Cognitive Science and Artificial Intelligence: An Interwoven Approach

Supervisor: Francesco Bianchini

Paolo Marzolo

July 2, 2021

Abstract

Due to the complex nature of the matter discussed, the relationship between Artificial Intelligence and Cognitive Science is often misrepresented. This is, in part, due to ignorance of the disciplines considered, which range from Mathematics, Statistics and Computer Science to Philosophy, Psychology and Neuroscience. In this text, we attempt to outline the shared history of the study of intelligence, intelligent behavior and rationality. We do this with an especially attentive focus to where paradigms and theories fall on the symbolic-connectionist spectrum. To achieve this, we include technical explanations for some of the theories mentioned. In the final section, we take a cross-historical approach to help identify trends and conclude with a brief overview of symbolic-connectionist integration proposals.

Contents

Chapter 1

Introduction

This is Paolo Marzolo's bachelor thesis, written as part of the three-year program in computer science at University of Bologna. The stated objective of this document is to analyze the history of Cognitive Science and Artificial Intelligence and identify how influences between the two disciplines and others led to a partially shared evolution in the overarching research topics throughout their lifespans. Other similarities will be pointed out. Some of the algorithms and concepts contained throughout the sections will be explained in detail, in order to give the reader a complete understanding.

The structure of the document will be as follows: after this introduction, a brief glossary will introduce some of the terms that will be used in this document with a short definition; this was included to avoid having "foundational" terms be constrained by a specific philosophy or line of research. Then, the rest of the document will develop parallel to the history of the disciplines. In the final section, a bird's-eye-view will provide additional insight, and a brief discussion of the roles of symbols will conclude the contents.

Chapter 2

Terms and Definitions

Before definining our glossary, it is important to understand the reasoning behind why we chose to include it. When discussing researchers' understanding of human thought, it is nearly impossible to avoid using terms that have a strong history. As an example, simply the mention of "thought" could already be considered too far from a behaviorist perspective. A further example is a recent discussion that took place after a controversial paper by Nunez was published (Núñez et al. 2019), questioning the multidisciplinarity of Cognitive Science as a discipline (and journal) and declaring "The prospect launched by the cognitive revolution of a unified and coherent interdisciplinary seamless cognitive science did not materialize". This questioned the existence of the field of study itself: we will discuss our usage of the term in the first entry of this glossary.

Cognitive Science. As we will see in following sections, saying "definitions of Cognitive Science have evolved throughout the years" would be a massive understatement. Nonetheless, its multidisciplinary nature is clear in what the International Encyclopedia of Social & Behavioral Sciences (*International Encyclopedia of Social & Behavioral Sciences - 1st Edition* 2021) reports 'may have been the first published use of the term cognitive science':

> The concerted efforts of a number of people from ... linguistics, artificial intelligence, and psychology may be creating a new field: cognitive science.

At the same time, even the "essential original features" identified by Gardner in 1987 (Gardner 1987) (summarized here as (1) necessity to speak about mental representation as a separate layer of analysis from the biological, (2) faith that the computer is central to the understanding of the human mind and (3) deemphasizing factors such as emotions or cultural factors) would be completely or partially thrown out by contemporary scholars.

In a more recent publication (Boden 2008), Cognitive Science is characterized as

> ... better defined as the study of 'mind as machine' ... More precisely,

> cognitive science is the interdisciplinary study of mind, informed by theoretical concepts drawn from computer science and control theory.

Not only was its definition cloudy and unstable ("cognitive science is ... a perspective, rather than a discipline in any conventional sense" (Sheehy and Chapman 1995)), but as Nunez points out, its disciplines have varied frequently throughout the years, and a variation in how represented they are in the Cognitive Science enterprise followed. Because of the reasons outlined here, far removed from the subject of this document, we will avoid using the term "Cognitive Science", and prefer the acronym "DCS", for Descriptive Cognitive Sciences.

Descriptive Cognitive Sciences (DCS). As we mentioned, the disciplines which make up Cognitive Science are not only multiple, but subject to interpretation as well. Since the objective of this work is to compare it to the history of Artificial Intelligence, we will from this point on use the acronym "DCS", for Descriptive Cognitive Sciences, as an alternate approach to the Constructive one taken by Artificial Intelligence researchers. This is not to say that there cannot be *constructive, psychological* approaches to the explanation of cognition: the only reason we chose this is because we found it to be an intuitive use of the term.

Mind. Once again, although we take notice of the history of the term, we have to select a few terms to use in our language. Hereafter, we consider the mind as the non-physical correlate of human brains: "the complex of faculties involved in perceiving, remembering, considering, evaluating, and deciding. Mind is in some sense reflected in such occurrences as sensations, perceptions, emotions, memory, desires, various types of reasoning, motives, choices, traits of personality, and the unconscious." (*Mind* 2021).

Chapter 3

A History of Influences

As mentioned in the introduction, our approach will follow the historical sequence of events, although some references or explanations may be anachronistic for clarity. In order to give a general view, we split the histories of these disciplines into broad periods: one for (more or less) every substantial shift in approach and views. Generally, every time period will mention two sides of the story: one will focus on DCS, and the other on AI and Computer Science. At the same time, the two fields are in close relation: because of this, it becomes hard to neatly define the two fields, so some paragraphs will be somewhere in the middle (as they should be!).

3.1 Landscape before 1950

Although the official birth of the "Cognitive Science" institutions is in the late 1970s, reasoning about thought has been a staple in philosophical research for centuries. Because of the scope of this document, we will focus on a few critical concepts, and use them to set the stage for the first significant shift of ideas.

3.1.1 Mathematics and Computer Science

Some of the most relevant contributions to the "reasoning as computation" line of research come from Mathematics and what would later become Theoretical Computer Science. We will outline some of them here, while we trace part of the history of conceiving of thought as computation, first, and computers as devices for computation, second. In this respect, the following step is to be expected: can we use devices for the computation that thoughts "work" with?

Laws of thought modeled in mathematics, using algebra

Boole's Laws of Thought and Boolean Algebra. To avoid going too deep in mathematical concepts for our purposes, we can think of Boolean algebra as the branch of algebra where the variables can be either true or false (1 and 0), and the main operations on its variables are conjunction (and, $\wedge$), disjunction (or, $\vee$), negation (not, $\neg$). Through these, logical operations can be

described. In "An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities", one of the author's two monographs on algebraic logic, George Boole, then mathematics professor in Ireland, introduces Boole's algebra as an extension to Aristotle's logic. In it, Boole provides Aristotle's algebra with mathematical foundations, and expands it from two-term to any-term. Boole's algebra differs from modern Boolean algebra (in Boole's algebra *uninterpretable* terms exist) and cannot be inteprepted as set operations; still, its introduction marked a step towards the formalization of laws of thought and a possible bridge between mathematical research and thinking processes (even the title of the book it was introduced in gives a very clear direction). Boolean algebra would instead be developed by Boole's successors (Jevons, Peirce, Schroder and Huntington in particular); this work allows boolean algebra to now be defined by the Stanford Encyclopedia as

> the algebra of two-valued logic with only sentential connectives, or equivalently of algebras of sets under union and complementation.

Modeling algebra, so thought, is possible through some computational structure

Automata theory. The study of how automatic calculators (more properly, abstract machines or automata) can be used to compute and solve problems is a part of theoretical computer science research. The history of Automata Theory is especially interesting, as it will let us meet some important researchers: it features two neurophysiologists, Warren McCulloch and Walter Pitts, and is thus born from the desire of modeling human thought itself. The first model was proposed in 1943 (McCulloch and Pitts 1943), in a seminal paper that also introduced other research themes we will come back to later. A little over twelve years later, two computer scientists, Mealy and Moore, generalized the theory to more powerful machines, "finite-state machines". The general idea behind them is this: starting from an input and a set of states, a *transition function* maps the current state and an input to an output together with the next state. They do not have any memory, and as such can only "solve" simpler problems: if used to recognize languages, they can only recognize regular ones.

More powerful abstract machines had already been proposed: Turing had introduced "Turing machines" in 1937 (Turing 1937), as part of his proof of the Entscheidungsproblem. The relationship between automata "expressive power" and language complexity is outside the scope of this document.

Study of feedback is subject-agnostic, self replication, artificial neural networks

Cybernetics. Although in recent years the term "cybernetic" has been used to mean futuristic/sci-fi technology, Cybernetics is a transdisciplinary discipline that studies regulatory systems. The core of the discipline is feedback loops (or circular causality), where the result of actions is taken as input for (choosing) future actions. Cybernetics is not bound to any particular usage, so its applications include biology, sociology, computer science, robotics and many others. Its flexible approach led to many different definitions: two early ones are the one used in Macy cybernetics conferences, "the study of circular causal and feedback mechanisms in biological and social systems" (Steer 1952), and the definition by Norbert Wiener, considered the originator of cybernetics, "the scientific study of control and communication in the animal and the machine"

(Wiener 1961). Although Plato used the word itself to signify the governance of people, our interest resides in contemporary cybernetics, born in the 1940s. Before the paper mentioned above by McCulloch and Pitts, the study of feedback was considered by Anokhin in 1935 (Anokhin 1935) (physiologist). In the same year as the McCulloch-Pitts paper was published, Wiener, together with Rosenblueth and Bigelow, published "Behavior, Purpose and Teleology" (Rosenblueth, Wiener, and Bigelow 1943): these three researchers, together with McCulloch, Turing, Grey Walter and Ross Ashby, would go on to establish the discipline of cybernetics. Wiener coined the term to denote "teleological mechanisms".

An important addition to the field would be the Von Neumann cellular automata: these are another model of computation part of automata theory. A cellular automaton is a grid of cells (of any dimensions, but for clarity, consider a 2-dimensional one first), where each cell has a finite number of states it can be in; the cellular automata evolve by moving from generation zero ($t = 0$) to the next generation ($t = 1$) following mathematical rules: the state of every cell is determined by its past state and the surrounding cells. Without going into the specific rules Von Neumann determined, this is relevant to us because it introduces two fundamental concepts: self replication, soon to be adopted by cybernetics as a core concept, and the formal study of evolutionary mechanisms in simulation. Another fundamental contribution from cybernetics is the creation of Artificial Neural Networks, introduced in the same McCulloch-Pitts paper we mentioned earlier.

Information theory and technical advances. As we have seen, theoretical advances were many and varied, but the technical advances were what drove the ability to put those into practice. Among those, we have to mention the move from electromechanical devices to vacuum tube-based computers, which gave birth to a device for controlling the connections between telephone exchanges, thanks to Flowers, in 1934. The record for the first general-purpose stored-program (as in, controlled by wires, the opposite of a stored-program computer) went to Konrad Zuse, with the Z3 machine. This machine also used a binary system, but it was not a universal computer. In 1944, the Bletchley Park cryptanalysts started using Colossus. The first Turing-complete (i.e. with the same computing ability as the Turing machine) computer was completed in 1945. It used over 18.000 vacuum tubes. The first stored-program computer, built as a testbed for new technology and design, was the Manchester Baby, ran in June 1948 (*Computer Resurrection Issue 20* 2012).

As part of the advances of this period, we must mention the birth and development of Information Theory. Information Theory encompasses the study of quantification, storage and communication of information, in digital form. After being introduced by Nyquist and Ralph (Nyquist 1928), the field was firmly established by Shannon's "A Mathematical Theory of Communication" in 1948. Without going into details, its main influences include the bit as a unit of informationa and the necessity of redundancy of a source when using unreliable communication channels.

Lastly, we note that neuroscience had new tools at his disposal: electrophysiological techniques, such as brain stimulation, single cell recording and EEG recording (*International Encyclopedia of Social & Behavioral Sciences - 1st Edition* 2021) were instrumental to the research into localization studies (such as deficits derived from brain lesions) approached by Geschwind in the 1950s.

3.1.2 DCS

The DCS landscape around 1950 was strongly rooted in Behaviorism, with hints of the revolution that was soon to come. Some of the larger influences from the Computer Science side, such as the McCullough Pitts artificial neural network we mentioned, would in fact have a relatively small impact and be re-discovered later.

Psychology as a science

Behaviorism. Behaviorism emerged as the dominant school in Western psychology as a reaction to depth psychology and other forms of psychology that did not fit well with scientific experimental verification. That is not to say it was unprecedented: Thorndike presented the law of effect (using consequences to strengthen or weaken behavior) in 1898. Still, behaviorism was introduced as "methodological behaviorism" by a 1924 publication by John Watson (Watson 1924), and then further expanded by many researchers, among which B. F. Skinner.

Behaviorism, more than a way to impose empirical constraints on studying psychology, is a doctrine of how to do behavioral science itself. The Stanford Encyclopedia identifies three claims as the roots of behaviorism (as a doctrine):

- Psychology is the science of behavior. Psychology is not the science of the inner mind – as something other or different from behavior.
- Behavior can be described and explained without making ultimate reference to mental events or to internal psychological processes. The sources of behavior are external (in the environment), not internal (in the mind, in the head).
- In the course of theory development in psychology, if, somehow, mental terms or concepts are deployed in describing or explaining behavior, then either (a) these terms or concepts should be eliminated and replaced by behavioral terms or (b) they can and should be translated or paraphrased into behavioral concepts.

These fundamental truths identify three of the various flavours behaviorism is studied in. Skinner, mentioned above, was the first to suggest that covert behavior, such as cognition and emotions, is governed by the same controlling variables as observable behavior: although focused on the third "truth", his philosophy combines all three mentioned pillars, and is described as *radical behaviorism* by skinner himself (B. F. Skinner 1974).

One can easily see how the philosophy itself forced the practitioners into a state of absolute experimental dependency, which constrained the concept

explored to the scientific realm. At the same time, its complete rejection of mental processes (or at least their relevance to scientific study) is the complete opposite of the assumptions that were made on the "CS" side of comprehension. Other behaviorists, though, were less radical: Clark Hull was willing to put drive inbetween stimulus and response, but only to create a corresponding theory that explained it in terms of behavior (Hull 1931); Edward Tolman, instead, proposed rats navigate a maze following a mental map (Tolman 1948).

Psychophysical isomorphism

Cognitive signs. Just like Tolman, other cognitive-leaning psychologists proposed ideas that did not fit with the behavioral narrative. Among them, we mention some relevant ones. The Gestalt psychology refused the behavioristic assumption that conscious experience could be considered by reducing it to the sum of it parts, and proposed the principle of totality. It also proposed the principle of psychophysical isomorphism, which meant the cerebral activity was correlated to conscious activity (Wagemans et al. 2012). Vygotsky and Luria pioneered "cultural-historical psychology", which noted the role of culture and language in the development of higher psychological functions; Luria, alone, also published research on individuals' thought processes as his doctoral dissertation.

