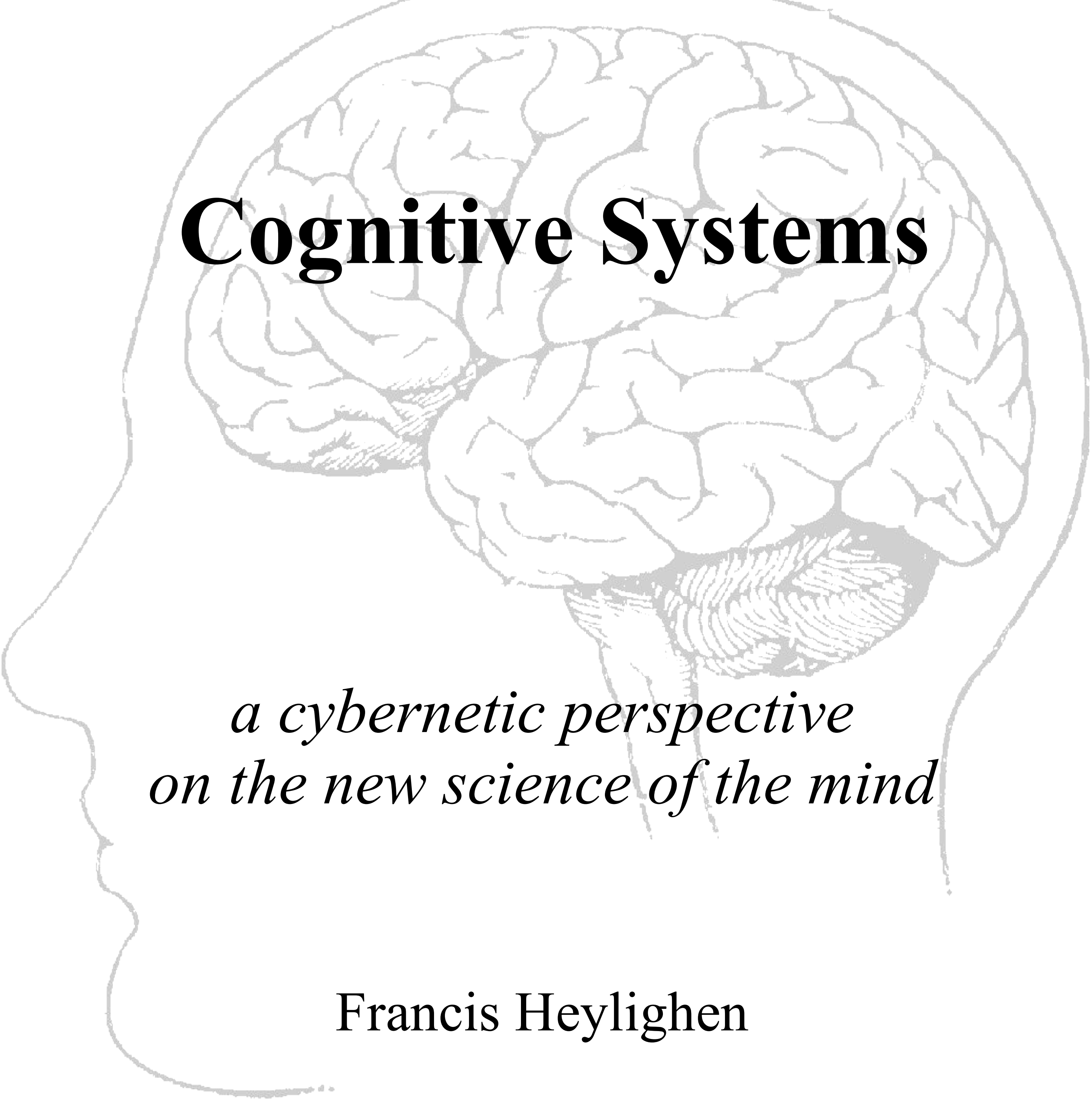


Cognitive Systems

a cybernetic perspective
on the new science of the mind

Francis Heylighen

Lecture Notes 2014-2015

ECCO: Evolution, Complexity and Cognition - Vrije Universiteit Brussel

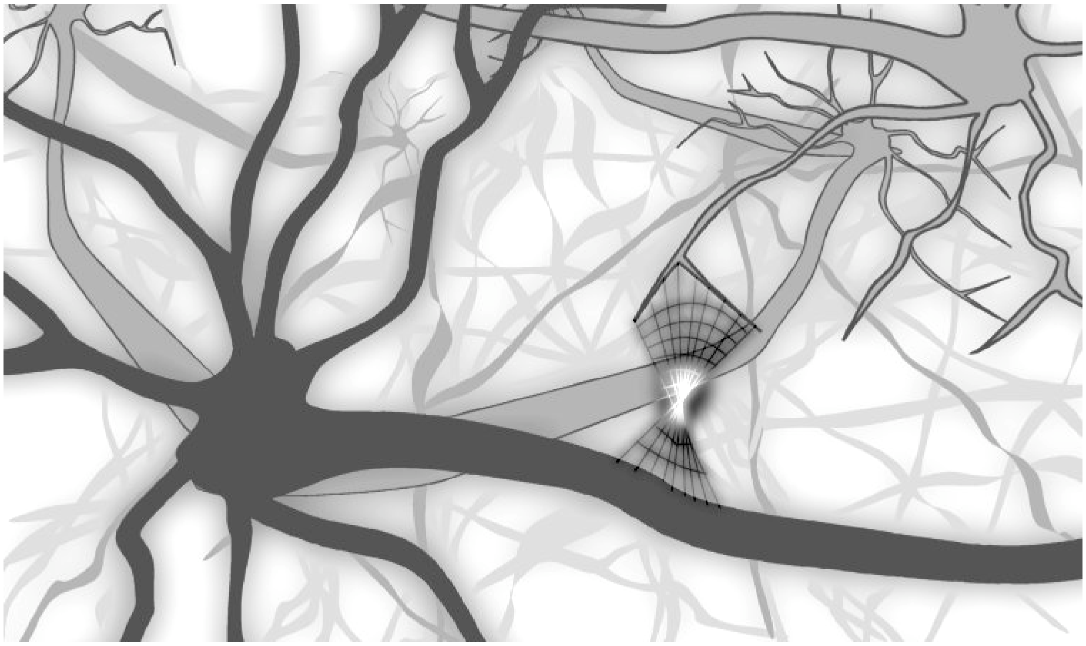


TABLE OF CONTENTS

Introduction

What is cognition?

Cognitive Science is the modern science of the mind. **Cognition** derives from the Latin verb *cognoscere,* which means “get to know”. This means that cognition focuses on knowledge, albeit not as a static substance or “thing”, but as a *process*. More generally, when we speak about cognition we are focusing on *the mind as an information processor*, i.e. a system that acquires, uses and transforms information. As such, the science of cognition typically studies issues such as the following:

Knowledge

- What is knowledge?
- How is knowledge organized or structured?
- How can we distinguish true (good) and false (bad) knowledge?

Perception and learning

- How do we acquire new knowledge?
- How do we interpret incoming information?
- What are perception, learning, and discovery?
- What is the difference between knowledge and memory?

Intelligence

- How do we use knowledge?
- How do we solve problems, make decisions, and plan actions?

However, it is important to note that cognition is not just about the kind of explicit knowledge and rational thinking that we typically find in scientific or philosophical reasoning. Cognition also includes subconscious, intuitive, and affective experiences and feelings, since these too are based on the processing of information. For example, emotion, consciousness, and behavior are all cognitive phenomena.

More generally, we can say that cognition investigates the functioning of the brain at the higher level. It is not so much interested in the details of neurophysiology or brain anatomy, although it may draw inspiration from them if they illuminate higher order mechanisms. It rather focuses on the function of the brain and its components: what, how and why does it do?

Cognitive Science (CS) as a scientific domain emerged in the 1970's, inspired by computer simulations of cognitive processes. It is a very multidisciplinary field, which includes at least the following domains:

- (cognitive) psychology
- artificial intelligence (computer simulation of cognition)
- epistemology, logic, and philosophy of science
- linguistics
- (cognitive) neuroscience
- cultural anthropology or ethnography (study of beliefs and behaviors in different groups)
- ethology (study of animal behavior)

However, the CS program soon encountered a number of conceptual and practical problems. The implementation of cognitive science theories in artificial intelligence programs was not as successful as expected. This was mainly due to a too reductionist or mechanistic view of the mind. Traditional CS sees the mind as a kind of computer program, composed of information processing modules that manipulate symbols on the basis of explicit inference rules. This mechanistic philosophy is sometimes critically referred to as “cognitivism”. These difficulties led to a countermovement in the 1980's and 1990's, which emphasized the holistic, interactive and self-organizing character of cognition. This included alternative approaches such as connectionism, constructivism, situated and embodied cognition, distributed cognition, dynamical systems, and studies of consciousness.

As yet, there is no integrated theory of cognition. The present approach seeks to find such integration by applying the conceptual framework of general systems theory and cybernetics. Therefore, I have called this approach “Cognitive Systems”, thus emphasizing the *systems philosophy* that is its foundations. The simplest way to show the need for such a holistic approach is by considering the fundamental problems caused by the traditional, analytic or reductionist view.

The naive view of cognition

The best way to explain the difficulties that cognitive science faces is by starting with the simple, intuitive view of the mind that is implicitly held by most people, including many scientists and philosophers. This perspective has fundamental conceptual problems and must therefore be replaced by something radically different. However, it is very difficult to completely get rid of it because it is so intuitive. To detach ourselves from these intuitive preconceptions, it is worth investigating them in detail, pointing out their hidden biases, and making explicit the problems that these entail.

Dualism

Descartes was the first philosopher to address the problem of mind from within the new mechanistic worldview, which would later be developed by Newton as the foundation of classical mechanics. According to mechanics all the phenomena around us can be reduced to the movement of material objects, such as particles, as determined by the laws of nature. This mechanistic view poses an intrinsic problem since it does not seem to leave any space for mental phenomena. Descartes solved this problem by proposing *two* independent realms: *mind* and *matter*. While matter follows the laws of mechanics, mind has a logic of its own that cannot be reduced to mechanical principles. This philosophy is known as **dualism**. It is essentially outdated, although a few philosophers and even brain scientists still hold on to it.

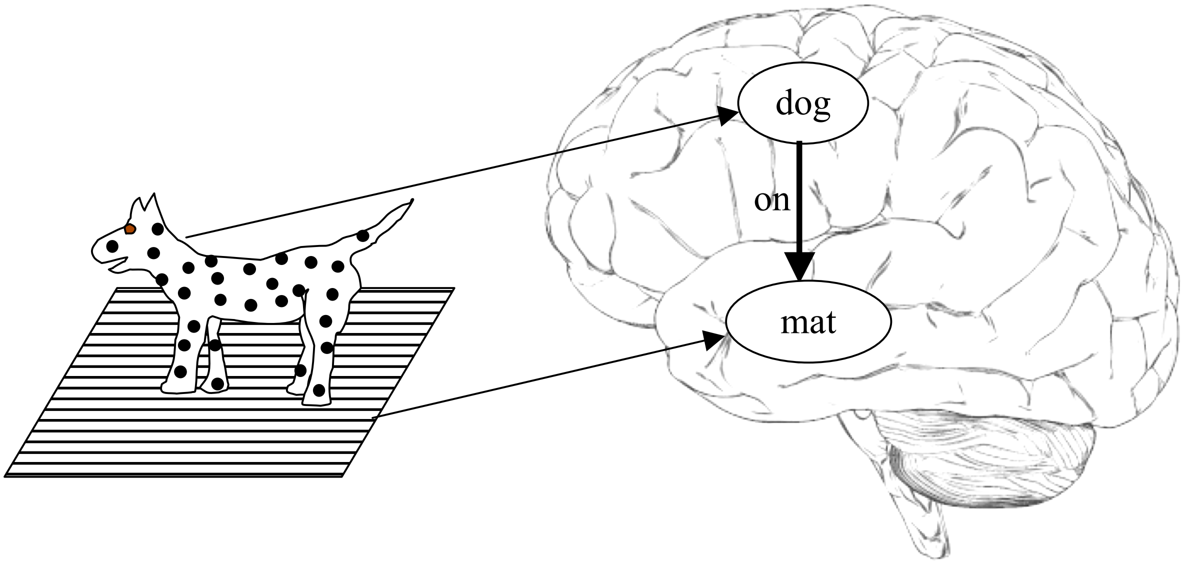
The assumptions of dualism are simple. Outside, we are surrounded by material reality. This consists of hard, indivisible particles or pieces of matter, which obey the deterministic, mechanical laws of nature. Such determinism leaves no place for free will, intention or agency: since all material events are already fully determined by the laws of nature, there is no freedom to intervene or change the course of events. The atomic structure of matter leaves no place for thoughts, feelings, consciousness, purpose, or other mental phenomena. Therefore, we need to assume that there exists another reality inside: the mind, which reflects about external reality as perceived through the senses. Descartes conceived this mind as an immaterial soul, having a free will. To explain how this mind could still affect the body, which obviously is made out of matter, he assumed that the mind communicates with the body through the pineal gland, a small organ in the brain stem.

While simple and intuitive, dualism creates a number of fundamental problems. First, adding the independent category of mind to the one of matter obviously makes things more complicated. More fundamentally, as pointed out by the 20th century philosopher Gilbert Ryle, Descartes' mind functions like a "ghost in the machine"—similar to the *Deus ex Machina* that suddenly drops from the sky to solve all problems when the plot in a novel or play has become too complicated. The body behaves like a mechanical, deterministic machine. Yet, it is inhabited by some spooky "ghost" that pulls the strings, and that performs all the tricks that are too complicated for us to understand mechanically. Indeed, we have no scientific theory of mind as a separate category, unlike our very reliable and precise theories of matter. Finally, if mind can affect matter beyond what matter would already do on its own, then it must contravene the deterministic laws of mechanics, implying that these otherwise very reliable laws cannot be trusted.

In spite of these shortcomings, Descartes' dualist philosophy remains simple and intuitively attractive. It is still (implicitly) used nowadays by scientists and lay-people, albeit most often in a "materialist" version, which we will now investigate in more detail. This more modern reformulation of dualism tries to avoid the notion of mind as a kind of non-physical, ethereal entity similar to a soul, by sticking as much as possible to material mechanisms that can be observed and analysed into their components. However, as we shall see, this approach does not succeed in overcoming the fundamental separation or duality between the (material) world and the (material) mind that observes it.

The reflection-correspondence theory of knowledge

The naïve mechanistic or materialist view of the mind is based on the idea that knowledge is merely a mirror image or reflection of outside reality. The assumption is that for every object or phenomenon in reality there is a corresponding concept or idea inside the mind. For example, a dog (external) is represented by the concept "dog" inside the mind. Concepts are typically represented by words, but could also be visualized as images, or represented using some more abstract "language of thought". The relations between objects are similarly represented by relations between concepts. E.g. when the dog stands on a carpet, the relation is represented by the relational concept "on" (see Figure). The whole of such concepts and their relationships produces a map, model, or image of reality.



This simple philosophy of knowledge produces a very straightforward notion of *truth:* true knowledge means that the network of relationships in the mind accurately corresponds to the actual relationships between objects in outside reality. Mathematically, we can say that there is an isomorphism (structure preserving mapping) between outside objects and inside concepts. This correspondence can be checked by direct observation: is there really a dog standing on the carpet? This view is sometimes called **naive realism.** It assumes that our mental contents are simply representations or reflections of the reality outside the mind, and that perception is nothing more than a process mapping external onto internal components.

This is comparable to the process of a camera taking a snapshot of a scene. The resulting photo can then be seen as a map of the environment—the way satellite photos are often used as maps—since it is isomorphic to that environment. Memory then is nothing more than the set of photographs and sound recordings made via perception that are stored in some kind of warehouse inside the brain.

In this reflection-correspondence view of cognition, thinking or reasoning is simply an exploration of the inside map in order to deduce features of the outside world. For example, by investigating the map in front of me—assuming it is accurate—I can infer that if I turn left on the next crossing, and then take the third side street on the right, I will arrive in front of the church.

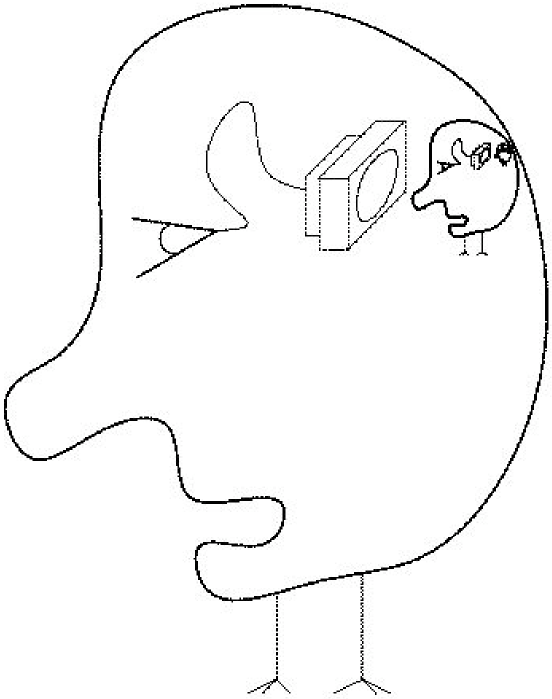
Problems of the reflection theory

Although simple and attractive, this philosophy leads to a range of fundamental problems. First, reality is much too complex to map in detail: we can only register and remember the tiniest fraction of the potentially available information. Moreover, why would we need such an accurate reflection if we have the world itself? Too detailed maps are essentially useless: just imagine a 1/1 scale map of a city, where every stone, weed or broken bottle is reproduced in full detail. On the other hand, a classic example of a simple and useful map is the London underground (subway) map, which reduces a tangle of thousands of streets, railways, and crossings to a small number of distinctively colored, straightened lines, representing the underground lines with their stations. Simplifying a map may *seem* obvious, but the problem is that there is no objective way to decide what to leave out and what to include in the map. All maps, models and representations are strongly determined by the purpose for which they are used. For example, a bus map will look completely different from an underground map, even though they cover the same terrain. Both in turn will look complete different from a geological map indicating water basins and elevation.

More fundamentally, as Kant taught us, we have no access to the “Ding-an-sich”, i.e. the objective reality outside of us, only to our very simplified and distorted perceptions of it. We cannot compare our mental contents to reality, only to our perceptions—which are themselves already part of our mental contents. Therefore, there is no absolute way that we can make sure that the reflection is accurate. This forces us to abandon accurate reflection as the ultimate criterion of truth.

Yet another problem with the reflection view of mind is that it does not explain abstract or affective ideas. For example, how can you perceive compassion, the number zero, causality, or democracy? Which concrete objects are mapped onto these abstract concepts? Even for the phenomenon that initially inspired this philosophy, imagery, it turns that out that there is no true isomorphism between the mental image and the thing it represents. For example, try to imagine a picture of the Parthenon before your mind’s eye: can you count the number of columns in the front? If you cannot, it means that there is no exact correspondence between object and mental representation.

Most fundamentally, the reflection view does not explain the active role of the mind. Indeed, it does not tell us what happens to these internal maps: who or what is using them for what reason and in what way? Trying to answer that question merely leads us into another conundrum, that of the **homunculus**.



The Homunculus problem

Cartesian materialism is an attempt to keep the mechanistic metaphysics of Descartes while getting rid of the idea of an immaterial soul. In this philosophy, the mind is seen as a (material) component of the body (e.g. the brain or some component of it) that interacts with the world via the sensory organs and muscles. The philosopher Daniel Dennett has proposed the term "Cartesian theater" to sketch the picture that results when this idea is combined with the reflection-correspondence perspective: the mind somehow sits in a theater where the incoming perceptions are projected as images onto a screen; it looks at them, interprets them, and decides what to do; it sends its decisions as commands to the muscles for execution. In a more modern metaphor, we would describe the situation as if the mind acts as a *control center* for the body, the way an air traffic controller keeps track of the incoming planes on a radar screen, analyzing the situation, and issuing directions to the pilots.

While this picture may seem more satisfying to a scientifically trained mind than Descartes' ghostly soul, it merely shifts the difficulty. The fundamental problem with the mind as control center is that it is equivalent to a **homunculus** (diminutive of the Latin "homo" = human being): a little person watching the theater inside our brain, and reasoning like an intelligent being in order to deal with the situation it observes. However, the point of the exercise was precisely to explain how a person reasons! We have explained the mind simply by postulating another, "smaller" mind (homunculus) within the mind.

Such reasoning leads to an infinite regress. Indeed, to explain how the homunculus functions we must assume that it has a mind, which itself implies another homunculus inside it, which must contain yet another homunculus, and so on. It is as if we are opening a series of Russian dolls the one nested into the other one, without ever coming to the last one. Another way to illustrate the circularity of such reasoning would be to consider a recipe for making cake where one of the ingredients is cake: how can you ever prepare such a cake if you don’t already know how to do it? To evade this paradox, we need to make a radical break with the way of thinking that produced it.

The need for a systems view

To stop such an infinite regress, we need to posit a place where it ends: the component of our brain where consciousness resides and where rational decisions are made. But unless we go back to Cartesian dualism, and postulate a mysterious, ungraspable soul, we will not find such a place where the outside world ends and the true mind begins. Indeed, trying to pinpoint the place where decisions are made, we still come to the conclusion that that place must be able to perceive what is going on outside of itself, and therefore that it must have a seat in the theater, bringing us back to the homunculus reasoning.

The attempt to situate the mind in a specific place or separate component is a remnant of **reductionism**, the philosophy that explains all phenomena by analyzing them into separate parts, and then determining the properties of the parts. We should understand the mind not as a collection of parts, but as a *whole,* which is *distributed* over many components. It is not located in any one of them, but in the network of their interconnections. Different parts of cognitive processes take place in different parts of the network, but there is no single part where everything comes together, no “seat of the soul”. We should also accept that there is no one-to-one correspondence between mental and physical components: the mind as a whole stands in a complex relationship to the world as a whole. Mental components do not behave like static, independent objects. They are part of a dynamic network of relationships: a process. Such a holistic and dynamic perspective requires a new scientific worldview, which we may call *systems thinking*.

But before introducing the philosophy of systems, it is worth reviewing the ideas of traditional cognitive science. These can be seen as elaborations of the naïve reflection-correspondence view and its implied homunculus towards increasingly sophisticated and realistic theories. However, the true move away from the underlying reductionism will only come in a second stage, becoming most visible in the 1980’s, which we will call the “new cognitive science”.

Classical Approaches to Cognition

A brief history of epistemology

Before addressing cognitive science proper, we will quickly review the philosophy of mind that preceded the more scientific approaches, and that in part inspired them. Traditional **epistemology** (philosophy of knowledge) and the related philosophy of science and philosophy of mind have two major historical roots: rationalism and empiricism.

Rationalism

Plato—who can be seen as the founder of classical, Western philosophy—saw knowledge as apprehension of ideal, eternal forms. For example, a triangle is an abstract, geometrically defined shape that can only truly exist in the mind. Any concrete, observable triangle, perhaps as drawn with a piece of chalk on a blackboard, is merely a very imperfect shadow or distorted reflection of that ideal form. For Plato, an idea is more fundamental than its material realization. This defines the ontological position of idealism. Therefore, if we truly want to understand the nature of reality, we can only do so by abstract *reasoning*, not by observing nature. This defines the epistemological position of **rationalism**.

To do that, we must start with foundational principles, such as Euclid's axioms of geometry, and by deduction derive the rest of all possible knowledge—the way Euclid showed how a variety of theorems can be deduced from the axioms. Different rationalists proposed different foundational principles on which to build their cognitive edifice. The most famous one is Descartes' *Cogito ergo sum* (I think therefore I am). While it is hard to find agreement on specific foundational principles, all rationalists by definition agree that knowledge is developed using reason or reflection. However, it is difficult to explain how reason alone can help us to discover such concrete, apparently contingent facts as that swans are white or that ice is cold.

Empiricism

The empiricists, in contrast, see knowledge as essentially the registration of perceptions. To develop knowledge, we should base ourselves on sensations, on the data we receive about the outside world through our senses. To avoid the problems of the naïve reflection view of the mind, however, the **empiricists** must also explain how more abstract knowledge is produced. They assume that this happens through the **induction** of general rules from recurrent, concrete observations. For example, the repeated observation that the sun rises morning after morning allows me to extrapolate that the sun will rise *every* morning. As such, the mind can induce causal laws out of the repeated co-occurrence of a sensed cause with a sensed effect.

However, as the philosopher David Hume noted, no number of observations, however large, can prove that the sun will *always* come up in the morning. More concretely, even after having seen thousands of swans that all were white, you may discover to your surprise that black swans do exist. Thus, while empiricism may seem a more practical and realistic philosophy than rationalism, it still has fundamental questions to answer.

The Kantian Synthesis

Immanuel Kant is considered one of the greatest philosophers of all time, in part because he was the first to propose a synthesis of rationalism and empiricism. First, he noted that we have no access to the *Ding-an-sich*, the thing-in-itself as it exists outside us, only to our necessarily imperfect perceptions of it. Therefore, there is no way to prove theories on the basis of observations alone. Some truths must be assumed a priori, i.e. on the basis of reason. For example, the axioms of logic and mathematics, such as the law of contradiction (A and not A cannot both be true), cannot be derived from observation. Others, though, such as the fact that the sun comes up every morning, must be observed a posteriori. Even then, observations alone are insufficient to induce general concepts. They must be supported by pre-existing cognitive structures which Kant called a priori intuitions or “categories” of thought. These are very abstract and general concepts, which include quantity, negation, possibility, existence, causality, time and space. They are necessary conditions for the receptivity of our mind, i.e. its ability to organize experience into an intelligible form.

Logical empiricism

In the beginning of the 20th century, the philosophers of the Vienna Circle developed a very influential epistemology known as **logical empiricism**, logical positivism or logical atomism. It too is a kind of synthesis of rationalism and empiricism, although it is more explicit and “scientific” than the one of Kant, and more leaning towards empiricism. They start from the idea that individual observations function as the atoms of knowledge, i.e. the smallest, objective units. No other, more abstract concepts should be used. This rather strong assumption is the positivists’ way to get rid of unproductive philosophical discussions, which typically center on ill-posed questions or ambiguously defined ideas, such as “What is the nature of God?” or “What are good and evil?” In this, they follow one of the former members of the Vienna Circle, Ludwig Wittgenstein, who famously remarked that: “What one cannot speak about, one must remain silent about”.

However, since knowledge is more than a collection of disconnected facts, these cognitive atoms must be assembled into a coherent theory or model of the world. This is achieved by means of the operators of formal logic (conjunction, disjunction, implication, quantifiers, etc.). The logical relations between propositions (which represent atomic observations) make it possible to make inferences about things as yet not observed. For example, the statement On_table(cat) &

Under_table(dog) allows us to infer Higher_than(cat, dog). As such, a theory can always be verified by observing whether its predictions are realized. The aim of the logical empiricists was to develop a universal language for science, so that all scientific theories could be expressed in the same explicit, unambiguous manner.

Pragmatism

Pragmatic philosophers base their epistemology on very different assumptions, which give priority neither to observation nor to reason. For them, there is no absolute criterion of truth, neither empirical nor rational. The only criterion that counts is that theories should be good at solving problems. Therefore, they do not mind using abstract, unobservable concepts, as long as they are useful.

In mathematics, this philosophy was formulated most clearly by Henri Poincaré under the label of *conventionalism*. For Poincaré, mathematical theories are chosen by convention, because they are simple and practical, not because they are “true”. For example, in geometry there exist both Euclidean theories (based on Euclid’s axiom that two parallel lines never intersect) and non-Euclidean ones (where this assumption is not valid). Both are logically consistent. As abstract mathematical theories, we cannot decide between the two by making observations. For most everyday situations, the simpler Euclidean theory is perfectly applicable, but in certain more advanced situations, like in Einstein’s theory of general relativity, we need to use a non-Euclidean one. It is up to us to decide which theory is most useful in which context.

Evolutionary epistemology

Karl Popper started his career influenced by the Vienna Circle, but became dissatisfied with its philosophy. His basic criticism was that logical empiricism fails to solve the problem of **induction**. He noted that a good theory makes an infinite number of predictions, while you can only make a finite number of verifications. For example, no matter how many times you verify the prediction that “a swan is white”, this will never prove the theory that “all swans are white”. On the other hand, the observation of a single black swan will disprove that theory.

Therefore, Popper proposed to replace the criterion of verification by the one of **falsification**: you can only prove that a theory is false, not that it is true. If at least one of its predictions is contradicted by observation, then the theory is refuted. This is a general method to eliminate bad theories. However, there is no general method to generate good theories: you can only make plausible hypotheses, or what Popper calls “conjectures”, and check whether they survive attempts at falsification. The more attempts the theory survives, the more trust you can have in it—although you will never be absolutely certain that it is true.

Popper’s falsificationism can be easily generalized to a more pragmatic theory of knowledge. First, we must note that a single contradiction with observation is in general not sufficient to reject an otherwise reliable theory. After all, observers can make errors too, if only because they use instruments (such as telescopes or microscopes) that are themselves based on theories (e.g. the laws of optics) that are not absolutely reliable. The core of Popper's philosophy, however, remains: theories are generated by trial-and-error. A variety of novel hypotheses are generated; after extensive testing, the bad ones are rejected and the best ones kept. This is remarkably similar to the mechanism of evolution described by Charles Darwin, where new types of organisms are generated by blind variation of existing designs, after which the bad ones are eliminated by natural selection.

This insight led Donald T. Campbell to found the domain of **evolutionary epistemology**. His fundamental thesis is that all knowledge (not just scientific theories) is a product of blind variation and the selective retention of those knowledge structures that most adequately represent reality, in the sense that they are as little as possible contradicted by experience. For example, the belief that putting on wings like a bird allows you to fly will quickly be eliminated (together with its carrier) once you do the test and jump from a tower. If such testing goes on long enough, the remaining knowledge will be generally adapted to the reality or environment in which its carrier lives.

Campbell and other evolutionary epistemologists applied this reasoning not only to theories and beliefs developed via reason or observation, but also to our in-born cognitive mechanisms, such as a bird's instinctive knowledge of how to fly, and our instinctive fear of heights, which tells us not to attempt flying. Such instincts, perceptual mechanisms, and other inherited "categories" of thought are the product of evolution at the biological level. Natural selection across the generations has made sure that cognitive mechanisms that made too many errors have been eliminated. Therefore, our sense experiences and categories are basically reliable. However, they are not absolutely so, since evolution is fallible, as illustrated by the fact that we still sometimes fall prey to perceptual illusions and cognitive biases.

From behaviorism to cognitive psychology

The need for scientific observation

The start of psychology as a science in the 19th century came with the observation that reflection and introspection—the methods used until then by philosophers—are not sufficient to study the mind. To start with, their results are far too subjective. This is obvious when we note how different people, including very intelligent ones such as Plato, Descartes and Hume, come to very different conclusions while using the same method of introspection. Moreover, some of their more sophisticated argumentations have shown how seemingly obvious, intuitive impressions about the mind can often be plain wrong—as we have illustrated with the homunculus fallacy. To make theories about the mind more reliable, we therefore must make mental phenomena objectively measurable.

This was the assumption that drove Wilhelm Wundt, who can be seen as the first scientific psychologist. To achieve this, Wundt designed clever experiments to quantitatively estimate different properties of the mind. He initiated one of the most popular paradigms in experimental psychology, the measurement of **reaction time**. This basically asks the question: how many milliseconds does it take to perform a particular mental operation, such as reading and understanding a particular word, or adding two numbers? Comparing the duration of different operations then allows us to confirm or falsify various hypotheses about how these operations are supposed to take place. For example, if our hypothesis states that operation A is more complex than operation B (e.g. because it encompasses an operation of type B as a subprocedure), then the observation that A requires less time to perform would contradict this theory.

Behaviorism

In its most extreme version, this reaction against introspection produced the psychological doctrine of **behaviorism**, which dominated academic psychology over the first half of the 20th century. It can be seen as a direct application of positivism to psychological research. Its basic assumption is that *only observable behavior* is worth investigating; abstract mental phenomena, such as ideas, feelings, or consciousness, have no role to play in a scientific theory.

