fact, is that a large number of simple computational units can achieve intelligent behaviour when networked together. This insight applies equally to neurons in biological nervous systems and to hidden units in computational models. The representations and

problem of commonsense compositionality in symbolic systems was proposed in Lieto and Pozzato (2020) and applied to both cognitive modelling problems (e.g., the PET FISH) and in the context of computational creativity applications. Nonetheless, modelling commonsense reasoning (including commonsense compositionality) in a human-like fashion and with human-level performances remains an open problem in the context of symbolic systems.

20 The Heuristic search hypothesis has been very influential in AI since many algorithms (e.g., from the "hill climbing" to the "beam search" to the notorious A* algorithm) that have been developed to improve the efficiency of finding optimal or suboptimal paths in problems represented as a graph-like structure have been developed by starting from this hypothesis. For an introduction to these classical algorithms, we refer the reader to introductory books on AI (see e.g., Russell & Norvig, 2002). One of the first successful and convincing implementations of such "search-based" approaches (e.g., the A* algorithm) was in the robot Shakey, developed in 1966 by Nilsson and colleagues (see Hart et al., 1968; Nilsson, 1971; Fikes et al., 1972).

21 The expression "emergentist approaches" is determined by the fact that the class of modelling frameworks of this tradition assume that the information to be processed is learned from the environment in a bottom-up way and intelligent behaviour (if any) is assumed to be an emergent property coming from this interaction. Within emergentist frameworks we can include dynamical systems (using differential equations to model the dynamic of a system and its change over time, caused by the interaction with the environment) and enactive approaches (usually employing both connectionist and dynamical frameworks and assuming embodied agents). We refer to Vernon (2014, Chapter 2), for an introduction to such frameworks.

algorithms used by this approach, therefore, were (and are) more directly inspired by neuroscience rather than psychology. As a consequence, differing from the PSSH, in this modelling framework (and in general all the so-called emergentist modelling frameworks) the "physical hardware" (e.g., the body) instantiating the actual computation is assumed to play an important role.

From a historical perspective, the connectionist movement took inspiration from the functional models of nervous cells, introduced in the pioneering work by Warren McCulloch and Walter Pitts (developed during the pre-cybernetic period and heavily influencing cybernetic research), showing how every "net" of formal neurons – if furnished with a tape and suitable input, output, and scanning systems – is equivalent to a Turing machine[22] (McCulloch & Pitts, 1943). Such initial insights were later enriched by research from Donald Hebb (Hebb, 1948) about the learning processes in the nervous system[23] and further studies of learning and classification processes in networks, à la McCullock and Pitts, lead to the development of the first artificial neural network (ANN) known as Perceptron (developed by Rosenblatt in 1958).[24]

After these pioneering works, during the 1960s, research on neural nets seemed to take a step back once a notorious book by Minsky and Papert (1969) showed the limitations of the then-existent Perceptron in discriminating very simple visual stimuli. Despite such limitations, however, various researchers continued to work on this framework and the "renascence of neural nets", that took place in the 1980s, happened in ground that was still fertile. Nevertheless, "this renascence was marked by at least two crucial events, accompanied by the development in those years of computers with great computing power, allowing them to simulate neural nets of increasing complexity" (Cordeschi, 2002: 213). In particular, in 1982, John Hopfield proved that symmetrical neural nets necessarily evolve towards steady states – then interpreted as attractors in the dynamic system theory – and that they can function as associative memories (Hopfield, 1982). In 1985, James MacLelland, David Rumelhart, and their collaborators introduced the approach known as parallel distributed processing (PDP) of information by starting a number of investigations on natural language acquisition by emphasizing the role of artificial neural networks and of parallel computation

22 In 1936, Turing introduced the abstract computing machine bearing his name and explicitly construed a universal machine that could simulate, with appropriate encoding, any computation carried out by any Turing machine (including, of course, the universal one) (Turing, 1936–37).

23 Roughly speaking, so-called "Hebbian learning" consists of the evidence that when the axon of a given cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, this kind of associative connection A→B leaves a trace in the nervous system that learns this simple associative rule.