Lastly, we mention Miller, then a trainee at Stevens's Psychoacoustic Laboratory at Harvard: he will soon become relevant, as part of the 1956 cognitive revolution.

In our exploration of the state of disciplines around 1950, it is clear that Computer Science was firmly en route to a first attempt at thought modeling though mathematical "symbols": if, as they suspected, thought was to be considered a use of (or better yet, possible to model with) algebra, then once physical computers were capable enough they would be capable of thought. On the other side of the fence, DCS was still firmly rooted in behaviorism: in their view, the entire discussion would be based on false premises which were in turn based on wrongful research; the roots of human behavior were to be found in human behavior itself, and assuming otherwise was not only useless but unscientific, as it would lead to unprovable theories and impossible experiments. At the same time, cognitive suggestions were starting to appear, challenging the general (or at the very least American) current view.

3.2 1956: A Pivotal Year

As we have mentioned in the previous section, there were various lines of research into thought modeling: automata theory was focused on what problems were possible to model, cybernetics took (analog and biology-based) feedback and self replication as founding pillars, while information theory dedicated itself to information storage and transmission. Instead, DCS was still mostly led by behaviorist views, but cognitive-oriented proposals started to emerge. This trend would continue in 1956, and spike in 1957 with a publication by Noam Chomsky that would change the field.

3.2.1 CS

The most relevant event of 1956 (and quite possibly of the history of AI research) is the Darmouth College Workshop, a sort of convention that connected researchers from diverse fields interested on similar topics. This is also the context in which the term "Artificial Intelligence" was attached to the field.

Birth of AI

Dartmouth College Workshop. As we said, at the start of the 1950s thinking machines were being inspected by a few different disciplines; in 1955 John McCarthy, an Assistant Professor at Dartmouth, proposed a conference to organize and fertilize such disciplines. He proposed the name "Artificial Intelligence" because, unlike today, it was still neutral; Wikipedia reports that avoiding cybernetics was partly due to 'him potentially having to accept the assertive Norbert Wiener as guru or having to argue with him'. The project was formally proposed in September, by four of those who would become (if they weren't already) prominent researchers in the field: McCarthy himself, Marvin Minsky, Nathaniel Rochester and Claude Shannon. Among the extraordinary attendees we mention: Minsky (who will become very relevant in the next section), Bigelow (co-author of the seminal paper "Behavior, Purpose and Teleology." on cybernetics), Solomonoff (inventor of algorithmic probability), Holland (pioneer of genetic algorithms, was invited but did not end up attending), Ross Ashby (psychiatrist and cybernetics pioneer), McCulloch (who we've already mentioned), Nash (prolific mathematician, also known for his work on game theory), Samuel (creator of what is considered the first AI program, a checkers program) and finally Allen Newell and Paul Simon, who presented their recently completed "Logic Theorist". Although the discussions were not directed, many of the topics would have a long-lasting impact on the field, like the rise of symbolic methods and limiting domains (which would lead to expert systems).

Reasoning as search

Logic Theorist. The Logic Theorist was created in 1955 by Newell and Simon, helped by the systems programmer John Shaw. In order to prove a theorem, the simplest strategy is to start from the theory's postulates and create new theorems by combining them; then continue by combining every theorem with every postulate and every theorem again, exploring the entire truth spectrum. Although this may seem obvious, this is part of the first important concept introduced by the Logic Theorist: seeing the truth space as a tree, that started with the hypothesis and aimed at the proposition to prove; envisioning *reasoning as search*. Of course, exploring the entire tree is impractical, because of the time it takes to explore the entire truth space (as it grows exponentially); when considered from the point of view of "modeling the human thought process", this solution would not be useful even if it was practical, because this is not how human theorem-provers work. In order to solve this problem, the Logic Theorist introduced the second important factor: employing *heuristics* to ignore branches that were unlikely to lead to the goal. The last important factor is technical: in order to implement the Logic Theorist, the authors implemented IPL, a programming language that used symbolic *list processing* in the same

way as the following, fundamental, Lisp.

3.2.2 DCS

This section, will talk about some of the important findings that seemed difficult to integrate with behaviorism and the important actors behind them.

Memory study

Miller. George Miller, before becoming one of the founders of cognitive psychology, was of the behaviorist school (although he later wrote of one of his works on language 'By Skinner's standards, my book had little or nothing to do with behavior' (**millerCognitiveRevolutionHistorical2003**)). After slowly moving to the cognitive side, driven by similar thinkers ('Peter Wason, Nelson Goodman and Noam Chomsky had the most influence on my thinking at that time'). In 1956 he published a paper that had a sizable impact: "The magical number seven, plus or minus two". In it, he observed tht various experimental findings revealed that, on average, human can hold seven items in short-term capacity. We note that it is not the finding that goes against behavioral philosophy and psychology, but the framework in which it is put in general: such attention to mental processes would be irrational, when seen from a behavioral point of view, who disregard mental processes as a whole.

Categories as understanding

Bruner, Goodnow, Austin, and the basis of cognitive science. Another important book published in 1956 is "A Study of Thinking", by Bruner, Goodnow and Austin. The book is focused on using categories for concept formation, or how human beings group the world of particulars into classes, together with the results of relevant experiments. Before such experiment on cognition, Bruner had dedicated himself to the study of perception: two relevant studies we report are the one on estimating the sizes of coins or similarly sized wooden sticks (the first were significantly overestimated), and another one on slowing reaction times while playing cards in connection with reversed suit symbols. These two experiments are relevant because of the focus on the internal interpretation of external stimuli. Other foundational ideas of cognitive science, developed in the years following the Miller publication, include the application of the scientific method to human cognition (if anything was to come after behaviorism, it could not avoid its history of scientific "rigor"), the interest towards information processing and storage, and as we will see in the next paragraph, a degree of possible innateness.

Innateness, productions, syntax

Chomsky and the final departure. In this last section, we move further than 1956. Nonetheless, it is extremely relevant to the subject discussed, and represents the most decisive blow (in purely historical terms; this document has no psychological authority to express an evaluation of *any* theory) to behaviorism. In 1957, Skinner published "Verbal Behavior". In it, he describes the controlling elements of verbal behavior, and attempts to form a hypothesis about the behavioral framework with which verbal behavior is to be understood. In it, he uses specific terminology for his analysis, using both existing words and neologisms; in his own words: 'The emphasis [in Verbal Behavior] is upon an orderly arrangement of well-known facts, in accordance with a for-

mulation of behavior derived from an experimental analysis of a more rigorous sort. The present extension to verbal behavior is thus an exercise in interpretation rather than a quantitative extrapolation of rigorous experimental results' (Burrhus Frederic Skinner 1957).

In the same year, Chomsky proposed a different model for understanding language; in "Syntactic Structures" he argued two important points that would have a large impact on the field of linguistics:

1. **Syntax vs semantics.** The first point he makes is the clear distinction between syntax and semantics: '...such semantic notions as reference, significance, and synonymity played no role in the discussion.'

2. **Generative grammars.** His approach to syntax was formal, and followed both his teacher's (Zellig Harris) and notions advanced by Danish linguist Louis Hjelmslev: language was to be understood as a generative grammar, which bound by "phrase structure rules" (producing new sentences) and "transformations" (modifying exising sentences).

This seminal paper would soon be interpreted as an argument for a mentalistic, *innate* view of language production. However, this intepretation was not originally put forth in the book itself: '[Chomsky's generative system of rules] was more powerful that anything ... psycholinguists had heretofore had at their disposal. [It] was of special interest to these theorists. Many psychologists were quick to attribute generative systems to the minds of speakers and quick to abandon ... Behaviorism' (Steinberg, Nagata, and Aline 2013).

Two years later, Chomsky published a scathing (this time both in historical terms and considering the tone of the paper) review (Chomsky 2013) of the book which had a widespread effect of the decline of behaviorism's influence. In it, one of the points he argued was that children are not taught the rules of language, and the amount of input they receive is not sufficient to derive them. This argument would later be called the "Poverty of the Stimulus" argument, and to this day represent a very controversial issue of linguistics and language acquisition.

In the words of Newmeyer (Newmeyer 1986):

> Chomsky's review has come to be regarded as one of the foundational documents of the discipline of cognitive psychology, and even after the passage of twenty-five years it is considered the most important refutation of behaviorism. Of all his writings, it was the Skinner review which contributed most to spreading his reputation beyond the small circle of professional linguists.

The review has been criticized by other writers, such as MacCorquodale (MacCorquodale 1970), but its effect cannot be ignored.

As we have explored in this section, 1956 was both the culmination and start of a cognitively-inspired revolution across the DCS. As behaviorism grew less popular, cognitive findings and research drew more interest. At the same

time, one of the very first AI programs was presented, and it tackled a purely symbolical problem with a purely symbolical approach: the trend was clear, and it was pushing towards a cognitive approach.

3.3 1960-1970: Great Promise

The years after 1956 are considered by many the "golden years" of AI research: thanks to considerable successes and a general wave of optimism, money was poured into the field, which thankfully generated more results and increased the hopes again. Although some interest was generated towards neural networks, this was completely shut down by a Minsky critique in 1969 (analogous to the Chomsky critique of "Verbal Behavior"). Psychology saw the rise of research into representations, categories and memory, as we will briefly overview in this section.

3.3.1 Computer Science and Artificial Intelligence

For what concern Computer Science and Artificial Intelligence, this period saw various directions, inspired by some of the previous research we've touched on in the previous sections. Most of them were focused on symbolic AI; at the same time, the "perceptron" proposal from McCulloch and Pitts saw some interest, before being shut down for more than ten years.

Difference from current to goal, self play

Reasoning and the General Problem Solver. As we mentioned, an important paradigm was introduced with the Logic Theorist: seeing reasoning as search. In this respect, we present the work of Samuel and Newell and Simon, both presented in 1959. Newell and Simon worked on what they hoped could become a general version of the LT, the General Problem Solver. Although the paradigm of reasoning as search was maintained, the GPS did not prune paths that were unlikely to lead to the goal, but used *means-ends analysis* to limit search. When following MEA, a system chooses, given a current state, an action that reduces the difference between the current state and the goal state. By focusing on the difference between current and goal state, MEA improves on brute-forcing all possible choices. In addition, if knowledge about the relative importance of differences is available, the goal-seeking system can follow the path which decreases the difference most, further pruning the possible choices. The correspondance difference-action, also called operator, must be given as an input, and represents "a priori knowledge" of the problem. This separation between problem-specific knowledge and strategy of how to solve it is a relevant feature of the project, and an important point when compared with the following paradigm, expert systems. Samuel, instead, presented a checkers playing program, with several fundamental ideas: the program worked by exploring a search tree of the reachable board positions, while scoring each position to prune the search tree. Samuel also had the program play against itself to become a better player, and memorize positions and evaluations to effectively extend the search depth in those positions.

Knowledge-based systems

Symbols and successes. Following the GPS, other symbol and knowledge-based systems led to great successes. In 1958, the same year in which he invented Lisp, McCarthy published "Programs with Common Sense" (McCarthy 1960); in it, he described a hypothetical program that used general knowledge to search for solutions to problems (such as generating a plan to drive to the airport). To be called the Advice Taker, it also allowed for additional knowledge (axioms) to be introduced during the course of operation. As such, it embodied an important principle of knowledge representation: manipulating a formal representation of the world and its workings as a mean to solving problems. For later purposes, we mention the Shakey project at Stanford, which used subgoals (like GPS) and logic to control a robot. Minsky, who moved to MIT in 1958, started supervising students who tackled limited problems that seemed to require intelligence to solve: these would become known as microworlds. Two of these were Daniel Bobrow's STUDENT (1967), which could solve high school algebra word problems and Tom Evans's ANALOGY (1968), that solved geometric analogy problems from IQ tests. Research based on microworld continued throughout the 1970s.

Finally, we mention the different perspective taken by Joseph Weizenbaum, then MIT professor: between 1962-1964 he created ELIZA, a natural language processing program that mimicked conversational ability while following a simple script, the most popular of which was the "Rogerian" DOCTOR. Although the creation of it was meant to show how superficial interaction between machines and people really is, it gave (although briefly) the impression of an intelligent interaction; it did not even have any storage, so links between sentences were impossible. Its relevancy is now both historical, as it was the first attempt at creating the illusion of intelligence through human-machine interaction, and ethical, as its creation led to some important (and intended!) ethical questions regarding its usage as a therapeutic tool.

Solving without predetermined algorithms

Genetic algorithms and perceptrons. Between the late 1950s and early 1960s (Friedberg 1958), Friedberg started researching machine evolution (later called genetic algorithms), with scarce success. The basic idea was to make a series of small modifications to a program, then select the best-performing variant and repeat the process until the result was good enough. Unfortunately, due to how immature representation research and because of computing power constraints, these showed very limited success and the program was dropped.

Following the work of McCulloch and Pitts, research on neural networks picked up. Bernie Widrow researched his adalines (Widrow and Hoff 1962), while Rosenblatt researched perceptrons. In addition, in 1962 Block showed, with the perceptron convergence theorem, that if a pattern of connection strenghts that matches a certain input data exists, then the learning algorithm can always adjust the strengths correctly (Russell and Norvig 2002).

3.3.2 DCS

DCS research in this period was mainly focused on modeling higher-brain function, such as memory, within the new framework of cognitive psychology. At the

same time, behaviorism shifted from the strict theoretical research and found itself evolving into Applied Behavior Analysis as a scientific discipline used in therapy.

Behaviorism shifts closer to its current form. In a study from 1959, "The psychiatric nurse as a behavioral engineer" (Ayllon and Michael 1959), the authors demonstrated the effectiveness of using a token economy to reinforce adaptive behavior for patients with schizophrenia and intellectual disability. The practical application of behavioral research grew throughout the years: a journal, the "Journal of Applied Behavior Analysis" was founded in 1968; the "Behavior Analysis" subdivision in the American Psychological Association was introduced in 1964; the "Applied Animal Behaviour Science" was founded in 1975.

Memory as a tripartite system

Memory research. A series of studies on memory by different researchers helped clear the picture: Sperling focused, throught the 1960s, on sensory memory, starting with his PhD thesis at Harvard in 1959, and papers like "The information available in brief visual presentations" (Sperling 1960) and "Successive approximations to a model for short-term memory" (1967). Peterson worked on short term memory, with "Short-term retention of individual verbal items." (L. Peterson and M. J. Peterson 1959), and Waugh studied the difference between short-term memory and long-term memory in "Primary memory." (Waugh and Norman 1965). This allowed Atkinson and Shiffrin to propose, in 1968, the Atkinson-Shiffrin memory model: it viewed memory as a tripartite system, split between sensory memory, short term memory and long term memory. Although the idea of tripartite systems wasn't novel (James et al. 1890) and some of the concepts it included, like rehearsal as the transfer mechanism, have been criticized by later research, it sparked additional interest in the area of memory.