The basic paradigm underlying behaviorist theories and observations is that of **stimulus-response**. It drives the following generic experiment:

- repeatedly subject a person or animal to an event they are bound a perceive: a *stimulus* (this could be a flash of light, a noise, or a word they have to read)
- precisely observe each time the behavioral reaction: the *response* (this could be a movement, a sound made, something said)
- try to find the *(cor)relation* between the two, i.e. when the experiment is repeated, in how far does the same stimulus tend to elicit the same response?

The stimulus-response relations (S → R) that are induced in this way are the foundations (or "logical atoms") of any behaviorist theory. In fact, for a behaviorist the mind is little more than a huge collection of S → R pairings.

The biggest success of behaviorism, the demonstration of **learning** by conditioning, used a slightly more sophisticated version of such experiment. The classic example is Pavlov's dog. The Russian psychologist Pavlov repeatedly exposed a dog to the sound of a bell (stimulus) shortly before feeding the dog. After a few experiments, the dog starts to salivate (response) as soon as it hears the bell: it is conditioned to expect food after hearing the bell; it associates the bell sound with being fed. The more often this happens, the stronger the **association** becomes. This creation of expectations or associations between perceived events is called "classical conditioning".

Another basic form of learning, termed “operant conditioning”, was investigated extensively by B. F. Skinner [1938]—mostly using animals such as rats and pigeons. The animal is confronted with a particular stimulus, such as a lever or button. Given that the animal generally does not know how to react, the response it produces—such as pushing down the lever—will initially be more or less random. However, if that response is followed by a reward (e.g. food), the behavior is **reinforced**: the next time the animal is confronted with such a lever, it will push it more quickly and eagerly. If the response is followed by a punishment (e.g. an electric shock), the behavior is suppressed or inhibited, and the next time the animal will be less likely to push down the lever. In this way, animals (and to some degree people) can be efficiently taught to exhibit almost any not too complex behaviors. For example, pigeons can be taught to peck in a particular order on a specific subset of differently colored or differently shaped buttons.

Information processing

Starting in the 1950’s, behaviorism was gradually overtaken by a more sophisticated, “cognitivist” approach that focused on more complex internal operations—which we would associate more with humans than with animals. This initial version of cognitive psychology was inspired by information theory and the first computers that had just made their appearance, and the input → processing → output paradigm on which they are based. Stimuli are now interpreted as information that enters the mind: **input**. Responses are seen as the corresponding **output** of the system “mind”. The focus, however, has shifted to the **processing**, i.e. the mental operations intervening *in between* stimulus and response.

The mind is seen as similar to a computer performing some program that transforms and interprets the information. This means that it must contain one or more processor and memory components to process and store that information. Like in a computer program, cognitive processes are typically decomposed into different stages or subroutines—such as perception, pattern recognition, storing, inference making, retrieval, and evaluation—that are performed by specialized modules. The incoming information interacts with information that already resides inside the mind, e.g. as stored in one of the memory units. This internal state depends on previous experiences. This means that there is no longer a simple S → R relationship, since the same stimulus can lead to different responses, depending on the information stored in memory. For example, the stimulus question “Do you believe me?” will get different responses depending on what the person who asked the question said previously.

Memory

One of the biggest advances brought by the information-processing approach is an elucidation of the phenomenon of memory. The computer analogy suggests an immediate distinction between two essentially different types of memory:

- Short-term or **working memory** is similar to computer RAM, in which on-going operations are performed. It contains the present state of the process, dependent on just the last few inputs and operations. In humans, it turns out that working memory has a very limited capacity: only 7± 2 items can be actively kept in mind without taking note, e.g. when trying to remember a shopping list. The cognitive psychologist George Miller, in a classic paper discussing these limitations, called this the “magical number”. Working memory has more important uses than storing shopping lists, though: you need it to keep in mind and evaluate different possible combinations of concepts while reasoning about a problem. Therefore, the limitations of working memory impose strict limitations on your capacity for conscious information processing. For example, to perform a multi-step calculation, such as $(2 \times 4) + (3 \times 3)$, in your head, you need to store the provisional results ($8 = 2 \times 4$, and $9 = 3 \times 3$) in your working memory before you can compute the final result ($17 = 8 + 9$).

- **Long-term memory** (LTM) is similar to a computer hard disk. It is used to store possibly useful data for an indefinite duration so that they may be reused later, whenever they turn out to be relevant. For example, to perform the above calculation you need to know the tables of multiplication, which include such facts as $3 \times 3 = 9$. LTM has a virtually unlimited capacity, containing at least millions of facts. Unlike a hard disk where data are sequentially stored, however, it is organized as a network of concepts connected by associations. This means that information retrieval from LTM works in a way very different from a computer, which automatically locates and reproduces the exact bits that it registered on its hard disk. There are two basic processes:

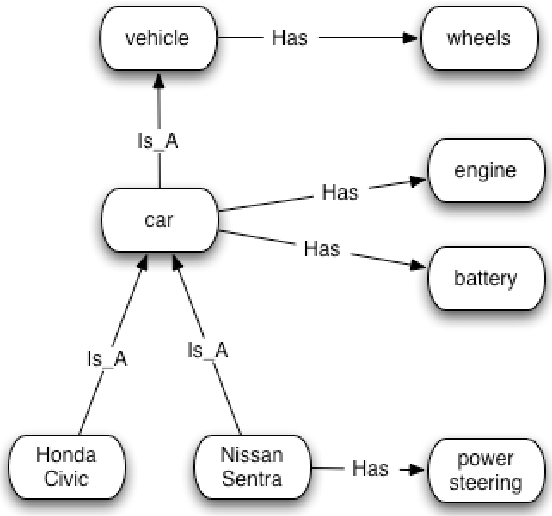
    - **recognition** is the matching of an input pattern (perception) to a pattern stored in LTM. This in general works automatically, quickly and reliably. The perception (e.g. of a person’s face) activates the associated memory (e.g. who the person is)

    - **recall**, on the other hand, is the reactivation of a pattern in LTM without matching input pattern, e.g. when trying to remember the name of your primary school teacher. This is much more difficult, requires conscious attention, and often fails, even when the pattern is safely stored. This happens e.g. when people say that the name is “on the tip of their tongue”.

Semantic Networks

A further investigation of the way knowledge is structured inside LTM led cognitive psychologists to the theory of semantic networks. A **semantic network** is a network consisting of concepts, represented as nodes, that are connected by links. Both concepts and links belong to different types or categories. Concepts may include cognitive representations of events (such as WWII, or your_wedding), of objects (such as your computer, your dog Fido, your boss John), of classes of objects and events (such as the class of all weddings or all dogs), and properties or features of objects or events (such as the property of being brown, or long-haired). The network is called "semantic" because the meaning of a node is determined by the whole of its links with other nodes.

The most important type of link connecting concepts is called Is_A , as in "Fido Is_A Dog", or "a Dog Is_A Mammal". Is_A connects instances (specific objects or events) or classes to the more encompassing classes that they belong to. The Is_A relation induces a hierarchical ordering on concepts, from more specific to more general. Other common semantic relations or link types include Has_Part (e.g. Dog Has_Part Tail) and Has_Property (e.g. Dog Has_Property Barks). But in fact, any type of relationship (e.g. Brother_of, Larger_Than, Has_Color, …) can in principle be used as a link type in a semantic network. This unlimited freedom is both the main strength (flexibility) and the main weakness of semantic network representations: cognitive scientists have not been able to agree on which link types should or should not be included in a representation of LTM. Yet, it seems unlikely that the brain would use a different type of link for any conceivable type of relationship.

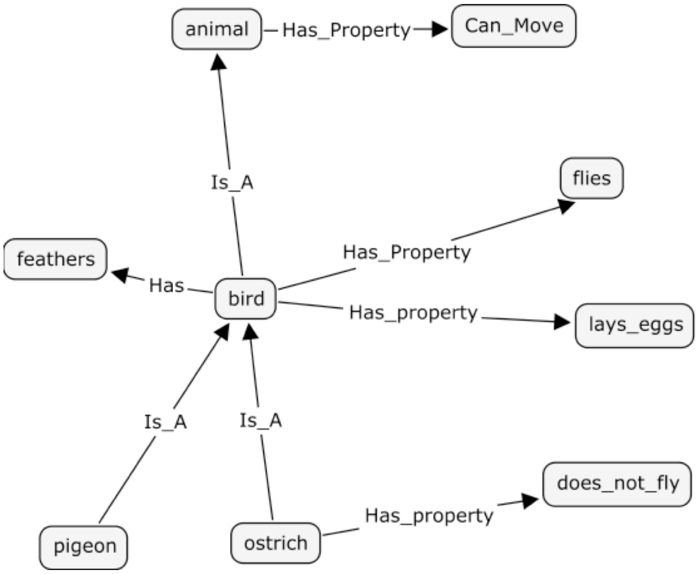


Default reasoning

A major strength of semantic network models is that it is easy and intuitive to make inferences by following sequences of links. For example, a pigeon is a bird and a bird is an animal, therefore a pigeon is an animal. Similarly, if we know that a Honda Civic Is_A Car, that a Car Is_A Vehicle, and a Vehicle Has Wheels, then we can deduce not only that a Honda Civic is a vehicle, but that it has wheels. In semantic network terminology, the Honda and Car categories have **inherited** the property of having wheels from their vehicle supercategory. This seems like a pretty straightforward visual representation of logical inference. Yet, psychological research has demonstrated an important cognitive mechanism, **default reasoning**, that does not follow the strict rules of logic.

First, it has been observed that in the mind concepts are not defined as logical categories obeying the principle of the excluded third (i.e. either something belongs to category A or it does not belong). Instead, concepts should be viewed as "families" or *clusters* of items that are more or less similar, but such that there is no strict boundary between the items that are inside and those that are outside the category. For example, while Car and Truck definitely belong to the "Vehicle" category, this is much less obvious for Wheelbarrow or Toy_Truck, which have some of the typical properties of vehicles (such as wheels), but lack others (such as the ability to transport people).

Such a cluster is centered on a **prototype**, which is a kind of generalized, “most typical” instance of the category. For example, a prototypical bird is small, lays eggs, has feathers, and flies. The properties of the prototype will normally be linked to the node representing the category.



Yet, an ostrich is large and cannot fly, but it is still a bird. In a semantic network, it is assumed that subcategories and instances inherit their properties *by default* from the higher category. “By default” means: unless there is evidence to the contrary, i.e. unless the instance is somehow an exception to the general rule. In that case, the exceptional properties are linked directly to the instance, overriding the properties that are indirectly inherited. For example, ostrich inherits “lays eggs” and “has feathers” from bird, but the direct link from ostrich to “cannot fly” overrides the inherited property “can fly”. Inheritance by default makes the memory network much more efficient, since links or properties only need to be stored at the level where they are most likely to be needed. Thus, if we learn about a new type of bird, we do not need to explicitly memorize that it lays eggs, has feathers, can fly, etc. since it inherits these properties by default from the bird category. We only need to learn its truly distinguishing features, e.g. that it lives in the rain forest and has a long purple tail.

Problem solving

The essence of intelligence

So far, the topics we have discussed have little to do with **intelligence** as it is conventionally understood. This is no longer the case with problem solving: intuitively, the more complex the problems you can solve, the more intelligent you are supposed to be. Problem solving as a domain was first investigated scientifically by Newell and Simon in the 1950s and 1960s. To do this, they developed the technique of *protocol analysis*: after proposing a problem to a person, they let him or her solve it while “thinking aloud”, i.e. putting into words every idea that comes to mind while reflecting on the problem. The experimenters write down all these reasoning steps, and afterwards analyze this “protocol” to see which steps and assumptions the subjects have made.

A classic problem analyzed in this way is the following puzzle. In the sum below, numbers have been replaced by letters. Given that D = 5, find the corresponding numbers for all the other letters so that the sum is correct:

```
   DONALD
 + GERALD
 = ROBERT
```

A protocol may then start with "5 + 5 = 10, therefore T must be equal to 0. L + L is an even number but since you need to add the 1 of 10 to their sum, R must be uneven…"

Other examples of the kind of problems considered by Newell and Simon are proving a mathematical theorem (for which they built a computer program that could prove theorems in formal logic) and winning a game of chess. But problems do not need to be of the intellectual or puzzle type: they can also be very practical—like getting your car to start on a cold morning. Inspired by their psychological observations and computer simulations, Newell and Simon developed a general theory of problem solving that is in principle applicable to any problem, leading them to design what they called the General Problem Solver. This theory was later used and elaborated by many others, and is centered on the concept of **search**.

Any problem is characterized by:

- an **initial state**, i.e. the situation you start from but that is unsatisfactory (such as a car that does not start, or an unsolved puzzle);
- a **goal state**, i.e. a conceivable situation that would satisfy your criteria for a problem solution (such as a car that drives, or a puzzle where all the pieces have fallen into place).

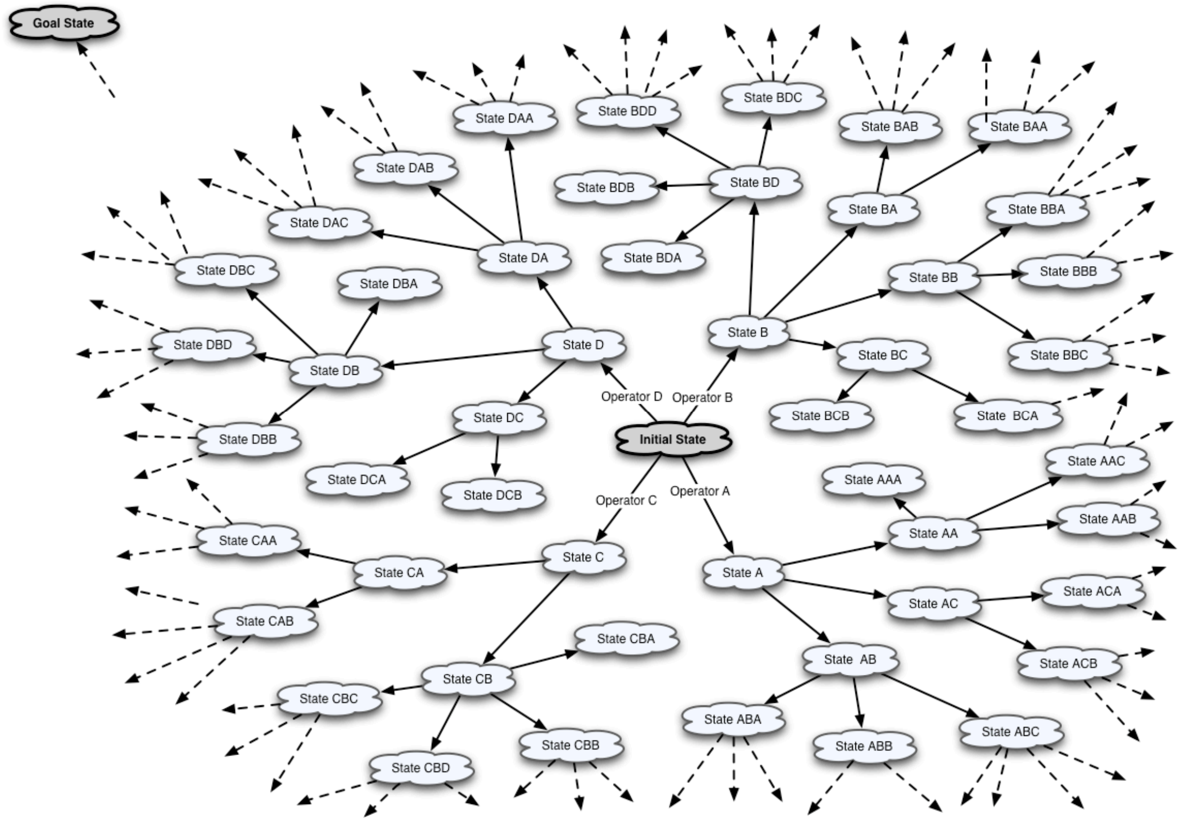
The problem can then be defined as: *how can you find a goal state starting from the initial state?*

According to Newell and Simon, the most general method to search for a goal state is **generate-and-test.** First, you need to generate a new state starting from the present state. This happens typically by applying a predefined "operator", i.e. a certain type of move that is possible according to the constraints or rules of the situation. For example, in chess certain moves are allowed depending on the type of piece (e.g. queen or knight) and the configuration on the chessboard. With a car, a possible move is to check the fuel meter or to open the hood to check the batteries. Such moves in general need to be performed in the right order: it is not possible to check the batteries before first opening the hood. Once you have moved to a new state, you need to *test* it: is it a solution or goal state (e.g. is there a checkmate)? If not, then you again need to generate a new state. This procedure is repeated until a solution is found. Note that *generate-and-test* is essentially equivalent to *trial-and-error* or to the *variation-and-selection* that we find in Darwinian evolution: trial or variation = generate a new state to be tried out; error or selection = after testing, reject all "errors", i.e. states that are not solutions, and start again. This rudimentary method can be studied in more detail by introducing the concept of a **search space**.

Search spaces

All the possible states that can be generated by applying a sequence of possible operators together define an abstract problem space. Problem solving can then reformulated as:

> finding a path or trajectory through the problem space that leads from the initial state to a goal state. This path is preferably as short as possible.



The difficulty is that the number of possible trajectories or states to be tested increases *exponentially* with the number of moves. Suppose that in each state there are 10 operators applicable to generate new states. This means that for one move there are 10 possibilities, for two moves 10×10 = 100, for three moves 10×10×10 = 1000, etc. Because of this exponential explosion in the number of possibilities to be considered, in practice it is impossible to systematically explore the search space for more than a few moves away from the initial state: we cannot consider all the possible states to check whether one of them is a goal state. For a more concrete example, consider chess: in each round of the game there are about 1000 moves and countermoves possible. Suppose that you would try to think 8 steps ahead. This means that you must consider $1000^8 = 10^{24}$ possibilities, an astronomical number! According to this reasoning, even the simplest problems quickly become unmanageable. The conclusion is that generate-and-test must be supplemented by methods to reduce the size of the search space.

Generate-and-test or trial-and-error is an essentially *blind* method of search: the searcher has no foreknowledge of where to look in the search space. In practice, **knowledge** or **expertise** is needed to effectively solve problems. Cognitive scientists have observed again and again that the true experts, such as chess grandmasters, car repair specialists or medical doctors, know plenty of specific methods to reduce search in their domain. These methods can be divided in two categories:

- **algorithms** are systematic, deterministic procedures guaranteed to produce a solution after a certain time. Examples of algorithms are the rules that we use to perform a division, or to solve a set of linear equations.
- **heuristics** are rules of thumb that may or may not work, but that generally lead to the solution much more quickly than blind search. They do this by focusing the attention on the (small) part of the search space that is most likely to contain the solution. An example in chess is trying to capture the most powerful pieces (e.g. the queen) of your opponent, rather than immediately searching for a checkmate.

General heuristics

Most algorithms and heuristics are very domain-specific: they require specialized knowledge, e.g. about algebra, chess strategies, medical diagnosis, or car maintenance. However, there are a few heuristics that are in principle applicable to any domain, and that therefore are of general use in problem solving, even when the problem-solver has as yet no knowledge about the domain.

Analogy

Analogy is a method we all use, consciously or unconsciously. The principle is to try to reduce a new, unknown problem to a problem already solved earlier, so that we can apply the specific methods that have been successful in the previous case. For example, if the problem is that a motorboat doesn't start, then you can try to find the components in the boat that appear equivalent to the components you know in a car, and try the same approach as when a car doesn't start. When trying to calculate the energy levels for a helium atom, a physicist will typically start by remembering how these are calculated for the (much simpler) hydrogen atom.

This is why complex problem domains (such as cognitive science) are typically explained by means of concrete examples. By seeing the analogy between a new type of problem, concept or theory and a phenomenon you already know and understand, the new idea will become much more easy to grasp. This explains why scientific approaches (“paradigms”) are rarely formulated in the form of abstract principles, and more commonly through a collection of exemplars: typical phenomena, cases, or experiments that are easy to understand and remember but that can be generalized to more universal rules when needed. The exemplars function like prototypes from which other cases inherit most of their properties by default. The only thing that needs to be added are the distinctive features of the cases, i.e. their specific properties that deviate from the default expectation. Problem solving can then concentrate on these distinctive features, while applying well-known algorithms or heuristics to the default features.

Divide-and-conquer

If a problem is too complex to systematically search for possible solutions, you can make it much simpler by dividing or "factorizing" it into subproblems. Starting from the goal (end), you try to determine intermediate steps (subgoals) to get there. The principle is that, if you have one by one solved each subproblem (reached its goal state), then you have solved the problem as a whole. The subgoals act as "stepping stones" towards the final solution, making a seemingly insoluble problem easy.

Example: assume your car is out of fuel and you need to refill it while being unable to drive. You can factorize that problem into the following subproblems, each of which is relatively easy to solve:

- find container
- locate petrol station
- go with container to petrol station
- fill container
- return with container to car
- fill tank from container

The big benefit of **factorization** or **divide-and-conquer** is that it stops the exponential explosion in the number of states that you need to explore. Indeed, you only need to look ahead to the next subgoal, which is typically only a few moves away. A smaller number of moves implies an exponentially smaller number of trajectories to consider.

Example: chess: assume that you can split up an 8-step look-ahead into 4 subgoals, each 2 steps ahead. The number of moves to consider is then reduced from 1000^8 to 4×1000^2.

Means-Ends Analysis

Newell and Simon in their General Problem Solver have developed a more sophisticated version of factorization, which they called Means-Ends Analysis. The idea is to see each subproblem (or "end") as a specific *difference* between initial state and goal state that needs to be reduced. Solving the problem then means finding (sequences of) operators ("means") to eliminate each of the differences in turn, thus transforming the initial state into the desired end state. In the car out of fuel problem, an example of an operator or means to tackle the "locate petrol station" subproblem is to "look on map"; the "go to petrol station" subproblem may be solved by "walking", "hitchhiking", or "calling a taxi".

Hill climbing with backtracking

Hill climbing is perhaps the most flexible general heuristic. Its basic assumption is that you can estimate the "closeness" to the goal of any given state. This corresponds to a measure of the "goodness", quality or fitness of that state. In chess, you could estimate your closeness to the goal of winning the match by counting the difference between the chess pieces that you captured and those that you lost. Visualize this measure now as "height", i.e. elevation above the plane that represents the search space. The goal state or solution then is the highest point in the landscape, i.e. the top of the hill. To reach that point, you can use the following **hill-climbing heuristic**: of the different possible moves/operators available in the present state, choose the one producing the "highest" or best next state. In other words, follow the path with the steepest ascent. Since you only need to look one step ahead, there is a maximal reduction of search. You repeat this climbing until a state with satisfactory fitness is obtained.

This method has a major shortcoming, though: there may be local maxima, i.e. small hilltops that are lower than the main hill, with unsatisfactory fitness. On such a local maximum, all next moves can only reduce the height, so the hill-climbing rule is no longer applicable. One solution is to **backtrack**, i.e. go back to a previous state and choose the next best option there, in the hope that it will not end up in a local maximum. You may need to backtrack repeatedly before getting back on a path leading to high quality states.

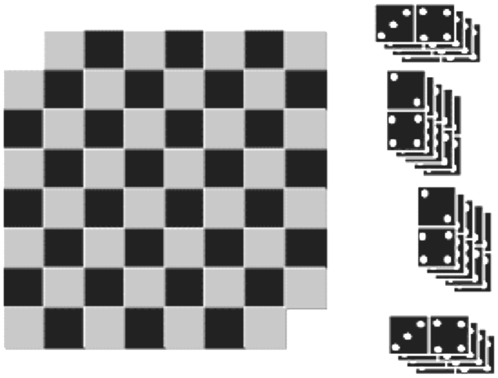
Representation change

Here the principle is that by redefining or reformulating states, operators and/or fitness function, an apparently unsolvable problem may become much easier, and perhaps even trivial. This happens by reducing the search space or number of operators that need to be applied.

<u>Example:</u> Roman numerals vs. Arabic numerals.

Multiplication is relatively easy with Arabic numbers, but practically impossible with Roman ones. Addition, on the other hand, is very easy with Roman numerals: XXII + XV = XXXVII. Thus, by changing the representation of numbers, we can make a very difficult problem much easier.

<u>Example:</u> mutilated checkerboard problem.



Consider a checkerboard that lacks two squares on the opposite sides of the diagonal. Given that one domino covers two adjacent (horizontally or vertically) squares, can you find a way to place dominoes so that all squares are covered? At first sight, this problem is very difficult: there are hundreds of possible combinations of dominoes to try out. Now, redefine the main operator: one domino covers not just two adjacent squares, but 1 black and 1 white square. Since the checkerboard lacks two squares of the same color (black in the drawing), it now becomes immediately obvious that it cannot be covered in this way.

Symbolic Artificial Intelligence

Computer simulation as method

Artificial Intelligence (AI) can be defined as the simulation of human intelligence by means of computer programs. The computer appears as a perfect tool for simulating abstract reasoning processes. Assume that a computer program gets the same input as an intelligent person, such as a question or problem description. The simulation works if the program produces the same or equivalent output as the person, such as the desired answer or problem solution. If we manage to write a program that systematically achieves that, then we can say that we have really understood what intelligence is. The philosophy underlying computer simulation as a method to study cognition can be summarized as “I'll understand it only when I can build it”.

This method has several advantages over earlier approaches:

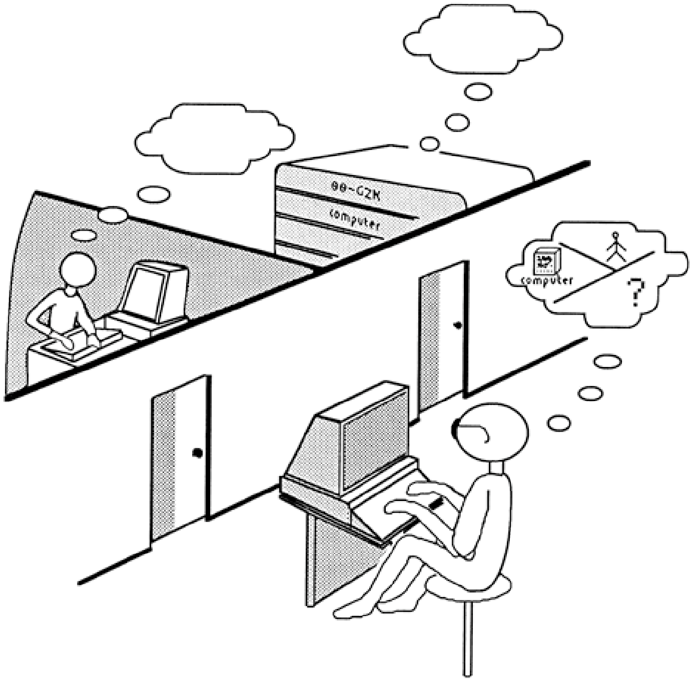
- it avoids the vagueness and subjectivity of introspection
- it avoids the limited access to the mind of empirical observation
- it forces researchers to be explicit about all components and processes of intelligent behavior, since a computer can only understand completely formal descriptions. Thus, it avoids theories (such as “homunculus-type” theories) that intuitively seem OK, but that in practice lack crucial components.

How do we recognize an intelligent computer program?

One of the first issues in AI is the success criterion: when can we say that we have truly designed an intelligent program? An intuitive criterion is implied by the definition, “Intelligence is the ability to solve *complex* problems”. However, what we consider *simple*, for example recognizing a friend in a crowd, often cannot be done by a computer. On the other hand, what is a complex problem for us may be simple for a computer, for example making a long calculation.

A commonly used assumption is that “Intelligence is everything a computer cannot do”. This is an implicit criterion used by people critical of the possibility of AI: since they a priori assume that computers cannot be truly intelligent, any advance made by AI researchers is interpreted by them merely as evidence that the behavior that the AI program now simulates does not require real intelligence. For example, critics of AI have maintained for decades that computer programs could never win a chess match against a human Grandmaster. But this objection had to be abandoned when the IBM computer Deep Blue convincingly beat the world champion Gary Kasparov. Nowadays, these same critics will be inclined to say that playing chess is merely the mechanical application of rules, which does not demand true, human intelligence. Obviously, this is not very fruitful as criterion. This brings us to the most generally accepted criterion.

The Turing test



The criterion advanced by the computer scientist Alan Turing is the following: you have achieved real Artificial Intelligence when a person conversing freely with a computer program (via some text terminal so that the person cannot see who is producing the sentences) cannot distinguish it from a true human. While this seems more objective, intuitive and easy to measure than the previous criteria, it too suffers from some serious shortcomings:

In some ways, success in the **Turing test** may be much easier to achieve than it seems. Indeed, already in the 1960's the Eliza program, which reacted to questions by restating them, and then formulating related questions, fooled people into believing that it was a human psychotherapist. It is not difficult to mislead people, as illustrated again and again by illusionists.

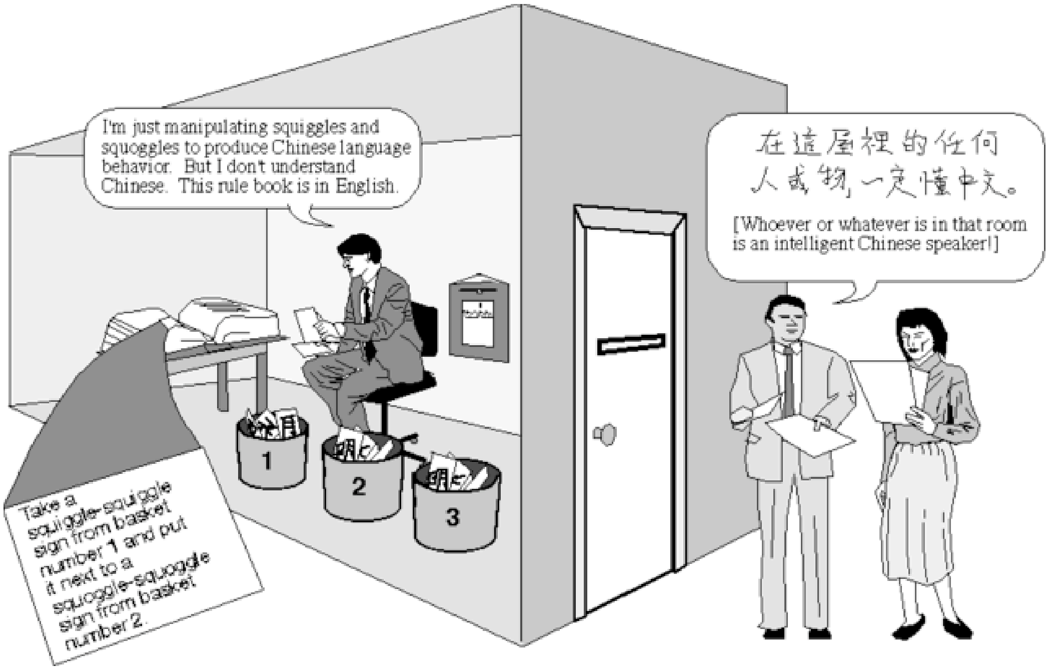
On the other hand, people such as children, mentally handicapped persons or people from a different culture, may be intelligent in advanced ways without being able to hold a sophisticated conversation about common topics.

In addition to these practical problems, there is a more fundamental objection to the assumptions behind the Turing test.

The Chinese room thought experiment

This thought experiment was proposed by the philosopher John Searle [1980] to demonstrate that even the most sophisticated computer program cannot exhibit real intelligence. Imagine the following situation: a (non-Chinese speaking) man in a room receives pages with Chinese characters. Following elaborate written procedures, he responds to each page by assembling another set of Chinese characters. If the procedures are sufficiently sophisticated, a Chinese person outside the room may get the impression that these are intelligent responses to his questions. In other words: the room passes the Turing test for intelligence. Yet, neither the man in the room (processor) nor the written procedures (program) have any understanding of Chinese. In Searle's view, an AI program is nothing more than an automated, electronic version of a Chinese room. Therefore, we must conclude that whatever seemingly intelligent behavior it appears to mimic, it lacks the fundamental understanding of the world that is required for intelligence.