24 The Perceptron was one of the first neural network architectures. This simple form of neural network consists of a first layer, corresponding to the sensory system (an analog for a retina), which is randomly connected to one or more elements in a second layer of nodes: the association system. The latter consists of association cells, or A-units, whose output is a function of the input signal.

in the study of cognitive phenomena. They showed how a learning algorithm based on error correction, known as "backpropagation",[25] made it possible to overcome the main limitations of neural nets described by Minsky and Papert (Rumelhart, McClelland, & the PDP Research Group, 1986).[26] Back then, the achieved results had a strong echo since they were also considered the first example countering the predominant (in both linguistics and AI) Chomskian view of language processing, which took moves from the book "Syntactic Structures" (Chomsky, 1957), declaring the primacy of syntax and grammars. Since these pioneering works, connectionist systems have been widely adopted in a variety of applications in both the cognitive modelling and AI communities. Connectionist systems (and emergent systems in general) have been important in the AI landscape since they have provided more suitable solutions (with respect to the symbolic approach) able to deal with the environment and with the processing of the perceptual aspects of sensory input. In particular, they have fought the tendency of (early) symbolic AI to consider, in an isolated way, perceptual systems, motor systems, and high-level cognitive functions etc.[27] On the other hand, they have targeted the close interaction between the "mind" (natural or artificial), the body (i.e., the "hardware"), and environment.[28] This has led, in some cases, to radical

25 The backpropagation rule intervenes to change the weights of the connections between the hidden units, going backward from the error, which is calculated at the output units. Rosenblatt had anticipated the formulation of various aspects of this rule that, however, was fully formalised by Geff Hinton, winner of the Turing Award Prize in 2019 for, among the other things, the invention of the backpropagation algorithm.

26 The work and its assumptions were not free from criticisms. See, for example, Pinker and Prince (1988) and the subsequent debate that dominated the late 1980s and 1990s.

27 This tendency was, in a later period, contrasted also within the cognitivist/symbolic approach by Allen Newell. While Simon, in fact, continued his development of "microtheories" or "middle-range" theories (see Cordeschi, 2002) by focusing on the refinement of the analysis of verbal protocol, Newell didn't consider the construction of single simulative microtheories a sufficient means to enable the generalisation of "unifying" theories of cognition (the original goal of Information Processing Psychology). Therefore, diverging from Simon, he proposed building simulative programs independent from single cognitive tasks and able to include invariant structures of human cognition. In this way, he started the enterprise of studying and developing integrated and multi-tasking intelligence via cognitive architectures that would have led to the development of the SOAR system.

28 It is worth noticing, however, that in classic "cognitivist" tradition as well the importance of the environment in the deployment of intelligent behaviour was somehow recognised. Herbert Simon, in fact, in his lecture series on "The sciences of the artificial" (later published as a famous book with the same title), introduced the so-called "Ant metaphor", which would later come to be known as "Simon's Ant metaphor" and which can be described as follows: "An ant, viewed as a behaving system, is quite simple. The apparent complexity of its behaviour over time is largely a reflection of the complexity of the environment in which it finds itself". Simon then applies this consideration to human beings by suggesting that the apparent complexity of human behaviour is also largely a reflection of the complexity of the environment in which we live. Therefore he suggests that the environment should play an important role in building simulative models of cognition since "the behaviour takes on the shape of the task environment". Despite these relevant insights, however, early AI systems assuming the PSSH did not succeed in integrating such aspects in their models and were severely criticized by proponents of the