Neurons and thresholds

Milestones and Neuroscience. In addition, 1960 was the inauguration year for Miller-Bruner center for Cognitive Studies. At the same time, neuroscience was developing: in 1962, the FitzHugh-Nagumo model was presented, as a semplification of the previous Hodgkin–Huxley model. These models gave a formal background to the activation behavior of neurons (once the stimulus reached a threshold, the system is briefly excited before going back to resting state). In the same period, Katz modeled neurotrasmission across synapses (Katz and Miledi 1967) (KATZ 1969), and from 1966 Kandel and others started examining biochemical reactions to learning and memory in Aplysia (a genus).

Lastly, we mention a line of research by Shepard and his student Cooper and Metzler, in which they showed that reaction time in subjects asked to determine whether a transformation was a rotation or a reflection increased linearly with rotation degree: this suggested an internal image that was being rotated.

Throughout the 1960s, cognitive research started to appear, and was soon to be institutionalized. Artificial Intelligence programs focused mainly on symbolic systems, but research on neural networks increased, with some important theoretical findings supporting the hopes of the ideators. This was to come to

a screeching halt, with, once again, a critical (and later partly controversial) review of the literature by Minsky and funding issues.

3.4 1970-1985: Symbols and Knowledge

During this period, AI left the connectionism world behind, partly following a Minsky literature review in 1969. Meanwhile, symbolic research continued, and found its new paradigm: using domain-specific knowledge to solve bigger, more complicated tasks in *narrower* fields. The DCS domain expanded on their previous views, and some important points of contact with AI were explored.

3.4.1 Computer Science

As we mentioned, research in this period mainly focused on symbolic systems. The addition of context and domain specific knowledge shifted interest from imitating human or semi-human intelligence to strictly solving problems as well as possible. Although these *expert systems* worked well, a series of companies and projects that overpromised advances were born, which lead to what is now commonly called "AI winter", when companies and nations realized their hope was, sometimes, unfounded.

Minksy and perceptrons. As we mentioned in the last section, research on connectionist models, although less popular than symbolic models, kept going. One of the most relevant techniques for training neural networks was introduced in 1969 (Bryson and Ho 1969), although the original research was about optimal control instead of machine learning. This is considered the first description of modern back-propagation (LeCun, Touresky, et al. 1988), but their version was never applied to machine learning, where backpropagation would be rediscovered in the 1980s. Instead, research on perceptrons slowed down considerably after 1969. In the words of the standard textbook on the topic "The subsequent demise of early perceptron research efforts was hastened (or, the authors later claimed, merely explained) by the book Perceptrons, which lamented the field's lack of mathematical rigor ... and noted the lack of effective learning algorithms for multilayer networks". The book in question, Perceptrons (Marvin Minsky and Papert 1969), also noted other theoretical limitations of perceptrons, some of which may have been misinterpreted and so contributed to the general feeling towards perceptrons in general. In the last chapter the authors mention (shortsightedly, in hindsight) that multilayer neural nets would be a "sterile" extension.

Expert systems

Integrating knowledge. The approach so far (consider for example the efforts behind the GPS) had been to solve problems by modeling human thinking processes in their most general form, to be then applied to a specific problem. With DENDRAL, a research project that started in 1965, the creators explored a different approach: it used a large number of special-purpose rules, extracted from analytical chemists, to infer molecular structure of molecules from mass

spectrometer data and the elementary formula of the molecule. As such, it became the first knowledge-intensive system. It was soon to be followed by other expert systems, such as MYCIN (1972 and onward (*MYCIN — artificial intelligence program* 2021)), which diagnosed blood infections, and gave rise to a new, very successful paradigm. Its successes brought AI into the commercially viable technologies, and the increase of demand brought an increased interest in knowledge representation schemes: in particular, the two main approaches were logic-based and frame-based. Frames were a Minsky proposal (M. Minsky 1975) to organize facts about objects and events into a large hierarchical taxonomy.

Microworlds

Natural language understanding. In the previous section, we introduced research on microworlds. The most famous microworld was the block world, a set of blocks on a tabletop (real or virtual), which had to be rearranged, one block at a time, according to instructions. The success of this microworld (SHRDLU was the name of the successful program (Winograd 1971)) derived from the cooperation of many different researchers, as "Artificial Intelligence: a Modern Approach" reports:

> The blocks world was home to the vision project of David Huffman (1971), the vision and constraint-propagation work of David Waltz (1975), the learning theory of Patrick Winston (1970), the natural-language-understanding program of Terry Winograd (1972), and the planner of Scott Fahlman (1974).

(Russell and Norvig 2002) As a fitting consequence to such successes, in 1976 Newell and Simon formulated the Physical Symbol System Hyposthesis (Newell and Simon 1976). It states that "a physical symbol system has the necessary and sufficient means for general intelligent action"; which means that any system possessing intelligence must operate by manipulating symbols. This statement would later be challenged by many researchers.

Natural language understanding was another area in which domain knowledge carried great importance: the success of the block world-natural language program was in fact due to its specificity, and a series of programs followed it (R. C. Schank and Abelson 1977) (Wilensky 1978) (R. Schank and Riesbeck 1981), which all focused on understanding natural language by reasoning with the knowledge required.

3.4.2 DCS

During the 1970s, Cognitive Science went from a collection of studies with similar intentions to a true discipline, complete with relevant courses, journal and grants for research. Points of contact with existing AI research became more common, as did researchers working in both fields. Some philosophers, possibly accusing the weight of the wrongful predictions of AI research, raised critiques towards the field in general, while others began developing collective theories of mind, from Fodor's functionalism to the Computational Theory of Mind.

ELIZA and therapy

Points of contact. In this paragraph, we will highlight two relevant con-

tact points during this period. The first is the result of a collaboration between Gordon Bower and one of his students, John Anderson. Bower had gone from learning theory and animal testing, to mathematical models of learning to, finally, cognitively oriented work about mental representation, such as the study of chunking for short-term memory usage (*Gordon H. Bower. - PsycNET* 2021). Their cooperation would give rise to a semantic network model named HAM, later described in their 1973 boook Human Associative Memory (Anderson and Bower 1973). Anderson would keep working on it, later adding a production system, increasing the types of nodes and links between nodes, and explaining the time it takes to perform a task as due the matching for the production system, until he would publish the ACT-R architecture, still in research today.

The second important connection was a clash between the author of ELIZA, mentioned in the last section, and the psychiatrist Kenneth Colby. Colby expanded on the work of Weizenbaum (Colby, Watt, and Gilbert 1966), and wrote what he considered was a "computer program which can conduct psychotherapeutic dialogue", with which Weizenbaum clearly disagreed. Later, in 1976, Weizenbaum published "Computer Power and Human Reason" (Weizenbaum 1976), where he declares that computers should never be allowed to make important decisions as they would always lack compassion and wisdom.

Mind as information processing system

Putnam, Fodor, and the Computational Theory of Mind. In this section, our aim is to give a brief introduction to the philosophical path around the Computational Theory of Mind (CTM), a family of theories and views which hold that the human mind is an information processing system, and as such cognition and consciousness (sometimes not both, according to the specific variant) are a form of computation. Although it was introduced in 1961 by Putnam (Horst 2003), it was developed by Fodor throughout the following decades. The Stanford Encyclopedia defines the CTM as combining "an account for reasoning with an account of the mental states". Of these, the second one (Representational Theory of Mind), argues that intentional (i.e. that refer to something) mental states, such as beliefs, are relations between "a thinker and symbolic representations of the content of the states" ('I believe there is a book on the table' would be the functional *belief* relation between me and the mental, symbolic representation of 'there is a book on the table'). The first, instead, maintains that reasoning involves the symbolic representations *only* in their non-semantic, syntactic properties. As such, this process can be considered a formal symbol manipulation, which qualifies as computation.

The relational character of mental states was initially introduced by Fodor in 1978 (J. A. Fodor 1978), where he identified mental states as a three-way relationship between the individual, representations and propositional contents.

We believe that, seen in this context, the Computational Theory of Mind is a coherent theory for a few different viewpoints in the history of AI. At the same time, as we have seen in this section, AI research steered away from such open-ended questions in favor of technical and engineering achievemnts, based on restricted domains, background knowledge and specific rules: a far cry from the cognitive model the CTM would imply, and detached from the connectionist

possibilities it considered just a few years prior.

3.5 1987-1993: Bodies as the Key to Minds

Unfortunately, the enthusiasm for expert systems did not last long. As it had happened before, the hype behind AI research was too great for its own good, and towards the end of the 80s the market suffered a few serious blows. As always, though, a new wind was blowing, with several new paradigms. One in particular was very closely related to DCS research: embodied cognition.

3.5.1 Computer Science, AI and engineering

The shortcomings of expert systems, serious as they may be, had their main effect on the economic side, as research continued. Here, we will go over the reasons behind the economic crash and the new perspective in research.

Expert systems issues

Expert Systems and the Hype. The downfall of expert systems was foreseen by some in the research community: from a 1984 article (University, Stanford, and 94305 1984),

> Yet Minksy and Schank contend that today's systems are largely based on 20-year-old programming techniques that have merely become practical as computer power got cheaper. Truly significant advances in computer intelligence, they say, await future breakthoughs in programming.

Although their argument was a theoretical one, the practical implications had a large impact on the market: in the late 1980s, desktop computers were slowly overtaking specialized and expensive Lisp machines. In 1987, the reasons to buy them simply ended, and a large industry fell overnight.

The expert systems themselves started to show their flaws: they couldn't be updated, could not learn, and made large mistakes when given unusual innputs. Some issues with them had been shown years earlier, like the qualification problem (the inability of listing all the necessary preconditions for an action in the real world to have its intended effect). They worked in very specific scenarios, but were not as successful a recipe as they had been presented. Some of the initiatives launched were retracted, and funding dwindled (McCorduck 2004).

Robotics. As a direct consequence of the "lowering the mind into the body", robotics went back to the forefront of AI research. In 1990, Brooks published "Elephants Don't Play Chess" (Brooks 1990), in which he argued that symbols are unnecessary for cognition, because

> the world is its own best model. It is always exactly up to date. It always has every detail there is to be known. The trick is to sense it appropriately and often enough.

This is obviously against the Physical Symbol System Hypothesis, and represents a general awakening towards robotics-based approaches. In fact, symbol-sustaining researchers such as Minsky felt similarly, for what concerns focusing on lower-level processing; in 1986 Minsky writes "In general, we're least aware of what our minds do best, [...] we're more aware of simple processes that don't work well than of complex ones that work flawlessly" (Marvin Minsky 1986). The comparative difficulty of sensorimotor skills compared to reasoning is considered the Moravec's paradox:

> it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility

(Moravec 1988) This general feeling led to research into Behavior-Based Robotics (robots that exhibit complex behavior while having little internal modeling state) and Nouvelle AI, pioneered by Brooks himself, working on situated robots (robots interacting with their sensors and their environment) with intelligence close to one of an insect.

Connectionism returns! Around the mid-1980s, following the 'downfall' of expert system and symbolic approaches in general, the connectionist approach based on *neural networks* made a strong resurgence. A new algorithm for training such networks was rediscovered by at least four different groups (Russell and Norvig 2002): it was the same algorithm found in 1969 by Bryson and Ho. We presented this research as it is usually presented: as if in antithesis to symbolic approaches. As we will see in a following paragraph, though, research was beginning to shift to a cooperative view, using complementary approaches to explaining cognition. The field went on to bifurcate into a CS and engineering-focused strand (exploring possible neural architectures, determining their properties) and a neuroscience/empirical-focused strand (modeling biological neurons as accurately as possible).

Dealing with uncertain data

Probabilistic reasoning. We mentioned how brittle expert systems were found to be when they were applied in the real world: this drove researchers to a more scientific approach, aiming at reproducible experiments instead of philosophical claims, building on existing theories instead of constantly introducing new approaches, and including probability instead of Boolean logic. This has been called the victory of the *neats*, but it would not stand forever; the recent interest in deep learning has shown impressive results from an overall *scruffies*-based philosophy (new ideas are proposed, tested and evolved quickly, without always completing the mathematical background). Probabilistic reasoning was pushed forward by a 1988 piece by Judea Pearl (Pearl 1988), who introduced an efficient formalism for dealing with uncertain data as well as practical algorithms.

3.5.2 DCS: between philosophy, psychology and neuroscience

As we said, the disillusion towards expert systems led some researchers to advocating for a new approach, based on the physical world and robotics. They considered abstract thinking to be the least interesting human skill, and argued for "lowering" the mind into the body. The approach was not new: we remind the reader about the impact of cybernetics on the birth of the field of AI.

Levels of analysis

Marr and computational neuroscience. David Marr had a similar approach towards vision, about a decade earlier. With papers in 1969, 1970 and 1971 he proposed computational theories on cerebellum (David Marr 1969), neocortex (D. Marr and Brindley 1970), and hippocampus (which he called 'archicortex') (D. Marr and Brindley 1971). Afterwards, he focused on vision, together with the Italian researcher Tomaso Poggio. To them, vision was to be understood 'bottom-up', focusing on the physical level before any symbolic processing. They considered vision an information processing system, to be analyzed at three levels (D. Marr and Poggio 1976):

- *Computational level.* What the system does and why.
- *Algorithmic level.* How the system does what it does, and with which processes.
- *Physical level.* How the system is implemented.

The system may seem very simple; nonetheless, the idea of levels of analysis and its similarity to computational approaches signal the resurgence of interest in computational neuroscience that was on the horizon.

Encapsulated, innate modules

Fodor and modularism. The interest toward lower level systems wasn't new, especially in the philosophical side of research. Fodor had been advocating for a different notion to understand the mind, also very reminiscent of technical paradigms in Computer Science: modularism. As it was introduced in 1983 (Jerry A. Fodor 1983) and developed in the decades since, it considered a system (i.e. the mind) modular if it was at least partially composed by subunits, innate neural structures with distinct, evolutionary-developed functions. The definition of module changed, but the initial proposal contained 9 features that characterize such systems; of these, we mention:

1. **Information inaccessibility and encapsulation**. The direction of information flow is restricted; for example, althuogh you may be aware of perceptional issues while watching an optical illusion, the perception will not change.
2. **Speed and superficiality**. Modular systems are mostly fast (Fodor considers roughly half a second) and concern superficial concept: in Fodor's book this may be interpreted as computationally simple (few calculations)

or informationally simple (general); both may be true, and Fodor generally excludes the possibility of modules working with 'theoretical' concepts such as "turbine" or "proton".