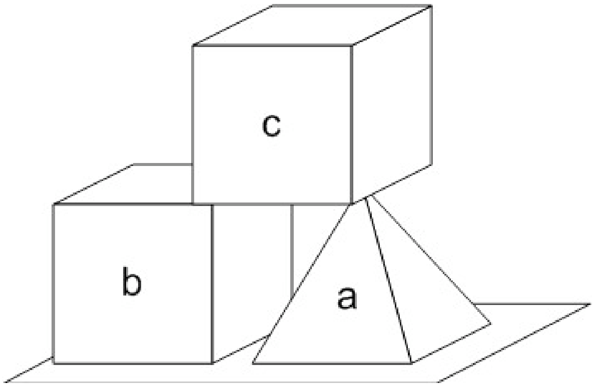
The reply of AI proponents is that Searle's analysis is too reductionist: while neither component may understand Chinese on its own, the room as a whole (man interacting with procedures) does.


I'm just manipulating squiggles and squoggles to produce Chinese language behavior. But I don't understand Chinese. This rule book is in English.
Take a squiggle-squiggle sign from basket number 1 and put it next to a squoggle-squoggle sign from basket number 2
1
2
3
在這屋裡的任何人或物一定懂中文。
[Whoever or whatever is in that room is an intelligent Chinese speaker!]

Toy worlds

Another criterion for intelligence is that an intelligent computer program should be able to deal with the real world. But since the world as a whole is too complex to simulate in a computer, we can build a simplified, "toy" version of it, and see how well an AI program can deal with it.

A traditional example of such a simplified environment is a "Blocks world", which consists of a virtual space in which blocks of different shapes and colors are strewn around. The program must find ways to manipulate the blocks in order to solve problems by reasoning about the situation: e.g. "block c lies on top of



block b, therefore to reach b, I should first remove c". If this works, the physical equivalent of a simulated toy world may then be explored by a robot using the same AI program.

Symbol systems

The fundamental units of AI simulations are **symbols**. These are abstract patterns inside the computer, such as b, that represent or symbolize an outside phenomenon, such as a block. An example in human cognition is the word "cat" which symbolizes or stands for a catlike animal. Symbols are very easy to manipulate: they are formal or abstract, in the sense that they are decoupled from the phenomena they represent. Therefore, they can be manipulated independently of these phenomena. At the same time, they can be implemented in a very concrete, physical manner, e.g. as spoken words, letters on paper, or electromagnetic configurations in a computer memory. This makes it easy to process them mechanically, according to fixed rules—whether by a human or a computer.

The basis of an AI program is a **knowledge representation**. This is a system of symbols connected by relationships or rules, such as a semantic network or a collection of formulas in predicate logic. It functions as a formal model of the knowledge that the system has about the phenomena represented by the symbols. As we noted when discussing problem solving, different representations may be useful for different domains or types of problems.

This general approach to modeling cognition is based on the **physical symbol hypothesis**. It was formulated by Newell and Simon as:

> *a physical symbol system has the necessary and sufficient means to produce general intelligent action.*

More precisely: any system (human or machine) exhibiting intelligence must operate by using elementary physical patterns (**symbols**), combining them into structures (**expressions**), and manipulating these expressions (using rules or procedures) to produce new expressions (inference). Solving a problem simply means:

- representing the problem's initial state as a combination of symbols (*input*)
- applying operators, rules and heuristics to combine and recombine these symbols (*processing*), until
- a new combination is produced that can be interpreted as a goal state or problem solution (*output*).

Criticism: While this assumption appears adequate for higher-level, formal types of reasoning such as playing chess, or proving theorems, it seems less appropriate for everyday intuitive action or low-level information processing such as interpreting visual stimuli. More fundamentally, the physical symbol hypothesis makes abstraction of the relation between a symbol (the *signifier*, in semiotic terminology) and the phenomenon it represents (the *signified*): it assumes that manipulating the signifiers alone is all that it is needed to produce intelligence. By that it ignores the processes that produce the signifiers in the first place, i.e. the interactions between the outside phenomena and their inside representations. While this abstraction makes modeling intelligence much simpler, it leads straight into the **symbol grounding problem,** as we will detail shortly.

Generative Grammars

Generative grammars, introduced by the cognitive linguist Noam Chomsky to describe the foundation of human language, are rules to combine symbols (words) into expressions (phrases or sentences). Since these rules can be applied recursively (meaning that the same rules can be applied again and again to the same initial expression in order to generate ever more complex expressions), generative grammars can generate an infinite number of possible expressions. Yet, this infinity is still only a fraction of the infinite number of all possible combinations of symbols. Only the expressions generated according to the rules are grammatically correct, and therefore meaningful.

Parsing is the process of analyzing an expression in order to find the sequence of rules that has generated it. When an expression or sentence has been parsed, its logical structure has been revealed. This is necessary for natural language understanding (e.g. determining which are the subject, verb and object of a sentence), but also for understanding an expression in any symbolic knowledge representation, such as predicate logic. Expressions that cannot be parsed are meaningless for an AI program.

Inference

The core mechanism in AI is the process of **inference**, which uses the knowledge stored in symbolic form to derive new symbolic knowledge. This is modeled on logical or mathematical deduction. Expressions initially assumed to be true play the role of the *axioms* of the formal system. They represent the primitive knowledge or basic assumptions of the system. These expressions are usually inserted by programmer—although they may in principle also result from perception of the outside world. They are usually formulated by means of predicate logic or a similar formalism, as illustrated by the following expressions.

Examples:

- Bird(sparrow) : Predicate logic representation of "the sparrow is a bird"
- (forall x): Bird (x) ⇒ Has_Feathers(x): representation of "all birds have feathers"
- Larger_than (ostrich, blackbird): "the ostrich is larger than the blackbird"
- Larger_than (blackbird, colibri)
- Larger_than (x, y) & Larger_than (y, z) ⇒ Larger_than (x, z)

Using deduction rules that have been built into the program, the system can then derive new expressions from known expressions.

Example: if A and (A $\Rightarrow$ B) are both true, then B is true: this is the well-known *modus ponendo ponens* deduction rule.

Expressions inferred in this way then correspond to the *theorems* of the formal system, i.e. true expressions that are derived from the axioms.

Example: from the axioms in the previous example, we may deduce the following expressions:

- Has_Feathers (sparrow)
- Larger_than (ostrich, colibri)

A problem or question that the system must answer then simply corresponds to an expression to be proven or disproven. For example, animal (penguin)? This expression is considered to be true if we can find a sequence of deductions from the axioms that produces this proposition as a theorem, and false (or indefinite) otherwise. Alternatively a problem or question can be stated as: find the values of the variables (x, …) for which the expression would be true, e.g. bird(x) & Cannot_fly (x) ? Given this question, the system will try to find all x that fulfill this condition, i.e. all birds that according to the stored knowledge cannot fly.

Prolog is a programming language used in AI that is based on a simplified form of predicate logic (similar to the examples we gave). Entering an expression like the one above in Prolog will automatically produce an answer. The algorithm it uses is based on backward chaining: starting from the expression to be proven, it searches for expressions that imply it, filling in the variables when needed, until it gets down to one of the "axioms" or expressions assumed to be true. In AI, there exist various such algorithms and systems of deduction rules for deriving true expressions from given axioms. Together, they constitute what is called an **inference engine**. Different inference engines have typically their own strengths and weaknesses, the one working better on one type of problems, the other on another type.

Expert systems

In practice, a purely deductive logic is not sufficient for real-world problems: knowledge is not that clearly defined or consistent. Instead, human experts use specific heuristics or rules of thumb to produce "likely" answers, rather than answers that can be proven to be true. We already saw default reasoning as an example of such less than logical inference. In addition to semantic networks, another common format to represent such "approximate" reasoning are production rules, which have the form:

IF A is true, THEN B is true with probability or confidence x,

(where x is commonly less than 100%.)

The inference engine must determine which rules are relevant in a given situation and choose which one(s) to apply. AI systems built on such practical knowledge rules for specific domains, are commonly called "expert systems" or **knowledge-based systems**. The design of such systems is called “knowledge engineering”. They have many practical applications, such as medical diagnosis (MYCIN, one of the very first experts systems helped with the diagnosis of lung diseases), or the repair of televisions or other complicated technical systems that require a lot of expertise.

The symbolic paradigm for cognitive science

The ideas underlying classical artificial intelligence, as founded on the **physical symbol hypothesis**, can be seen as a paradigm for the whole of cognitive science. Symbols, knowledge representations, generative grammars, and inferences have also been postulated as the building blocks of knowledge in domains like linguistics, psychology, logic, and philosophy of science. AI has shown how powerful, detailed and explicit models of cognitive processes can be built out of these elements. Moreover, these models have produced some impressive practical applications, such as expert systems, theorem provers and chess-playing programs.

In addition, symbolic models are simple and intuitive. They can be seen as extensions of the **reflection-correspondence theory** of knowledge: each mental symbol reflects or represents an aspect of the external situation. Yet, they avoid **naïve realism**: there is no simple correspondence or mapping from material objects to mental symbols. The reason is that symbols not only represent concrete objects or instances, but also abstract categories and relationships. They also sidestep the homunculus problem: the role of the homunculus interpreting the representations and making decisions is played by the inference engine. Like **logical empiricism**, symbolic cognitive science integrates empiricism (because most symbols stand for observable features) and rationalism (because inference engines can produce new knowledge purely by reasoning). Symbolic CS has moreover assimilated the lessons of **pragmatism**: symbolic representations are chosen not for their truthfulness or their correspondence to objective reality, but for their efficiency in problem solving. In conclusion, the symbolic paradigm appears like a flexible, elegant and general foundation for cognitive science.

Shortcomings of the symbolic paradigm

In spite of these advantages, symbolic CS has encountered several fundamental problems, which we will now review.

Failure to simulate real-world intelligence

Most obviously, in spite of half a century of research, AI has failed in its main aim: the simulation of a human level intelligence that is general, i.e. that can adapt to a broad variety of tasks and domains. The only successes are very specialized, “expert” programs and simulation of toy worlds. There has never been a program passing a satisfactory Turing test. At this moment, there exist no intelligent robots. Present-day robots are either extremely specialized machines used in industry, e.g. to assemble small components, or more general-purpose “electronic pets” with an intelligence level comparable to an insect.

Even the much less ambitious early goal of AI, automatic translation between different human languages, has failed to come near to any human level of reliability (just try the different automatic translators available on the web). Moreover, those translating programs that are most effective have abandoned the true AI approach of trying to *understand* words and sentences in favor of more low level approaches such as gathering a lot of statistical material about the occurrence of different expressions in different languages, and seeing which expressions are most likely to occur in a given context.

This does not imply that the many critics who claim that AI is in principle impossible are correct, but only that the traditional symbolic approach to AI has not been able to live up to its promises. Let us try to understand more fundamentally which obstacles this approach has encountered.

The knowledge acquisition bottleneck

A first observation made by AI researchers is that formal inference capabilities are useless to tackle real-world situations without sufficiently detailed and concrete knowledge. To reach a level comparable to adult humans, an AI program would need to know millions of facts. However, it is very difficult to exteriorize and formalize the knowledge that people use intuitively. Knowledge engineers, who develop expert systems, have called this the **knowledge acquisition bottleneck**: it is much easier to enter symbolic knowledge into a computer system than to get it out of a human expert’s head. The reason is that humans do not reason logically, and do not store their knowledge in an explicit, symbolic form. Most knowledge is fuzzy, subconscious and intuitive. It is therefore very difficult to express in the form of logical statements.

As a result, AI programs have always lacked common sense or “real-world” knowledge. Initially, researchers thought that this was only a problem of quantity of knowledge, and that it would suffice to collect an extensive amount of expressions that describe the world as we know it. This led to the CYC project, which was started in 1984 by AI pioneer Douglas Lenat [1995] with a large team of collaborators to collect all “common-sense” knowledge that the average human possesses. At this moment, the knowledge base contains over a million human-defined assertions, rules or common sense ideas. Typical pieces of knowledge represented in the database are “Dogs are animals” and “Animals die eventually”. When asked whether dogs die, the inference engine can draw the obvious conclusion and answer the question correctly. However, after more than 20 years of development, the CYC project still has not produced any truly impressive applications.

The knowledge representation problem

As we noted when discussing problem solving, different representations of the same problem domain each have their own advantages and disadvantages. No representation is adequate for all possible cases. In practice, this means that whatever representation you are using, you will encounter plenty of cases in which this representation is highly inefficient, i.e. where searching for solutions to the problems takes far too long to be practical. This observation has led to attempts to automatically transform one representation into another one, but except for some specific cases, this turned out to be too ambitious for representations in general. The reason is that symbolic representations are intrinsically rigid: their elements or "atomic units", the symbols, are a priori given, and therefore cannot be changed. At most, we can change some of the grammatical rules by which symbols are combined into expressions, but if the best conceptualization of the problems cannot be expressed by the given symbols, this will not be of much help.

Closely related is the problem of **learning** or knowledge discovery: how can you program a computer to extract or induce new concepts and rules from data? The AI-related domain of "machine learning" and its more application-oriented descendant of "data mining" have produced quite a number of useful techniques. Yet, most of these are not based on the symbolic paradigm with its qualitative, logical reasoning, but on various more quantitative and trial-and-error based algorithms, including statistical clustering and factor analysis, genetic algorithms, and neural networks (see further).

Lack of autonomy

The problem runs even deeper. Even with plenty of knowledge and an adequate symbolic representation, logical or probabilistic deductions are insufficient for modeling common-sense understanding of the world. One of the problems is simply: when should the system stop making inferences? An infinite number of theorems (derived expressions) can be deduced from a finite number of axioms (basic expressions). However, most of those have no real-world relevance or utility, and the computer program will just get bogged down producing ever more trivial or meaningless observations. For example, after deducing from "Fido is a dog" and "a dog is a carnivore" that "Fido is a carnivore", the program may continue deducing that "Fido is a mammal", "Fido is a vertebrate", "Fido is an animal", "Fido is a living being", "Fido is a thing", "Fido is an entity", "Fido has weight", "Fido has mass", "Fido has energy", "Fido has volume", "Fido has length", "Fido has height", "Fido exists", etc., etc.

There is nothing in such a symbolic, logic-based formalism to determine which expressions are meaningful or relevant: in logic all true expressions are equally important. In practice, this is of course not the case: we are only interested in a limited number of things at any moment. This means that we need heuristics and adapted problem representations to reduce the search space, and focus on the things we find important. However, these heuristics are very context-dependent: what works well in one domain, generally does not work in another domain. The problem then is: how can we fully specify all necessary and sufficient conditions to come to a meaningful conclusion? In the domain of planning action, this has become known as the *frame problem*: which are all the relevant facts that we should consider when planning our actions, and when can we stop deducing further facts?

More fundamentally, the problem is that AI programs are designed as question-answering systems: it is the human user who introduces the problem or question that the system is supposed to tackle. This is obvious in the Turing test, where the intelligence of the system is probed by asking questions. Human beings, on the other hand, are **autonomous**: they decide for themselves what problems to tackle, which issues to pay attention to, or which, if any, questions to answer. As such, they do not have to wonder about how many inferences to make, or which axioms are relevant. What is relevant depends on their own goals and preferences. These are more fundamental to their cognitive functioning than any questions artificially introduced by some experimenter.

The symbol grounding problem

Traditional AI is based on the manipulation of abstract symbols. However, although symbols are supposed to represent outside reality, it is never specified *how* they do this, i.e. how they are **grounded** in reality. Given the way symbols are used inside the program—undergoing combinations according to formally specified rules—they might as well correspond to meaningless concepts with no counterpart in reality. This may remind us of the way medieval philosophers were pondering deep questions such as "How many angels can dance on the head of a pin?", when no one had ever observed an angel. This might not be a problem if there were a simple, unambiguous correspondence between symbols and objects in the outside world. But this is what we have called the reflection fallacy: knowledge is not simply a reflection of an objective reality.

To check what a symbol means we need a process of **perception** that compares external situation and internal representation. By starting with symbols, AI has neatly sidestepped the problem of perception. But attempts to build robots, which cannot function without some kind of perceptual apparatus, have shown that human-like perception is a much more complex process than anybody had anticipated. It turns out to be extremely difficult to interpret a pattern of pixels as sensed by a camera as the depiction of a particular concept or thing, such as the dog "Fido", or the property "liquid". Insofar that present-day programs manage such recognition, they do this by using non-symbolic methods, such as neural networks. The most cogent arguments brought forth by critics of AI, such as John Searle [1980, 1992] and Hubert Dreyfus [1992], focus on the fact that AI programs have no **experience** of the phenomenal world: they reside in a purely abstract, symbolic realm where there are no such things as sensation, feeling or perception.

Conclusion

We can summarize the above observations by noting that symbolic AI, and with it most of traditional cognitive science, is too rationalist and logic-based. Logic has turned out not to be a sufficient foundation for modeling cognition. It leaves no space for the fuzziness and subjectivity of experience, intuition or feeling. Moreover, starting with symbolic representations cuts out any interaction between an AI system and the real world: it cannot perceive or experience phenomena.

The symbolic paradigm is moreover too mechanistic: *everything* needs to be explicitly programmed according to fixed rules. The system breaks down as soon as some knowledge, some specification of the problem or situation, or some part of the program is missing. It cannot easily learn new facts on its own, discover new insights or adapt to different circumstances, and it cannot decide on its own what is important or what to do next. A true intelligence should be able to self-organize, to autonomously develop and grow smarter, in the same way as a human mind does.

We will now discuss a number of different approaches that have taken this criticism to heart, and attempted to develop a fundamentally different theoretical framework to address these problems.

New Approaches to Cognition

Connectionism

Probably the best-known alternatives to symbolic representations of cognitive systems are connectionist or “neural” networks. These are directly inspired by the working of the brain, with its neurons connected by synapses, rather than by the formalisms of logic and mathematics. They are no longer based on independent, discrete *units* of meaning (symbols), but on the *connections between* units. They do not distinguish strict categories, where something is either *A* or *not A*: everything is fuzzy, fluid and changing. The meaning resides in the connections between otherwise meaningless elements. It is “grounded” in input received through perception. There is no logical formalism or set of fixed deduction rules, but a continuous adaptation of the system to experience. As a special bonus, there is no need for detailed programming: the system organizes itself.

Connectionist models were originally proposed by the cyberneticians McCulloch and Pitts [1943], and then elaborated by the AI researchers Minsky and Papert in the 1960s, until it turned out that the existing networks were unable to solve an important class of problems. After this obstacle was surmounted with the introduction of the backpropagation algorithm, their final return to fashion came with the work of the psychologists Rumelhart and McClelland [1986] in the 1980s.

Components of a Neural Network

Connectionist representations are based on the concept of a **neural network** (sometimes called “artificial neural network” to avoid confusion between computer simulations and the actual neurons that exists in our brain). It consists of the following fundamental building blocks.

The role of the neurons in the brain is played by **nodes** or units x_i, i = 1, ...N (the index i ranges from 1 to the number N, the total number of nodes in the network). Nodes function very roughly like concepts, i.e. cognitive units. They can be “activated” to a variable degree. (In the brain, activation corresponds to the intensity of impulses or electrical activity passing through the neuron). The “mental state” or “working memory” of the network can be represented as a (N)-dimensional vector $A(x_i)$ of **activation** values over the nodes.

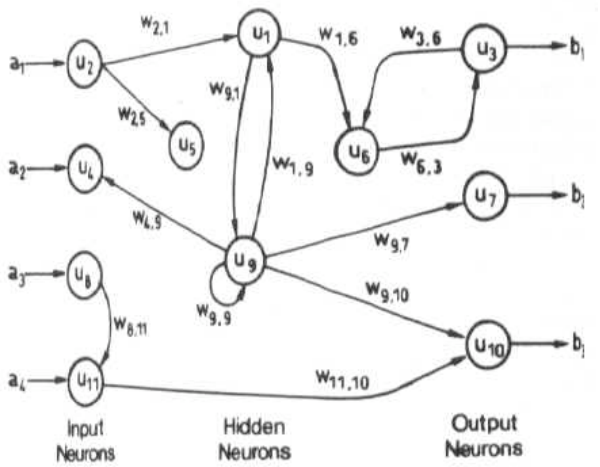
The activation function takes values between 0 (more rarely -1) and 1:

$$0 \leq A(x_i) \leq 1, i= 1, ...N$$

This entails a generalization of binary logic: assertions can take on more than two truth-values (yes-no or 1-0); they can be true to a certain degree.

The role of the synapses is played by **connections** between the units. These connections too exist to a certain degree, which is called their *strength* or *weight*. Connections function as "associations" between the concepts. They can be seen as a generalization of implications in logic or of links in a semantic network. The long-term memory or knowledge of the system is determined by the strength of all the different connections. These connection weights $w(x_{i,}\ x_j) = w_{ij}$, are generally represented as a (N×N)-dimensional matrix:

$$-1 \leq w_{ij} \leq 1,\ i, j = 1, \ldots N$$



Spreading activation

The most basic process in a neural network is the **propagation of activation**. Activation is passed on from an activated node to all nodes it is connected to. This happens simultaneously or in parallel, unlike traditional AI systems where typically only a single step or inference is performed at a time. The flow of activation passing through a link is proportional to the link strength: stronger links transfer more activation. The activation entering a node x_j from different previous nodes x_i at time t is summed to determine the new activation of the node at time t+1:

$$A(x_j, t+1)= a_j\ (t+1) = \sum_i w_{ij}.a_i\ (t)$$

However, usually this sum of incoming activations must surpass a **threshold** for the node to become activated itself. If it remains below the threshold value, the node remains non-active (A(x) = 0). The overall process of **spreading activation**, propagating from neuron to neuron, is the connectionist representation of a process of "thought" in the brain.

Learning

The second fundamental process in a neural network is **learning**: the changing of connection strengths (and therefore of stored knowledge) as a result of experience. There exist different learning algorithms, of which the most basic are:

Hebbian learning

Here a link is strengthened each time the two nodes that it connects are co-activated (i.e. activated simultaneously, or, in some cases, the second activated shortly after the first). This is directly inspired by the behavior of real synapses. The increase in weight is proportional to the product of the two activations:

$$w_{ij}\,(t+1) = w_{ij}(t) + c.a_i(t).a_j(t)$$

$0 < c \leq 1$ plays here the role of a *learning parameter*. The higher c, the bigger the influence of the most recent co-activation on the weights (and therefore the weaker the influence of earlier rounds of learning). In other words, increasing c makes the system learn new experiences more quickly, but also forget older experiences more quickly.

Hebbian learning is a form of unsupervised learning. This means that no one (not even the environment) is telling the system what it should learn: it derives its knowledge only from correlations between its experiences. As such, it creates associations between experiences that tend to co-occur, like Pavlov's dog, which learned to associate the sound of a bell with the appearance of food. It also allows the system to cluster similar experiences into the same category, while distinguishing it from other categories. Thus, Hebbian learning finds structure in what initially seemed like a stream of incoherent experiences, so that the cognitive system becomes capable of anticipating further experiences by extrapolating from what it already perceived. For example, Pavlov's dog learned to anticipate food whenever it heard the sound of a bell.

Error correction or delta learning

This is a form of supervised learning, where the experimenter or the environment tells the system how it should behave, and the system tries to match that externally imposed target. The delta rule takes into account the difference (error) between target and internal activation. The target activation a_{ext} represents the desired output of the neural network. The actual output, i.e. the internally generated activation a_{int}, is merely an attempt to produce the target, and therefore may need to be corrected. This correction takes place by adjusting the link weight so that the internal activation it generates comes closer to the external activation it received as target. Otherwise, the formula is similar to the one for Hebbian learning:

$$w_{ij}\,(t+1) = w_{ij}(t) + c.\ a_i(t).\ (a_{ext\ j}(t) - a_{int\ j}(t))$$

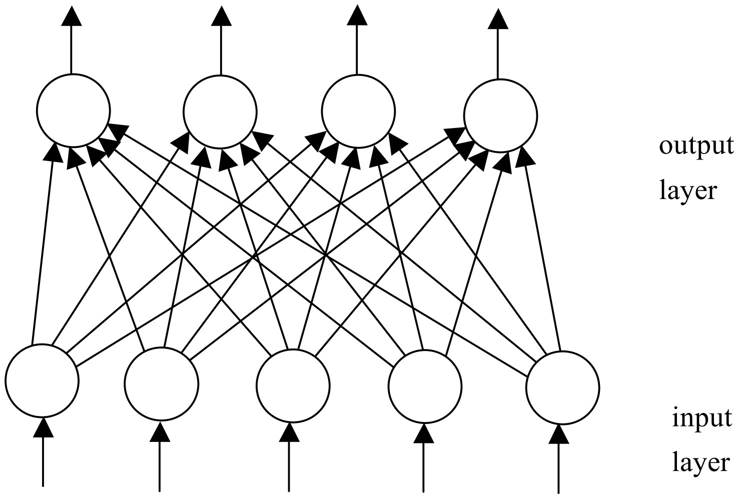
Note that the network stops learning as soon as internal activation a_{int} equals target a_{ext}.

A network learning according to the delta rule can be trained to produce a certain type of behavior. This happens by repeatedly performing the following training session:

- subject the input nodes to an example pattern
- compare the activation a_{int} of the output nodes to the desired activation a_{ext}

- correct the link strengths so as to bring the actual output closer to the desired output according to the above "delta" (difference) formula

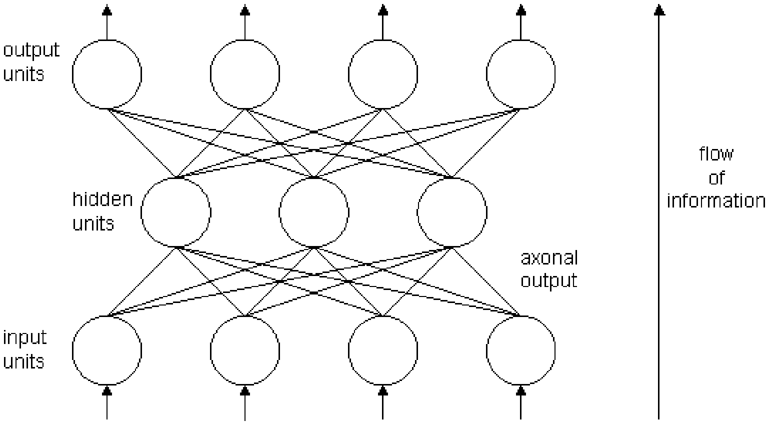
Correction is a form of **reinforcement learning** (as used in operant conditioning): connections that produced desired output are reinforced or strengthened (rewarded); connections that produce undesired output are weakened (punished). This is how Skinner trained his pigeons or rats to press certain buttons, and not to press other ones.



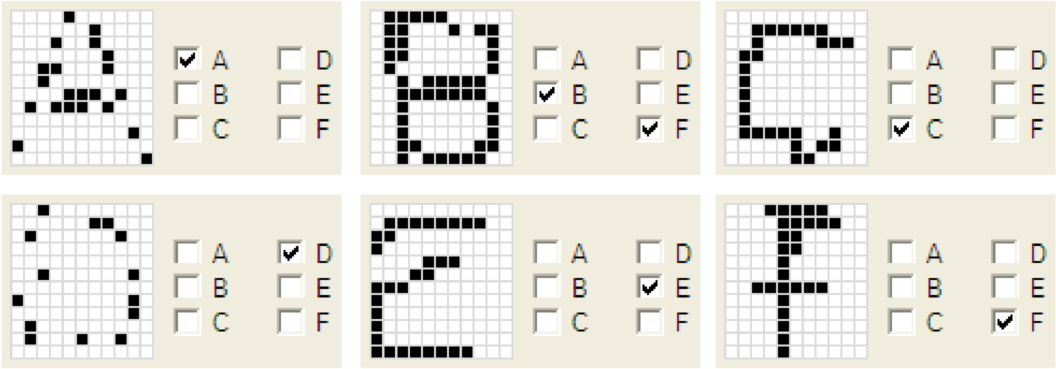
Feedforward networks

The simplest type of neural network is a single layer *perceptron*. The name derives from the idea that a perceptron is a very simplified model of the processes that interpret sensory stimuli, i.e. **perception**. It consists of one row or "layer" of input nodes (representing sensory input) and one row of output nodes (representing the network's interpretation of that input). Each input node is connected to each output node. The activation is propagated "forward" in a single step, from input layer to output layer. The activation of the input layer represents the problem to be solved; the activation of the output layer represents the network's proposed solution. Such a network can be used for simple classification tasks, where the input is some sensed pattern (e.g. black and white pixels within a square space), and the output consists of the activation of one or more categories that the pattern belongs to (e.g. "line", "circle", "square", etc.).

To get a multiple layer feedforward network, you can add one or more "hidden" layers of nodes in between the input and output layers. These hidden layers have incoming connections from input or previous hidden layer nodes. Their outgoing connections go to output or subsequent hidden layer nodes. A multilayer network can learn more complex classification and discrimination tasks than a single layer one. However, to achieve this it needs a mechanism to propagate corrections of link strengths from the output layer (where the target pattern is applied) step by step to earlier layers. This mechanism is called *backpropagation*.



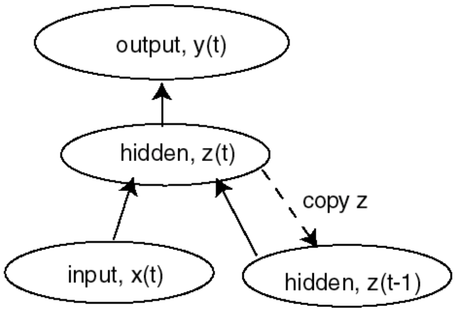
Example application: a feedforward network can learn to recognize a letter from a pattern of activated pixels, e.g. a 11 × 12 rectangle of pixels, some of which are white (not activated), others black (activated). These pixels (132 in this example) correspond to the input nodes. The nodes of the output layers can then represent different possible letters (A, B, C, …) represented by the pixel patterns (26 in this example). Initially, the connections in the network will have small random values, leading to random activation for the output layer. This output activation can then be compared to the desired activation (e.g. when the pattern corresponds to the letter F, the output node corresponding to F is fully activated, all others are not activated). This comparison leads to a correction of the connection strengths leading to the output nodes. This correction is then propagated backwards to the previous layers of connections. After many such training sessions, the network will have learned to correctly recognize most letters. Moreover, randomly changing the value of a few pixels is unlikely to change the result, since the activation from these "noise" pixels will have little effect on the final output activation. Thus, the network can learn different but similar versions of the letter F and classify them all correctly.



Other applications of feedforward networks include handwriting recognition, voice recognition, and "reading aloud" texts (i.e. mapping combinations of letters onto the correct sounds to pronounce the words).

Recurrent networks

In contrast to forward propagating networks, recurrent networks allow connections “back” to nodes (“layers”) activated earlier. Thus, they also implement feedback, not just feedforward.



A *simple recurrent network* is similar to a multilayer feedforward network except for one complication: the state of the hidden layer is copied to a “context layer”. This context layer then propagates its activation back to the hidden layer, providing it with an additional input. This is a simple implementation of short-term memory: the context layer remembers the state of the network one step earlier. Such simple recurrent networks are useful to predict patterns in time, such as sentences, where the preceding inputs (previous words) help to guess the next state (following word).

In a general recurrent network, all nodes are connected to all other nodes, so that we cannot distinguish layers, inputs or outputs. This means that activation can propagate backward as well as forward. The result or “output” of the network is the activation pattern achieved when activation has “settled down” into a stable configuration. General recurrent networks are in principle more powerful than the simpler varieties we have discussed before, but therefore also much more complex to investigate.

Advantages and shortcomings of neural networks

Neural networks have a number of clear benefits compared to symbolic AI programs:

- They can make “fuzzy”, “intuitive”, non-rational decisions. These are not based on explicit logical criteria, but on ever changing experience.
- They are flexible and fault-tolerant: small variations in input make little or no difference in output.
- They are *robust*: when the network is damaged by randomly removing nodes or links, it continues to function, although it gradually makes more errors as more damage occurs. This is called *graceful degradation.*
- They are *self-organizing*: there is no need for a detailed program or plan to tell them what to do. They learn the required behavior on their own from the examples provided.
- They are *decentralized*: processing happens in parallel, distributed over all the different nodes. There is no need for a “central executive” to coordinate the activities.