assumptions that have proposed the complete elimination of the notion of "representation" (intended in the cognitivist/symbolic sense) from the vocabulary of the cognitive and artificial sciences. This movement was led by the roboticist Rodney Brooks through the proposal of the so-called "Subsumption Architecture" (Brooks, 1986, 1991). This proposal consists of a layered, decentralized, robotic control architecture that does not make any use of internal representation of the world (i.e., the motto of this view is "use the world as a model"), where the relevant parts of the control system interact and activate each other through sensing the world. Subsumption architecture has been very influential from an engineering point of view since a vast variety of effective, implemented robotic systems use it.[29] It is based on the so-called "creature hypothesis", according to which the most important part in the design of an intelligent artificial system can be reduced to the difficulty of building a machine that act as smart as an insect. In other words, the underlying assumption of such a hypothesis is that once the perceptual/reactive part of a "creature" (natural or artificial) is built, then building the rest of the intelligence features is an easy task to achieve. The figure 1.2 below shows the characteristics of this kind of architecture. Each layer, programmed by using finite state machines of problem solving was assumed to deal with specific tasks (e.g., the task of avoiding obstacles, wandering, seeking, etc.) and higher levels of the hierarchy subsume the actions of lower levels. The design of successive task-achieving layers is stopped once the overall desired task is achieved (Figure 1.2).

Such radical proposal, however, has also shown significant limitations. In fact, even if they lead, through the development of innovative architectures for decentralized action control, to the ability of acting in non-structured environments in real time, these systems nevertheless showed their limitations when asked to deal with more high-level cognitive tasks, such as planning, reasoning, multi-agent coordination, and so on. Such tasks, on the other hand, were dealt with in a more satisfactory way via the symbolic approach, thus suggesting the practical utility of the notion of "representation".

The classical move, in this case, was the adoption of hybrid approaches trying to connect low-level and high-level faculties by integrating neural and symbolic approaches. Investigations of the integration between "symbolic" and "subsymbolic" in AI have coexisted during recent decades, but despite the realization of

---

"emergentist" paradigms. Emergentist modelling approaches, in fact, have proven to be more efficacious in modelling the environment and its intervention in the emergence of intelligent behaviour.

29 The first implementation of such an architecture was executed in robots like "Allen" and "Herbert", developed by Brooks and his group at MIT in the late 1980s. In particular, Herbert, a soda-can collecting robot, was able to exhibit the following capabilities (uncommon at that time): moving around in a real environment without running into obstacles; detecting soda cans using a camera and a laser; using an arm that could extend, sense, and evaluate whether or not to pick up the soda can, etc. Nowadays, Subsumption Architecture is employed in the most successful robotic platform so far: the Roomba robot!

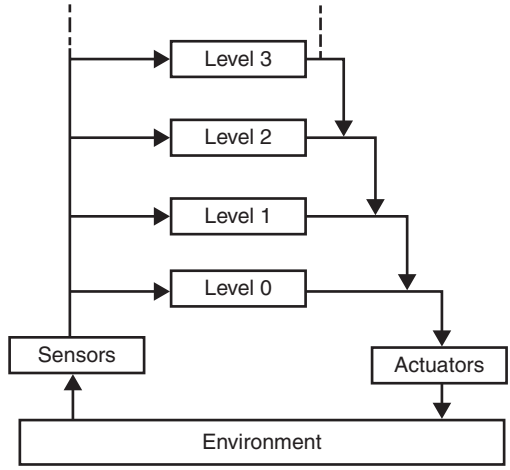

FIGURE 1.2 Brooks' Subsumption Architecture (adapted from Brooks, 1999).

many hybrid systems, a general solution to the problem of the *ad-hoc* integration of such heterogeneous components does not yet exist. In particular, connectionist models have continued to achieve the best results in handling activities like pattern recognition and classification or associative learning. They have failed, however, in handling higher cognitive functions, like complex inference-based reasoning, which are better modelled by symbolic approaches.[30] A well-known problem of these connectionist representations, for example, concerns the difficulty of implementing compositionality in neural networks (Fodor & Pylyshyn, 1988). Finally, another classical problem of artificial neural networks is represented by their "opacity": a neural network behaves as a sort of "black box" and specific interpretation for the operation of its units and weights is far from trivial. Despite such foundational problems, today neural networks are used in a variety of fields that range from machine vision to natural language processing to autonomous cars, due to the success of the new generation of deep learning architectures. On the other hand, symbolic approaches also suffer from a number of problems, other than the above-mentioned ones of dealing with commonsense reasoning and commonsense compositionality; these range from the "frame problem" (McCarthy & Hayes, 1969) to the "symbol grounding" one (Harnad, 1990). In short, the frame problem consists of a difficulty in formally