3. **Dissociability**. A system is functionally dissociable if it can be damaged without significant impairment to other systems. This is reminiscent of studies on aphasia and other brain injuries which leave other capabilities perfectly untouched.

4. **Innateness**.

Fodor considers relatively low-level systems of the mind to be modular (like perception or language), while high-level systems are not to be understood by modularism. Following thinkers would go on to expand the idea to "Massive Modularity", arguing all elements of the mind are in fact modular.

As we saw in this section, the crisis of expert system, while it crippled the overly excited AI market, coincided with a renewed interest towards embodied and situated systems. This was a common thread between AI and DCS research. Still, expert system did not disappear, but coexist with other approaches: as an example, paradigms introduced with them are still the basis for modern knowledge representation techniques. We close this chapter by mentioning that a companion theory to situated robotics (closely related to Nouvelle AI) in the DCS was Situated Cognition, which slowly emerged at the end of the twentieth century. Because it argues for learning as and individual's increasingly effective performance across situations, with cognition inseparable from the context, it is closer to Skinner's behavior analysis than previous storage-and-retrieval-based theories.

3.6 1993-2010: Agents and Cooperation

Most of the paradigms we introduced in the last chapter were developed throughout this period as well, so we won't get into them again here. Still, this newfound period of success had two important features: the rekindled interest in the general problem of AI, and the availability of very large data sets. On the DCS side, the 90s brought fundamental new technologies, extensions of previous research and new all-encompassing frameworks.

3.6.1 Artificial Intelligence

As we mentioned, the (academic, first) success of the new paradigms pushed researchers to solving the "whole agent" problem again. In particular, the new context in which agents had to learn to operate was the Internet: AI algorithms started to act as the foundations behind, for example, search engines and recommender systems. Clearly, the process of merging previous results in separate tasks had its own share of issues, but the ideas we mentioned in the previous section allowed for a more complete picture: sensory systems (whether

that was speech recognition or vision) were known to provide imperfect information, so planning systems had to handle them accordingly, using probabilistic approaches. Examples of this are the two challenges set by DARPA for autonomous driving, respectively 135 miles along an unknown desert trail, completed in 2005 by STANLEY, and 22 miles in an urban environment, completed in 2007 by BOSS. This approach is known as the "Intelligent Agents" approach; researchers hoped that a complete agent architecture (like Newell's SOAR (Press 2012)) would give researchers the tools to build intelligent systems from the interaction of agents.

Data over models

Big Data. Still, the largest impact on the world of research was probably the new availability of very large datasets, thanks in no small part to the pervasive effects of the Internet. Researchers Banko and Brill (Banko and Brill 2001) argued that the increase in the size of the dataset (two or three-fold) would outgrow any advantage that was to be found by tweaking the algorithm. This sentiment, which is not by any means a formal proof, is echoed throughout the machine learning industry: another article by Norvig et al mentions, in the context of learning from text,"But invariably, simple models and a lot of data trump more elaborate models based on less data" (Halevy, Norvig, and Pereira 2009).

Back-propagation in deep networks

Multi-Layer Perceptrons and further. This paragraph contains technical terms: for the interested reader, Appendix A is available; reading Appendix A is strongly recommended before the next section. Multi-Layer Perceptrons were not one of the main direction of research, at this point. Nonetheless, research continued: LeCun applied backpropagation to a deep (i.e. with multiple hidden layers) network to recognize handwritten ZIP codes in 1989 (LeCun, Boser, et al. 1989). From this, a previous method to recognize 3D objects (matching a handcrafted 3D object model with 2D images) was adapted by Weng in 1992 (Weng, Ahuja, and Huang 1992) to learn how to combine the 2D images to recognize 3D objects (in cluttered scenes) without supervision: the features that were once hand-merged were converted to convolutional layers. This paper also introduced max-pooling. Following research includes multilayer boolean networks (Carvalho, Fairhurst, and Bisset 1994), slowly training six fully connected layers (G. E. Hinton et al. 1995), extending the feed-forward approach to include lateral and backwards connections (Behnke 2003), but both shallow and deep learning Artificial Neural Networks never outperformed Hidden Markov Models (HMM); note that these were using generative models of speech, pronunciation dictionaries and acoustic models. A good overview of HMMs and their application to speech recognition can be found in (Gales and Young 2007a). In 1997, Hochreiter and Schmidhuber introduced long short-term memory cell architecture (Hochreiter and Jürgen Schmidhuber 1997), still in use today. Still, as a 2007 paper reports (Gales and Young 2007b), "almost all present day large vocabulary continuous speech recognition (LVCSR) systems are based on HMMs".

Throughout the 2010s, AI research racked up a series of wins; apart from the ones already mentioned, we also note: autonomous planning and scheduling

in space exploration, by REMOTE AGENT(2000, generated plans and monitored their execution), MAPGEN (2004, the previous one's successor, planned NASA's Mars Exploration Rover), MEXAR2 (2007, mission planning fo the European Space Agency's 2008 Mars mission); game playing, by DEEP BLUE in chess and Watson in Jeopardy!; logistics planning, by DART (DARPA's logistics planner for the 1991 Persian Gulf crisis). We note that DARPA mentioned the deployment of DART more than paid back their 30-year investment in AI.

3.6.2 DCS

Throughout the end of the twentieth century, important technical advances allowed the resurgence in interest towards neuroscience to spike: the tools that were developed during this period would go on to become a staple in neuroscientific research. Theoretical research, instead, now included models of cognition that explored cooperative work and genetic influence; philosophical models went in a similar direction. In this section, we will then explore the difference in approach between Grand Unified Theories and specialized, expansive theories.

New technology. There were four technologies that would go on to be instrumental in the study of the brain: fMRI, TMS, PET and NIRS. The functional Magnetic Resonance Imaging measures brain activity by monitoring blood flow. The insight of its relevance belongs to the 1890s with Angelo Mosso, and the theory behind it is based on a discovery in 1936 of the different reaction of oxygen-rich and oxygen-depleted blood with Hb (hemoglobin), but the technical usage, based on works on rodents (Thulborn et al. 1990) (Ogawa et al. 1990), was only available from 1990. The first usage on humans belongs to 1992 (Kwong et al. 1992). The TMS, instead, can be used to both monitor and stimulate; although the first stable devices appeared around 1985, the FDA approval came in 2008 (Horvath et al. 2011). Near-infrared Spectroscopy (NIRS) uses the near-infrared region for spectroscopy, and its first clinical application was seen in 1994 (Ferrari and Quaresima 2012). The PET-CT scanner, based on techniques in use since the 70s, was the first to use a cylindrical array of sensors, and was named by Time as the medical invention of the year.

Radicalizing the Computational Theory of Mind. A couple chapters ago, we briefly went over the traditional Computational Theory of Mind. The CTM as we described it attempts to keep processes of reasoning and symbols entirely inside the mind itself. Externalist views, though, point out that, for example, unknown property of objects are external to the mind and cannot be constrained in representations existing within it (to this, Fodor responded by including in the CTM a causal account for mental content: a mental representation R only stands for a real-world object X if Rs are reliably caused by Xs). A more radical externalist thesis holds that cognition is both *embodied* and *embedded*. Embodied, in the sense that perception, action, and even reasoning use tissues and material that goes beyond the neurons in the brain, and with it they involve non-representational, non-computational bodily skills and processes. Embedded, not only as in "interacting with the environment

driven by inputs and outputs", but also in the sense that things outside the organism, whether that's books, prostethes, or the Internet, are a fundamental part of cognition itself. This view was put forward mainly by Clark (Clark 2003) and Chalmers (Clark and Chalmers 1998), and is the backbone behind the Extended Mind Thesis. As such, these externalist views can be considered an extension of 'modest' CTM (meaning that not *all* aspects of the mind have to be computational, so the 'offloaded' portion isn't) or a new framework in itself.

Deriving vs explaining intelligence

Grand Unified Theories and expansive theories. In this last paragraph, we will briefly mention three unified theories and two 'expansive' projects that started in this period. By Grand Unified Theory we mean a theory that starts with some relatively simple concept, and derives the behavior of the brain without ad-hoc measures for every portion or process; normally, GUTs are symbolic, as a model of the brain that simulates brain activity by simulating real-world neurons would not be considered a GUT.

The first GUT we present is the Free Energy Principle. First introduced by Karl Friston in 2006 (Friston, Kilner, and Harrison 2006), it views the mind (and systems in general) as minimizing the difference between its internal model of the world and the real-bounded perception. Because of its *very* complex nature, we won't get into its formal definition here; the two important features we wish to note is that it has later been acknowledged by its creator as not falsifiable, and the relevancy of the interest towards the backward pathways, from the signal-processing areas of the cortex back to the sensory ones. According to this theory, these pathways would carry predictions, and by comparing the two directions the brain would be able to calculate its error.

The second GUT we mention is the Integrated Information Theory: introduced by Tononi in 2004 (Tononi 2004), it takes the existence of consciousness as certain, and reasons about the properties that the physical substrate needs in order to implement it. Others axioms include the compositional nature of consciousness but also its irreducibility (as in, inability to be subdivided) to its components, and how conscious experience is definite, both in content and in spatio-temporal unit.

Lastly, we mention the Global Workspace Theory. GWT was proposed by Baars in 1988 (Baars 1988), but is still being developed actively. It is often described using the theater metaphor: among the people in the theater, consciousness only lights up a few actors, while screenwriters, the director and the audience sit in the dark (although they still shape the play). This is of course just a superficial view of GWT.

In contrast to GUTs, let us consider the specialized, not interpretability-oriented, expansive view: instead of reasoning about cosciousness or how to 'force' the mind into a theoretical framework, these theories (or better, projects), their aim is to simulate it in order to then understand it. Two relevant projects we mention is the Blue Brain Project, founded in 2005 at the École Polytechnique Fédérale de Lausanne, which runs a simulated brain made up of biologically realistic models of neurons. Interestingly, its initial goal, reached in 2006,

was to create a simultaed rat neocortical column, considered by some the smallest functional unit of the neocortex and of great interest for Kriston. Another similar project is the Semantic Pointer Architecture Unified Network, a cognitive architecture pioneered at the University of Waterloo. Consisting of 2.5 million simultated neurons, its capable of recognizing numbers, memorizing them and even writing them down with a robotic arm. Its subsystems are organized to resemble relevant brain regions.

Ethics of AI. As AI progressed, the debate of its ethics has gotten more heated. A full history of issues, principles and challenges is outside the scope of this document. Nonetheless, we mention a few of the ethical challenges raised by intelligent systems acting on their own. We considered Weizenbaum's position on this in 1976, which highlighted the worry of AI threatening human dignity. The weaponization of AI is, by now, a very relevant topic, but the discussion is not new, as this "Call for debate on killer robots" (*BBC NEWS — Technology — Call for debate on killer robots* 2021) from 2009 highlights. Other issues also include biased AIs deriving from biased data, bad actors using AI to influence society negatively, or the possibility of an intelligence explosion (or singularity) causing an AI takeover (Bostrom 2003).

In this section, we explored some of the new paradigms proposed by researchers for the comprehension of the mind, and put forward what we believe is the single largest impact on AI of the period: the availability of huge datasets. In the following section, we will explore where this led us, and what the state of the art is now across disciplines and problems.

3.7 2010-now: Deep Learning and New Perspectives

Entering the 2010, machine learning and non-symbolic approaches had begun to capture the market, both in research and in commercial applications. In the previous section, we stressed how the absolute most relevant, impactful and impressive advance was in fact the *availability* of large-scale datasets. Here, we will see what this caused, and what the current state-of-the-art is able to do. We will also use this opportunity to consider what the level of expertise required is, and briefly describe the two main frameworks for creating and training deep learning networks.

Before reading on, if you do not have experience with basic machine learning and neural networks, we ask you to read Appendix A. This content was moved to an appendix not because of its secondary importance, but only to avoid making this chapter too long. In what follows, Appendix content won't be re-introduced.

3.7.1 AI

Throughout the recent years, the AI field has boomed, in no small part thanks to deep learning's success. Wikipedia points to the "big bang" of deep learning as

early as 2009, when researchers started training deep learning neural networks on Nvidia GPUs. Others (Parloff 2016) point to the ImageNet victory in late 2012, or a related paper a couple months prior (Cireşan, Meier, and Juergen Schmidhuber 2012), but by now deep learning has become one of the areas of Computer Science with the highest research output. We will now consider two fields in which it has obtained significant advantages, and a recent development.

Computer Vision. Research in Computer Vision is varied, so a complete description is impossible for us. A few of the categories are:

- *Image Classification.* It entails assigning a label to an image. The standard architecture has been a Residual Neural Network, which is a convolutional neural network in which some layers are skipped, resembling a structure seen in the brain. Recent works involves making ResNets more efficient, and with less parameters for equivalent state-of-the-art performance (about 86.5%) (Tan and Le 2021) (Brock et al. 2021).

- *Image Segmentation.* It consists of partitioning an image into sets of pixels with a label assigned. Recent efforts include using the encoder of an EfficientNet and the decoder of a UNet (a 2015 architecture (Ronneberger, Fischer, and Brox 2015)), and running the CNN obtained on unstructured data (Baheti et al. 2020).

- *Object Detection.* The name is fairly self-explanatory. Recent work includes experiments with new bounding box shapes and loss functions (Zhang et al. 2020).

Natural Language Processing. The most impactful recent NLP model is GPT-3. *It is a model that is trained via the Generalized Progressive Transformer (GPT) framework. GPT is a transformation-based neural network that has the advantage of requiring fewer parameters than ResNets. GPT-3 is a model that is trained with TensorFlow. The resulting model is significantly more efficient than ResNets (around 70% of the parameters), but it is not as efficient as ResNets when making discrete predictions.* In fact, the italics section was generated by GPT-3, after giving "What is Image Classification" and the previous "Image Classification" description as prompts, and asking "What is GPT-3?". The reader may now be able to more easily understand why the paper in which it was presented contained a section warning of the model's potential dangers (Brown et al. 2020). Its full version has 175 billion parameters.