On the other hand, neural networks also have disadvantages:

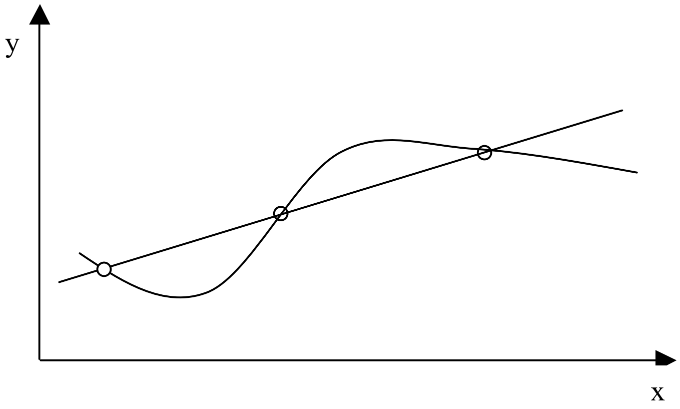
- They do not provide higher-level reasoning: language, logic, planning...

- They can only work when they receive input from the designer or trainer. In that sense, they are not autonomous.
- It is difficult to understand for the designer why or how a network comes to its decisions. Its "reasoning" is *distributed* over all its links and nodes, and therefore it is difficult or impossible to reconstruct what "arguments" led it to its conclusion.
- It is difficult for the same network to perform very different functions without the one interfering with the other. Learning a new function (e.g. recognizing numbers) generally makes the network "forget" how to perform a previous function (e.g. recognizing letters).

Constructivism

The underdetermination of theories by data

To introduce the constructivist perspective on cognition, we need to go back to the problems of epistemology and philosophy of science. For a given set of observations or data, there is always an infinite number of possible interpretations, theories or models that can explain those data. This is because a theory normally represents an infinite number of possible events, of which the observed events are only a finite subset. For example, the theory that "the sun always comes up in the morning" is based on a finite number of observations, although it makes an infinite number of predictions. The non-observed cases are "induced" by postulating general rules covering both actual and potential observations.



For example, through a couple of data points in a diagram you could draw a straight line, and induce that all further observations will lie on that line. However, you could also draw an infinite variety of the most complicated curves passing through those same points, and these curves would a priori fit the empirical data just as well. There is no objective criterion to prefer the "straight" model.

Therefore, we must first *build a theory on our own* and then try to fit the perceptions into it. This is what Popper meant when he said that science advances by conjectures (proposed new theories) and refutations (elimination or falsification of the theories that do not fit the observations). The observations or data alone are not sufficient to determine a good theory.

Not instruction, but construction!

In the words of Ernst von Glasersfeld [1984], constructivism can be defined in the following way:

> *Knowledge is not passively received either through the senses or by way of communication, but is actively built up by the cognising subject.*

In other words, the world does not tell us how to interpret it: sensory data are merely bits and pieces, without clear organization. This may be clarified by another constructivist slogan:

> Perception is not *in–formation* (formation of internal cognitive structures).

There also is not a programmer or instructor to put all the knowledge into the system. Good teachers do not *instruct* pupils, i.e. tell them in detail what to remember; instead they help the pupils to *construct* their own understanding of the concept, i.e. let them find out on their own how the different bits and pieces fit together into a meaningful idea.

The mind is not a recording apparatus, like in the reflection-correspondence view, but an active, autonomous living being. We construct an interpretation ourselves and project it onto our perceptions in order to make sense of the world. The role of knowledge is not to reflect a hypothetical objective reality or Ding-an-Sich, but to helps us to "fit in" with our local environment as we subjectively experience it ("Umwelt"). Different organisms have different Umwelts, and will therefore experience the world in a completely different way.

Another famous constructivist, the cybernetician Heinz von Foerster [1981], noted that *the brain cannot absolutely distinguish a perception from a dream or a hallucination* [see also Maturana, 1974]. These are all merely patterns of neural activation. What we call a perception is merely a particular interpretation of that pattern, which attributes its origin to the outside world, even though much of it is internally generated. Let us illustrate this process of internal construction using some examples from memory, the cognitive process that most resembles an objective registration of perceptions.

The construction of memory

Here is a little experiment. A list of some 20 words is read to the subjects, e.g.

> winter, icy, Siberia, warm, cooling, penguin, frozen, flu, chilly, ice, wind, hot, Antarctica, wet, fresh, breezy, igloo, cool, snow, Pole, glacier, frost, sleet

When trying to remember as many words of the list as possible, people will typically remember the word "cold", even though it is not part of the list. This is because "cold" is strongly associated to all other words. Therefore, the brain tends to "fill in" or "induce" the missing piece that it expects to be there. The conclusion is that memory is not an objective registration of perceptions: it is a "reconstruction" on the basis of vague, selective and subjective impressions, to which plausible assumptions are added in order to provide continuity and coherence.

There is a famous psychological experiment to demonstrate the unreliability of memory. People are shown a short movie in which a passer-by is attacked by a white man and afterwards helped by a black man. When retelling what they saw, many people claim the attacker was black. What happens is that they don't clearly remember the attacker, but they remember a black man being involved, and they implicitly assume that blacks tend to be criminal. Therefore their mind fills in the observation of the skin-color of the attacker, which it didn’t actually register. This experiment shows that witness reports, however confident, are not sufficient proof to convict a criminal [Loftus, 1974].

The psychologist Elizabeth Loftus carried out another surprising experiment to demonstrate the creation of **false memories**. The subjects were told a number of events that supposedly happened when they were children, and asked which ones they remembered. Most of the stories were collected from their parents or siblings and checked for their accuracy. However, some of the stories were invented by the experimenters. These described events that might plausibly have happened to the subjects, such as getting lost in a shopping mall, or accidentally spilling a bowl of drinks on someone's clothes during a party. When these stories were presented to the subjects as if they were true, a sizeable percentage of the people claimed to effectively remember the events. When prodded by the experimenters to remember more accurately, they came up with concrete details, such as the people present or the weather that day, which were not part of the original story. The more often they were interviewed about these childhood anecdotes, the more their “memories” became intense and detailed.

This and other evidence led Loftus [Loftus & Ketcham, 1996] to postulate the existence of a False Memory Syndrome. This may explain some dramatic court cases, in which therapists were convicted for making their patients falsely believe that they were abused as children. If a psychotherapist is convinced that his or her client's emotional problems stem from some repressed childhood experiences, he or she can by suggestive questioning make the client construct a number of detailed “remembrances”. Without independent evidence, it is virtually impossible to distinguish true experiences from falsely remembered ones.

<u>Example</u>: The existence of the false memory syndrome can be illustrated by the following true story. After a series of therapy sessions, a young woman became convinced of having been raped several times by her parents, and forced to undergo two abortions using a coat hanger. Her father, a clergyman, had to resign from his post when the allegations became public. However, a medical examination revealed that she not only never had been pregnant, but in fact still was a virgin. The patient sued the therapist for implanting false memories and received a $1 million settlement. For another example, I refer to the sensational stories of abuse “remembered” by the so-called X-witnesses after the Dutroux pedophile scandal in Belgium.

Coherence epistemology

The constructivists claim that knowledge is generated largely internally. Even our perceptions themselves are shaped by pre-existing knowledge. For example, seeing a planet through a telescope assumes belief in the laws of optics that govern the telescope. Hearing someone say something assumes the ability to understand spoken language: someone with poor hearing or poor language may "hear" something that was not said. Seeing a particular shape, such as a square, assumes familiarity with the concept square.

Therefore, perceptions cannot have priority in establishing the "truthfulness" of beliefs, in contradiction to what the empiricists argue. On the other hand, theoretical assumptions just as much lack objective foundations, in contrast to what the rationalists would argue. How then can we decide which theory is the best? The constructivists solve this problem by using the criterion of **coherence** [Rescher, 1974]. The best theory is the one that has most overall support from both other theoretical assumptions and from "observations", i.e. the system of assumptions and perceptions that is maximally coherent with each other. Two beliefs (assumptions or observations) are coherent if they are not only not inconsistent or contradictory, but also mutually supportive, i.e. if the one makes the other one more plausible.

The philosopher of science Paul Thagard [2002] has implemented this principle in a recurrent connectionist model of historical scientific discussions and the resulting revolutions in knowledge. Beliefs (assumptions, observations, arguments...) are represented as nodes in a network. Mutually inconsistent (contradictory) beliefs are connected by inhibitory links (i.e. links with a negative weight): when the activation of the one goes up, the activation of the other is pushed down. Mutually supporting beliefs are connected by excitatory links (i.e. with a positive weight): when the activation of the one goes up, the activation of the other is pushed up as well. The network is then started by giving "accepted" beliefs positive activation. This activation spreads along positive and negative links. When equilibrium is reached, the final positive activation pattern determines which beliefs have "won the argument". These usually corresponded to the ones that were eventually accepted by scientists. The beliefs that end up with zero or negative activation have lost, and are rejected.

Illustration: the construction game

The process of construction can be illustrated by an interesting party game.

A volunteer is sent away to another room while the rest of the group discuss their plan of action. When the volunteer comes back, the others tell him that they have invented a "dream" for him. This is a story in which the volunteer takes part, and which supposedly takes into account his experiences and personality. Being a "dream", however, it does not need to be based on any real or even physically possible events. The game then consists in the volunteer trying to guess the story that the others have made up for him. He can ask questions, but the others are only allowed to answer “yes” or “no”. For example, he might ask: “Does it happen here?” or “Do I take part as my real self?”

The volunteer is prodded to let his imagination take over, and not try to be too analytic, because otherwise he would never be able to guess something as fanciful as a dream. In the beginning, his questions will be very vague and general, but as he gets more answers, his picture of the dream sequence will get more concrete, and so the questions become more specific. For example, “Does the car arrive after the lady has left?” If he cannot think of any more questions to ask, the game is over, and the volunteer is asked to tell the dream as he has reconstructed it from the answers he got. Depending on the inventiveness of the volunteer, the resulting story can be very involved and detailed indeed. However, when he then asks whether his reconstruction of the original story is accurate, the group tells him that there never was a story!

The only thing the others discussed was a rule for answering his questions. The basic rule is very simple: every question that ends with a certain letter, say “s”, is answered with “yes”; all others are answered with “no”. The only exception to the rule is that answers should be consistent. Thus, if the volunteer asks the same question with different words, the answers should remain the same, even if the last letters of the questions are different. Since all groups members give the same answers (though they sometimes may get slightly confused because they heard or understood the question differently), and since the different answers confirm each other, the victim gets the impression that they effectively have agreed on a coherent story. Each further answer gives him one more bit of information, which he uses to further develop the story he has in mind. Yet, the end result, however coherent, detailed and involved, is purely a creation of his imagination, with no counterpart in any reality except his own mind.

Although this example is rather extreme, it helps us to understand the process of mental construction of ideas, beliefs and theories. Observation of the outside world does provide us with information, in the form of “bits”, which helps us to decide whether we should accept or reject a certain hypothesis. But we must generate these hypotheses ourselves, starting from our pre-existing theories about how the world around us is organized.

Social constructivism

In addition to coherence, there is another criterion for accepting a belief as “true”: **consensus**. The more people confirm an observation, interpretation or belief, the more we take it seriously. For example, one person seeing a UFO is much less credible than twenty people seeing it. However, this criterion is problematic when people influence each other. In that case, they can all come to believe the same, even though only one (or none) actually “perceived” it, as we will discuss later in the section on collective cognition.

In practice, most of our beliefs/knowledge come from others. These include religion, language, culture, worldview, morals, and common sense. In most cases, it is difficult to determine what is the precise origin of these beliefs. Such a belief has been transmitted from person to person, each time accumulating some small changes. Through these changes, the belief adapts so as to be maximally coherent with other beliefs people already have. In the end, a complex system of beliefs may be constructed by the group, without any individual having control over it. Since there is no objective, outside criterion to which beliefs must adapt, separate groups will typically construct different and independent belief systems. For example, Western astrology is completely different in its assumptions and predictions from Chinese astrology.

These shared beliefs determine what we assume to be true or real: this is called the **social construction of reality**. Since perceptions are affected by ideas and beliefs, the way we perceive reality will depend on our cultural and social background. For example, when Galileo tried to convince the church authorities of his observation that the planet Jupiter has moons, they simply refused to look in his telescope, since they anyway “knew” there were no moons, and had no trust in this bizarre instrument. As another example, before Harvey's theory that the heart works like a pump, nobody seems to have noticed that the heart beats (or at least no mention of it is made in the literature). It seems that we first needed a theory of why the heart should beat before we could truly observe that it beats!

The conclusion is that different cultures or groups live in different “realities”, and that we have no objective criterion to say which one is right and which one is wrong. The social construction of reality is a basic thesis in “postmodern” social science. It is used to argue that Western science or philosophy is in no way superior to the one of other, supposedly more “primitive” cultures. However, postmodernism has recently come under more and more criticism, and seems no longer to be so fashionable as it used to be a decade ago.

Shortcomings

- Constructivism—and particularly postmodernism or social constructivism—tends to overemphasize the relativity of knowledge. There *is* an objective difference between astrology and astronomy: the one makes reliable predictions, the other does not.
- Constructivism does not sufficiently distinguish between external and internal sources of information. In practice, most of the time we *can* distinguish between perception and imagination.
- Constructivism lacks formal models of how knowledge is organized. It also does not explain precisely how knowledge is constructed.

Situated and Embodied Cognition

The Brain in a Vat problem

Most of traditional cognitive science (and part of constructivism) tends to see the mind as separated from the outside world. At best, it is merely passively observing what happens outside. For example, AI reduces cognition to the logical manipulation of abstract symbols. Even connectionism tends to reduce the influence of the outside world to a simple set of inputs that the cognitive system itself has no control over. This attitude is a remainder of **dualism**: the mind is seen as something intrinsically different and separate from matter. Even when modern science admits that the mind cannot exist independently of the matter in the brain, the assumption is that the brain alone is sufficient to produce intelligent behavior.

I'm walking outside in the sun!!

This leads us to envisage a theoretical disembodied intelligence as a "brain in a vat": a brain artificially kept alive by providing it with the right nutrients, and perhaps some electrodes that stimulate it, but with no body attached to it. Critics of traditional cognitive science say that it tries to build the equivalent of a brain in a vat. However, they claim that such a brain would not be able to exhibit intelligence, because intelligence evolved specifically in order to allow our bodies to interact more efficiently with the world, and because all the knowledge we have is derived from that interaction. The critiques on disembodied cognition have been formulated in different, but related ways, which we will now briefly review.

Embodiment

The main argument of the proponents of **embodied cognition** is that cognitive systems need to have the equivalent of a body through which they can interact with their environment [Ziemke, 1991; Clark, 1999]. This includes at least:

- **sensors** through which information about a real environment (not controlled by the programmer or experimenter) can enter the system. Examples are eyes and ears. These are also called *afferent* channels, as they bring data into the system.
- **effectors** or actuators (muscles, motor system) through which the cognitive system can act upon the outside world. Examples are hands and vocal chords. These are also called *efferent* channels, since they bring data out of the system.

- a **feedback** loop connecting the two, so that the effects of actions can be perceived, and the system can establish a connection between the two. For example, a child trying to speak must hear its own voice in order to learn to articulate. That is why deaf children who have not received specific training in general are unable to speak.

Additional body elements that may be important include the glands that produce hormones, which affect emotion and mental functioning, and internal organs that sense the internal state (e.g. hunger is triggered by the emptiness of the stomach and a low glucose level in the blood). Since the body is an imperfect system, subject to various physical constraints (e.g. weakness, inaccurate perception, inertia, cold, hunger, ...) the mind cannot afford to be busy only with purely rational, abstract reasoning: it must constantly adapt to the practical situation and correct for unforeseen problems.

Situatedness

The closely related philosophy of **situated cognition** focuses on the fact that in order for us to understand its functioning a cognitive system must be situated in a realistically complex environment, not in an idealized world of logic and abstraction, a psychological laboratory, or a simulated “toy world”. The real world is indeed much more variable, ambiguous and unpredictable than the simplified “environments” that have been used in cognitive science. Yet, a truly intelligent system by definition is able to cope with these complexities. Moreover, part of that intelligence directly derives from the environment by way of learning from it, interacting with it, or being directed by it.

One of the biggest problems encountered by symbolic cognitive science is how to represent the world in a sufficiently detailed way (the knowledge acquisition and representation problem), and how to make inferences in such a complex model (the frame problem). Yet, there is no need to make detailed representations when you can directly access the world itself. This insight can be summarized by the slogan: *The world is its own best model* [Brooks, 1991].

For example, why calculate the precise force and the angle with which you would have to push a block in order to get it out of the way (as if often done in “blocks worlds” models of robotic action), if you can just try it out and adjust your movement according to the feedback you get through your senses? Why plan ahead all the different steps of a problem-solution, if you can just perform some actions and see whether they bring you closer to the solution?

Another implication of situatedness concerns the way we learn. The concepts we learn always appear in a concrete **context**: a specific situation to which the concept applies. It is impossible to fully specify this context by means of a formal definition stating the necessary and sufficient conditions for a phenomenon to belong to that category. For example, the concept “he is an animal” cannot be explained by “an animal is a living organism that can autonomously move”. Yet, the expression is easily grasped when used in the right context (e.g. after a night of wild sex). Practical experience shows that trying to teach concepts via dictionary definitions is very inefficient. Students make mistakes, such as “my parents and I correlate well, since most of the time we are present together”. Teaching by using the concepts in appropriate contexts leads to much quicker understanding and remembering.

Conceptual metaphors

If the foundation of cognition is concrete interaction with the world, then how can we learn to reason abstractly? According to the cognitive linguist George Lakoff (in collaboration with Mark Johnson [1991]), we do this by using **conceptual metaphors**: analogies between the abstract concept, and a more concrete, “embodied” concept that we intuitively understand because of our experience in interacting with the world. This is best illustrated with some examples:

Desire Is Hunger: the abstract concept of “desire” is made more intuitive by comparing it to the bodily sensation of hunger. This metaphor is found in various English expressions:

- He is sex-starved
- She thirsts for recognition
- Sexual appetite
- He hungers for her touch.

Existence Is A Location (here): the very abstract notion that something exists is “embodied” by comparing it to the notion that something is present in the vicinity. Again, Lakoff has collected plenty of English expressions that illustrate this usage:

- It came into existence
- The baby is due any day
- The baby is a new arrival.
- No answers were forth-coming.

Some other conceptual metaphors are: Life is a Journey, Change Is Motion, or Change Is Replacement.

In further work, Lakoff [Lakoff & Nunez, 2001] has applied his theory even to mathematical concepts. For example, the mathematical notions of element, vector space, operator, transformation and set (collection) all refer to physical objects, the space in which they move, or the way we manipulate them. Similarly, the basic concepts of logic can be understood metaphorically: proposition = expression = spoken sentence, deduction = demonstration = pointing out. The latter can be illustrated through an expression such as “let me walk you through the different steps of the proof”.

Enaction

Enactive cognitive science, as proposed by Varela, Thompson & Rosch [1991], is again a closely related conception of cognition: thought or knowledge only becomes meaningful when it is implemented, “acted out”, or *enacted* via interaction with the environment. The cognitive system is "structurally coupled" to the environment. This means that changes in the environment systematically affect the system, and vice-versa. Therefore, mind and world, or system and environment, cannot really be separated. Out of this interaction, reality is constructed. This perspective can be situated in between objectivism (knowledge as reflection of outside world) and subjectivism (knowledge as internal construction).

As suggested by the name, this approach emphasizes the role of action. Even perception can be seen as perceptually guided action. A classic illustration is found in saccades (quick eye movements): the eye is constantly moving its gaze so as to explore the most interesting parts of the scene. Out of the relationships between the different aspects thus sensed, the brain constructs a complex picture of the surroundings. More generally, cognitive structures emerge from recurrent sensori-motor patterns, i.e. the correlations between perceptions and actions, which enable action to be perceptually guided.

Implementing Situated Cognition

Autonomous robots

According to the situated and embodied perspective, symbol-based computer simulation is not a good way to understand cognition, since it lacks a realistic mind-environment interaction. Instead, AI researchers should strive to build autonomous *robots* that can independently interact with the real world through sensors and effectors.

Such robots should not have a detailed program of what to do, but only a broad goal. A typical example of such a robot will have as its goal to find an electric contact to plug into when its battery runs low, and explore the room otherwise. Rather than planning ahead, the robot should solve any problems as they arise. For example, it should find a way around when there is an obstacle in front of the contact. Such robots should start with very simple, yet vital, behaviors, e.g. move around without bumping into walls and furniture (this is called "wall-following behavior"). The underlying philosophy is that we should first learn to simulate the behavior of insects, not the logical reasoning of humans! For example, researchers have built robots with six legs that learn to walk around efficiently the way a cockroach would walk.

Rodney Brooks [1991], one of the founders of this approach, has developed a so-called *subsumption architecture* to control autonomous robots. The principle is that goal-seeking behavior is implemented in a hierarchy of levels. At the lowest level, there is direct sensory-motor interaction. For example, when the robot detects an obstacle (perception via the senses), it immediately stops moving (control of movement). This results in quick reactions and the avoidance of basic problems. The higher levels control the lower levels. For example, after the movement has stopped, the higher control may direct the robot to resume movement in the general direction of the electrical contact, but with a deviation so as to avoid the obstacle. Such subsumption architecture allows more complex planning without getting stuck in immediate but unexpected problems, as would happen with a robot using traditional AI programming.

Software agents

In addition to building robots (hardware agents), the embodied approach to AI also uses **autonomous software agents**. Since robots are very dependent on (expensive and unreliable) hardware, it may be more fruitful to simulate their general behavior in software. However, such simulations should be more realistic than the older "toy worlds". This requires the following steps. First, define a virtual environment obeying (realistic) physical constraints. Then, define agents equipped with sensors to perceive (imperfectly) that environment, effectors (typically ability to move) to act upon that environment, an autonomous goal (e.g. avoid dangers and gather enough virtual food) to steer their actions, and feedback from sensors to effectors. If you add the possibility for unsuccessful agents (that fail to reach their goals) of "dying", and for successful ones of "reproducing" with variations, then you have the basic paradigm for Artificial Life, a domain originally inspired by the computer simulations of AI but otherwise quite different in its philosophy.

In such a virtual environment, researchers can test out different cognitive architectures for the agent to see which ones work best. A nice illustration of such a simulation was performed by my former PhD student Carlos Gershenson [2004]. He built a "virtual laboratory" that enabled him to compare the behavior of different agents in the same virtual environment. These agents were programmed according to different cognitive paradigms, including connectionist, symbolic, rule-based, and Braitenberg vehicles (see later). Somewhat surprisingly, the different agents were about equally successful overall, although some were better in certain aspects…

Dynamical systems

Dynamical systems, as reviewed by Port & Van Gelder [1995], is a new modeling paradigm for cognitive processes that goes beyond the reduction of cognition to symbol manipulations. This perspective notes that the environment, the cognitive system, and their interaction each can be seen as physical processes in time. Mathematically, such processes are modeled as dynamical systems.

A *dynamical system* is determined by a set of quantitative variables s_a, s_b, s_c, …, together determining a state $s = (s_a, s_b, s_c, \ldots)$, that change simultaneously and interdependently over time. The state of the system follows a trajectory $s(t)$ through the state space under the influence of some abstract "force" or dynamics (which is typically described by a differential equation). Such processes can come to equilibrium in an attractor state or region, or periodically cycle around it, or even exhibit chaotic behavior.

The basic implication is that we should take (continuous) time into account while modelling cognitive processes. This approach is in principle more general than older models of cognition:

- logical deduction takes place outside of time: all axioms and theorems are supposed to be simultaneously true.
- computation, which is the basis of AI, takes place in discrete time steps, one for each basic operation.
- spreading activation in a feedforward net takes place in only a few time steps (one step to go from one layer to the next). When the output layer is reached, the process stops.

It must be noted, though, that the reason older approaches approximate time by discrete steps is simply because continuous models, such as those used in differential equations, are much more difficult to deal with, and therefore the dynamical systems approach can as yet only handle extremely simple processes with just a few variables.

Distributed cognition and the extended mind

The **extended mind** hypothesis, proposed by the philosopher Andy Clark [1997; Clark & Chalmers, 1998], notes that the mind not only interacts with the environment: it actively *relies* on the environment to perform certain cognitive activities. The mind thus extends outside of the brain and into the physical environment. This philosophical position is sometimes called "active externalism": cognitive elements (symbols, concepts) are not only defined by their relation to the outside phenomena they represent (a position known in philosophy as externalism), they also take part in this outside world.

There are good practical reasons why the mind should use the environment. We have already noted that the mind is not so good at accurately storing or mechanically manipulating information, because of the "magical number" restriction on working memory. Material systems in the environment can be more reliable for this. This can be illustrated by a variety of examples where people use material tools to facilitate cognitive tasks:

- writing in a notebook to remember agenda items and telephone numbers
- drawing diagrams on a piece of paper to better understand logical relationships
- counting sheep by adding pebbles to a bag
- putting up signs in the neighborhood to help people find locations (e.g. the bathroom, a city)
- using fingers, an abacus or a computer to make calculations
- using a magnifying glass to perceive tiny objects

The implication is that we cannot separate mind and world, since the outside world itself performs cognitive functions for us, such as memorizing, perceiving, organizing and processing information. In fact, a number of complex cognitive processes such as computation or mathematical reasoning may actually be impossible without support from external tools, such as pen and paper. It seems extremely unlikely that advanced scientific theories, such as quantum mechanics or molecular biology, could ever have been developed without these cognitive tools. Moreover, in our everyday life we are constantly relying on written notes, signs, maps, etc. In that sense, our minds truly extend into the environment.

Cognition can extend not only across material objects, but across social organizations. This is the perspective of **distributed cognition**, as investigated by the ethnographer Hutchins [1995]. A complex organization, such as a Navy ship, consists of many individuals that communicate with each other using various channels, such internal phones, signs, or shouting. Moreover, they collect and process information using tools such as compasses, notebooks, and nautical instruments. People and tools together solve complex problems, such as controlling the ship's course while maneuvering into a port. The information about the ship's precise position, speed, direction, as well as information about the port layout, position of other ships, wind direction, etc. is collected and processed by all individuals and instruments together, working as a single cognitive system that is distributed over many interacting components.

A similar perspective inspires distributed artificial intelligence. Here, the cognitive system is formed by a group of software agents or robots, each specialized in a particular aspect of the problem. By communicating their results and thus collaborating, they may together solve problems that are too complex for any agent individually.

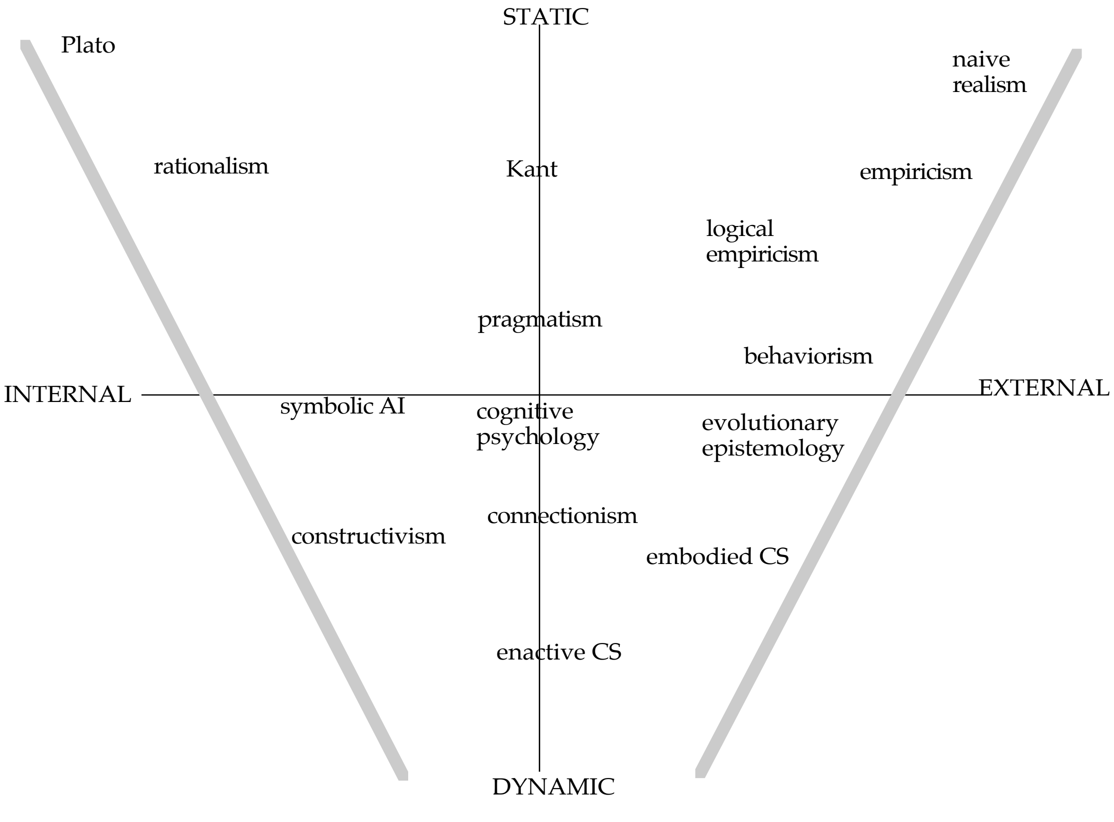
The Systems View of Cognition

Summary of previous developments

When we compare the different "schools" or paradigms in cognitive science, we can situate them on two basic dimensions:

(1) **Internal vs. external**: More rationalist approaches (e.g. Plato, constructivism, symbolic AI) emphasize that knowledge is primarily generated *internally*, by the mind itself, independently of the environment. More empiricist approaches (e.g. logical empiricism, behaviorism, situated and embodied CS) insist that knowledge originates *externally*: it is produced by information coming via perception from the environment. In between, we find approaches that accord similar importance to internal and external influences (e.g. enactive CS, connectionism, cognitive psychology).

(2) **Static vs. dynamic**: Historically, the first approaches tended to see cognition as apprehension of absolute, objective truths. Later, the emphasis is on knowledge as something that is constantly under construction and evolving, without ever approaching a time-independent "truth".



In the longer term, the trend is towards increasingly dynamic views that emphasize relations between inside and outside. These are what we might call "interactivist" approaches: enactive cognitive science, connectionism. This is clearly seen in the picture above where the more recent approaches tend to be situated more towards the bottom and center of the scheme. In addition there is a trend to pay more attention to the internal relations between the cognitive units, rather than to the units on their own, e.g. connectionism, semantic networks. All in all, models of the mind have become increasingly *complex* and *dynamic*, shifting from a reduction to independent parts (the soul, ideas, observations, symbols, …) to a focus on the interactions (connections, feedbacks, collaboration, …) that assemble the parts into an organized whole.

The need for integration

Different paradigms are good in modeling different aspects of cognition, for example AI for modeling deduction, connectionism for learning, and embodied CS for autonomous behavior. Yet, they each have their own shortcomings. Unfortunately, there is no obvious way to combine the benefits in a single theory, since the paradigms tend to be opposed in their philosophies and basic assumptions, e.g. symbolic vs. subsymbolic, or rationalist vs. embodied.

In other sciences, the most successful attempt at integration came from *general systems theory,* which is more recently being revived under the label of *complex adaptive systems* or *complex systems science*. Such systems thinking moreover has the benefit of focusing intrinsically on interactions. A systems approach to the problems of mind can be found in **cybernetics** [Wiener, 1948; Ashby, 1964; Heylighen & Joslyn, 2001]. This is a paradigm slightly older than cognitive science that emerged in the 1940s. Cybernetics had great ambitions for understanding intelligent behavior. It moreover had some immediate successes in both theory (e.g. W. R. Ashby's [1952, republished 1960] book "Design for a brain", or G. Bateson's collection of essays "Steps towards an ecology of mind" [republished 2000]), and practice (e.g. the neural networks of McCulloch and Pitts [1947] and the autonomous "tortoise" robots built by W. G. Walter [1950, 1951]).