30 The novel generation of connectionist models based on deep learning have also recently gained attention for the results obtained in tasks like automated machine translation (Jean et al., 2015). However, this success in language based tasks seems to be mainly obtained because that task has been treated as a machine vision task, where the structure (i.e., the patterns) of a source language had to be mapped and compared with the one of a target language. Despite these new achievements, deep learning language models still provide poor results, compared to other approaches, in high-level cognitive tasks, ranging from Question Answering and Narrative/Story Comprehension to Commonsense Reasoning.

representing, in logic-based representational languages, changes in an environment in which an agent (e.g., a robot) has to solve some tasks without having to explicitly resort to an enormous number of axioms to also exclude a number of intuitively obvious – for humans – non-effects. For example, if a robot places a cup on a table, it is necessary not only specify that the cup is now on the table, but also that the light remains on, that the table is still in the same place, that the robot is still in the same room, etc. The symbol grounding issue, on the other hand, concerns the problem of how to obtain the grounding between symbolic representations and the corresponding entities that they denote in the external world. This is notoriously hard for symbolic systems and is alleviated in connectionist systems, since the data they directly take in input (e.g., images, signals, etc.) are closer to the perceptual "real world" sensory data. Summing up: the last 65 years of applied research have shown that both the main modelling approaches developed in the context of cognitive modelling and AI communities have different strengths and limitations. In any case they are not able to account, if considered in isolation, for all aspects of cognitive faculties.

Death and rebirth of a collaboration

As showed in the previous sections, AI pioneers were explicitly inspired by research on human cognition, and the cognitive approach was considered – without any doubt – to be the best strategy to pursue, so as to build intelligent machines (see Lake et al., 2017). Schank (1972), in the journal *Cognitive Psychology*, declared, "We hope to be able to build a program that can learn, as a child does, how to do what we have described in this paper instead of being spoon-fed the tremendous information necessary". A similar sentiment was expressed by Minsky (1975):

> I draw no boundary between a theory of human thinking and a scheme for making an intelligent machine; no purpose would be served by separating these today since neither domain has theories good enough to explain or to produce enough mental capacity.

This initial (excessive) enthusiasm, however, started to vanish (a fierce critique on the over-the-top optimism of that period was given by Hubert Dreyfus, 1972) and, after the first few decades of pioneering collaborations, starting from the mid-1980s, AI and the new-born interdisciplinary field of Cognitive Science started to produce several sub-fields, each with its own goals, methods, and evaluation criteria. On one hand, this divorce from considering human or nature-inspired heuristics has led AI to achieving remarkable results in a variety of specific fields (by focusing on quantitative results and metrics of performance, and on a machine-oriented approach to the intelligent behaviour). On the other hand, however, it has significantly inhibited cross-field collaborations and research efforts targeted at investigating a more general picture of what natural and

artificial intelligence is, and how intelligent artefacts can be designed by taking into account the insights coming from human cognition.

In the last few years, however, the cognitive approach to AI has gained renewed consideration, both from academia and the industry, in wide research areas such as Knowledge Representation and Reasoning, Robotics, Machine Learning, Bio-Inspired Cognitive Computing, Computational Creativity, and other research fields that aspire to human-level intelligence. Nowadays, in fact, artificial systems endowed with human-like and human-level intelligence (McCarthy, 2007) are still far from being achieved and, using the words of Aaron Sloman, "the gap between natural and artificial intelligence is still enormous" (Sloman, 2014). This sort of "cognitive renaissance" of AI still considers the "cognition in the loop" approach as a useful one to detect and unveil novel and hidden aspects of cognitive theories by building properly designed computational models of cognition, which are useful to progress towards a deeper understanding of the foundational roots of intelligence (both in natural and artificial systems).