Neuro-Symbolic Reasoning. As one can imagine, merging the fields of symbolic and connectionist AI is not a new idea. For example, in a 1997 book by Alexandre and Sun (Alexandre and Sun 1997), they identify the possible strategies for neuro-symbolic processing. In this distinction, *unified* strategies attempt to attain neural and symbolic capabilities using neural networks, while *hybrid* approaches combine neural networks with symbolic systems, such as expert systems or decision trees. We want to discuss this approach further, so we will dedicate a separate section in the following chapter on representation

h

Figure 3.1: Strategies for neuro-symbolic integration (Alexandre and Sun 1997)

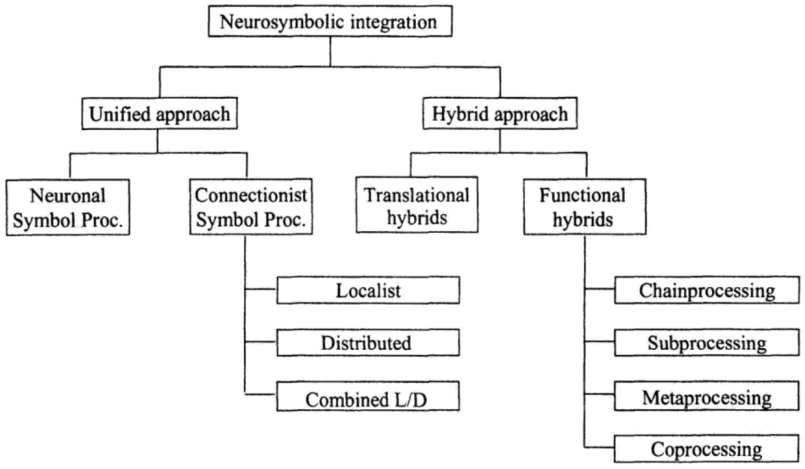


Figure 1 Classification of integrated neurosymbolic systems.

across symbols and neurons, after which we will have the means to consider some current directions. For now, suffice it to say that the amount of published papers on neuro-symbolic integration has seen a marked increase in the last few years, and many different perspectives have emerged.

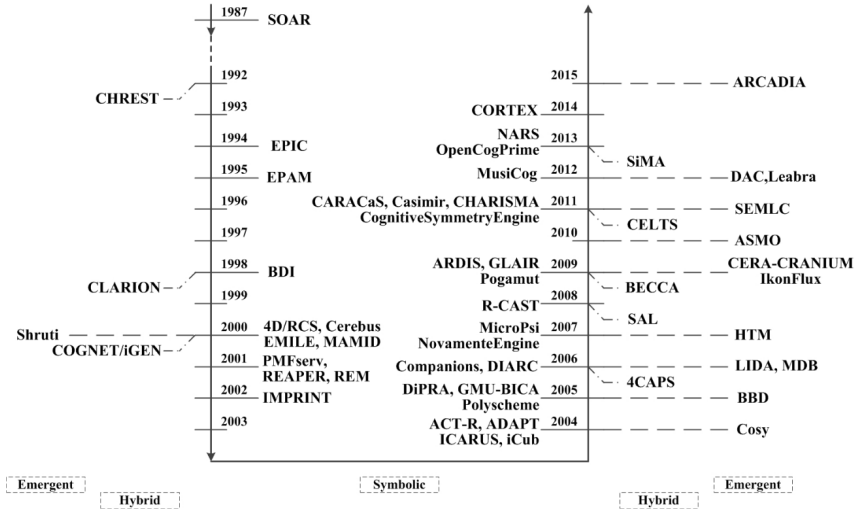
3.7.2 DCS

As we mentioned in the introduction, attempts at integrating existing theories have been recently put forward: as an example, in 2020 Safron proposed a novel model "Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework" (Safron 2020).

Thanks to technical advances and renewed interest in the field, neuroscience made significant advances. Due to its biological root, when seen from an outsider's point of view its research seems more homogeneous, and even recent advances are clearly understandable. In addition, most papers published focus on narrow, specific issues and areas. This specialization of research is shared by psychology papers, but their less constrained nature makes for a more diverse array of theories, practices and explanations. In fact, if we take consciousness as a general indicator for studies in high-level cognition, a recent survey (Michel et al. 2018) on practitioners ("249 participants completed the survey, among which 80% were in academia, and around 40% were experts in consciousness research") found that most perceived getting funding for it was more difficult than other subfields of neuroscience, and work that was done was perceived as less rigorous. Now, complete cognitive architectures are mostly proposed in an

h

Figure 3.2: Cognitive Architectures, in a temporal view (Ye, T. Wang, and F.-Y. Wang 2018)



AI context, while past proposal from DCS are further explored and completed. In that same survey, most non-experts found the IIT (described in the last section) most promising, while overall the global workspace theory was considered the most promising. Still, the sheer amount of existing CAs makes them intractable for this document: for a complete survey, see (Ye 2018) (Ye, T. Wang, and F.-Y. Wang 2018); in this paper, the CAs examined are arranged in a temporal arrangement.

Neuroscience and advances. As we mentioned in the last section, interest in neuroscience rose significantly: the 2014 Nobel prize in Physiology or Medicine was awarded to Keef, Moser and Moser neuroscientists who discovered place and grid cells (the first are neurons that fire frequently when the subject is in a specific location in the environment, while the second are neurons that fire at regular intervals as the subject navigates an open area), although the relevant papers were published earlier (O'Keefe and Burgess 2005) (E. I. Moser, Kropff, and M.-B. Moser 2008). Large projects included the Human Connectome Project (Ltd 2010), which maps entire brains, the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative (*The NIH BRAIN Initiative — Science* 2013) in the US and the Human Brain Project (*A Countdown to a Digital Simulation of Every Last Neuron in the Human Brain* 2012) in Europe. The last interesting data point we mention is that in a bibliometric study from 2006 to 2015 (Yeung, Goto, and Leung 2017) the most frequently reoccurring high impact yearly term was "autism".

Overall, the recent landscape of AI research is diverse, but not quite as

sprawling as it has been in the past; with increase in computing power, positive results shifted from expert system to connectionist architectures, and in particulare Deep Neural Networks showing the largest results. In the next chapter, we will see how, although they may be a great tool for solving problems, they are sometimes unwieldy for *understanding* problems. In this way, they have not been a panacea for understanding the brain, also due to the specific biological characteristics that are still being investigated. Nonetheless, brain research continued, with some large projects that were recently started and have yet to be completed. Psychology is, by now, more practice oriented, and neuroscience practitioners report difficulty in finding funding for consciousness research. Finally, we would like to mention three resources for academic research in AI: distill.pub, for their incredibly clear explanations and interactive articles, stateoftheart.ai for the impressive community-driven visualization of trends in research and models, and paperswithcode.com for the always up-to-date repositories and the focus on open sourcing research.

Chapter 4

Perception shifts

4.1 Visualizing trends

In this next page, we include a *qualitative* chart for visualizing trends in research as we explained them. We want to stress this is not a bibliometric paper, nor is the chart supposed to represent fixed, agreed upon values. Instead, the intended aim is to show how trends in DCS and AI research intertwined and influenced each other throughout the decades, as we did in the History chapter, from a bird's eye view.

The chart is organized as follows: when holding the sheet of paper sideways, the vertical axis represents how symbol oriented every piece of research is. The vertical axis is not a strict symbol-network distinction, but (a) organized as a spectrum, as most research papers are not completely on one side or the other, and (b) representative of the research focus more than the strict content. As an example, let's consider the paradigm of embodied cognition: its interest towards lower level processes and beliefs about their relative complexity does not directly mean they would use Neural Networks to implement decisional processes; in fact, for most robotics-oriented work of the period neural networks would not have been a good choice, due to the limited computing power. Still, the insurgence of such a paradigm shows an interest towards processes that are not limited to high-level rational thoughts or symbolic reasoning, so Brooks' research was not placed strongly in symbolic territory.

The horizontal axis, instead, represents time. Once again, for illustrative purposes, time placements are not exact, as they aren't meant to be. Colors were used to distinguish DCS research, in red, and research that was Computer Science - Mathematics - Artificial Intelligence - Engineering oriented. Lastly, we mention that, of course, only a portion of the works mentioned in the text were represented, itself a portion of published research.

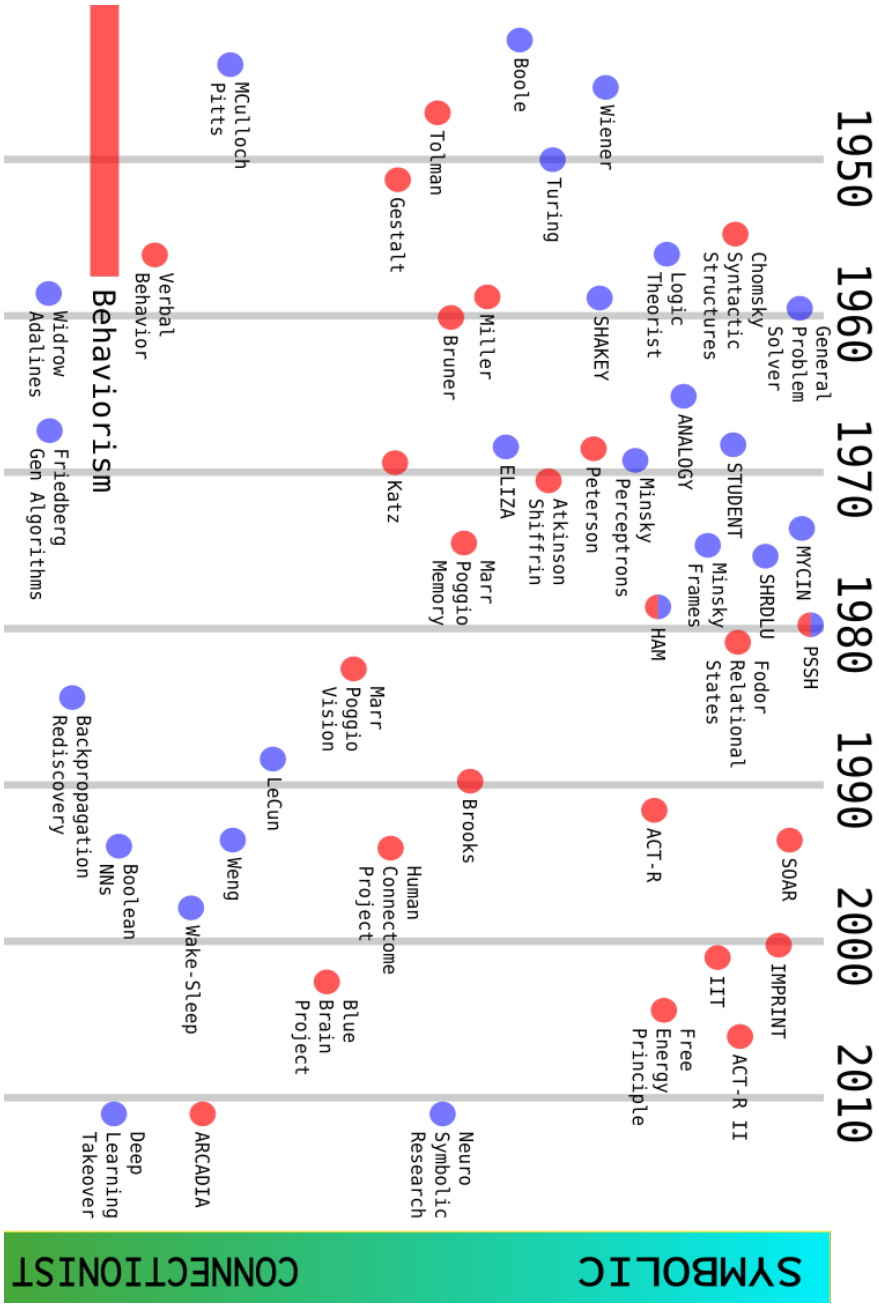


Figure 4.1: Research trends, visualized.

4.2 Symbols, Subsymbols and their Integration

In this section, we will analyze the differences between symbolic and connectionist approaches, and investigate what the possible avenues for integration are. This subject is being heavily researched, and the community's interest in it has increased massively in the last few years; still, since we will need some cross-historical notions, we decided to include it as a separate section instead of the final section of the History chapter.

4.2.1 Semantic differences

Throughout the History chapter, we did not give a strict, overarching definition of the two: at the same time, we explored researchers' opinions, from Boole's modeling of thought and the Physical System Hypothesis to the uninterpretability of Neural Networks, which gives you a clear picture of the two sides of the spectrum. For the sake of clarity, we report a section from (P. Smolensky 1987):

> [...] goals, beliefs, knowledge, and so on are all formalized as symbolic structures. [...] Thus, in a medical expert system, we expect to find structures like (IF FEVER THEN (HYPOTHESIZE INFECTION)). These symbolic structures are operated on by symbol manipulation procedures composed of primitive operations like concatenating lists, and extracting elements from lists. According to the symbolic paradigm, it is in terms of such operations that we are to understand cognitive processes. [...] The symbolic level that implements knowledge structures is alleged to be exact and complete. That means that lower levels are unnecessary for accurately describing cognition in terms of the semantically interpretable elements.

He then goes on to note that this paradigm, called by Hofstadter the 'Boolean dream', has (at least by itself) proven to give little insight into how the brain works, and tends to build brittle, rigid systems.

The largest difference between the two, he notes, is then the semantic interpretation of the formal models: while in symbolic systems symbols are used to denote the concepts themselves, semantically interpretable, in connectionist models the semantically interpreted entities are *patterns of activation*. This leaves us with a spectrum of possible representation paradigms; the two ends are fully local or localized representations (symbolic) and fully distributed representations (connectionist). The means by which these distributed representations are handled cannot be the symbol manipulation procedures, but are instead differential equations on the dynamical system implemented by the network, which uses continuous variable as opposed to discrete ones.

4.2.2 Neuro-symbolic approaches

Clearly, the brain works with networks of biological neurons propagating activation and strengthening and weakening connections. At the same time, symbolic-

driven thought *is* possible, as humans are able to conduct symbol manipulation procedures on concepts. This means that somehow, somewhere the two are integrated one way or another. Although, for length reasons, we were unable to give proper discussion to Neural Networks issues, their uninterpretability (because of distributed representations of both symbols *and* the rules governing them, 'if you open them up and peer inside, all you can see is a big pile of goo' (Mozer and Paul Smolensky 1989)) and vulnerability to adversarial attacks (i.e. manufactured examples trick networks) has been leading practitioners to the same conclusion as researchers: it would be beneficial to attempt hybrid systems, with both paradigm's advantages and none of the issues.

To classify hybrid approaches, we will follow Henry Kaytz's taxonomy, presented at AAAI 2020 (Kautz 2020): he distinguishes four types of integration:

1. **Type 1.** The first type is deep learning itself, i.e. considering symbolic manipulation as an emergent behavior, when symbols constitute input (text, questions, images) and output (text, categories).

2. **Type 2.** The second type are hybrid systems like DeepMind's AlphaGo, where a neural network is coupled with a symbolic problem solver (in this case, Monte Carlo tree search).

3. **Type 3.** The third type is a hybrid system where a NN solves one task, then interacts with a symbolic system specialised in a complementary task. An example is NS-VQA (Yi, Wu, et al. 2018), where NN tackle vision and language while reasoning is left for a symbolic system.

4. **Type 4.** This type includes those systems in which symbolic knowledge is compiled into the training set, with (Lample and Charton 2019) brought as an example.