However, this approach was eclipsed by the popularity of symbolic AI in the 1960s and 1970s. The reason was that the designs of cybernetics were analog, and therefore clumsy and difficult to build, whereas the AI approach was digital, and therefore easy to implement with the new computer technology. Moreover, cybernetics focused on apparently low-level functions, such as movement and adaptation, while AI immediately tried to tackle high-level cognitive tasks, such as proving theorems or playing chess. Yet, most of the basic concepts of cybernetics, such as feedback, autonomy and self-organization, have now come back as part of the new approaches to cognition, such as constructivism, connectionism, situated and embodied CS and dynamical systems. In hindsight, it seems that cybernetics had gotten more things right than AI. Still, it needs to be updated with more recent ideas that have proven their utility. That is what we will try to do in the remainder of this text.

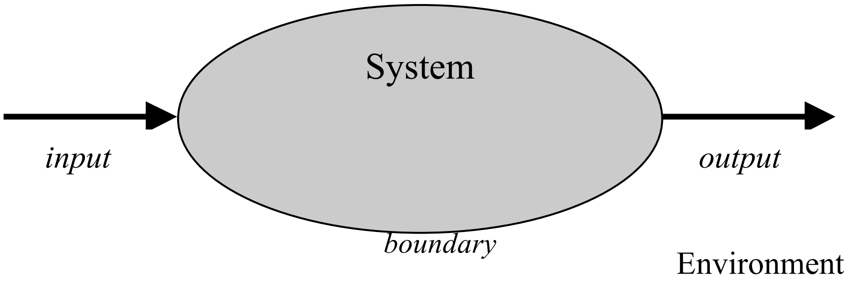
Basic concepts of systems theory

The systems philosophy can be summarized by the well-known saying that “the whole is more than the sum of the parts”. A whole possesses **emergent** properties, i.e. properties that are not properties of its parts. For example, an organism has the property of being alive; the same cannot be said of the atoms and molecules that constitute it. A song has the properties of melody, harmony and rhythm, unlike the notes out of which it is composed. A **system** can be defined as a number of parts *connected* by relations or interactions. The connections are what turn a collection of parts into a coherent whole.

Systems theory thus makes abstraction of the substrate or material out of which a system is made: what counts are the relations between the elements, not the elements themselves. For example, transistors in a computer may play the same role as the neurons in the brain, as long as their pattern of interconnections is similar enough. In that way, systems theory transcends the matter-mind dichotomy that gave rise to dualism. Neither matter nor mind are a priori categories; the essence is *organization*, i.e. the pattern of connections and the information that is passed on along them, which give the system its coherence.

The most basic distinction in systems theory is the one that separates a system from its **environment**. The environment is defined as everything that is not part of the system, but that directly or indirectly interacts with the system. This distinction is represented explicitly by the system **boundary**, which designates the separation between the inside and the outside of the system. It is important to note that the distinction system-environment is always to some degree subjective: we can draw the boundary differently, including or excluding different phenomena, depending on our goals or focus of interest.

<u>Example:</u> clothes, hair, glasses, symbiotic bacteria... may or may not be included in the system “person”. Alumni, part-time students, visitors, externally employed cleaners may or may not be included in the system “university”.

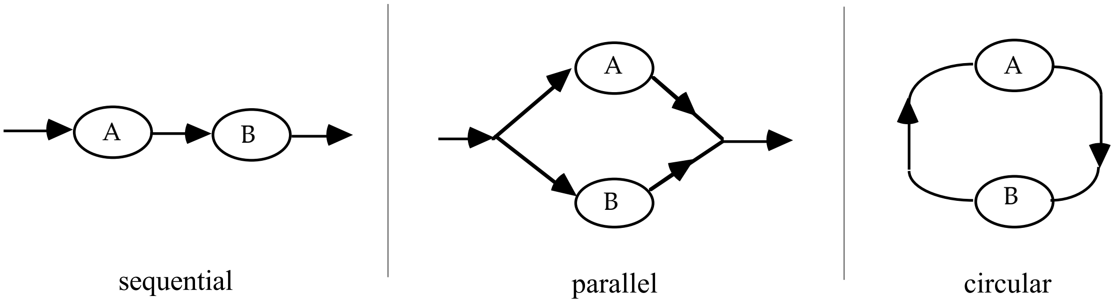


System and environment by definition interact: they exchange matter, energy and/or information, and thus mutually affect each other. This leads us to define two other basic systems concepts:

- **input**: what enters the system from the environment
- **output**: what leaves the system to end up in the environment

Two systems A and B are **coupled** if they share some of their inputs and/or outputs. There are three basic types of system couplings:

- **sequential** or serial: output of A = input of B
- **parallel**: input of A = input of B, output of A = output of B
- **circular**: input of A = output of B, output of A = input of B



sequential parallel circular

A number of coupled systems together can form a new, larger system: a **supersystem**. The original systems are now the **subsystems** of the new, encompassing system. Generally, each system contains a number of subsystems, and is contained in one or more supersystems. Subsystems and supersystems form a hierarchy.

Example: atoms → molecules → organelles → cells → organs → organisms → societies → Earth → solar system → galaxy ...

Control systems

Living systems

Organisms require a constant input of matter and energy (food, resources) in order to maintain their state of being alive. This input is necessary to provide energy for metabolism and movement, to rebuild worn tissues, such as hairs, skin and red blood cells, to grow, and to reproduce the system. Living systems are intrinsically goal-directed: they try to maintain and (re)produce themselves, in spite of perturbations from the environment. Their implicit goal or value is **fitness**: survival, growth and reproduction. This goal has been built into them by evolution via natural selection: organisms that were not good at achieving fitness have lost the competition with those that were better, and have thus been eliminated. This is the evolutionary principle of the "survival of the fittest".

Agents

Definition: an **agent** is a goal-directed system that tries to achieve its goals by acting on its environment.

Agents are typically organisms, such as animals or people. However, they can also be artificial systems, such as robots or software agents, with preprogrammed goals. They can even be organizations or other social systems, such as a firm, a football team, or a country, that consist of coordinated individual agents with a shared set of goals (e.g. making profit for a firm, winning for a football team).

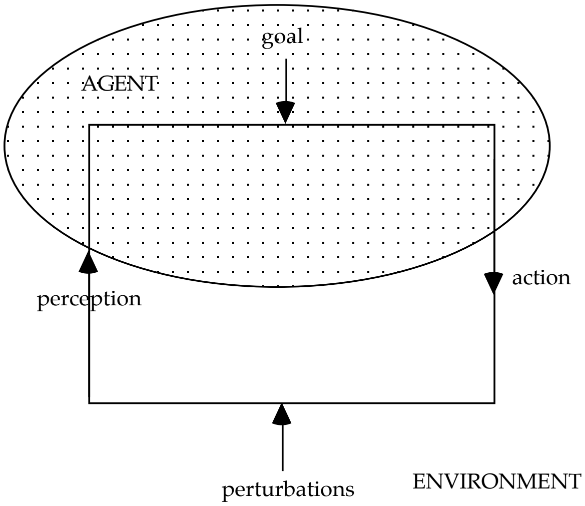
Control

Cybernetics has shown how goal-directedness is achieved via control [Ashby, 1964; Heylighen & Joslyn, 2001].

Definition**: control** is the successful reduction of deviations from the goal by appropriate counteractions.

Example: hunger = deviation from state of sufficient energy; counteraction = find and eat food. The agent is in control if it manages to eat sufficient food not to stay hungry.

A deviation thus triggers an action, which produces a reduced deviation, which in turn produces a further action, and so on, until all deviations have been eliminated. Thus, control is characterized by a negative feedback loop. A **feedback** loop is a circular coupling of a system with itself via the environment. The feedback is called **negative** when it reduces deviations, **positive** when it increases deviations.



The feedback loop characterizing a control system or agent has the following basic components (see picture):

- **perception**: information enters the agent from the environment, representing the situation as experienced by the agent
- **goal**: internal representation of the ideal or preferred situation for the agent
- **action**: the agent affects the environment in order to bring the perception closer to the goal
- **diversions/perturbations**: changes in the environment that affect the situation independently of the agent (i.e. that are not under control of the agent), making it deviate from its present course of action.

In cybernetics, such unexpected changes are traditionally called **disturbances** or **perturbations**. However, their effect can be negative, like in the case of the appearance of a storm or a predator, but also positive, like in the case of the appearance of food or a tool. Instead of disturbances, such unexpected opportunities are better called **affordances**, i.e. they afford the agent an additional possibility to achieve its goals. For example, the presence of an apple affords you the opportunity to eat. Disturbances and affordances together may be called "**diversions**", since they redirect or divert the agent's course of action, either in a positive or in a negative way.

Let us delve a little deeper into the different processes that make up the control loop.

Cognitive aspects of control

For an agent, a **problem** can be defined most generally as a difference (deviation) between the experienced situation (perception) and the desired situation (goal). If there would be no difference, the agent would be perfectly satisfied and would have no reason to act. A problem in this sense is not necessarily negative or unpleasant: it is sufficient that the agent can conceive of some way to improve its situation and is motivated to seek such improvement. For example, if I feel like drawing, then my "problem" is defined as the difference between an empty page and a page with an esthetically pleasing sketch on it. A problem should also not be seen as something purely intellectual: if the cup I am holding tilts a little bit too much to the left, so that coffee may leak out, this defines a problem that I must resolve by restoring the balance.

The agent's task is to solve the problem, i.e. to select and perform one or more actions that together eliminate (or minimize) that difference. Achieving this requires the cognitive functions below.

- **perception:** the agent needs to sense as precisely as possible what deviations there may exist, and in how far previous actions have affected these.
- **representation:** perception produces an internal representation of the outside situation, such as a pattern of activation across neurons. Note that this representation is not an objective *reflection* of external reality as it is, but a subjective *experience* of how the agent's personal situation may deviate from the preferred situation. There is also no reason to assume that a representation consists of discrete units (symbols) that represent separate objects or aspects of

the situation. Fundamentally, the agent's cognition does not represent objective phenomena, but subjective sensations that depend on the agent's goals.

- **information processing:** the agent needs to process or interpret the information in the representation, and in particular determine precisely in what way it differs (or may start to differ) from the goals, and which actions could be used to reduce those differences. This requires some process of inference.

- **decision making:** the agent needs to select an action to perform. In general, only one action can be performed at a time; if several actions seem appropriate, the best one needs to be determined.

- **knowledge:** to make adequate decisions, the agent has to know *which action* is most appropriate to reduce *which deviation*. Otherwise the agent would have to try out an action at random, with little chance of success, and thus a high chance of losing the competition with more knowledgeable agents.

- **intelligence:** if the problem is complex—so that solving it requires more than one interpretation and/or action—the agent may need to look ahead at likely future situations by making inferences, exploring their consequences, and developing a plan to deal with them, i.e. by designing a sequence of well-chosen, coordinated actions that as much as possible take into account the intricacies of the situation.

Note that these distinct functions do not necessarily correspond to distinct components in the cognitive system: the same component (e.g. a nerve connecting a sensor to an effector) may perform more than one function (e.g. perception, representation, knowledge, etc.).

Physical aspects of control

Let us follow through the control process outside the agent, noting how the environment too participates in solving (or worsening) the problem.

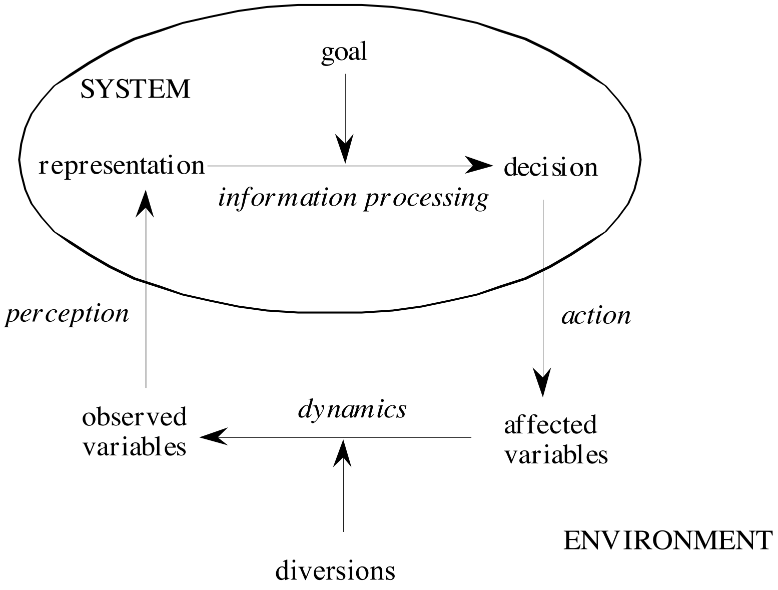
- **action:** the agent should be able to perform a sufficiently broad repertoire of actions to affect the environment in the needed way: the more variety there is in the diversions, the more variety there must be in the actions to deal with them (this is Ashby's [1959, 1964] "Law of Requisite Variety"); this requires sufficiently powerful and flexible effectors.

- **affected variables:** only certain aspects of the environment are affected by the agent's actions: for example, the agent cannot change the weather, but may be able to find or make a shelter against the rain.

- **dynamics**: changes in the environment, whether produced by the agent or by diversions (i.e. all events not produced by the agent), generally lead to further changes, according to the causal laws or dynamics governing the environment. For example, a stone pushed over a cliff by the agent will fall down to the bottom, where it may break into pieces. This dynamics may help or hinder the agent in achieving its goals. It may even perform some of the required information processing, like when the agent adds stones together to perform calculations (*calculus* = Latin for "small stone")

- **observed variables**: the agent cannot sense all changes in the environment, whether caused by its own actions, diversions or dynamics; the variables it can perceive should ideally give as much information as possible relevant for reaching the goal; irrelevant variables are better ignored since they merely burden the cognitive system

When we put the different internal and external components of the control process together, we end up with the following more detailed scheme [Heylighen & Joslyn, 2001]:



This scheme is a feedback cycle with *two* inputs: the goal, which stands for the system's preferences, and the diversions, which stand for all the processes in the environment that the system does not have under control but that can affect these variables. The system starts by observing or sensing the variables that it wishes to control because they affect its preferred state. This step of *perception* creates an internal representation of the outside situation. The information in this representation then must be processed in order to determine: 1) in what way it may affect the goal; and 2) what is the best reaction to safeguard that goal.

Based on this interpretation, the system then decides on an appropriate action. This action affects some part of the environment, which in turn affects other parts of the environment through the dynamics of that environment. These dynamics are influenced by the set of unknown variables that we have called the diversions. This dynamical interaction affects among others the variables that the system keeps under observation. The change in these variables is again perceived by the system, and this again triggers interpretation, decision and action, thus closing the control loop.

Uncertainty

In general, the agent cannot be certain which action is appropriate. This is because the environment is infinitely complex: every phenomenon in the universe has potentially some influence on what can happen here and now. Moreover, every situation is unique: even seemingly identical situations can produce different outcomes, as illustrated by the phenomena of chaos and quantum indeterminacy. Although the agent may not be able to sense and distinguish all these unique situations, even with relatively simple sensors the number of possible combinations of sensed values is astronomical.

Example: consider a primitive eye with 100 pixels that can each sense only black or white (2 values). Such an eye can in principle still distinguish 2^{100} possible combinations of the values for these pixels in a single view. Consider now a sequence of 10 views. This produces $(2^{100})^{10} = 2^{1000}$ which is about 10^{300} possible combinations (a 1 followed by 300 zeros!)

We may conclude that the agent cannot have perfect knowledge of what to do for each possible situation. This implies that some of the actions the agent performs will not be optimal, or not even adequate. However, this is not grave because errors can generally be corrected by subsequent actions. The only real requirement is that actions must be more likely to improve than to worsen the situation. If that condition is met, a long enough sequence of actions will eventually bring the agent close to its goal.

Given that the agent cannot act with certainty, it will need to make “informed guesses” about what action to take. This means that it should maximally reduce the uncertainty, taking into account any information it can get. To do that, the agent must have good heuristics to simplify the situation: learning to ignore the less relevant aspects; learning to distinguish the most relevant aspects; selecting the actions most likely to have positive effects. Moreover, the agent needs to be ready to immediately correct the undesirable effects of any action. This requires sensitivity to small changes, ability for quick action, and immediate sensory-motor feedback.

Cognition as vicarious selection

The cybernetician Ashby [1960, 1964] noted that *appropriate selection is the essence of intelligence*. Indeed, we have defined intelligence as the ability to solve problems, and we have analyzed problem solving as selecting the right sequence of actions that lead from the initial state to the goal state. Anything that helps us to make the right selection—such as heuristics or simply additional information about the problem situation—will therefore increase our power to solve problems, and thus amplify our intelligence. In essence, a cognitive system is therefore something that helps us to make the right selection from a variety of possibilities, i.e. the right decision. The larger the variety (i.e. uncertainty), the more difficult it becomes to find the solution by trial-and-error, and the more need there is for some cognitive mechanism to reduce complexity and support adequate selection.

This creates what evolutionary theorists call a strong “selective pressure” for cognition. Since we are all alive, we have adapted to that pressure. In other words, all the presently living agents have been selected to survive. On the other hand, agents with poor choice of actions have been eliminated by natural selection. The surviving agents must have evolved some mechanism(s) to choose adequate actions. Donald T. Campbell [1974], the founder of evolutionary epistemology, called these internal mechanisms “vicarious selectors”. They select appropriate actions from the myriad possibilities. In that way, they “stand in” for, substitute, or represent natural selection by the environment. (Note the meaning of the term “vicar”: delegate, representative, substitute).

If the vicarious selector had not eliminated the wrong actions, natural selection would have eliminated the agent itself. For example, we have an instinct for spitting out berries that taste bitter: the berries may be poisonous, and not spitting them out might have led to death. There also is an instinct for pulling back from a ravine or precipice: not pulling back may have resulted in a lethal fall. There exist many different types of vicarious selectors at different levels of complexity. These include “instinctive” knowledge inherent in the genes, perceptions, which “stand in” for the external situation, knowledge learned from experience, and information received via communication from others. From the point of view of evolutionary epistemology, *all* cognitive mechanism function essentially as vicarious selectors: they all help us to select the right actions, and do so in order for us to survive natural selection.

Reactive Agents

Condition-action rules

The simplest types of agents are called **reactive**. They directly or immediately (re)act on the phenomena they perceive. This means that they have no internal storage (memory) for information, and they do not "process" information in the sense of manipulating it internally. They also do not anticipate or reflect about what might happen. They simply associate perceived situations with the actions appropriate for that situation. Thus, they follow the stimulus → response scheme. They respond automatically, by "reflex", to a sensed condition or stimulus, and that in always the same manner.

Their knowledge is in the form of **condition-action rules**. These have the following form:

IF a certain *condition* (perceived phenomenon) is present

THEN perform a certain *action*, appropriate to that condition

For every condition (type of situation experienced) that the agent can distinguish, the rule specifies or selects the proper action. We will from now on represent such a rule using the following short notation:

condition → action

<u>Examples</u>: banana → eat, tired → sleep, tiger → run away, button → push

Such a reflex-like way of acting can be implemented by a direct connection transmitting activation from a sensor (condition perceiver) to an effector (action performer).

For the agent to be fit or "in control", its actions should be such that they bring the situation (generally) closer to the goal. This implies a negative feedback relation that reduces deviations from the goal.

<u>Example:</u> consider a simple sea-living creature that needs to stay in the right temperature zone to survive. Its required behavior can be implemented with 3 condition-action rules:

- too cold → go up
  (this could be implemented by a cold sensor that activates an effector for moving upward)
- too hot → go down
- just right → don't move

This last rule could possibly be replaced by: just right → go up or just right → go down. The small error that this produces will immediately be corrected by negative feedback, so that the agent will oscillate a little around the right temperature zone.

Evolution of reactive knowledge

Given the existence of a set of perceptions $P = \{p\}$ and actions $A = \{a\}$, knowledge k is in the mapping from the first set to the second set: $k: P \rightarrow A: p \rightarrow a = k(p)$

There exist |P| × |A| possible condition-actions rules (where |P| stands for the number of elements in the set P)

Example: for the sea-living creature, there are 3 × 3 = 9 possible rules that select actions. Some of these rules will endanger survival, e.g. too hot → go up, or too cold → don't move. If by mutation such a rule appears, the creature with this mutated rule will be eliminated by natural selection. Thus, bad rules are eliminated together with their carrier. Other rules improve survival, e.g. too hot → go down. These will be favored by natural selection: creatures that exhibit them survive better than those that do not. Some rules, while not being dangerous, have no immediate value, e.g. just right → go up. These will be eventually replaced by better rules, such as just right → don't move. As a result, the rules that are eventually retained will be true vicarious selectors: they will select actions for the creature just as if natural selection had selected them, but so that the creature no longer runs the risk of dying if it would choose the wrong action.

More generally, variation and selection will not only affect the connections between conditions and actions, but the conditions and actions themselves. Perceivable conditions vary when sensors change. For example, a light sensor that was initially sensitive to red light may mutate so that it becomes sensitive primarily to yellow light. Or a receptor in a cell for a particular type of molecule (e.g. hormonal signals or smells) may change shape so that it recognizes a different type of molecule. Possibilities for actions evolve when new or variant effectors appear, such as new muscles, or genes producing a different type of enzyme. These sensors and effectors will be selected if they offer better possibilities for interaction with the environment, i.e. if they make it easier for the agent to reach its goals because it can intervene more directly in the different diversions.

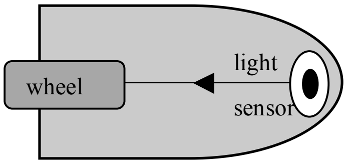
In this way, variation and selection of conditions, actions, and rules will eventually produce a well-adapted system of rules, or what we may call adequate knowledge. This implements Campbell's [1974] notion of knowledge as a system of vicarious selectors, in this case rules that select the right type of action. Note that the environment does not *instruct* the agent: it does not tell them which are the correct rules; it merely makes sure that incorrect rules are eliminated ("falsified" in Popper's terminology). The agent has to discover ("**construct**") the correct rules on its own. Thus, evolution of knowledge fits in with the philosophy of **constructivism**.

Braitenberg vehicles

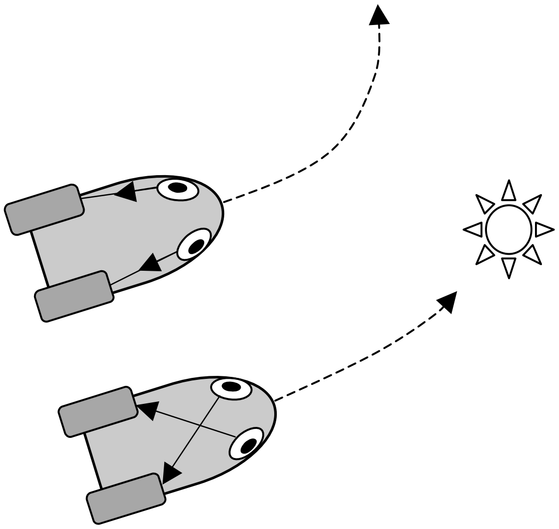
It is possible to design very simple, robot-like agents to illustrate the capabilities of condition-action rules. For this, we can find inspiration in the thought experiments proposed by the cyberneticist Valentino Braitenberg [1984] in his book “Vehicles”. The agents (vehicles) conceived by Braitenberg have primitive sensors that merely sense the amount of light. They use wheels, each driven by its own motor, as effectors. A sensor is directly connected to an effector, so that a sensed signal immediately produces a movement of the wheel. Depending on how sensors and wheels are connected, the vehicle exhibits different, goal-oriented behaviors. This means that it appears to strive to achieve certain situations and to avoid others, changing course when the situation changes. Let us illustrate this with the following examples, which are small variations on Braitenberg's simplest designs.

A first agent has one light-detecting sensor that directly stimulates its single wheel. This implements the following condition-action rules:

- more light → faster forward movement
- less light → slower forward movement
- darkness → standstill



This behavior can be interpreted as a creature that is afraid of the light and that moves fast to get away from it. Its goal is to find a dark spot to hide. This will work better if the terrain or the wheels are irregular, so that it doesn't move in a straight line, but somewhat randomly changes direction.



A slightly more complex agent (top in the picture) has two light detectors (left and right) each stimulating a wheel on the same side of the body. It obeys the following rule:

more light right → right wheel turns faster → turns towards the left, away from the light

This is more efficient as a behavior to escape from the light source, since the creature can move in different directions, and tends to orient towards the direction from which least light comes.

An alternative agent (bottom) is the same but with the connections from sensor to effector crossed (left sensor → right wheel). This implements the following behavior:

more light left → right wheel turns faster → turns towards the left, towards the light

This agent seems to first “attack” the light, and then run away from it.

In another variation, the connections are negative or inhibitory: more light → slower movement. In this case, the agents move away from the dark and towards the light.

In a complex environment with several—possibly moving—sources of light and shadow, Braitenberg vehicles will exhibit a complex and dynamic behavior. They will describe a zigzagging trajectory with accelerations and slowdowns, moving as fast away as possible from any strong light sources while apparently exploring the surroundings, until they find a deep pocket of shadow where they can "rest". Suddenly changing the conditions—as when a light is switched on in an otherwise dark room—will make them all move, like a bunch of cockroaches scurrying away to hide. This behavior is undoubtedly goal-directed, efficient and adaptive. We might even see it as intelligent, the way we attribute some minimal intelligence to a cockroach.

Yet, the functioning of the agent is purely mechanical, without any information processing or other processes that we normally associate with cognition. This illustrates the power of the cybernetic paradigm: it is sufficient to have causal connections (here interpreted as condition-action rules) that produce a negative feedback (here reducing any deviation from the state of darkness) to produce goal-directed behavior.

Such causal connections will spontaneously evolve through natural selection, on the sole condition that their implicit goal (i.e. the attractor of the dynamical system defined by these connections) corresponds to a state that is fit for the system. For example, if the agent (e.g. a worm or cockroach) is effectively more likely to survive in dark conditions (e.g. because it is less likely to be eaten by a predator), then variation and selection will automatically produce this type of light-evading rules. This again illustrates how from the systems perspective there is no fundamental separation between mind (intelligent, goal-directed behavior) and matter (sensors connected to effectors).

Example: bacterial movement

Let us discuss a related example, but this time of a real organism: the movement of the bacterium E. Coli. The bacterium senses the concentration of food and poison molecules in its immediate surroundings (condition) while it moves. (Note that it cannot sense the *direction* from which these molecules come, unlike a couple of light sensors that can determine in which direction the light source lies). If the concentration of food molecules increases and the concentration of poison decreases, it will keep moving in the same direction. If the concentration of food decreases or poison increases, however, it will randomly change direction, until it again finds a direction in which concentrations evolve in the positive sense. The result is that on average it moves toward food and away from poison, even though it has no idea in which direction these lie. What is experienced as "food" and what as "poison" is the result of natural selection eliminating all bacteria that moved towards a type of molecule that their metabolism was not able to handle.

Stigmergic coordination between rules

Stigmergy

Definition: an activity is **stigmergic** if the trace left in the environment by an action stimulates an agent to perform a subsequent action

The concept of **stigmergy** was introduced by the entomologist Grassé [1959] to explain the activity of social insects, such as termites, ants and wasps. This activity (e.g. building a nest) is apparently complex and coordinated, even though the individuals are very dumb. The word derives from the Greek “stigma” (mark, stimulus, sign) and “ergon” (work). The principle is that the activity performed by an agent produces a perceivable change in the environment. The perception of this mark or trace functions as a condition that stimulates the same (or a different) agent to perform the next step of the activity [Parunak, 2006]. In that way, the environment is used as a memory, to create continuity and coordination between independent condition-action rules. Grassé focused on the fact that this memory can be shared by a collective of agents (as we will discuss in the chapter on collective cognition), thus helping them to work collaboratively. Here, we will use the concept to explain the coordination of rules within a single agent’s mind.

Coordinating rules

At first sight, reactive agents behave purely mechanically or automatically:

- they react in always the same way to the same stimuli
- they have no memory of previous actions
- they cannot integrate perceptions from different sensors since each immediately produces an action

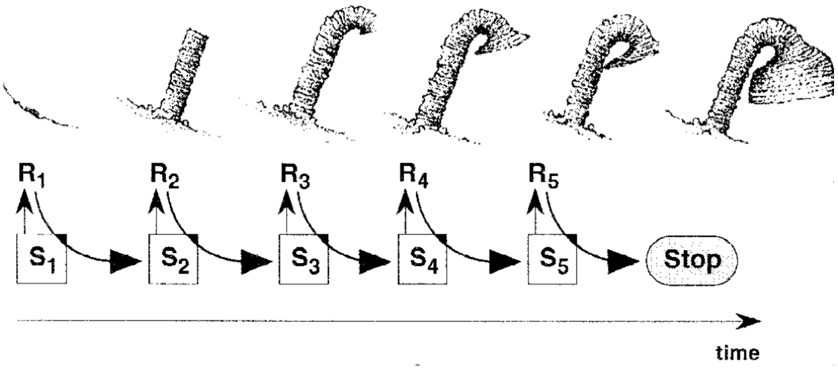
Yet, stigmergy provides them with a “working memory” that keeps track of what they have done. This memory is external and registered in their environmental situation. Each action changes this situation and thus the perceived conditions. As such, it can trigger new condition-action rules. Thus, at first sight independent rules can “collaborate” on a complex problem: when the one has done its duty, but only partially solved the problem, another one is ready to come into play. While the rules are implemented independently, their activities are causally connected via their effect on the situation. Since each tries to reduce a particular difference between the perceived and the desired situation, all together can tackle a problem with many dependent or independent differences. The one simply corrects the problem left unsolved (or created as a side effect or mistake) by the other. (This can be the seen as an application of the heuristic of Means-Ends Analysis.)

Example: a bacterium has to find food but avoid different types of poison: each rule (condition part) recognizes one type of food or poison, and tells the bacterium to change course whenever that condition is getting worse (and stay on the course otherwise). As a result, the bacterium will steer a complex path avoiding poisons but targeting the nearest and highest concentrations of food.

Two rules can be in conflict, e.g. when food and poison are present in the same place. In this case, the “food” rule says to move forward, the “poison” rule to move away. Suppose both stimulate the same effector. The effector will perform the average of the two actions: e.g. move away, but not so fast. The net effect is that behavior will be a compromise, or in the worst case inaction, as when rule 1 says “move forward”, but rule 2 says “move backward”. However, such inaction is unlikely to continue since any small difference in the strength of perception will produce net overall movement, however small. This change will in turn change the situation, and thus the relative strength of the two rules. This is likely to further reduce the balance between the rules.

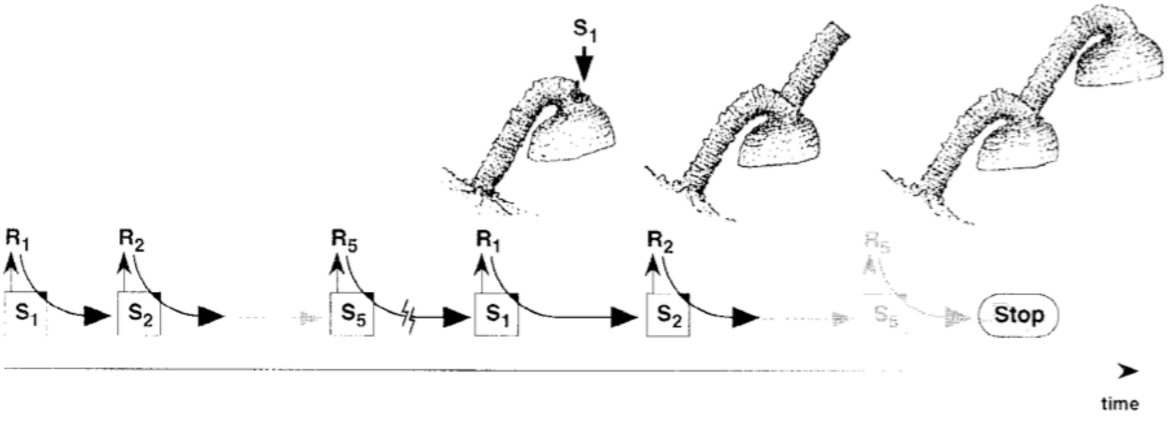
Complex activities

The effects of different actions are simultaneously or subsequently applied to the agent’s situation with respect to the environment. This results in a complex activity, taking into account many sensed variables and their interactions. While the behavior as a whole appears complex, it consists merely of a sequence of simple reactions to an environmental state that changes in a complex way. To adopt an example from Herbert Simon [1970], the path of an ant crawling over a beach may look very complicated, but it is controlled by very simple rules. Stigmergic coordination between rules similarly explains how the instinctive behavior of animals can produce very complex results, such as elaborate nests, spider webs, seduction rituals, hunting strategies, etc.