5. **Type 5.** Type 5 networks are the tightly-coupled, distributed neural-symbolic systems, with symbolic logic rules used as templates for structures in the neural network. Examples include Tensor Product Representations (McCoy et al. 2019) and Logic Tensor Networks (Serafini and Garcez 2016).

6. **Type 6.** Finally, Type 6 systems would be able to complete symbolic reasoning inside a neural engine, and enable combinatorial reasoning. He notes the objective of such an architecture would be expert reasoning, instead of commonsense reasoning, and writes 'a step toward superintelligence, not human intelligence'.

This ends how far we're going to go with our exploration into hybrid systems. We wish to conclude by noting that, although interest is high, research into them is still in its infancy: as a counterpoint to hybrid systems' effectiveness, we bring a recent impressive result from DeepMind (Ding et al. 2020), where the authors manage to surpass neuro-symbolic state-of-the-art proposals (a) on a task designed *specifically* to focus on reasoning and expected to favour neuro-symbolic approaches (Yi, Gan*, et al. 2019), while (b) using less than 60% of

available labelled data, artificially inflating the dataset by masking part of the image and implementing self-supervised learning.

Chapter 5

Conclusion

Throughout this paper, we traced the history of Computer Science, Mathematics, Artificial Intelligence, Psychology, Neuroscience, Philosophy and a host of other disciplines united towards a deeper understanding of intelligence and the mind. All mentioned disciplines have gone through various phases, but our approach mainly focused on their relationship to symbolic and connectionist models. After the historical perspective, we used the papers we mentioned to trace a cross-historical view of research trends. In a dedicated section, we finally explored recent avenues for symbolic and connectionist integration. This document is meant to serve as both an introductory path through the history of the disciplines, as many insights are to be gained by considering past proposals in relationship to present paradigms, and as a general picture for how trends move through discoveries, critiques and research.

Appendix A

Learning and Neural Networks

A.1 Learning

Here, we will give a very brief overview of learning and its types. We will not use a historical approach: many of the algorithms and mathematics entailed are a staple of statistics, and it would be an unreasonable (and misplaced) effort to recount it here.

Machine learning is "is a field of computer science that aims to teach computers how to learn and act without being explicitly programmed (*Machine Learning* 2019)". "Artificial Intelligence - A Modern Approach" identifies four deciding factors that determine the improvements to an agent's component and how to make them:

1. Which component of the agent will be improved.
2. What prior knowledge the agent already holds.
3. What representation is used for the data.
4. What feedback is available to learn from.

The three types of learning are determined by the feedback available to learn from:

- **Unsupervised learning.** The task in unsupervised learning is to learn patterns in the input without any explicit feedback. Common types are clustering, which is grouping items into categories that share some degree of similarity, and principal components analysis, mainly used for dimensionality reduction of data.
- **Reinforcement learning.** In reinforcement learning, the agent learns from rewards or punishments depending on its performance. Still, the

decision of which action was most responsible for the feedback is up to the agent.

- **Supervised learning.** Supervised learning entails the agent observing input-output pairs and learning an approximation of the function between them, or more properly, learning the function that maps an input to an output.

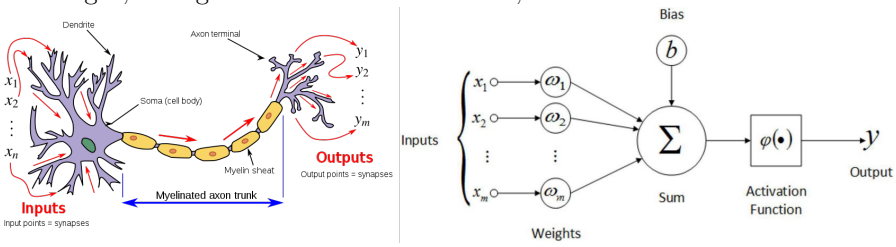
A.2 Neural Networks

In this chapter, we will highlight some of the theory behind neural networks, in the simplest and most streamlined way we can find. We will attempt to move as much math out of the way, but do expect some simple notation. We felt this was necessary to understand the state-of-the-art and compare it to symbolic approaches.

A.2.1 Basic structure and perceptrons

The structure of an artificial neuron, the unit of simple artificial networks, *resembles* that of a biological neuron. There are two main differences: the number of outputs (the artificial neuron has only one, which can at most be replicated), and the function computed by the "body" of the neuron. From this, the sim-

Figure A.1: On the left, an illustration of a biological neuron, from wikipedia. On the right, a diagram of an artificial neuron, from Towards Data Science

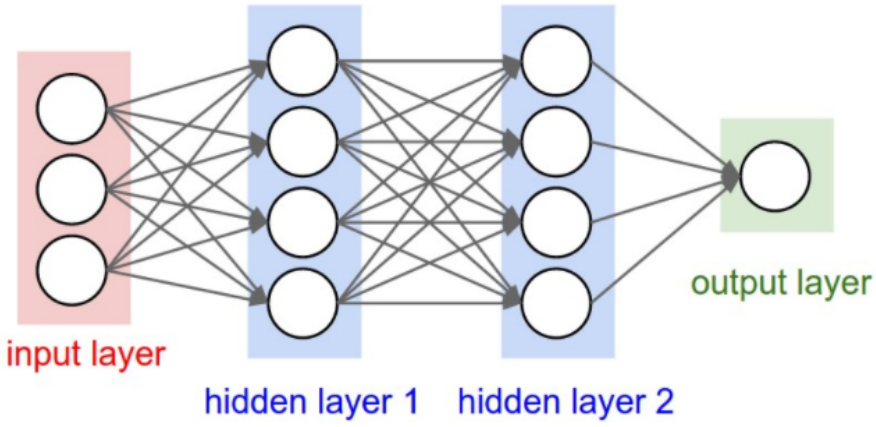


plest artificial network are just direct acyclic graphs, from an n-dimensional input to an m-dimensional output. In an artificial network, the neuron usually computes a weighted sum of the inputs (the weights are what the network learns in training), 'squishes' the result with a function (most often the sigmoid function, which constrains the output between 1 and 0), and shoots the result in the output. In equations, the result y of the computation of a single neuron is calculated like this:

$$y = \phi(\sum_{i=0}^{i=n} x_i * \omega_i + b) \tag{A.1}$$

Where ϕ is the 'squishification' function (it can also be used to only consider the neuron active if its inputs are above a certain threshold!), ω_i are the weights,

Figure A.2: Image from CS231n



x_i are the inputs and b is the bias.

This already gives us enough information to understand simple multilayer perceptrons, and maybe even to imagine why the computational effort gets too great with hundreds of neurons interconnected. This is one of the two main reasons behind NN's 'late bloom'; the other one is the available data.

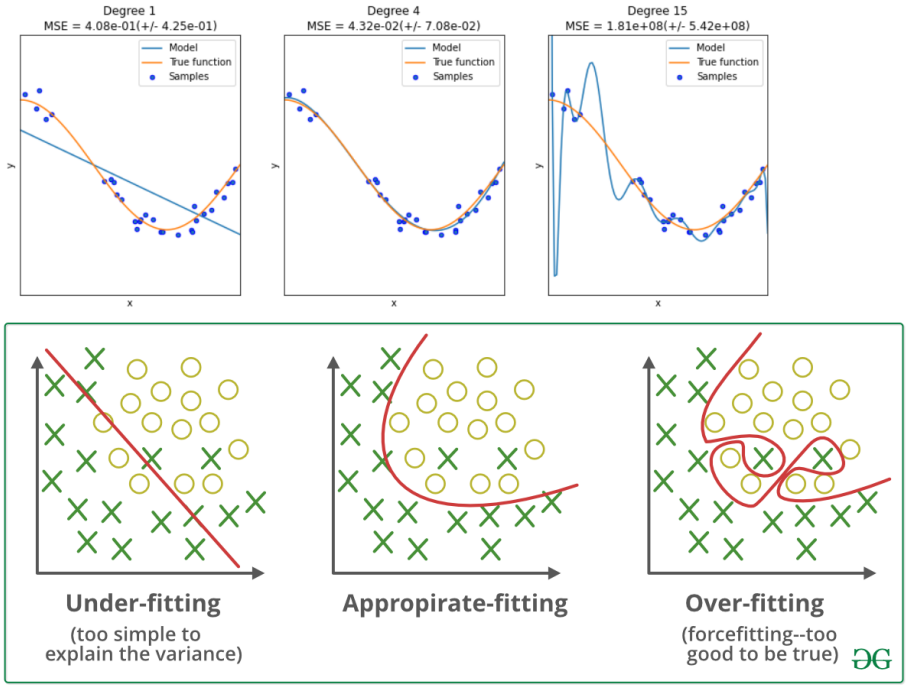
A.2.2 Underfitting and Overfitting

Let us now consider two of the main issues all Neural Networks can have, and a couple ways to mitigate them: if a NN is not expressive enough, it will *underfit* data. This means the function it learns to represent will be a far approximation of what it would take to recognize if a new input matches the training data, and is normally easily fixed by either increasing training time or deepening the network. The second, opposite issue common to NN is overfitting. Overfitting occurs when the network adapts too much to the training data, becoming too specific to it and refusing any input that does not match the training data exactly. This is a trickier problem to solve: simple solutions involve drop-off layers (layers in which sometimes some connections are "severed", so the network can't rely too much on any single neuron) or regularization (constraining the degree of freedom the network has).

A.2.3 Additional concepts

There are three more concepts relevant to us, that will help us understand network architectures: pooling, convolutions and recurrence.

Figure A.3: Underfitting, fitting, and overfitting, from GeeksForGeeks and DataScience Foundation



Pooling

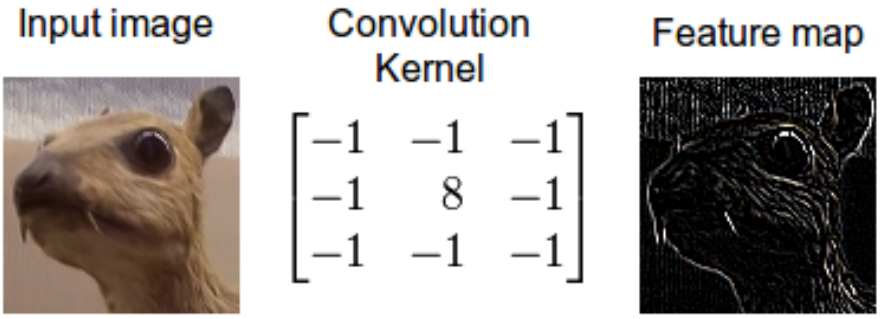
Pooling is a simple operation that reduces the dimension of the input: for every window of size $n \times n$, either the average value or the maximum value is taken. This has two effects: it regularizes the network, as it can't learn by using any specific input, and it reduces the number of free variables the network has to learn, as the final matrix is smaller than the input.

Convolutions

Assuming the reader is familiar with matrix multiplication, a convolution is a function that takes as input a kernel of size $n \times n$ and a matrix (in the case of NN, of neural activations), and only consists of computing the dot product between every 'window' of size $n \times n$ with the kernel.

As it turns out, this simple operation is quite powerful: it allows to, for example on images, extract relevant features from an activation matrix, or sharpen or blur an image. When used in NN, kernels are not hard-coded to perform such tasks: they are learned by the network as it decides which features are relevant to solving its task. Convolutional layers are used in networks because of three important features:

Figure A.4: An example of convolving an image, from Wikipedia



1. They take advantage of locality: because of their nature, they're able to easily express relationship between items that are close in the input, while perceptron treat all inputs as equidistant to each other.

2. They reduce complexity: in multi-layer perceptrons, all inputs are connected to each other. In convolutional layer, they aren't. In addition, the filter is the same for the whole image: this drastically reduces the number of free parameters the network needs to learn.

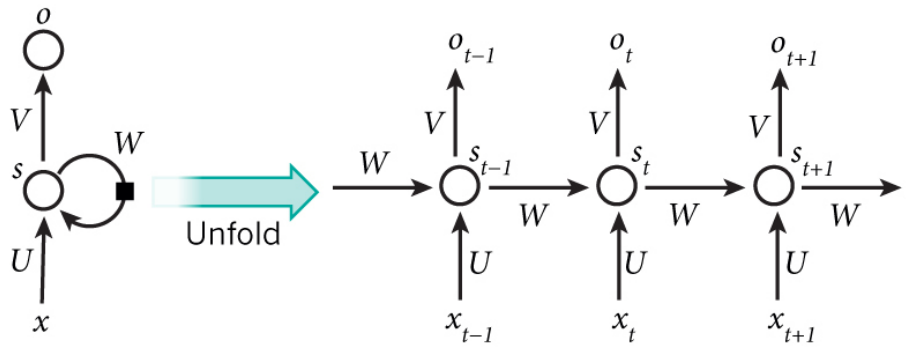
3. They can be used in conjuction with pooling layers: this grants them a degree of translational invariance, as the same result (or an average of the results) accounts for a region of the input, instead of a single point.

As can be easily gleamed, convolutional networks are primarily used in visual recognition tasks: their characteristics and features are both apt to the task and were specifically developed for it; in fact, the original inspiration for them comes from the neuronal architecture of brain regions dedicated to visual processing.

Recurrence

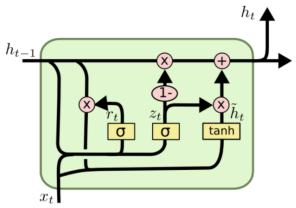
Now, for recurrent networks. Recurrent networks were designed to tackle time-sequence problems, or tasks that require temporal context, often with long sequential data. As such, they process one element at a time (every time-step), and then pass the result of their computation to the next time step to facilitate the progress. This type of recurrent neuron is only capable of limited time-context sizes: consider the phrase "John went to pick up his truck and a bag of chips [...]. Later, *he decided to...*"; as the size of [...] grows, the network's ability to predict the italics part dwindles to none, as it has to not only encode the information about the subject, but transmit it through multiple timesteps before using it. This is absolutely possible as hand-picked parameters, but networks have a hard time learning them.

Figure A.5: A recurrent unit, with its unrolled version. The three s_{t-1}, s_t, s_{t+1} are the same unit in three different time steps (LeCun, Bengio, and G. Hinton 2015).



To tackle this, more complicated recurrent architectures were created. In the text, we mentioned LSTM (Long Short-Term Memory) cells. A much more complete introduction to them can be found here, a blog post so well-written it has become the unofficial standard introduction to LSTMs. What is relevant to us is that they do not have the long-term dependency issue we mentioned before, thanks to a gated *cell state* and some simple forget and remember operation. Innumerent variations exist: as we mentioned in the text, this period of Deep Learning research does not stem from a focused and methodic mathematical search for optimized structures, but from successive approximations and intuitions.

Figure A.6: An LSTM cell, from Christopher Olah's blog



A.3 Additional Architectures and Features

In this section, we will briefly go over some of the architectures in which networks are organized.

Encoder-Decoder. These models have one general characteristics: they are divided into two functional parts; the encoder forces the network to model the input into a (normally) lower-dimensional vector, which trains it to avoid noise; the decoder then translates this vector into a useful output. The specific characteristics of encoders and decoders don't really matter: with recurrent cell-based encoder and decoder, the network can approach variable-length tasks; with convolutional layers, the network becomes simple and quick to train (and can be further slimmed down by additional improvements like quantization);

removing the decoder leaves the possible problem space mapped into a fixed-dimension vector, that can then be exported as a map in, for example, natural language processing tasks.

Attention. For a detailed explanation of some attention-based architectures, see distill.pub. A simple explanation is that attention is a mechanism to let neural networks interact with other mediums, while keeping the interaction differentiable (hence learnable), whether the medium is a fixed memory (Neural Turing Machines), another RNN (Attentional Interfaces), or itself, by choosing how many times to 'think' about a given input (Adaptive Computation Time). An example of an attention-based architecture is a Transformer.

Generative Adversarial Networks. GANs utilize two different networks: a generative network, which generates possible candidates, and a discriminative network, which evaluates them. This causes both of them to learn at the same time, and training is stopped when the discriminative gives the wrong answer about half the time. They were born as a way to train networks in unsupervised tasks, but they are now used in unsupervised, semi-supervised and supervised tasks.

Bibliography

A Countdown to a Digital Simulation of Every Last Neuron in the Human Brain (June 2012). en. DOI: 10.1038/scientificamerican0612-50. URL: https://www.scientificamerican.com/article/human-brain-project-digital-simulation-neuron/ (visited on 06/29/2021).

Alexandre, Frederic and Ron Sun (1997). *Connectionist-Symbolic Integration: From Unified to Hybrid Approaches*. Lawrence Erlbaum Associates.

Anderson, John R. and Gordon H. Bower (1973). "Human Associative Memory". English. In: (visited on 06/21/2021).

Anokhin, P. K. (1935). "Problems of centre and periphery in the physiology of nervous activity". In: *Gorki, Gozizdat*.

Ayllon, Teodoro and Jack Michael (Oct. 1959). "The psychiatric nurse as a behavioral engineer". In: *Journal of the Experimental Analysis of Behavior* 2.4. tex.pmcid: PMC1403907, pp. 323–334. ISSN: 0022-5002. DOI: 10.1901/jeab.1959.2-323. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1403907/ (visited on 06/20/2021).

Baars, Bernard J. (1988). *A cognitive theory of consciousness*. eng. Open Library ID: OL2391521M. Cambridge [England], New York: Cambridge University Press. ISBN: 978-0-521-30133-6.

Baheti, Bhakti et al. (June 2020). "Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment". en. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, pp. 1473–1481. ISBN: 978-1-72819-360-1. DOI: 10.1109/CVPRW50498.2020.00187. URL: https://ieeexplore.ieee.org/document/9150621/ (visited on 06/29/2021).

Banko, Michele and Eric Brill (2001). "Scaling to very very large corpora for natural language disambiguation". In: *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pp. 26–33.

BBC NEWS — Technology — Call for debate on killer robots (2021). URL: https://web.archive.org/web/20090807005005/http://news.bbc.co.uk/2/hi/technology/8182003.stm (visited on 07/02/2021).

Behnke, Sven (2003). "Neural Abstraction Pyramid Architecture". en. In: *Hierarchical Neural Networks for Image Interpretation*. Ed. by Sven Behnke. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 65–94. ISBN: 978-3-540-45169-3. DOI: 10.1007/978-3-540-45169-3_4. URL: https://doi.org/10.1007/978-3-540-45169-3_4 (visited on 06/27/2021).

Boden, Margaret A. (June 2008). *Mind as Machine: A History of Cognitive Science*. English. Clarendon Press. ISBN: 978-0-19-954316-8.

Bostrom, Nick (Jan. 2003). "Ethical Issues in Advanced Artificial Intelligence". In.

Brock, Andrew et al. (Feb. 2021). "High-Performance Large-Scale Image Recognition Without Normalization". In: *arXiv:2102.06171 [cs, stat]*. URL: http://arxiv.org/abs/2102.06171 (visited on 06/29/2021).

Brooks, Rodney A (1990). "Elephants Don't Play Chess". English. In: p. 12.

Brown, Tom B. et al. (May 2020). "Language Models are Few-Shot Learners". en. In: URL: https://arxiv.org/abs/2005.14165v4 (visited on 06/29/2021).

Bryson, Arthur E. and Yu-Chi Ho (1969). *Applied optimal control: optimization, estimation, and control*. A Blaisdell book in the pure and applied sciences. Waltham, Mass: Blaisdell Pub. Co.

Carvalho, A. de, M. C. Fairhurst, and D. L. Bisset (Aug. 1994). "An integrated Boolean neural network for pattern classification". en. In: *Pattern Recognition Letters* 15.8, pp. 807–813. ISSN: 0167-8655. DOI: 10.1016/0167-8655(94)90009-4. URL: https://www.sciencedirect.com/science/article/pii/0167865594900094 (visited on 06/27/2021).

Chomsky, Noam (2013). *A Review of BF Skinner's Verbal Behavior*. Harvard University Press.

Cireşan, Dan, Ueli Meier, and Juergen Schmidhuber (Feb. 2012). "Multi-column Deep Neural Networks for Image Classification". In: *arXiv:1202.2745 [cs]*. arXiv: 1202.2745. URL: http://arxiv.org/abs/1202.2745 (visited on 06/28/2021).

Clark, Andy (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

Clark, Andy and David Chalmers (Jan. 1998). "The Extended Mind". In: *Analysis* 58.1, pp. 7–19. ISSN: 0003-2638. DOI: 10.1093/analys/58.1.7. URL: https://doi.org/10.1093/analys/58.1.7 (visited on 06/25/2021).

Colby, Kenneth Mark, James B. Watt, and John P. Gilbert (Feb. 1966). "A Computer Method of Psychotherapy: PRELIMINARY COMMUNICATION". en-US. In: *The Journal of Nervous and Mental Disease* 142.2, pp. 148–152. ISSN: 0022-3018. URL: https://journals.lww.com/jonmd/Citation/1966/02000/A_COMPUTER_METHOD_OF_PSYCHOTHERAPY__PRELIMINARY.5.aspx (visited on 06/21/2021).

Computer Resurrection Issue 20 (Jan. 2012). URL: https://web.archive.org/web/20120109142655/http://www.cs.man.ac.uk/CCS/res/res20.htm#d#d (visited on 06/17/2021).

Ding, David et al. (Dec. 2020). "Object-based attention for spatio-temporal reasoning: Outperforming neuro-symbolic models with flexible distributed architectures". In: *arXiv:2012.08508 [cs]*. arXiv: 2012.08508 version: 1. URL: http://arxiv.org/abs/2012.08508 (visited on 07/01/2021).

Ferrari, Marco and Valentina Quaresima (Nov. 2012). "A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application". en. In: *NeuroImage* 63.2, pp. 921–935. ISSN: 1053-

8119. DOI: 10.1016/j.neuroimage.2012.03.049. URL: https://www.sciencedirect.com/science/article/pii/S1053811912003308 (visited on 06/25/2021).

Fodor, J. A. (Oct. 1978). "Propositional Attitudes". In: *The Monist* 61.4, pp. 501–524. ISSN: 0026-9662. DOI: 10.5840/monist197861444. URL: https://doi.org/10.5840/monist197861444 (visited on 06/22/2021).

Fodor, Jerry A. (1983). *The modularity of mind*. MIT press.

Friedberg, R. M. (Jan. 1958). "A Learning Machine: Part I". In: *IBM Journal of Research and Development* 2.1, pp. 2–13. ISSN: 0018-8646. DOI: 10.1147/rd.21.0002.

Friston, Karl, James Kilner, and Lee Harrison (July 2006). "A free energy principle for the brain". en. In: *Journal of Physiology-Paris*. Theoretical and Computational Neuroscience: Understanding Brain Functions 100.1, pp. 70–87. ISSN: 0928-4257. DOI: 10.1016/j.jphysparis.2006.10.001. URL: https://www.sciencedirect.com/science/article/pii/S092842570600060X (visited on 06/25/2021).

Gales, Mark and Steve Young (2007a). "The Application of Hidden Markov Models in Speech Recognition". en. In: *Foundations and Trends® in Signal Processing* 1.3, pp. 195–304. ISSN: 1932-8346, 1932-8354. DOI: 10.1561/2000000004. URL: http://www.nowpublishers.com/article/Details/SIG-004 (visited on 07/02/2021).

— (2007b). "The Application of Hidden Markov Models in Speech Recognition". en. In: *Foundations and Trends® in Signal Processing* 1.3, pp. 195–304. ISSN: 1932-8346, 1932-8354. DOI: 10.1561/2000000004. URL: http://www.nowpublishers.com/article/Details/SIG-004 (visited on 06/27/2021).

Gardner, Howard E. (June 1987). *The Mind's New Science: A History of the Cognitive Revolution*.

Gordon H. Bower. - PsycNET (2021). URL: https://content.apa.org/record/2007-00058-003 (visited on 06/21/2021).

Halevy, Alon, Peter Norvig, and Fernando Pereira (Mar. 2009). "The Unreasonable Effectiveness of Data". en. In: *IEEE Intelligent Systems* 24.2, pp. 8–12. ISSN: 1541-1672. DOI: 10.1109/MIS.2009.36. URL: http://ieeexplore.ieee.org/document/4804817/ (visited on 06/24/2021).

Hinton, G. E. et al. (May 1995). "The "wake-sleep" algorithm for unsupervised neural networks". en. In: *Science* 268.5214. Publisher: American Association for the Advancement of Science Section: Reports, pp. 1158–1161. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.7761831. URL: https://science.sciencemag.org/content/268/5214/1158 (visited on 06/27/2021).

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735 (visited on 06/27/2021).

Horst, Steven (July 2003). "The Computational Theory of Mind". In: URL: https://stanford.library.sydney.edu.au/archives/fall2015/entries/computational-mind/ (visited on 06/22/2021).

Horvath, Jared C. et al. (Mar. 2011). "Transcranial magnetic stimulation: a historical evaluation and future prognosis of therapeutically relevant ethical concerns". en. In: *Journal of Medical Ethics* 37.3. Publisher: Institute of Medical Ethics Section: Clinical ethics, pp. 137–143. ISSN: 0306-6800, 1473-4257. DOI: 10.1136/jme.2010.039966. URL: https://jme.bmj.com/content/37/3/137 (visited on 06/25/2021).

Hull, C. L. (1931). "Goal attraction and directing ideas conceived as habit phenomena". In: *Psychological Review* 38.6, pp. 487–506. ISSN: 1939-1471 (Electronic), 0033-295X (Print). DOI: 10.1037/h0071442.

International Encyclopedia of Social & Behavioral Sciences - 1st Edition (2021). URL: https://www.elsevier.com/books/international-encyclopedia-of-social-and-behavioral-sciences/smelser/978-0-08-043076-8 (visited on 06/14/2021).

James, William et al. (1890). *The principles of psychology*. Vol. 1. 2. Macmillan London.

KATZ, B. (1969). "The release of neural transmitter substances". In: *Liverpool University Press*, pp. 5–39. URL: https://ci.nii.ac.jp/naid/10009658302/ (visited on 06/20/2021).

Katz, B. and R. Miledi (Aug. 1967). "Ionic Requirements of Synaptic Transmitter Release". English. In: *Nature* 215.5101. tex.copyright: 1967 Nature Publishing Group, pp. 651–651. ISSN: 1476-4687. DOI: 10.1038/215651a0. URL: https://www.nature.com/articles/215651a0 (visited on 06/20/2021).

Kautz, Henry (Feb. 2020). *AAAI2020 Talk Slides*. URL: https://www.cs.rochester.edu/u/kautz/talks/Kautz%20Engelmore%20Lecture%20Directors%20Cut.pdf (visited on 07/01/2021).

Kwong, Kenneth K. et al. (1992). "Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation." In: *Proceedings of the National Academy of Sciences* 89.12. Publisher: National Acad Sciences, pp. 5675–5679.

Lample, Guillaume and François Charton (Dec. 2019). "Deep Learning for Symbolic Mathematics". In: *arXiv:1912.01412 [cs]*. arXiv: 1912.01412. URL: http://arxiv.org/abs/1912.01412 (visited on 07/01/2021).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (May 2015). "Deep learning". en. In: *Nature* 521.7553, pp. 436–444. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14539. URL: http://www.nature.com/articles/nature14539 (visited on 06/27/2021).

LeCun, Yann, Bernhard Boser, et al. (1989). "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4. Publisher: MIT Press, pp. 541–551.

LeCun, Yann, D. Touresky, et al. (1988). "A theoretical framework for backpropagation". In: *Proceedings of the 1988 connectionist models summer school*. Vol. 1, pp. 21–28.

Ltd, BMJ Publishing Group (Dec. 2010). "News". en. In: *Journal of Investigative Medicine* 58.8. Publisher: BMJ Publishing Group Limited Section: News, pp. 929–935. ISSN: 1081-5589, 1708-8267. DOI: 10.2310/JIM.0b013e

3182025955. URL: https://jim.bmj.com/content/58/8/929 (visited on 06/29/2021).
MacCorquodale, Kenneth (1970). "On Chomsky's review of Skinner's Verbal behavior". In: *Journal of the experimental analysis of behavior* 13.1, p. 83.
Machine Learning (May 2019). URL: https://deepai.org/machine-learning-glossary-and-terms/machine-learning (visited on 06/28/2021).
Marr, D. and Giles Skey Brindley (Nov. 1970). "A theory for cerebral neocortex". In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 176.1043, pp. 161–234. DOI: 10.1098/rspb.1970.0040. URL: https://royalsocietypublishing.org/doi/10.1098/rspb.1970.0040 (visited on 06/23/2021).
— (July 1971). "Simple memory: a theory for archicortex". In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 262.841, pp. 23–81. DOI: 10.1098/rstb.1971.0078. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.1971.0078 (visited on 06/23/2021).
Marr, D. and T. Poggio (May 1976). "From Understanding Computation to Understanding Neural Circuitry". English. In: URL: https://dspace.mit.edu/handle/1721.1/5782 (visited on 06/23/2021).
Marr, David (1969). "A theory of cerebellar cortex". English. In: *The Journal of Physiology* 202.2. tex.copyright: © 1969 The Physiological Society, pp. 437–470. ISSN: 1469-7793. DOI: 10.1113/jphysiol.1969.sp008820. URL: https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1969.sp008820 (visited on 06/23/2021).
McCarthy, John (1960). *Programs with common sense*. RLE and MIT computation center.
McCorduck, Pamela (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. AK Peters Ltd. ISBN: 978-1-56881-205-2.
McCoy, R. Thomas et al. (Mar. 2019). "RNNs Implicitly Implement Tensor Product Representations". In: *arXiv:1812.08718 [cs]*. arXiv: 1812.08718. URL: http://arxiv.org/abs/1812.08718 (visited on 07/01/2021).
McCulloch, Warren S. and Walter Pitts (Dec. 1943). "A logical calculus of the ideas immanent in nervous activity". English. In: *The bulletin of mathematical biophysics* 5.4. tex.copyright: 1943 The University of Chicago Press, pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259. URL: https://link.springer.com/article/10.1007/BF02478259 (visited on 06/16/2021).
Michel, Matthias et al. (2018). "An Informal Internet Survey on the Current State of Consciousness Science". English. In: *Frontiers in Psychology* 9. Publisher: Frontiers. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2018.02134. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02134/full (visited on 06/29/2021).
Mind (2021). English. URL: https://www.britannica.com/topic/mind (visited on 06/14/2021).
Minsky, M. (1975). "A framework for representing knowledge". In.
Minsky, Marvin (1986). *The Society of Mind*. URL: https://www.amazon.com/Society-Mind-Marvin-Minsky/dp/0671657135 (visited on 06/24/2021).