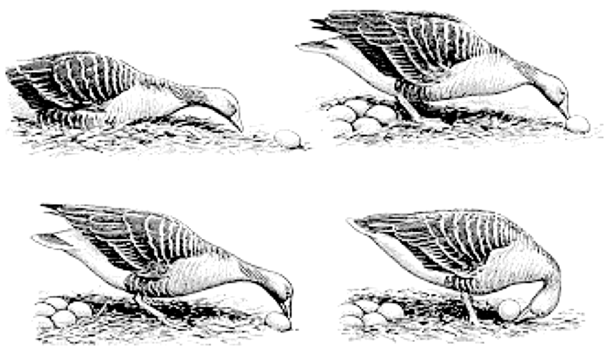


Example: wasp nest building: This mechanism can be illustrated by the wasp *Paralastor* sp. building its nest in the shape of a mud funnel [Théraulaz & Bonabeau, 1999]. The subsequent stages S_1, S_2, …S_5 (perceived conditions) of the partially finished nest trigger corresponding responses R_1, R_2, …R_5. The outcome of a building action R_i produces a new condition S_{i+1} that triggers the next action R_{i+1}, until the final stage (after R_5). The wasp does not need to have a plan for building such a nest, or to remember what it already did, since the present stage of the activity is directly visible in the work already realized.

However, the rigidity of the underlying rules becomes clear when the sequence is disturbed so that stages get mixed up. For example, the wasp's initial building activity is triggered by the stimulus S_1, a spherical hole. When at stage S_5 (almost complete funnel) such a hole is made on top of the funnel (indicated by an arrow), the wasp "forgets" that its work is nearly finished, and starts anew from the first stage, building a second funnel on top of the first one!



Example: stereotypic behavior of geese



The ethologist Konrad Lorenz, who studied the behavior of geese, observed a similarly "stupid" behavior. When an egg rolls out of its nest, the goose reacts with a stereotypic action in which it extends its neck, and pushes the egg back in with the lower part of its beak. Normally, this is perfectly adequate. However, when the experimenter suddenly takes away the egg while the goose is pushing, the goose simply continues its movement until its beak enters the nest. It seems to completely ignore the fact that there no longer is an egg to push! The explanation is that the condition-action rule specifies the initial condition (egg outside nest), and the appropriate action (push beak from egg back into nest), but without specifying any intermediate conditions (like egg halfway back in nest). If something goes wrong at an intermediate stage, no specific response is available. In normal circumstances (no experimenter making eggs disappear), an error, such as the egg rolling sideways, would simply restore the initial condition, and thus stimulate a new corrective action. With no egg visible, however, no new action is taken, and the initial action is simply completed—even though it has become completely useless.

While these examples from the animal world may seem to have little relevance for human cognition, we too often react in stereotypic ways. For example, when I drive my car to go to my office, I always take the same route, turning left when entering the city. However, I sometimes need to take another route to go to the city center. In that case, when I am not paying attention, it often happens that I turn left, as usual, instead of driving straight towards the center. The perception of the traffic situation at the entrance of the city automatically triggers my habitual reaction—turn left—, even when this is not appropriate. While this behavior is not instinctive, it has become pretty rigid because of continual reinforcement.

Reinforcement learning

In higher organisms, rules are not necessarily "hardwired" into the genes: they can *adapt* during the lifetime of the agent. Such adaptation or **learning** means that there is no need to develop new rules by the slow and painful process of variation and natural selection, which requires elimination of all the agents that obey unfit rules. The principle is simple: if a condition-action rule brings the agent closer to its goal, then the rule becomes stronger, i.e. its influence on the effector becomes stronger relative to other rules. We say that the rule is **reinforced** or "rewarded" for its positive contribution. This is the same mechanism that underlies operant conditioning as studied by Skinner [1938].

Example: *sensitization* is a stronger reaction to a particular type of stimulus that turns out to be very important. E.g. a rat that ate an unusual type of food and became ill afterwards will avoid that type of food later. Its initial distrust of the food is strengthened.

Alternatively, a rule that increases the distance to the goal is weakened. The rule is suppressed or "punished" for its negative contribution.

Example: *habituation* is a weakening of the reaction to a recurrent stimulus that turns out not to be so important. E.g. a snail that is repeatedly touched without negative effects will stop withdrawing into its shell.

This mechanism of **reinforcement learning** is a form of negative feedback control: average deviations from the goal are reduced by weakening the effect of unhelpful rules, and increasing the effect of helpful ones. Note that this control mechanism functions at a higher level: not at the level of conditions, like basic control mechanisms do, but at the level of condition-action connections. Thus, it can be seen as a meta-control system: controlling the control rules. This makes the cognitive system as a whole more flexible. Since all rules work together in determining the agent's course of action, learning will make behavior more efficient on average. There will be less "detours", "overshoots" or errors to be corrected later, and more direct steering towards the goal.

Learning is a more efficient mechanism than the trial-and-error of natural selection to improve rules: bad rules are not eliminated by killing the organism that holds them, but by adjusting or weakening their negative effect on the organism. However, the trial-and-error principle remains the same: when confronted with an unusual condition (e.g. a lever), the organism (e.g. a rat) tries out a potentially useful action (e.g. push the lever down). If this results in a punishment or error (e.g. the rat gets an electric shock), the connection between condition and action is suppressed, stimulating the animal to perform a different type of action (e.g. stay away from the lever). If the action results in a success (e.g. the rat receives food) the connection is reinforced, producing a potentially very useful new rule. The more often the rule is reinforced, the more deeply ingrained it becomes, and the more difficult it will be to “unlearn” it.

Anticipatory Agents

State-determined systems

Interiorization of stigmergic coordination

Although reactive agents are more flexible than might be expected because of the stigmergic interaction between their rules via the environment, they could become even more flexible by **interiorizing** this interaction. This means that rules would be triggered by other rules without need to perform actions in the environment. The advantages are obvious: no waste of time and energy, no need to correct actions that are misdirected, and therefore no risk of a potentially fatal error.

To keep the interaction wholly within the cognitive system, we need to introduce a new type of rules: **condition-condition rules**. Such rules, like condition-action rules, are triggered by the perception of their conditions. However, their effects do not immediately produce new actions, but new *internal* conditions. We will call these **conceptions,** to distinguish them from conditions arising directly from external input, which we will call **perceptions**. Conceptions may either produce actions or further conceptions. The process that produces a conception starting from a perception or another conception may be seen as an inference.

<u>Example:</u> banana (sensed condition = perception) → fruit (inferred condition = conception)
fruit → edible (another conception)
edible → eat (action)

The final action performed depends not only on the initial perception, but also on the **internal state** of the agent, i.e. the whole of other perceptions and conceptions. These conceptions may derive from various previous perceptions, functioning like internal "memories". The state can be seen as a working memory, keeping track of relevant previous perceptions and conceptions.

<u>Definition</u>: an agent that has an internal state or memory that influences its actions is called **state-determined**: its behavior is determined not only by observable external conditions, but also by "invisible" internal conditions, which are the result of previously experienced external conditions.

Working memory

Working memory can be seen as an internal medium or "message board" through which rules communicate with each other: condition-condition rules "post" their inferences on the message board. For example, the rule A → B, triggered by the perception of A, will post the conception B on the message board. Other rules check the message board. When they find a condition that fits their input condition, they add their own output condition to the board.

Example: assume A and B are already on the message board. Rule B → C will then add C to the board. Rule A & C → D can now also add D to the board.

This cooperation between rules via the board is similar to the stigmergic cooperation between rules via the environmental situation. Working memory in its role as message board performs the same function as the environment in the case of stigmergy. The only difference is that the external state (environmental situation) is replaced by an internal state (contents of working memory). The greater flexibility of internal memory allows the agent to perform various more complex cognitive processes. In particular, the agent can now "process" information, in the sense of reworking one or more initial perceptions step-by-step into a more advanced conception that is easier to interpret in terms of the required actions (like in the example where the perception of "banana" is processed into the conception "edible").

Some examples:

- The agent can compare or combine different perceptions:

  Examples: berry is small *and* light in color → berry is unripe

  light region *next to* dark region → boundary of object

- The agent can react on conditions that develop over time t, i.e. that have different values at different moments. A sequence of perceptions, e.g. A(t), B(t+1), C(t+2), ... can all be kept in working memory.

  Example: reduction of light (A > B > C...) signals nightfall

- The agent can remember several things to do or to pay attention to.

Most generally, working memory means that the agent becomes sensitive to the **context**. The context of a perception or conception can be defined as the whole of conditions that are different from the present perception/conception, but that still affect its interpretation. The same perception will in general mean different things depending on the circumstances. For example, the word "bank" means different things in the phrases "I brought my money to the bank" and "the ducks were sitting on the river bank". The same orange, striped, cat-like shape means something different when perceived in a toyshop or in an Indian jungle. The conditions defining a context are registered in working memory as simultaneous perceptions by other sensors, memories of earlier perceptions, or conceptions inferred from other earlier conceptions or perceptions. They correspond to all the messages simultaneously present in the "message board".

The context-dependent interpretation of the perception will determine subsequent conceptions and actions. For example, the perception of a tiger-like shape in a toyshop will trigger a very different reaction from the same perception in the jungle. Thus, the context will in general modulate the system's reaction, making it react differently to the same stimulus in different circumstances. This enables two important mental mechanisms: controlling goals, and cognitive preparedness.

Controlling goals

In reactive agents, there is no need for an explicit representation of the goal. The goal is the "attractor" of the dynamics produced by the condition-action rules, i.e. the set of situations that actions tend to end up in, or that the agent gets into but does not (through its own actions) get out of.

Example: for the Braitenberg vehicle that moves when there is light but stops moving when there is no light, the implicit goal is darkness. A dark spot acts as an "attractor" for its dynamics: once the agent gets into it, it will not move out on its own.

However, such implicit goals are rigid: they cannot be changed according to circumstances, but only by the higher order process of reinforcement learning (which assumes a higher order value system determining whether the effect of an action is good or bad).

In practice, goals depend on the context. For example, when hungry, a good rule is: food → eat. Otherwise, this rule is counterproductive. Therefore, it is good to be able to "modulate" rules according to the context. This requires an interaction between the rule in focus and other context-sensitive rules.

Example: in the context of hunger, it would be good to make eating a goal, i.e. to "activate" all rules that lead to eating.

This requires a working memory or message board where context can be registered. A condition-condition rule can then add the right condition to the memory. Further rules will be triggered only if the total state (directly perceived condition + context conditions) is right.

Example: low glucose level in blood (perception of hunger condition) → post "hunger" condition on message board. This together with the perception of food can now trigger the rule: food & hunger → eat. This rule would not have become active without the perception of hunger.

The controlled activation of goals in this way can lead to a hierarchy of goals, where some goals, implemented by more general rules, are more broad or abstract (i.e. active in a wide variety of contexts), while others are more specific. For example, the goal of "wanting to eat" is more general than the goal of "wanting to eat ice cream". The first will be triggered by the conditions "hunger" + "presence of food", the second by the more restrictive conditions "hunger" + "presence of ice cream".

Implementation as a connectionist network

State-determined agents require a physical structure to implement working memory. A direct connection *sensor* → *effector*, as in reactive agents, is not sufficient. Different sensory signals, including signals perceived at previous times, must come together internally and interact. This requires a network of signal-carrying connections coming together in nodes. In higher organisms, this function is performed by the nervous system and brain, which consists of neurons (nodes) and synapses (connections). In a single cell, the role is played by the chemical network of genes and proteins.

In a connectionist network, nodes represent conditions or actions, while connections represent condition-condition or condition-action rules. Connections, of the form A → B, can be stronger or weaker, where stronger connections bring more activation from A to B. When A and B are conditions, their strength can be interpreted as the conditional probability of B happening, given that A is the case: P(B|A). When B is an action, strength can be interpreted as the intensity or importance of the action.

Implementation of the message board

The chemical composition of protoplasm in a cell (e.g. a bacterium) functions like a working memory or message board where chemical signals (molecules) are "posted". These chemicals can activate or deactivate genes on the chromosomes. Activated genes produce proteins. These proteins can in turn activate or deactivate other genes. Thus, the gene-protein metabolism, just like the brain, is functionally equivalent to a **connectionist network**. However, it should be noted that the cell appears to have a single, global message board: all chemical signals are released in the same medium, and can in principle interact with all genes.

In the brain, on the other hand, message boards are local: only directly connected neurons can post a chemical "message" (via neurotransmitters and electrical activation) to a particular neuron. Other neurons can only indirectly—via the propagation of activation across intervening connections—affect a given neuron. Therefore, different brain regions can function in parallel, independently of each other. The advantage is that complex situations can be handled by many cognitive processes that work in parallel, without interfering with each other. The disadvantage is that there is no overall integration of information, no place where all the messages come together. Therefore, context conditions sensed in some brain regions may not be taken into account in other regions. We will see later how symbolic cognition and the mechanism of the "global workspace" allow the human mind to overcome this limitation of neural networks, and thus provide a full interiorization of stigmergy.

Conditions as categories

In reactive agents, a condition is typically implemented as a range of values for a sensor. For example, the condition "too cold" means that a temperature sensor senses a value below its reference value (e.g. 15° C). In state-determined agents, on the other hand, conditions generally take the form of conceptions, which emerge from a complex combination of different sensed and/or inferred values.

A condition fundamentally defines a category of perceived and/or conceived phenomena. This means that it separates the universe of all possible phenomena into two classes: those that fulfill the condition, and those that do not. For example, the condition “too cold” distinguishes all situations where the temperature is lower than 15° C from those where it is 15° C or higher. The more abstract conception “danger” distinguishes the category of all dangerous phenomena—such as fire, snakes, tigers, cliffs, etc.—from its complement of non-dangerous phenomena—such as butterflies, babies, clouds, trees, etc. The practical meaning or function of this condition lies in the implied action: danger → avoid.

Such more abstract conditions, or conceptions, largely correspond to what we have called “concepts” when discussing semantic networks. The difference is that a concept in a semantic network is defined by a fixed number of links to other concepts. All these semantic links together determine the necessary and sufficient conditions for some phenomenon to belong to the corresponding category. For example, a “car” could be defined as a “vehicle” (ISA link) that has “four wheels” and an “engine” (HAS_PART link). This means that if you perceive a vehicle with these two parts, then you can infer that it is a car.

In a connectionist network, however, there is no such clear separation between what belongs to a category and what does not. For example, it is difficult to draw the boundary between the category of cars and the one of trucks. There also is no fixed list of necessary or sufficient conditions: the conditions which in combination determine a conception are fuzzily defined and variable, and what appears necessary in one context may not be so in another context. For example, if you see a vehicle being examined in a garage, with two of its wheels detached, you would still immediately recognize it as a car. It is this inherent flexibility of connectionist networks, where the patterns of activation as well as the connections are continuously adapting, that makes them much more powerful—but also more difficult to understand—than semantic networks. Yet, the mechanisms by which this flexibility is implemented are themselves relatively simple, as we will now illustrate.

Thresholded spreading activation

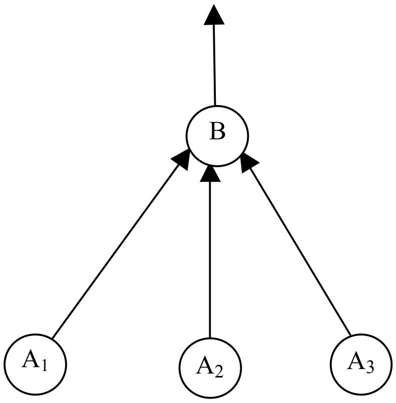
Signals from sensors activate nodes in the network. These nodes propagate the activation via connections to further nodes. The total incoming activation in a node is the sum of activations brought in by all incoming connections. This activation is propagated further via outgoing connections provided it passes a **threshold**. Each node has a threshold activation level. Propagation obeys the following rules:

- activation ≥ threshold → activation is passed on to all output synapses
- activation < threshold → no activation is passed on

Such a threshold is necessary in order not to let *all* nodes become activated, i.e. to make clear distinctions between actual and potential conditions. To understand what the threshold value means in practice, let us consider three special cases:

1) *High threshold*: in this case *all* incoming connections must be active to produce enough activation to propagate the activation. The node is equivalent to an “AND” gate, or conjunction of

all triggering conditions. This means that all these conditions are jointly necessary to activate their outgoing conception.



Example: $A_1 \rightarrow B$, $A_2 \rightarrow B$, $A_3 \rightarrow B$, then B is activated when all three A_1, A_2 and A_3 are active

2) *Low threshold*: here one connection is sufficient to activate the node. The node is equivalent to an OR gate, or disjunction of incoming signals. Each condition on its own is sufficient to conclude that B is the case.

Example: B activated when either A_1, A_2 or A_3 is active

3) *Intermediate threshold*: here some incoming connections must be active to activate the node. The node does not require a full conjunction, but a significant number of inputs to become active itself.

Example: B activated when A_1 and A_2, or A_2 and A_3, are active

This scenario allows for incomplete information: perhaps all three conditions are active, but only two could actually be perceived. In this case, the cognitive system implicitly assumes that the missing one could be active as well

Example: warm-blooded, beak, flies, lays eggs → bird. In general three of these conditions are sufficient to conclude that there is a bird, yet a platypus (rare Australian mammal) is warm-blooded, has a beak and lays eggs but is not a bird. This problem might be avoided by giving a relatively lower weight to the conditions "warm-blooded" and "lays eggs", since a bird shares them with many other animals, so that they would need two additional conditions to cross the threshold.

Analogical reasoning

Two intrinsically different phenomena may be analogous, in that they activate several of the same conceptions. For example, a motorboat and a car are both associated with the conceptions: moves, carries people, has motor, uses fuel, has steering-wheel, etc., even though they look very different and are used in very different circumstances. The recognition of such analogy or cognitive similarity is an automatic process of spreading activation activating the same conceptions. It does not require any reasoning or higher-order symbolic processing.

Since activation always propagates backwards to some degree (as we will shortly discuss under the label of "recurrent activation"), thinking about a motorboat tends to indirectly activate the conception of car via the activation of features that car and motorboat have in common. If then it turns out that your motorboat does not start, you will be quick to remember what you did when your car did not start. Analogy also appears at a more primitive, perceptual level, as when you observe that a particular cloud resembles a camel, since it has two humps and something that looks like a head. The perceived cloud and the conception of a camel here simple share some of their associated conceptions/perceptions, so that the one may remind you of the other.

In both cases, **analogy** is a source of creativity, in the sense that the association of A (e.g. a motorboat) with a very different phenomenon B (e.g. a car) will activate some features of B (e.g. spark plugs) that are normally never associated with A, but that may actually provide inspiration for solving a problem with A (e.g. the motorboat does not start) that you otherwise have no experience with.

Anticipation

The basic control mechanism is **feedback:** correcting deviations *after* they have been perceived. Its basic advantage is that there is no ambiguity about what deviation there is. The disadvantage is that it may be too late to correct. This can happen when the reaction demands more time to prepare and produce than the time that remains until serious damage (or even destruction) of the agent, for example when falling off a cliff, or being attacked by a tiger or poisonous snake. Therefore, it is better to complement feedback with a different control mechanism: **feedforward.**

Feedforward means predicting or anticipating a problem, and acting *before* it has truly occurred. This can be achieved with condition-condition rules.

Example: big striped cat → tiger, tiger → danger of being attacked. The perception "big striped cat" leads here to the conception (message in working memory) of "imminent attack likely". This in turn can now trigger the condition action-rule: attack → flee, before an actual attack happens.

Advantages of feedforward:

- less risk of being too late,
- more time to prepare efficient action,
- intervening when the problem is still far away generally requires less energy.

The disadvantage is that the prediction may never be realized, which would make the action useless or counterproductive. For example, the tiger may not be hungry, or the snake may not be poisonous. Therefore, feedforward must interact closely with feedback, and constantly correct anticipations on the basis of new perceptions (for example, stop running when the tiger moves away or shows lack of interest).

Expectation and preparedness

In general, the agent cannot make exact predictions about what will happen. However, this is not strictly necessary since feedback can correct errors as they become apparent. To make feedback more effective, it is sufficient that the agent would be **prepared** to take action. This means that it should have some degree of **expectation** that something is more or less likely to happen, so as to reduce the uncertainty about which action(s) may need to be performed. For example, in the context of a jungle the perceived play of light in a large bush may be caused by a tiger. Therefore, you should be prepared to run, even though the actual probability of a tiger being there may be small: the perception that triggered it makes it large enough to warrant preparedness. In the context of a park, the same ambiguous stimulus does not warrant any special preparedness, since the probability of it pointing to a tiger is negligible.

In this view, the task of cognition is to *maximally reduce uncertainty,* and thus optimally allocate preparedness. This is an extension of Ashby's and Campbell's view of cognition as appropriate selection: selection reduces the number of possibilities, and therefore uncertainty about what to do next. Uncertainty itself can never be wholly eliminated: in a complex, chaotic world, no prediction is absolutely reliable. However, reducing uncertainty may buy enough time to implement a solution before the problem gets uncontrollable. Physical preparation can happen e.g. by tensing muscles, moving into a certain position, or picking up appropriate tools, weapons or shields. Mental preparation can happen by already activating relevant neural circuits, so that they are ready to start processing information. Activating relevant conceptions implicitly means that other conceptions are not considered relevant, and therefore that it is not necessary to explore them in order to solve the problem. This internal selection of relevant conceptions strongly reduces the search space, and therefore the time needed to decide about the most appropriate action.

Example: a medieval encounter

Imagine a situation in which a swordsman approaches you. Assume that there is an initial uncertainty of 1000 actions that may possibly be appropriate, including greeting him, ignoring him, running away, etc. Now, the swordsman moves a hand to his sword. This perception triggers some anticipations of what the swordsman may be up to, producing a reduced uncertainty of 100 possible actions: he still may or may not attack, and even if he attacks, you don't know how, when or where, but the threat level has directly increased. The uncertainty is still very large, but reduced by a factor ten. Therefore, it takes on average 10 times less time to solve the problem. This is enough to make many potentially fatal problems controllable. For example, you could pick up a sword or shield yourself and hold it in a generally defending position. This prepares you for the most likely moves if the other were to attack.

Priming

The psychological mechanism of **priming** implements the creation of *cognitive preparedness*. (Cf. a "primer" is a (white) paint put as underlayer to prepare a material for a (colored) finish). Cognitive preparedness is sometimes called "set" (as in "mind set"): the mind is set or primed to perceive or interpret something in a particular way.

Example: at an examination the mind is set to expect difficult questions, not jokes, while the opposite applies during relaxed conversation with friends. The same remark may be interpreted as a joke in one context, as a question in the other.

Even ambiguous or incomplete perceptions can prime the mind for particular interpretations. For example, in an Indian jungle, you see a movement in a bush. In this context, your mind is prepared to perceive a tiger. Even when the probability of the movement being caused by a tiger is very small, it is worth being prepared, because the danger of a potential tiger attack is so large.

Priming is investigated in psychology through classical experiments that implement the following scenario. First, a priming stimulus ('prime') is given to the subject. For example, the word "striped" or "square" is shown. This word may even be shown for such a short time that it is not consciously registered (this is called a *subliminal* stimulus). Then, the main stimulus is given. This may for example be a word that can belong to either of two categories: animals or plants. The subject must react as quickly as possible to the stimulus by pushing a button, e.g. the red button for animals, the blue button for plants. The experimenter measures the reaction time, i.e. the number of milliseconds between main stimulus and button push.

This produces the following typical results: if the stimulus is the word "tiger", then the reaction time is slightly shorter when the prime word is "striped" than when the prime is "square". The same effect is found for primes like "lion", "dangerous", "jungle", i.e. all words associated with "tiger", compared to primes that do not have any association with it. Similar effects occur when stimuli are images, sounds, colors, etc.

Interpretation: even for such apparently simple tasks as recognizing a word, the mind requires some time for processing the information. This time is shortened when a previous stimulus "prepares" the mind to expect a word belonging to the right category: the stimulus will be recognized more easily. Thus, previous perceptions and conceptions (context conditions) prepare the mind to react more efficiently, as if by reducing the uncertainty about what will be the next stimulus.

Failure of Anticipation: Surprise, Curiosity and Mystery

We argued that the mind is constantly trying to predict what is going to happen. Given the very efficient learning mechanisms exhibited by connectionist networks, such anticipation is most of the time quite successful. That means that our expectations are usually fulfilled. However, the intrinsic unpredictability of the world implies that from time to time we will be confronted with a **surprise**, i.e. something novel that could not be anticipated with our existing knowledge and perception. Surprise is particularly important for a cognitive system.

First, lack of anticipation means that the agent is not prepared to deal with the unexpected phenomenon. If the phenomenon would turn out to be dangerous, this may cost the agent dearly. Therefore, a first reaction to surprise will be a general **arousal** or activation, both mentally and physically, so as to gather the energy that may be necessary for a fight or flight response. As we will discuss further, such arousal is the basis for emotion. Moreover, cognitive arousal leads to a strong focusing of **attention** or consciousness on the phenomenon that caused the surprise.

A concomitant behavioral reaction is the "**orientation response**", which is a general redirection of the senses towards the source of the unexpected event. This automatic reaction, which Pavlov called the **curiosity** reflex, is aimed at gathering information quickly, so as to assess what is going on. For example, when you suddenly hear a strange noise, your reflex will be to look around in the general direction from which the noise seemed to come, so as to possibly find out what caused it. Better understanding of the situation, together with a general readiness for action is the most general form of preparation to cope with any unexpected situation.

Surprises are not in general dangerous, though. Most opportunities are unpredictable as well. For example, an animal or a hunter-gatherer foraging for food generally does not know where a prey animal or a store of edible fruits or roots will be found. If food were predictably present in certain situations, then foragers would quickly learn how to most efficiently exploit that resource, thus exhausting it in the shortest time. For example, a carcass of a large animal left in the savannah will quickly be eaten by scavengers such as vultures and hyenas before other foragers may discover it. Foraging is in general opportunistic: agents **explore** their surroundings, looking in as many places as possible, driven by the desire to discover new opportunities (or what we have called **affordances**). Therefore, an unexpected phenomenon that is not clearly dangerous will be experienced as something attractive, something that needs to be explored for its possible usefulness.

Even neutral surprises, which are not accompanied by either benefits or dangers, are intrinsically interesting, because they offer an opportunity for learning. By definition, a phenomenon that we cannot anticipate implies that our knowledge of the situation is incomplete. The experience of such a knowledge gap defines a **mystery**. Finding out the precise content, meaning and features of this phenomenon will lead us to discover new information that may help us in the future to better predict, and thus control, the environment. Because they present an opportunity to acquire potentially important knowledge, mysteries are very attractive, succeeding in becoming the focus of our attention. For example, one of the most popular genres in literature and film is a "crime mystery", where the unexpected phenomenon is a murder and the hero of the story (together with the reader) is trying to find out how, why and by whom this was committed. This attraction for the unknown explains the innate drive of **curiosity** that leads us to explore unfamiliar situations, to try and fill gaps in our knowledge, to play, and to experiment with various objects and actions [Loewenstein, 1994]. As we will investigate further when discussing differences in intelligence, this motivation to learn strongly affects the overall cognitive competence that an individual achieves.

Bootstrapping of conceptions

Recurrent activation

An important recent insight, brought to the attention of cognitive scientists by brain theorists such as Jeff Hawkins [2004], is the fact that there is a *two-way* interaction between perceptions and conceptions (at least in the human cortex). Perceptions are external stimuli interpreted as conditions. We have defined conceptions as internally inferred conditions. In neural network terms, perceptions are situated in the input layer, conceptions in one of the "hidden" layers, while the triggers for actions are produced in the output layer. Normally, activation flows in the direction input → hidden → output, with the output feeding back via the environment to the input:

(1) Perceptions activate or prime conceptions.

  a. If the activation is strong enough to pass the threshold then the conception is activated.

  b. If the incoming activation is not strong enough, then we might say that the conception is **primed**, because it can more easily become activated if additional activation arrives so that the total activation now crosses the threshold.

(2) Conceptions activate or prime conceptions or actions.

(3) Actions change the external situation, with the changes feeding back into new perceptions and new conceptions.

This is the basic cybernetic control mechanism that allows the agent to correct possibly erroneous perceptions, conceptions and actions by sensing their external effects. However, a two-way interaction means that corrective feedback can also be *internal*. In neural network terms, this means that the network would be **recurrent**, i.e. containing loops going from later layers back to earlier ones. This can be interpreted as activation flowing back from higher hierarchical levels (more abstract or conceptual) to lower ones (more concrete or perceptual).

In particular, conceptions may activate or prime perceptions. For example, thinking about a tiger may prime the visual system for the perception of stripes. If stimuli are too weak or ambiguous to be clearly perceived, this priming will facilitate perception. For example, when you see a tiger in the dark its stripes may be initially unclear. Once you have inferred on the basis of other clues that this is a tiger, the unclear dark and light patches may now be recognized as stripes.

When the activation flowing back is strong enough, the conception may even activate a perceptual circuit, creating a perceptual impression independent of the sensory organs, as when you imagine a tiger before your mind's eye. This is the neural mechanism underlying mental imagery [Kosslyn et al., 2001] or **imagination**. It was demonstrated by an experiment in which people were asked first to look at a simple shape (e.g. a T), then to remember and visualize it. In both cases—seen and imagined—the same shape activated the same neurons in the perceptual layer.

The feedback loop does not end there, as new, imagined perceptions may activate new conceptions. These may in turn activate or prime new perceptions. This movement back and forth produces a recurrent process that we may call “bootstrapping”.

Bootstrapping

Definition: Two things A and B can be said to stand in a **bootstrapping** relation if A is used to develop, support or improve B, while B is used to develop, support or improve A.

In other words, A and B mutually produce each other, without need for external support or intervention. In terms of systems theory, A and B have a cyclical coupling: A → B, B → A.

The name derives from the “bootstraps”, which are handles on the back of your boots that you can pull upwards in order to supposedly lift yourself out of the mud. In this imaginary situation, the bootstraps and legs play the role of A, the arms and shoulders the one of B: A pulls up B, B supports A.

This is obviously impossible with real bootstraps. The reason is the physical law of momentum conservation: pulling up imparts upwards momentum on the boots B. However, since the total momentum of the body is constant, this implies an equal downward momentum on the shoulders A (action equals reaction, according to Newton's laws). The net effect is that the body (A + B) remains in the same vertical position. In order to move your body upwards (e.g. jump), you need something else to give downward momentum to: a solid floor, that acts as a base or foundation.

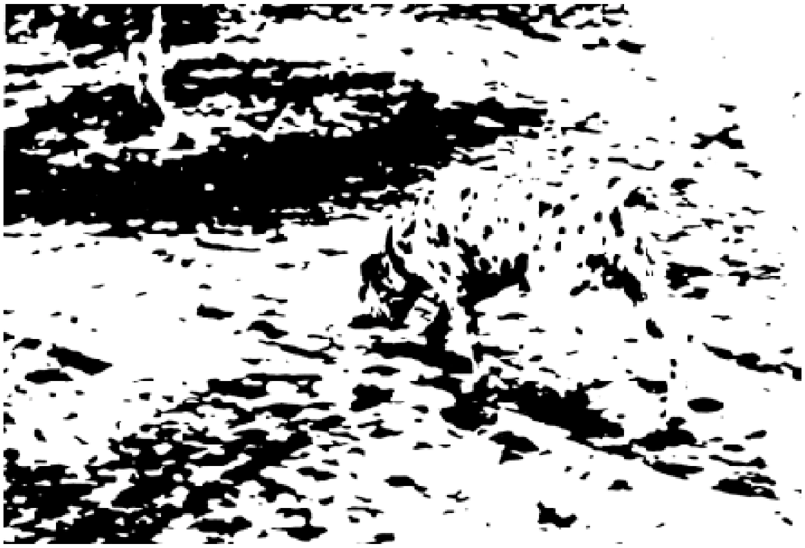
Metaphorically, however, bootstrapping is possible in many cognitive, informational, or self-organizing processes. The reason is that *there is no conservation law for information*: information can be created or destroyed. Therefore, increase in information for one component does not imply decrease for another. In fact, bootstrapping is a general method to improve the quality of information or knowledge. In computing, this is often implemented by an iterative or recursive algorithm, where a multi-step process is repeatedly applied to input information that consists of at least two components A and B, so that A is applied to process B, and B to A. This can gradually transform the information until the quality is “good enough”.

Implementing bootstrapping by spreading activation

Bootstrapping is useless for a single condition action rule. The loop A → B, B → A is just a movement back and forth along the same connection that changes neither A nor B. But when there are several perceptions and conceptions simultaneously active, the activation may *spread* while moving forward and backward, each time activating additional nodes that were not initially reached. On the other hand, the process may deactivate nodes that after one or more iterations do no longer receive enough activation to cross the threshold. After a number of iterations, the bootstrapping process is likely to reach an equilibrium state, i.e. a stable distribution of activation over percepts and concepts. This is typically a cluster of multiply and strongly connected nodes that mutually activate each other, and which is therefore quite stable. This distribution will "fill in" aspects that were initially poorly perceived or were missing in the conception. It forms the final "interpretation" by the cognitive system of the stimuli that initially triggered the activation.

Example: seeing a tiger under a bush

Bootstrapping will fill in for the parts obscured by the branches. Even though the stimuli are very patchy, our existing conception of how tigers should look like makes sense of the fragments and creates a coherent whole or *Gestalt*.



A nice visual illustration of the power of this mechanism of Gestalt perception can be found in the picture. At first sight (initial perception), it only consists of irregular black spots spread almost randomly over a white space. However, when we suddenly realize that part of these spots belong to the skin of a Dalmatian dog (conception), the whole ambiguous picture makes sense, and we recognize the outline of the dog, and infer that the other spots are merely shadows surrounding it. However vague the picture was at first, once the Dalmatian is recognized, it is difficult not to see it anymore: the interpretation of the visual stimuli has stabilized, and it is practically impossible to go back to the initial impression of random splotches.

Bootstrapping is not limited to perception-conception feedback. Conceptions may bootstrap other conceptions at different levels. For example, to a scientist a number of theories of specific phenomena may suggest a more abstract, encompassing theory. This global theory may suggest new local hypotheses. These hypotheses may be tested by observation, or just checked for their plausibility. The results may reinforce or correct both local and integrating theories. We have seen an example of how such a process of theory development can be implemented using spreading activation when we discussed Paul Thagard's [2002] work in coherence epistemology. There, the different observations (perceptions), arguments and hypotheses (conceptions) mutually reinforced or inhibited each other until a coherent pattern or theory emerged.

Associative learning

To understand how a state-determined agent can learn, we must extend the paradigm of reinforcement learning. Since conditions do not generally produce actions whose result can be perceived, simple reinforcement learning is not sufficient: in the rule A → B, where B is a conception, there is no a priori method to determine that B is good (brings the agent closer to its goal), and therefore that the rule should be reinforced or rewarded.

We can evaluate condition → condition rules when the second, predicted condition can be checked by perception: the rule A → B is to be rewarded if B is indeed perceived after the rule has been triggered, i.e. if B receives external activation that confirms the internal activation received from A. This can be implemented by the delta learning rule used in connectionist networks. The mechanism is illustrated by classical conditioning, like in the experiment of Pavlov's dog, where the learned rule is: bell → food.

B may not be directly perceivable, however. For example, B may be an abstract category, such as "fruit," which is used to support a concrete inference:

banana → fruit, fruit → sweet, sweet → eat

This rule using the abstract concept of "fruit" allows the agent to predict that a banana will be perceived as sweet even if it has never eaten one. If the banana is indeed sweet, the rule "fruit → sweet" is reinforced. But this does not tell us anything yet about the rule "banana → fruit". In such a case, we need to indirectly evaluate the quality of the connection. In feedforward networks, this is achieved via the mechanism of backpropagation: part of the reinforcement is propagated back to the previous rule "banana → fruit". However, this algorithm does not work in recurrent networks, such as those implementing bootstrapping: since rules are connected in cycles, backpropagation would cycle endlessly. Moreover, it is unclear how something like the backpropagation algorithm could be realized in an actual brain rather than in an artificial, computer controlled neural network.

Hebbian learning

The problems above can be solved by applying the very simple Hebbian learning rule: each time A and B are both activated (from internal and/or external sources), the connection strength A → B is increased with a fixed amount (say 1%), up to its maximum value (say 100%). Each time A is activated, but B is not, the connection is decreased with an equal amount, because it did not correctly anticipate the state of activation of B. On average the strength of the connection will converge towards the percentage of cases in which B became active when A was active, i.e. the conditional probability P(B|A). When A and B are perceptions of outside phenomena, the Hebbian rule will therefore teach the cognitive system to correctly anticipate or predict B, given A.

Note: This result is similar to the more precise "delta" or "error-correction" rule for learning, where the actual perception of B (external activation) is used to correct the anticipation of B, i.e. the degree of activation that B receives (internally) from A. But what happens if B is a conception, so that it does not receive external activation? In this case, we could still use the delta rule but only by generalizing it via a form of bootstrapping:

Compare the activation of B produced by A → B (direct or "internal" activation produced by the rule or connection) to the activation produced by all other connections coming to B (indirect or "external" activation). Some of this external or indirect activation may come directly from perceptions. Most of it will be the result of complicated sequences including feedbacks. The learning rule then is:

- if direct activation < indirect activation, then increase connection strength
- if direct activation > indirect activation, then decrease connection strength

This "generalized delta rule" is a negative feedback mechanism that reduces the differences between internal and external activation, and thus eventually makes both sources of activation equally strong on average. This implies that if one of the sources is wrong because of noisy perception, the other one will balance it out, thus bringing overall activation closer to what it is normally expected to be.

This is a form of bootstrapping because one source of information (indirect connections) is used to improve another source of information (direct connection A → B), and vice versa. Indeed, when examining A → C, the activation brought to C via A → B, B → C will be taken into account, so A → C partly determines how to change A → B. On the other hand, A → B partly determines how to change A → C.

The result is that the network as a whole becomes more **coherent**: different rules become more mutually supportive in their implications; inconsistencies between rules are gradually eliminated. This fits in with the **constructivist** view, which does not a priori distinguish between conceptions and perceptions.

Associations

Hebbian learning will create and reinforce **associations** between conditions (perceptions and conceptions) that are frequently experienced (activated) together (or the one shortly after the other).

<u>Example:</u> banana → sweet: the more often the two conditions co-occur, the stronger the association becomes, and the more activation will move from the one to the other.

However, learning cannot create associations between phenomena that have never been experienced (perceived or conceived) together.

<u>Example:</u> rabbit → hat, denoting the idea that a rabbit can wear a hat. Rabbit and hat will normally never be activated together. Therefore the delta rule will not increase the strength of the connection above 0.

This implies that "associative" agents cannot reason about situations that they have never encountered before (or at least of which they have encountered no similar situations before). Yet, people are able to reason about rabbits with hats or other situations unlike any they have ever experienced, for example, when thinking about creating a new cartoon character or circus act. To tackle this problem, we will need to consider a higher level of cognition: symbolic thought.

Episodic memory

A first step in the development of human-level cognition is the evolution of episodic memory. **Episodic memory** registers and stores "episodes", i.e. specific events or sequences of events.

Connectionist learning normally requires *repeated* activation to reinforce a connection sufficiently so that it becomes a reliable part of memory. This allows it to learn general, recurrent regularities or rules, like "bananas are (usually) sweet", but not specific occasions, like "the banana I ate yesterday was sweet". Yet, people (and apparently some higher animals) are able to remember individual episodes, such as an incident, a surprise, or even what they ate for lunch yesterday. These events have been experienced just once, yet their memory trace has somehow become strong enough to be repeatedly recalled, sometimes many years later.

This requires some internal amplification mechanism, where the connections for particularly interesting events are repeatedly activated until they are strong enough to store the necessary level of detail. The mechanism may be another form of iteration or bootstrapping in which activation recurrently cycles through the same connections, without additional stimulation from outside perception.

This happens probably in the hippocampus region of the brain, as confirmed by the following observations. Damage to the hippocampus (e.g. because of a stroke) usually results in profound difficulties in forming new episodic memories, although associative learning can still take place. In other words, people with a damaged hippocampus can learn new skills, but cannot remember learning them. For example, they may become familiar with the doctor that regularly visits them, but not remember ever having met the person.

Symbolic Thought

Extending working memory

Short term memory

In neural systems, working memory corresponds to a state of activation. This state is constantly *changing*:

- because of new perceptions (inputs) producing new activation
- because of inferences (propagation of activation) leading to new conceptions
- because of diffusion leading to the gradual disappearance of activation states
- most basically, because an activated neuron cannot *remain* activated. The neuron becomes fatigued or "tired", and will spontaneously lose its activation.

In that sense, "working memory" is not good at *memorizing*, i.e. storing and retaining information, but only at *working* with the information, i.e. processing it into new information. It is difficult for the brain to remember individual items in a "train of thought" without tracing back the whole sequence of activation spreading across associations. Individual perceptions or thoughts can only be stored via episodic memory, but this requires quite some time for consolidation in the hippocampus, and is only available for really important (typically emotionally loaded) "episodes". Moreover, only true *episodes* are stored, i.e. combinations or sequences of causally connected perceptions, such as "I ate a sweet banana yesterday after lunch". Episodic memory does not store separate items, such as "banana".

Yet, people have a limited storage capacity in working memory of approximately 7±2 items, according to Miller [1956], or 4 items, according to the more recent work of Cowan [2001]. This may for example contain a list of items to buy: bananas, carrots, potatoes, and tomatoes. This capacity is often called "short term memory".

A plausible mechanism to implement this in a neural network is activation cycling up and down along a perceptual-conceptual "bootstrap", e.g. from the concept "banana" to the visual image of a banana or the sound of the word "banana", and back. Because the activation moves away, neurons do not get "fatigued". Because it then cycles back, a neuron that lost its activation becomes activated again after a short period. Such a mechanism is very energy intensive, and requires constant regeneration of activation in a large array of connected neurons. Moreover, different "memories" (activation cycles) can easily interfere when activation spreads from the one into the other, thus perturbing the overall pattern. This may explain why our "magical number 7" (or 4) is so small.

Extending memory into the environment

We have seen how reactive agents use their external state (situation) as a substitute for their lacking internal state (working memory). However, like internal activation, this situation changes constantly with their actions or with changes in the environment. It is therefore not useful for reliably storing memories. Agents can use the environment in a smarter way, by "inscribing" their memory contents in a secure, external medium that is independent of themselves.

Examples:

- marking a terrain with pheromones (ants) or urine (dogs)
- making marks on wood to count sheep
- writing to store the contents of a shopping list,
- painting, drawing, sculpting, assembling ... to build up and retain the consecutive elements of a conceptual design

A stable, passive medium, such as paper, stone, or wood, retains the information until the agent itself changes it. Such medium can store an unlimited amount of information.

Changing the state of the environment can be done by manipulating objects: cognitive "tools". They have a state independent of the internal state of the agent. This object state can be stable or dynamic. If it is stable, it can be used for storing information, so that the agent can come back to read the object state at a later moment, while relying on the fact that the state should normally still be the same. If it is dynamic, it can be used for processing information, at least in so far that the agent has control over the process, i.e. that the changes in the state happen according to known, reliable rules.

For example, putting stones into a bag each time you see a sheep allows you to calculate how many sheep there are in total: the process of adding a stone to the bag is similar to the mathematical process of addition. Therefore, if you count the total number of stones, you can be sure that this number is the sum of all the sheep that you have seen. This process would no longer be reliable if stones could disappear from the bag, or break into pieces, because then the dynamics of the stones would no longer correspond to the cognitive process of addition.

More generally, manipulating such stable physical states his allows the agent to "try out" or explore possible actions by playing with objects.

Examples:

- puzzle pieces can be assembled in many different ways to see what fits together with what
- building blocks can be assembled in different shapes with different properties
- a ball can be thrown with different speeds or directions to see where it will end up

Symbolic representations

Symbols

The external trace or "mark" left by an agent in the environment functions as a symbol for the information or mental content that is stored. Perceiving the symbol reactivates the neural circuits that led to its creation, and thus recreates the corresponding working memory, even if other processes have in the meantime completely erased the original working memory. By repeatedly using the same type of symbol for the same type of mental content, the agent learns a strong association between the perceived symbol and the corresponding conception or perception:

- when perceiving the symbol, the agent is reminded of the concept;
- when thinking about the concept, the agent is reminded of the symbol.

This association may eventually lead to a condition-action rule of the form: concept → produce symbol. This rule works to *exteriorize* the mental content, making the inner cognition explicit and perceivable, while at the same time giving it a stable, easily manipulable form. The inverse association will produce a perception-conception rule: symbol → concept. This double rule symbol ↔ concept now implements a *correspondence* between symbol and concept. We can say that the symbol **represents** or denotes the concept, and that the concept is the **meaning** or signification of the symbol.

Symbols to aid thinking

A complex system of symbols can represent a complex process of thought. When different thoughts are exteriorized in sequence, an external collection of symbols is produced. Since earlier symbols are retained, until they are possibly erased, the collection grows more complex. Thus, the content of the external working memory develops as different internal condition-action or condition-condition rules collaborate to improve it. Since external working memory has a much larger capacity, the results can be much more complex than with internal working memory.

Examples:

- making a calculation using pen and paper: internal working memory would never be able to remember all the steps (numbers) in the calculation, but paper stores these numbers in a reliable and easily manipulable way. For example, it is easy to add or delete numbers and thus formulate subsequent steps in the process.
- drawing a plan for a building: the lines and shapes represent different components and aspects of the building (e.g. rooms, doors, windows), in their geometric relationships. Again, these symbols can easily be edited (added, erased, moved, …), where each change represents a new step in the cognitive process of designing the building

- writing a paper or essay to develop a complex scientific or philosophical argument: each word represents a concept, each phrase a particular relationship that is postulated between these concepts, while a sequence of sentences or paragraphs represents a process of inference between the corresponding propositions. Again, words, sentences and paragraphs can be edited in order to further develop, correct or refine the underlying train of thought.

A fundamental property of stable symbols is that they can be combined into more complex wholes without losing their initial meaning. This combination may now create a new insight that did not exist before and that could never have been attained without the process of symbolic representation.

For example, during the exteriorized reasoning process, a symbol representing A may be put together with a symbol representing Z, even though there is no direct association between A and Z in the mind. It is sufficient that at some stage in the process the concept A is activated and inscribed as a symbol (e.g. a word on a page). At another stage, possibly much later in the process, the same happens to Z. If both words are inscribed on the same page, they now present a potentially meaningful combination. The intermediate reasoning stages do not need to be inscribed, though they may be. The result is that a direct conjunction of A and Z now appears in the external, symbolic representation, while that conjunction never occurred in the internal reasoning process. The internal process merely propagates activation from concept to concept, following direct associations, e.g. A → B, B → C, C → D, … Y → Z.

Example: let us see how a conjunction could be created between a rabbit and a hat, via a sequence of such intermediate associations:

- rabbit (inscribed as symbol) → tame animal (internal conception)
- tame animal → circus
- circus → clown
- clown → funny hat (inscribed as symbol)

This ends end up in the symbolic representation of a rabbit with a funny hat, which then triggers the internal conception of a rabbit with a hat. “Reading” the combination of symbols will activate both concepts, which may trigger memories of their perceptual appearances. Thus, a person can imagine a rabbit with a hat, even though he has never seen both together. The external medium has been used as a “shortcut” to directly connect concepts that are only indirectly associated in the brain. This co-activation in working memory can now be stored in long-term memory by reinforcing the connection between the two concepts. In this way, complex thought processes using exterior symbols can be interiorized again.

Dynamic symbols

A symbol can also be “inscribed” in a medium without a stable structure, such as sound, light, smoke signals, or gestures. This is less useful to memorize concepts for the long term, because the sound or light signal will immediately disappear again. Still, perceiving that symbol will reactivate the internal conception and thus support internal working memory. The advantage of such unstable expressions is that they make it easier to communicate information to others: dynamic perceptions are easier and quicker to produce, and are more effective in attracting the attention of others. (In the chapter on consciousness we will analyze more deeply why changing phenomena capture the attention better than unchanging ones).

The same advantages apply to “communicating with oneself”, such as speaking to oneself or scribbling and erasing notes. This is merely an extension of internal cognitive processing during which the symbols support the dynamic cognitive processes.

Socially shared symbols: language

A symbol produced by an individual X as an exteriorization of a concept may be perceived by another individual Y. If person Y interprets the symbol in a way similar to person X, then we may say that the concept has been transmitted or communicated from X to Y. For this to happen, X and Y must “agree” about the meaning of the symbol: perceiving the symbol activates similar conceptions in their two brains.

This is a very non-trivial process: the neural network structures of X’s brain and of Y’s brain are a priori quite different, since they have different bodies, different sensory organs, and have undergone different experiences. They have therefore learned different associations and different ways of categorizing their experiences. However, we may assume that X and Y belong to the same species, which means that their bodies and brains are at least genetically similar. Moreover, we may assume that they have lived largely in the same environment, where they have encountered the same types of phenomena, such as animals, plants, rivers, or tools. Given that both natural and artificial phenomena tend to belong to categories of which the members are rather similar (e.g. animals of the same species, or tools with the same function), it is likely that X and Y will therefore have developed similar (but not identical) classifications, and therefore conceptions, of these phenomena. The problem for them is now to associate the same symbol A with their similar, but different, conceptions A_X and A_Y.

The AI-researcher Luc Steels [2005] has shown via computer simulations how a group of agents can learn to agree about the meanings of the symbols they use, i.e. start to use the same symbols or “names” for the same (or similar) concepts, via a process of self-organization. The basic interaction is called a “naming game”: agent X points to some phenomenon, and agent Y formulates a name for this phenomenon, i.e. the symbol that Y would use for this phenomenon. X then indicates agreement (if X would use the same symbol) or disagreement. If there is agreement, the association between the symbol and the underlying category is reinforced for both agents;

otherwise, the association is weakened. This game is repeated a large number of times for different phenomena, and with different pairs of agents. A symbol-concept association that is shared with little or no other agents will be weakened during these encounters so much that it eventually disappears. An association that is shared by several agents, on the other hand, will become stronger, and thus is likely to be adopted by an increasing number of agents. The simulation shows that eventually one symbol-concept association becomes dominant for each concept. The end result is the emergence of a shared vocabulary, i.e. a fixed set of non-overlapping symbol-concept correspondences. This vocabulary forms the foundation for a language that the agents can use to communicate symbolically.

Our focus, however, is not on communication but on individual cognition. The self-organization of a shared vocabulary corresponds to a process of **social construction** of a system of categories. In the naming game, only the symbolic representation ("name") of a concept varies across different social interactions. In reality, not only the symbols but also the concepts themselves are likely to evolve during these social interactions, as common associations between concepts tend to be reinforced, while idiosyncratic associations, which are shared with no other agents, tend to be weakened. Indeed, unusual associations will typically lead to misunderstandings, while common ones will facilitate communication.

Example: if your experience with the concept "dog" is limited to poodles, you will associate "dog" with "curly hair". When someone else points to a straight-haired dog while using the symbol for "dog", you may experience disagreement. Assuming that there is no other symbol for dog-like creatures available on which you both can agree, this disagreement will tend to weaken your association between "dog" and "curly hair" rather than the association between "dog" and the symbol that is used for it, while strengthening your association between "dog" and "straight hair". This will somewhat broaden your concept of "dog", thus making you more likely in the future to recognize dogs that do not resemble poodles.

The result of many of such naming games is that the concepts will become more similar between the interacting individuals. This has potentially a strong advantage for cognition. Indeed, this process may profit from the **collective intelligence** of the group of individuals (see further): collectively, the members of the group have a more complete and reliable experience of the common properties of phenomena. The process where idiosyncratic associations are gradually suppressed and replaced by more common ones will help to reveal this more accurate understanding. By pooling all these experiences via the process of social construction of a shared category, individuals in the group will acquire a broader and more reliable conception of this category.

Example: different people tend to have experience with different breeds of dogs that are superficially dissimilar in characteristics such as size, shape, length of hair, color, shape of ears, etc. Yet, these breeds are similar in their most essential characteristics, such as being mammals, carnivores, pets, social, and being able to bark. Therefore, they are best grouped in a single, clearly defined category: the “dog” species.

Conclusion: The evolution of a shared vocabulary of symbols will tend to produce a corresponding set of shared concepts that are more stable, more clearly defined, and richer in useful associations and combinations than concepts learned from individual experience. This makes language-based symbolic cognition potentially much more powerful than individual symbolic cognition.

From symbols to rational thinking

Interiorization of symbols

Assume that an agent learns to regularly exteriorize thoughts in order to reason more efficiently. This externalized reasoning use the following sequence of processes:

conception 1 (internal) → symbol 1 (externalizing action)

conception 2 → symbol 2

conception 3 → symbol 3

…

combination of symbols 1,2,3… (perception) → new conception (internal)

new conception → new symbol (external), etc.

If this happens often, the neural network will learn to associate the different steps in the reasoning process: associative learning will create strong connections between internal conceptions and memories of external perceptions that have been regularly co-activated. Instead of merely associating a conception with the action of producing the corresponding symbol, the mind will learn to anticipate the subsequent perception (internal) of this produced symbol (external), even when this symbol is not physically produced. This connection can now function as a shortcut, leading from conception straight to anticipated perception. This anticipated perception, i.e. a pattern of neural activation triggered by the conception, could be seen as an **interiorization** of the symbol. This interiorized symbol can now be associated straight to a new conception, without need to pass through an exterior symbol:

concept → memory of symbol's appearance (e.g. word, phoneme) → new concept

Thus, the interiorization of symbols can create shortcuts between concepts. Assume that concept A is connected to the interiorized symbol A' that represents it. Interiorized symbol A' may be associated with another interiorized symbol Z'. This is because linguistic symbols belong to a relatively small “vocabulary” of words, which are all related by grammatical rules and a rich set of common associations. Z' is connected to concept Z. This produces the following sequence:

A → A' → Z' → Z

This path is rather short for activation to follow. On the other hand, the shortest sequence of purely conceptual connections (without symbols) may be very long:

A → B → C → D → ... → Z

Note: such distances in connectionist networks can be calculated more accurately using Markov processes, as the average time for a random walker to reach Z from A, or as the amount of spreading activation reaching Z from A.

Inner speech means speaking to oneself without actually pronouncing the words. This requires activating the memories of sound perceptions of the words that represent the concepts. This is an example of a cognitive process that uses such a shortcut between concepts via interiorized symbols. Such a process may explain the emergence of symbolic thought. Indeed, conscious, rational thought is mostly linguistic using inner speech, as when you reason by forming sentences inside your head. But symbolic thinking can sometimes also be visual, using imagery or “inner visualization”, where your mind calls up memories of images it knows well, and then lets activation spread from those images to associated images. This process will be most efficient if it uses memories and associations of external symbols that have been extensively used, so that their properties are ingrained in the neural network.

This process of interiorization plays an important role in **cognitive development**. Babies are state-determined agents: they do not as yet use symbols. Their cognitive development is in the “sensory-motor” stage. That means that they react immediately to perceptions by actions, without inner reflection or reasoning. (Note that their actions are not merely reactive, since they depend on the baby's inner state). Small children quickly learn to talk, i.e. using external symbols and the corresponding concepts. When they then learn to think, they start by “thinking aloud”: speaking to themselves. As they grow older, they interiorize this language, turning it into inner speech. In that way, they learn to think symbolically.

Reflection

Interiorized symbols (words) allow “rational” thinking or reflection, i.e. considering different possible states of affairs in the mind without perceiving them or even having perceived them before, and making inferences on them. This includes anticipating what would happen in a particular situation, deciding how to act if a particular situation would occur, and choosing the “best” of the conceivable situations as a goal to aim for. Reflection allows the individual to plan actions long before they are performed—if they are performed at all. This is the essence of complex, internal problem solving.

Example: a chess player considering different sequences of moves

Such reflection uses existing, stored knowledge, some of which can be expressed in symbolic form, but also some of which is purely associative or intuitive (i.e. it cannot be expressed explicitly in the form of discrete symbols). Although it is tempting to express reflection purely formally or symbolically because it essentially relies on (internalized) symbols and on logical inference rules, in reality it cannot function without a huge amount of implicit, intuitive knowledge. This intuitive knowledge is needed to select the most relevant ideas from the astronomical amount of logically possible combinations and deductions, and to bridge the gaps in the argument for which symbolic knowledge is simply lacking. Thus, what we have called the “frame problem” for logical reasoning is avoided because of the associative knowledge in which the symbols are grounded.

Creativity

Language is a collection of symbols (words) organized by a generative grammar, i.e. a finite set of combination rules that allow us to form an infinite set of possible sentences. Symbols (external or internal) are not constrained to have direct associations in order to be activated together. This allows the mind to consider a much larger variety of possible combinations than what could be produced by activation spreading along associated perceptions. Thus, human beings have been able to conceive such counter-intuitive combinations like a rabbit with a hat, a boat with a motor, a vehicle that flies, or a negative number.

In this way, the human mind can be creative. It can imagine situations that no one has ever encountered, and reason about these situations as if they were real. This inner reflection already eliminates implausible or unworkable conceptions, thus producing realistic designs or plans to bring about imagined situations. This is the basis of invention, design and discovery. It is essentially what makes individual human cognition so much more powerful than animal cognition.

Moreover, the use of symbols to exteriorize knowledge not only creates a more extensive and reliable memory, it also allows different individuals to share that knowledge, by using a system of symbols (language) that is understood by all. This allows one individual to add insights to those of another individual. Thus, knowledge can develop across groups, societies and generations, accumulating ever more experience, creative insights and wisdom in the process. This is the origin of the human culture, science and technology that made us masters of our planet.

Consciousness and Feeling

Introduction

Consciousness is a very controversial and confusing topic, characterized by a lot of misunderstanding and a variety of bizarre, religious or spiritualist connections, e.g. with the soul, Buddhism, quantum processes, panpsychism, etc. Moreover, there exist little scientific methodology to study it. Therefore, until about 1990, consciousness was almost a taboo word in cognitive science and psychology. This changed in part because of new methods for brain imaging and detection of neural activation that allowed scientists to monitor what goes on during various forms of conscious or unconscious processing, and because of new insights proposed by theoreticians dissatisfied with the symbolic theories of cognition and their neglect of consciousness and experience. These innovations led to the emergence of the presently fashionable field of "Consciousness studies" that overlaps with cognitive science.

One clear step forward was the distinction by the philosopher Ned Block [1995] of two basic types of consciousness, thus removing a lot of confusion:

- **Access consciousness** is the ability to monitor and (to some degree) control one's own perceptions and conceptions. This includes the ability to make them explicit, express them in words, remember them, and decide which ones are most important. It typically requires strong, focused *attention*.
- **Phenomenal consciousness** is the *subjective experience* that we undergo when perceiving or thinking about phenomena, i.e. the "feel" of a phenomenon. It includes the fundamental sensed qualities of a phenomenon, or what the philosophers call *qualia* (plural of *quale*), such as the quality of "redness". We all know such qualities as a feeling or experience, but cannot really explain them in words to somebody who did not experience them (e.g. a blind person)

We will continue to study these phenomena from our connectionist-cybernetic perspective. This means that we interpret cognitive processes as activation spreading across a complex connectionist network that includes many feedback loops between perceptions, conceptions and actions. Consciousness will then in essence be a clearly experienced, focused type of activation, and not a strange, "spiritual" phenomenon that cannot be understood from within contemporary science.

Degrees of consciousness

In the following we will first focus on access consciousness, because this is the type of consciousness that is most well defined, easiest to observe, and most coherent with our intuition of what a “conscious state” is. Simply put, this consciousness is what disappears when someone is unconscious, e.g. because of distraction, sleep, or complete anesthesia. Since Freud, psychologists have understood that being unconscious of something does not mean that cognitive processing or mental activity stops: there exist a wide variety of subconscious processes. These are just ordinary forms of information processing based on spreading activation. The only thing they lack is that there is no separate, independently active brain region that is somehow monitoring what the process is doing.

Subconscious processes

We must first note that *most* cognitive processes happen subconsciously: they occur on “automatic pilot”, in a reflex-like, instinct-like manner. Examples include the control of breathing or walking, processing of light and sound, the recognition of familiar shapes, the anticipation of common movements and changes, and the understanding of language. These processes require little mental effort, even when they may require a lot of physical effort, as in the case of running. They are automatic, involuntary, without thinking: you do not need to explicitly want or set yourself a goal to perform these processes.

Example: pulling your hand away from a hot surface; anticipating where a thrown object will reach the ground

You also generally will not remember how you performed them, because the associated experiences and actions are not stored in episodic memory.

Examples: walking, driving for an experienced driver: you generally cannot remember when you have made a step or when you changed gear.

The reason is that genes (instinct) or experience (learning) have created strong and reliable connections along which activation can propagate quickly and without error. This means that implicit expectations and goals are fulfilled smoothly and reliably, so that there is no need to monitor, control or otherwise interfere with this neural activity.

Subliminal perception

When a word or image is projected on a screen for a very short period (200 milliseconds or less) and then followed by an image that remains for a longer period, subjects watching the screen will not consciously perceive the short-lived image, only the stable one. Yet, it can be shown through a variety of methods that they do process some of the information in that “subliminal” stimulus. For example, if you show them a number subliminally, they will claim never to have seen that number. But if you then ask them to guess whether the number was smaller or larger than 5, they will guess correctly with a probability (between 60 and 70%) much higher than chance.

Interestingly, if you ask them to perform an operation with the “subconscious number”, like adding 5 to it, they are no longer able to guess better than chance what the outcome of the operation is. This can be taken as evidence that the subliminal stimulus does not enter the working memory (or what we will later call the "global workspace") where operations are performed.

Implicit knowledge

We can even learn new associations and skills subconsciously. This is called **implicit learning.** This phenomenon can be illustrated by an experiment in which people were asked to predict the next letters in a seemingly random series of letters, which has, however, complex statistical regularities [Cleeremans & McClelland, 1991]. After people have seen many series, they tend to predict the next letters with a probability much better than chance. Yet, they have no idea why they pick out certain letters rather than others. Moreover, they even believe that they are guessing purely at random, so they are fully unaware that they have learned to anticipate the sequence.

In fact, most of the things people have learned have this nature: you probably know very well how to walk or how to drive a bike, but you cannot remember precisely how you learned it, and you would have even more difficulty explaining to someone else how to do it. While this is typical of the so-called "procedural" or "how to" knowledge that controls action, also the condition-condition rules that control anticipation are mostly implicit. For example, most of the time you would have no difficulty interpreting someone's emotions after seeing his facial expression. Yet, unless you are a trained psychologist, you probably would not be able to explain which movements of which facial muscles you paid attention to, or what precisely distinguishes an expression of disgust from an expression of fear. We constantly pick up such at first sight minor cues, such as someone's expression, position, arm movements, or intonation, and derive a lot of inferences from them, e.g. about the person's character, intentions or mood. We obviously must have learned how to do that, yet we cannot say when or where we have learned it, or what rules we use to come to such conclusions.

Such implicit (but very real) knowledge is what we call **intuition**. While seemingly mysterious, implicit learning and intuition are accurately simulated by neural networks, which learn the same kind of vague, indirect and implicit associations between a complex range of observations, allowing them to make reliable predictions. The fact that most knowledge is intuitive and subconscious explains the knowledge acquisition bottleneck that is encountered by knowledge engineers trying to codify someone's expertise in the form of explicit rules. It may also help to explain certain premonitions or the sometimes uncannily accurate guesses that a so-called "clairvoyant" or "psychic" can make about a person's present situation and thoughts, apparently without receiving any explicit information.