Minsky, Marvin and Seymour Papert (1969). "Perceptron: an introduction to computational geometry". In: *The MIT Press, Cambridge, expanded edition* 19.88, p. 2.

Moravec, Hans (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Moser, Edvard I., Emilio Kropff, and May-Britt Moser (July 2008). "Place Cells, Grid Cells, and the Brain's Spatial Representation System". In: *Annual Review of Neuroscience* 31.1. Publisher: Annual Reviews, pp. 69–89. ISSN: 0147-006X. DOI: 10.1146/annurev.neuro.31.061307.090723. URL: https://www.annualreviews.org/doi/10.1146/annurev.neuro.31.061307.090723 (visited on 06/29/2021).

Mozer, Michael C. and Paul Smolensky (Jan. 1989). "Using Relevance to Reduce Network Size Automatically". en. In: *Connection Science* 1.1, pp. 3–16. ISSN: 0954-0091, 1360-0494. DOI: 10.1080/09540098908915626. URL: https://www.tandfonline.com/doi/full/10.1080/09540098908915626 (visited on 07/01/2021).

MYCIN — artificial intelligence program (2021). en. URL: https://www.britannica.com/technology/MYCIN (visited on 06/29/2021).

Newell, Allen and Herbert A. Simon (Mar. 1976). "Computer science as empirical inquiry: symbols and search". In: *Communications of the ACM* 19.3, pp. 113–126. ISSN: 0001-0782. DOI: 10.1145/360018.360022. URL: https://doi.org/10.1145/360018.360022 (visited on 06/19/2021).

Newmeyer, Frederick J. (1986). *The politics of linguistics*. English. ISBN: 978-0-226-57720-3. URL: https://dialnet.unirioja.es/servlet/libro?codigo=605873 (visited on 06/18/2021).

Núñez, Rafael et al. (Aug. 2019). "What happened to cognitive science?" English. In: *Nature Human Behaviour* 3.8. tex.copyright: 2019 Springer Nature Limited, pp. 782–791. ISSN: 2397-3374. DOI: 10.1038/s41562-019-0626-2. URL: https://www.nature.com/articles/s41562-019-0626-2 (visited on 06/13/2021).

Nyquist, H. (Apr. 1928). "Certain Topics in Telegraph Transmission Theory". In: *Transactions of the American Institute of Electrical Engineers* 47.2, pp. 617–644. ISSN: 2330-9431. DOI: 10.1109/T-AIEE.1928.5055024.

O'Keefe, John and Neil Burgess (2005). "Dual phase and rate coding in hippocampal place cells: Theoretical significance and relationship to entorhinal grid cells". en. In: *Hippocampus* 15.7, pp. 853–866. ISSN: 1098-1063. DOI: 10.1002/hipo.20115. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hipo.20115 (visited on 06/29/2021).

Ogawa, Seiji et al. (1990). "Brain magnetic resonance imaging with contrast dependent on blood oxygenation". In: *proceedings of the National Academy of Sciences* 87.24. Publisher: National Acad Sciences, pp. 9868–9872.

Parloff, Roger (2016). *Why Deep Learning Is Suddenly Changing Your Life*. en. URL: https://fortune.com/longform/ai-artificial-intelligence-deep-machine-learning/ (visited on 06/28/2021).

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 978-1-55860-479-7.

Peterson, Lloyd and Margaret Jean Peterson (1959). "Short-term retention of individual verbal items". In: *Journal of Experimental Psychology* 58.3, pp. 193–198. ISSN: 0022-1015(Print). DOI: 10.1037/h0049234.

Press, The MIT (2012). *The Soar Cognitive Architecture — The MIT Press*. en. Publisher: The MIT Press. URL: https://mitpress.mit.edu/books/soar-cognitive-architecture (visited on 06/24/2021).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (May 2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *arXiv:1505.04597 [cs]*. arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597 (visited on 06/29/2021).

Rosenblueth, Arturo, Norbert Wiener, and Julian Bigelow (Jan. 1943). "Behavior, Purpose and Teleology". In: *Philosophy of Science* 10.1, pp. 18–24. ISSN: 0031-8248. DOI: 10.1086/286788. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/286788 (visited on 06/16/2021).

Russell, Stuart and Peter Norvig (2002). "Artificial intelligence: a modern approach". In.

Safron, Adam (2020). "An Integrated World Modeling Theory (IWMT) of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories With the Free Energy Principle and Active Inference Framework; Toward Solving the Hard Problem and Characterizing Agentic Causation". English. In: *Frontiers in Artificial Intelligence* 3. Publisher: Frontiers. ISSN: 2624-8212. DOI: 10.3389/frai.2020.00030. URL: https://www.frontiersin.org/articles/10.3389/frai.2020.00030/full#h5 (visited on 07/02/2021).

Schank, R. and C. Riesbeck (1981). "Inside Computer Understanding". English. In: *undefined*. URL: https://www.semanticscholar.org/paper/Inside-Computer-Understanding-Schank-Riesbeck/41a18aead56314c2eedd5efb5c7beef6c5212fbb (visited on 06/21/2021).

Schank, Roger C. and Robert P. Abelson (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Scripts, plans, goals and understanding: An inquiry into human knowledge structures. Oxford, England: Lawrence Erlbaum. ISBN: 978-0-470-99033-9.

Serafini, Luciano and Artur d'Avila Garcez (July 2016). "Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge". In: *arXiv:1606.04422 [cs]*. arXiv: 1606.04422. URL: http://arxiv.org/abs/1606.04422 (visited on 07/01/2021).

Sheehy, Noel and Antony J. Chapman (Sept. 1995). *Cognitive Science*.

Skinner, B. F. (1974). *About behaviorism*. About behaviorism. Oxford, England: Alfred A. Knopf. ISBN: 978-0-394-49201-8.

Skinner, Burrhus Frederic (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Smolensky, P. (1987). "Connectionist AI, symbolic AI, and the brain". en. In: *Artificial Intelligence Review* 1.2, pp. 95–109. ISSN: 0269-2821, 1573-7462.

DOI: 10.1007/BF00130011. URL: http://link.springer.com/10.1007/BF00130011 (visited on 07/01/2021).

Sperling, George (1960). "The information available in brief visual presentations". In: *Psychological Monographs: General and Applied*, pp. 1–29. URL: http://www.cogsci.uci.edu/~whipl/staff/sperling/PDFs/Sperling_PsychMonogr_1960.pdf (visited on 06/20/2021).

Steer, M. D. (Jan. 1952). "Cybernetics: Circular Causal and Feedback Mechanisms in Biological and Social Systems. Transactions of the Seventh Conference, March 23-24, 1950, New York. Heinz von Foerster, Ed. New York: Josiah Macy, Jr. Foundation, 1951. 251 pp. $3.50". English. In: *Science* 115.2978. tex.copyright: Copyright © 1952 by the American Association for the Advancement of Science, pp. 100–100. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.115.2978.100. URL: https://science.sciencemag.org/content/115/2978/100.1 (visited on 06/16/2021).

Steinberg, Danny D., Hiroshi Nagata, and David P. Aline (Oct. 2013). *Psycholinguistics: Language, Mind and World*. English. Routledge. ISBN: 978-1-317-90056-6.

Tan, Mingxing and Quoc V. Le (June 2021). "EfficientNetV2: Smaller Models and Faster Training". en. In: *arXiv:2104.00298 [cs]*. arXiv: 2104.00298. URL: http://arxiv.org/abs/2104.00298 (visited on 06/29/2021).

The NIH BRAIN Initiative — Science (2013). URL: https://science.sciencemag.org/content/340/6133/687 (visited on 06/29/2021).

Thulborn, Keith R. et al. (1990). "The role of ferritin and hemosiderin in the MR appearance of cerebral hemorrhage: a histopathologic biochemical study in rats." In: *AJR. American journal of roentgenology* 154.5. Publisher: Am Roentgen Ray Soc, pp. 1053–1059.

Tolman, Edward C. (1948). "Cognitive maps in rats and men". In: *Psychological Review* 55.4, pp. 189–208. ISSN: 1939-1471(Electronic),0033-295X(Print). DOI: 10.1037/h0061626.

Tononi, Giulio (Nov. 2004). "An information integration theory of consciousness". In: *BMC Neuroscience* 5.1, p. 42. ISSN: 1471-2202. DOI: 10.1186/1471-2202-5-42. URL: https://doi.org/10.1186/1471-2202-5-42 (visited on 06/25/2021).

Turing, A. M. (1937). "On Computable Numbers, with an Application to the Entscheidungsproblem". English. In: *Proceedings of the London Mathematical Society* s2-42.1. tex.copyright: © 1937 London Mathematical Society, pp. 230–265. ISSN: 1460-244X. DOI: 10.1112/plms/s2-42.1.230. URL: https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-42.1.230 (visited on 06/16/2021).

University, © Stanford, Stanford, and California 94305 (1984). *Why Computers Can't Outthink the Experts*. English. URL: https://exhibits.stanford.edu/feigenbaum/catalog/nr990gh3548 (visited on 06/23/2021).

Wagemans, Johan et al. (Nov. 2012). "A Century of Gestalt Psychology in Visual Perception I. Perceptual Grouping and Figure-Ground Organization". In: *Psychological bulletin* 138.6. tex.pmcid: PMC3482144, pp. 1172–1217. ISSN:

0033-2909. DOI: 10.1037/a0029333. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3482144/ (visited on 06/17/2021).

Watson, J. B. (1924). "The Unverbalized in Human Behavior". In: *Psychological Review* 31.4, pp. 273–280. ISSN: 1939-1471(Electronic),0033-295X(Print). DOI: 10.1037/h0071569.

Waugh, Nancy C. and Donald A. Norman (1965). "Primary memory". In: *Psychological Review* 72.2, pp. 89–104. ISSN: 1939-1471(Electronic),0033-295X(Print). DOI: 10.1037/h0021797.

Weizenbaum, Joseph (Jan. 1976). *Computer power and human reason: From judgment to calculation*.

Weng, John (juyang, Narendra Ahuja, and Thomas S. Huang (1992). "Cresceptron: a self-organizing neural network which grows adaptively". In: *In Proc. Int'l Joint Conference on Neural Networks*, pp. 576–581.

Widrow, Bernard and Marcian E. Hoff (1962). "Associative Storage and Retrieval of Digital Information in Networks of Adaptive "Neurons"". English. In: *Biological Prototypes and Synthetic Systems: Volume 1 Proceedings of the Second Annual Bionics Symposium sponsored by Cornell University and the General Electric Company, Advanced Electronics Center, held at Cornell University, August 30–September 1, 1961*. Ed. by Eugene E. Bernard and Morley R. Kare. Boston, MA: Springer US, pp. 160–160. ISBN: 978-1-4684-1716-6. DOI: 10.1007/978-1-4684-1716-6_25. URL: https://doi.org/10.1007/978-1-4684-1716-6_25 (visited on 06/19/2021).

Wiener, Norbert (1961). "Cybernetics : Control and Communication in the Animal and the Machine –2nd. ed". In: *212 p. Cambridge, Mass.: The MIT press, 1961. includes index*. CUMINCAD. URL: http://cumincad.scix.net/cgi-bin/works/Show&_id=caadria2010_000&sort=DEFAULT&search=series:caadria/Show?_id=4e2e&sort=DEFAULT&search=%2Fseries%3A%22CADline%22&hits=808 (visited on 06/16/2021).

Wilensky, Robert (1978). "Understanding goal-based stories." phd. USA: Yale University.

Winograd, Terry (Jan. 1971). "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language". English. In: URL: https://dspace.mit.edu/handle/1721.1/7095 (visited on 06/21/2021).

Ye, Peijun, Tao Wang, and Fei-Yue Wang (Dec. 2018). "A Survey of Cognitive Architectures in the Past 20 Years". In: *IEEE Transactions on Cybernetics* 48.12, pp. 3280–3290. ISSN: 2168-2267, 2168-2275. DOI: 10.1109/TCYB.2018.2857704. URL: https://ieeexplore.ieee.org/document/8424435/ (visited on 06/29/2021).

Yeung, Andy Wai Kan, Tazuko K. Goto, and W. Keung Leung (2017). "The Changing Landscape of Neuroscience Research, 2006–2015: A Bibliometric Study". English. In: *Frontiers in Neuroscience* 11. Publisher: Frontiers. ISSN: 1662-453X. DOI: 10.3389/fnins.2017.00120. URL: https://www.frontiersin.org/articles/10.3389/fnins.2017.00120/full#B26 (visited on 06/29/2021).

Yi, Kexin, Chuang Gan*, et al. (Sept. 2019). "CLEVRER: Collision Events for Video Representation and Reasoning". en. In: URL: https://openreview.net/forum?id=HkxYzANYDB (visited on 07/01/2021).

Yi, Kexin, Jiajun Wu, et al. (Dec. 2018). "Neural-symbolic VQA: disentangling reasoning from vision and language understanding". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., pp. 1039–1050. (Visited on 07/01/2021).

Zhang, Haoyang et al. (Aug. 2020). "VarifocalNet: An IoU-aware Dense Object Detector". en. In: URL: https://arxiv.org/abs/2008.13367v2 (visited on 06/29/2021).