Full consciousness

A few processes require full attention, awareness, or consciousness. This is because there is substantial uncertainty about whether their outcome will be as expected. Therefore, learned reflexes cannot be trusted on their own, and have to be closely monitored, so that they can be redirected if necessary. There are basically two types of causes for the focusing of attention or consciousness: an external event or an internal motivation.

1) **External**: when something unexpected happens:

Examples:

- stumbling while walking
- seeing something incongruous, such as an upside down car, or a man without arms
- hearing a sudden loud noise
- reading a phrase that does not make sense

In such cases, the mind anticipated incorrectly, and needs to focus its full attention in order to get back on the right track, i.e. understand what is happening and get back in control.

Examples:

- checking what made the noise, or was responsible for the upside-down car
- rereading the sentence and trying to understand the grammar or context

2) **Internal**: when the mind itself has decided that the task or situation is important enough to merit full concentration. More specifically:

a) when tackling a complex or unusual problem:

Examples:

- making a long division such as 23749/684
- standing on one leg
- walking for a 1 year old,
- driving for a beginning driver
- writing a scientific paper

Here the result by definition cannot be anticipated, and every small step must be checked via multiple feedbacks.

b) when it is important that nothing escapes the attention (e.g. because the risks associated with possible errors are too great, or because you consider this crucial)

Examples:

- walking on a high beam is not intrinsically more difficult than walking on a straight line on the floor, but because it is more important that you do not make a mistake and step wrongly, attention will be much more focused

- undergoing a job interview: answering the questions of the interviewer about your past is not more difficult than answering the same questions asked by a friend, but the outcome is more important and therefore deserves more attention

Conscious control is especially necessary with a combination of the above conditions, i.e. a situation that is unexpected, complex and critical—for example when suddenly being confronted by a tiger, or a man with a gun. In such situations, people will typically be extremely conscious of everything that is happening, and remember the situation later in extensive detail, even when it only lasted for a few seconds.

In situations of full consciousness a lot of activation is needed to explore different routes in the brain, e.g. to perform backtracking when previous attempts turn out to be misdirected, or to examine several aspects at once. This activation will leave many and strong traces, explaining the enhanced memory of the event. This activation can be measured via brain scans (functional Magnetic Resonance Imaging: fMRI). It turns out that when performing an unfamiliar task, the brain uses a lot of energy. However, when the task becomes more familiar after a few repetitions, the use of energy decreases dramatically. Interestingly, this decrease happens more quickly in more intelligent people [Haier, 1993]. The interpretation is that intelligent people simply learn new connections more quickly, and therefore they can shift more easily to subconscious, automatic processing.

Partially conscious processing

Many processes require an intermediate level of attention. This happens when the situation is not so routine that no attention is needed, but it is sufficiently safe and predictable so that we do not need to put in too much attention.

<u>Example:</u> chatting with a close friend, walking through a busy city, singing in the bath, ...

Such situations will typically remain for a while in short-term memory, but not in long-term memory, since they are not sufficiently "interesting" to activate episodic memory.

Consciousness of change

Summarizing the previous observations, we can note that consciousness is in a sense inversely proportional to the degree to which the cognitive system feels able to accurately determine what is going to happen or what needs to be done. The better the mind is at anticipating events and actions, the less consciousness or attention it needs. On the other hand, the more unexpected, difficult to control or novel a situation, the more attention it receives.

This principle is rooted in the cybernetic basis of cognition as the control of **diversions**: information processing is only necessary when the situation somehow deviates from the desired or expected situation. Thus, the origin of cognition is the perception and interpretation of *differences* or *changes*. This mechanism is so fundamental that the perceptual system tends to completely ignore unchanging phenomena.

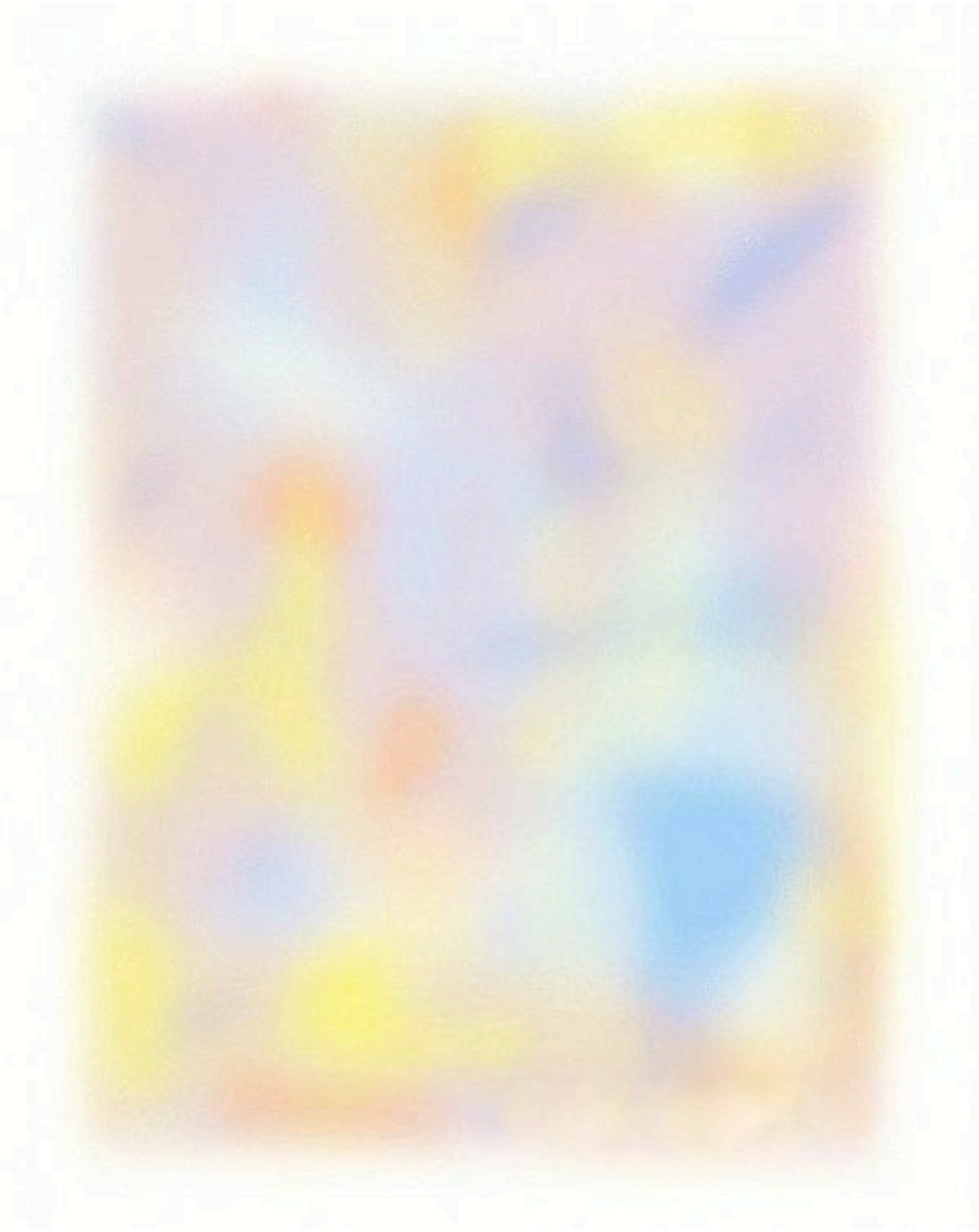
Example: Frogs can only see things that move, such as a fly or a bird approaching. They are effectively blind to static parts of the background, such as stones, branches or leaves. Functionally, this makes sense since frogs only eat moving prey, and only need to be afraid of moving predators.

Physiologically, the mechanism is simple: the neurons in the retina that sense incoming light get "fatigued" and thus stop propagating activation until a different type of stimulus comes in. This may be understood as a simple energy-saving device: why waste all that neural activation to endlessly repeat the message that things are still the same? The same mechanism is still present in human vision. This can be demonstrated via an experiment where a tiny projection apparatus is stuck to the eyeball, so that the image it projects moves together with the eye, and thus activates always the same cells in the retina. With this set-up, the image is no longer registered, and the person has effectively become "blind" to it. The only reason that we can see unmoving scenes is because our eyeballs are constantly moving, scanning different aspects of the scene, so that the stimulation of the neurons in the retina constantly changes. These fast, discontinuous moves are called "saccades".

For an illustration of the effect, consider the visual illusion on the next page. Choose any spot on the picture and stare at it without moving your eyes. After four or five seconds the periphery starts to blur as if mist is coming up and eventually the whole picture disappears: it seems as if you are staring at a blank sheet of paper! The explanation is simple: fixating a spot stops your eyes from saccading. Without these movements, the same cells in the retina always receive the same color stimuli. Because of fatigue they stop responding so that the stimuli seem to disappear. This does not work with a sharply contrasted mark (like a letter on the page), because you cannot completely eliminate microsaccades that move the focus of your eyes for not more than a millimeter or so. That movement is enough to vary the level of stimulation for the neurons exposed to the border between the mark and its background (and thus prevent neuronal fatigue), but not for the ones exposed to the very gradual changes in color in the picture.

Similar neglect of static stimuli occurs for the other senses. For example, you typically only become aware of an unchanging background noise, such as the hum of a refrigerator, when it suddenly stops. Similarly, you feel your shoes when you put them on, but while you are wearing them, you forget about them. You also quickly get used to a particular smell that hangs about a place, and stop noticing it.

At a higher, conceptual level too, people tend to be oblivious to circumstances that are always the same. This includes phenomena that are intrinsically moving or dynamic, such as the flow of a river or a repetitive sound, but where this change is perfectly predictable: while the phenomenon itself may change, some of its higher level properties, such as the speed and direction of the flow, or the rhythm of the sound, remain invariant. This means that neurons encoding these higher-level properties will stop propagating activation because of the same "fatigue" mechanism. Only when that property suddenly changes, such as the rhythm increasing or slowing down, will they start firing again, thus attracting our attention to a potentially meaningful change in the situation that demands cognitive adjustment and therefore conscious processing.



Causal Attribution

The same ignorance of invariant properties can be found in the cognitive process of **causal attribution**: the search for causes to explain observed events that were not anticipated. For example, if someone falls from a staircase, you will normally not attribute that fall to the force of gravity, the steepness of the staircase, or the weight of the person that fell. Physically, these are all necessary conditions for the subsequent effect: without them, no fall would have happened. Yet, they are not seen as causes, because they are invariant, i.e. present all the time. Causal attribution will typically single out the phenomenon that deviates from the "normal", i.e. from the default expectation of how things are supposed to be [Hilton & Slugosky, 1991]. In this case, the preferred cause may be a child's toy lying on the staircase, or a moment of dizziness, because these are necessary conditions for the fall that are *not* expected to be there.

Meditation

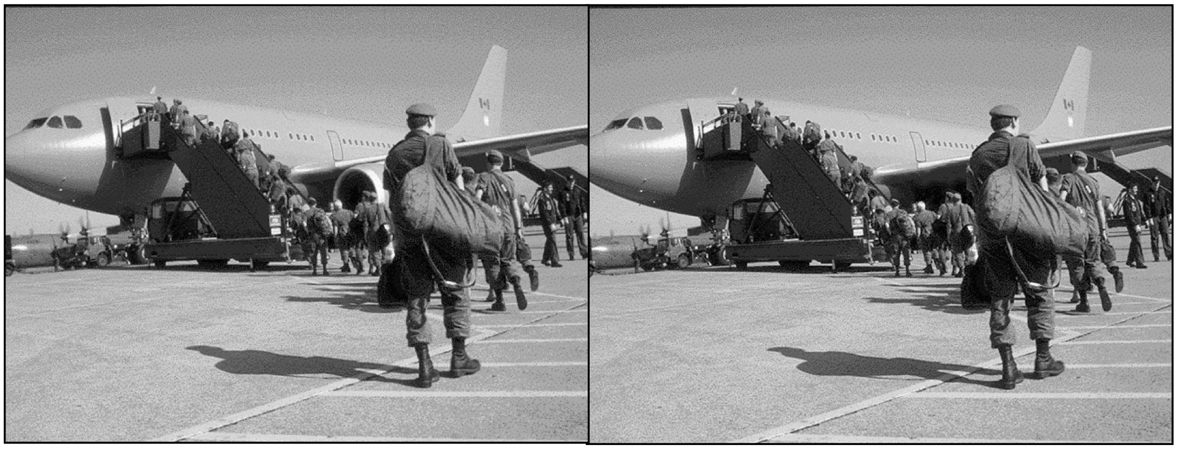
Meditation is an ancient technique to control and manipulate consciousness. It was developed in different religious traditions as a way to achieve a higher insight into reality. The principle is simple: while sitting without moving in a relaxed position, the meditator must concentrate as long as possible on an unchanging perception, or, somewhat more difficult, conception. Typical phenomena that are used to focus attention on are the flame of a candle, the air that enters the nose while breathing, or a word (a so-called mantra, such as "om") that is silently repeated. Because of the mechanism of neuronal fatigue, this is intrinsically very difficult: the candle flame quickly disappears from the perceptual field, and attention wanders to different perceptions (e.g. background noises, or pain in the legs from sitting too long in the same position) and conceptions (e.g. remembrances, or worries about work). Yet, the meditator is trained to as much as possible ignore these distractions and bring attention back to the phenomenon in focus.

When meditation is continued long enough, the effect is first a deep relaxation, and eventually a "**mystical experience**", where the subject loses the sense of being a specific individual located in a particular place at a particular time. This loss of identity is replaced by the experience of becoming one with the world, of dissolving into the All (the so-called "oceanic feeling"). Depending on the religious tradition, this ultimate state can be interpreted as a "union with God" (Christianity) or the reaching of Nirvana (Buddhism).

The psychiatrist Deikman [1966, 2000] has interpreted a mystical experience as a "deautomatization" of the cognitive system. Following our connectionist model, we can explain that in the following way. Normally, activation is automatically propagating along the strongest connections between neurons, reaching associated memories and thoughts, and thus situating the present experience within its broader context. The mechanism of neuronal fatigue ensures that such propagation processes never come to a halt: activation cannot stay in the same place, but must always move on. The continuous refocusing of attention, and therefore activation, on the same neural region, however, creates a kind of “overload” of these strong connections, so that they no longer can propagate activation. This region now effectively stops performing its normal functions. You can demonstrate this effect to yourself by always repeating the same word. After a while the word will lose all of its meaning, so that it does not even appear like a word anymore but like a meaningless sound. If the brain region that is deactivated contains the neural systems that monitors who, when or where you are, it may even appear as if you lose your sense of self, position, or time, as if you have become a timeless emptiness.

The loss of focused activation allows much weaker, more diffuse connections or associations outside of the main region to capture activation. The effect is that experience loses its sense of sharp distinctions and clear situation, instead producing a blurred, “oceanic” feeling, similar to the “mist” that comes up when you stare at the visual illusion, or to the loss of meaning for the word that you constantly repeat. Moreover, these weak connections are likely to produce various new associations that normally never get the chance to be activated because they are dominated by the strong, frequently reinforced associations. The result is a “fresh” experience, where the same phenomena are seen in a different light, possibly triggering novel sensations and creative insights. Similar mystical experiences can be produced by hallucinogenic drugs, such as LSD, which appear to facilitate the flow of activation along unusual connections.

Change blindness



Although our attention is especially attuned to detect change, sometimes the cognitive system can be fooled: if the mind thinks that it anticipated correctly, actual changes may not be noticed. This phenomenon is called **change blindness**. People may not become aware of a change when it is “masked” or camouflaged by another, more noticeable disturbance, such as a sudden interruption or flickering of the image. After the interruption, seemingly the same image is shown. Since the image at first sight looks identical, the brain assumes that it anticipated correctly and that the original image is still there. Yet, there is an important, but non-intuitive difference.

For example, a photo of an airplane is shown first with, then without, a motor on the wing, or a photo is shown in which some element of the background, such as a wall, has changed color. People will generally not notice the change, and find it hard to believe that there actually was a difference between the two pictures, although the difference becomes obvious when the two pictures are put side-to-side. In this case, although there was a failure of anticipation, the error was not perceived, and therefore consciousness was not triggered by it.

This example also illustrates the fact that perception does not create an accurate mental "image", but merely a more vague and abstract representation. Indeed, if such an internal image were registered, then it should be easy to notice the obvious difference between the images *before* and *after* the flickering. The fact that this does not happen demonstrates that perception merely pays attention to the aspects that the mind considers meaningful, in the sense that they are relevant—directly or indirectly—to the agent's goals. The presence or absence of a motor in the above photograph apparently is not considered sufficiently significant to be registered—except perhaps in the mind of a pilot. However, if the motor would suddenly disappear from the picture (without being masked by a more salient perturbation), then the change detection mechanism would kick in and immediately attract our attention to this incongruous phenomenon.

More generally, the change dependence of perception and awareness demonstrates how unrealistic the reflection-correspondence theory of cognition is. Cognition does not accurately reflect static objects: it only registers changes (deviations, distinctions) and relationships between changes (connections, causality).

Access consciousness

Controlling attention

Access consciousness implies the ability to control attention and therefore the propagation of activation. This requires directing activation to the task at hand, while inhibiting the spread of activation in unwanted directions.

A classic experiment to test this kind of control is the Stroop task [McLeod, 1991]: subjects have to look at a list of words denoting colors. The words are themselves colored, but the color of the letters and the meaning of the word are in general not the same. For example, "green" may be printed with red ink. The task is to as quickly as possible name the color of the ink (but not read the word) for each item on the list. Seeing an item triggers two contradictory perception-action rules:

word "green" → say "green"

color "red" → say "red"

In this task, the first rule must be suppressed or inhibited. Otherwise, there will be many errors. This is difficult, because we are trained to read words rather than consider the color in which they are printed. Normally, most activation would flow along the first neural connection, and not along the second one. Preventing this activation from following the wrong route demands a lot of effort or attention.

When the attention is strongly focused on a particular task, the resulting inhibition of other things can be so strong that phenomena that under normal circumstances would have reached consciousness, now fail to be noticed. This can be illustrated by another, more recent psychological experiment [Simons & Chabris, 1999].

Example: the gorilla experiment

In this experiment people had to watch a movie of people playing a ball game. They had to pay special attention to how many times the ball changed hands between the two teams—which was difficult given that there was a lot of movement. When a man in a gorilla suit very obviously passed in between the players, most observers did not notice him. When questioned afterwards, they were sure never to have seen such a crazy thing like a gorilla passing by.

Interpretation: the focus on the ball suppressed all rules that would normally be activated by stimuli irrelevant to the task, including something as unexpected as a gorilla.

Such suppression can even be accurately measured in monkeys. Sensors are inserted in the monkey's brain to detect the activity of different neurons. The monkeys are trained in a task (monitoring what happens on a computer screen) that demands full attention on one type of stimulus, e.g. green rectangles. As a result, the activation of the neurons responsible for perceiving other types of stimuli is suppressed. For example, when red circles appear on the screen, the neurons responsible for perceiving them are activated only 10% of their normal activation value without the task.

These experiments can be seen as further examples of the conception → perception feedback: a conception (in this case the specific task that demands the attention) changes the perception, so that stimuli relevant to the task are perceived more clearly, while the others are hardly perceived at all.

Self-awareness

Access consciousness is the ability to monitor and intervene in one's own cognitive processes, e.g. by refocusing attention on a task after you notice that you have been distracted. This means that the subject is aware that s/he is thinking, imagining, feeling or perceiving, i.e. aware not only of the phenomenon being perceived or thought about, but to some degree of the process of perceiving and thinking itself. Eventually this may lead to an awareness of the agent performing the process. Thus, the most developed form of access consciousness includes self-awareness or self-consciousness.

This type of consciousness seems typically human. For non-human animals, self-consciousness is measured via the mirror test: an animal is considered to exhibit self-awareness if it reacts when it sees in a mirror that a colored dot has been painted on its forehead. Only chimpanzees, dolphins and elephants have passed that test. Dogs and cats fail. It is not clear, though, in how far the mirror test really measures some form of self-awareness or self-concept.

Human self-awareness is probably supported by symbolic cognition. Access consciousness is typically accompanied by thought processes or reflection, which are themselves supported by interiorized symbols (words). One important symbol is the one representing the "self", for example the words "I", "myself" or my name. Children will typically first learn to speak about themselves in the third person, like "Johnny wants an ice cream". In this, they merely use the "symbolic trick" of referring to a stable phenomenon by means of a name they have learned through communication (with the parents). Later, they will start to understand that this phenomenon, the "self" is something special, that deserves a different way of addressing it, namely the first person "I". Interiorizing this symbol produces a self-concept that can be used in combination with other concepts to reason about themselves, producing thoughts such as "I can dance", "I am small", "I like dogs", … Through social comparison with other people, individuals will then learn which properties they share with others, and which properties make them unique. This is the basis of the sense of personal identity.

The global workspace model of consciousness

Interiorizing stigmergy

The brain consist of an array of many, largely independent modules that work in parallel, each specialized in a particular task, such as recognition of specific shapes, emotions, or control of specific movements. These brain modules have few direct connections that allow them to communicate so as to form a global picture of the situation. One way for them to pool their expertise is by exteriorizing the inferences made by some of the modules, so that the results can be perceived, i.e. re-entered into the brain and thus processed by the other modules. Exteriorizing cognition takes place through the creation of physical symbols, such as drawings, utterances or writings, that represent the mental contents. Typical examples of this process are talking to oneself, or taking notes and drawing schemas while thinking about a complex problem.

This is an example of **stigmergy,** i.e. indirect, environment-mediated interaction between the modules within one's brain: a module's outcome through action is converted into a change of the environment; this change is then perceived again, triggering new inferences by the same or other modules, that produce a new action, and a subsequent modification of the external symbols. In this way, an idea is step by step elaborated and refined.

As we saw, when the individual becomes experienced with this process, shortcuts are developed and symbols are interiorized again. Thus, children talking to themselves while thinking will soon learn to use inner speech, i.e. forming sentences in their head without actually vocalizing them. The "place" in the brain where these internal reflections are produced, and from where they trigger further reflections, plays the role of an internal stigmergic medium. We can see it as a kind of forum or meeting ground for the different more specialized brain processes, where they make their most important results "publicly" available, so that others can work on them. This idea is at the basis of the currently most successful theory of (access) consciousness, which is supported by several observations, experiments, intuitions and theoretical arguments: the global workspace theory.

The global workspace

The **global workspace** is the name given by the cognitive theorist Bernard Baars [1997] to this internal forum that globally interconnects all brain processes. A perception or conception becomes conscious only when it is "published" or "broadcasted" in the global workspace. In this view, consciousness is nothing more than an extended "working memory" within the brain where these conceptions are produced and combined, so that they can be submitted to the scrutiny of the various more specialized modules.

An intuitive analogy is the notion of a theater, where the action on stage (conscious thought) is brightly lit so that it is visible to all, while all other activity, like preparation of the décor or people sitting down to watch, remains in the dark (subconscious processes). Note that while this model uses the "theater" metaphor for mental processing, it does not presuppose an intelligent homunculus watching the theater. The theater is being watched by a crowd of simple modules or condition-action rules, none of which has truly a "mind of its own". In that way, the global workspace theory confirms our intuitive notion of a "theater of the mind" where incoming perceptions are projected for further monitoring, but without falling into the homunculus fallacy of assuming the existence of a mind within the mind.

As we saw, while neural networks can be very flexible, it is in general impossible for a single network to learn to perform different functions. Therefore, the brain needs a diverse collection of subnetworks or "modules", each with its particular specialization. The global workspace appears to be a shared internal environment that the brain has evolved in order to facilitate the coordination between these otherwise largely autonomous and automatically reacting modules. The global workspace is probably implemented as a network of long-range connections between different more specialized regions, connecting different parts of the neo-cortex (the most advanced part of the brain). Its function is similar to the one of working memory: keeping track of the presently most important perceptions and conceptions, and making them available for further processing.

To do this, activation needs to be sustained by periodic reactivation of the specific neurons that carry an "active memory", in some kind of self-reinforcing cycle. Thus, conscious thought is characterized by much stronger activity in the brain than subconscious or subliminal processing. This can be established by EEG or fMRI measurements that track the progression of a stimulus from subliminal to fully conscious. In the hundreds of milliseconds that this process takes, it looks as if the global neuronal workspace is "ignited" by an explosively growing activity that stabilizes once full consciousness is reached.

Once the activation has become self-sustaining, the corresponding perception or conception has achieved sufficient stability to be extensively examined by the different brain modules, and if necessary processed by them. Thus, consciousness works a bit like a computer, where the content of memory (RAM) is subjected to a controlled sequence of processes (programs stored on the computer's hard disk). Subliminal stimuli, on the other hand, merely spread automatically through a broad neural network while undergoing continuous transformation by that process. Therefore, no memory of the stimulus remains, and there is no way to control, reexamine or correct previous steps in the process. With conscious processes on the other hand, the activated pattern is sufficiently strong and stable so that it can be transferred to episodic memory if needed.

Sequential vs. parallel processing

The global workspace hypothesis explains a number of characteristic features of access consciousness. The first essential characteristic is that consciousness, like working memory, has a strictly limited capacity, forcing it to process information **sequentially**: you can truly pay attention to only one task at a time. People who claim to be multitasking are either switching their attention very quickly from the one task to the other (for example watching a movie while reading a book), or performing some tasks on automatic pilot, without conscious monitoring.

Subconsciously, your mind can be busy with many things in **parallel** (e.g. walking, talking, perceiving the surroundings, listening, feeling, …). The reason is that subconscious processes happen in different parts or modules of the brain that are specialized e.g. in processing sounds, producing language, interpreting visual stimuli, or preparing emotional reactions. Because they have few connections, these modules can be active independently, without interfering with each other.

However, for all these modules to be coordinated and controlled, their outputs must come together into a shared channel: the global workspace. Only one self-sustaining pattern can be active in that space at a given time. If more than one pattern would be active, activation from the one would spread and interfere with activation coming from the other, until both patterns of activation would merge and become indistinguishable. That is why people find it difficult to pay attention to more than one thing at a time. It explains phenomena such as not noticing a gorilla walking among people passing a ball: if full attention is already focused on the movement of the ball (meaning that observation of the ball fills up the global workspace), no attention (workspace) is left to watch out for gorillas or other incongruous phenomena… This is achieved by the suppression or inhibition of all activation coming from competing patterns.

Free will

Another feature that the global workspace theory explains is voluntary control or what is sometimes called "**free will**": the ability to consciously decide what you will think or do. This is one of the least understood features of the mind, which has given rise to a lot of controversy and confusion.

<u>Illustration</u>: the Libet experiment

> One of the most famous experiments about consciousness puts into question the intuitive notion of free will [Libet, 1999]. In this experiment, people are asked to move their arm, but they can choose themselves at what moment they do so. They are also asked to watch the very quickly rotating hand of a clock, and to remember the exact position of the hand at the moment they decided to move. This gives an estimate of when the conscious decision was made. However, a simultaneous registration of their brain waves shows that the wave of activation that prepares the movement already emerged some 400 milliseconds before the conscious decision to act (and about 600 ms before the actual movement), implying that the action was initiated subconsciously.
>
> This seems to imply that our feeling that we decide consciously what to do is merely an illusion, and that all cognitive activity is controlled by subconscious processes. However, the experiment also showed that people still have the ability to "veto" the impulse to act before it is executed, and therefore they retain some form of conscious control.

Let us try to explain will from the perspective of Baars's [1997] theory. The global workspace not only receives inputs from all subconscious processes, it also sends them its output. The activity in the workspace is "broadcasted" to all the modules of the brain, so that they can process it further and perhaps refine or complement its results. Because the conscious pattern is relatively stable, it can be systematically examined by the different modules, which thus have the time to "decide" what to do with it by examining different options and choosing the most attractive ones. Once the global workspace comes to a certain conclusion (e.g. that a particular thought should be expressed in language, or that a particular action should be executed), the specific modules responsible for implementing that decision (e.g. modules specialized in vocabulary, grammar and speech) will receive a strong input activation from the workspace that overrides whatever processes were going on in that module. That activation will be duly processed and passed on to further modules, until final execution of the decision. In that way, the global workspace is the final arbiter or decider, who controls what happens in the organism.

But how are choices made within the global workspace? This most likely happens via a process of *competition*: many initially subconscious perceptions and conceptions simultaneously stream into the workspace from different modules. Because of the limited capacity, however, only a single one, the "strongest" one, can be broadcasted. That dominant thought suppresses or excludes all the rival thoughts, becoming the thought that is "in control" of the mind and thus able to impose its focus and desires on the rest of the brain and body. The selection of what becomes dominant probably happens through a process of non-linear amplification, where stimuli that pass a certain threshold of activation start growing exponentially, while increasingly inhibiting the activation of competing stimuli, until only one wave of activation remains.

This amplification of a single perception or conception can be observed via brain imaging and EEG registration [Dehaene & Changeux, 2004; Dehaene, 2008]. We saw that when a stimulus is shown for a very short interval, it remains subconscious or **subliminal**: it affects the functioning of the brain, but is not available for conscious processing. This means that it only produces a weak and transient trace as measured by the different methods that track activation levels in the brain. When the interval is made longer, however, the perceptual activation has time to build up to the critical level where it becomes strong enough to win the competition for global workspace domination. Measurement of activation levels then shows a sudden explosive amplification of the signal, accompanied by a broadening of the area that is activated, until it fills the whole "workspace" (and thus excludes rival signals).

The homunculus illusion

We should not forget that the large majority of mental processes remain subconscious, meaning that they are not broadcasted via the global workspace, although they still tend to influence what happens there. Therefore, the feeling that *all* our actions or thoughts are controlled by consciousness (global workspace) is an illusion. This illusion is created by the fact that the origins of most actions or thoughts lie in subconscious processes that never reach the global workspace, so that it seems as if they do not exist.

Another illusion is the one of a unified "self" sitting somewhere in our brain that makes all the decisions. This illusion is at the basis of the fallacy of the **homunculus** controlling what is going on in the Cartesian theater. While the global workspace is similar to some degree to such a theater, what happens there emerges from many different subconscious processes that are competing for dominance. The strongest inputs (e.g. those caused by unexpected phenomena, driven by strong motivation, or with a high subjective or emotional value) tend to suppress the weaker ones, until one becomes strong enough to "fill" the workspace with activation. That input then temporarily determines what you are conscious of, giving it a privileged—but by no means unique—ability to influence your further thoughts and actions.

But the contents of the workspace, and thus your consciousness or "self", is likely to be modified shortly by the outcome of some other, as yet subconscious process, coming from a different part of your brain. Thus, there is no single "self" in control of your thoughts. Your mind is more like a "society" of different modules and inputs that are collectively solving a problem too complex for any one of them individually to solve. What appears to be your "self" talking is merely the module that is speaking out loudest at the moment, thus claiming the floor of the theatre, and making the other ones listen, until another one takes over the floor and speaks in turn, potentially dismissing everything the previous one proposed.

Phenomenal consciousness

The "hard problem" of consciousness

While the function or usefulness of the global workspace or access consciousness is rather clear, the same cannot be said about subjective experience: why do we have any sensations or feelings when we perceive or think about some phenomenon? Why cannot we just process the information automatically or mechanically, like a robot or a computer?

The philosopher David Chalmers [1996] has called this "the **hard problem** of consciousness". According to Chalmers, the problem cannot be solved by traditional scientific methods. He argued this by conceiving a "zombie". This is a being indistinguishable physically and psychologically from a normal human being, that behaves just as any other person, but that does not have any feelings while reacting to phenomena. Science per definition cannot make any observations that would allow you to identify a zombie, and therefore has nothing to say about the issue. Yet, Chalmers claims zombies are essentially different from real people capable of consciousness.

From a cybernetic perspective, this is actually a false problem. According to Leibniz's principle of the identity of the indiscernibles, zombies by definition must be the same as normal people, since there is no way to distinguish the one from the other. Still, we must explain what subjective experience is and why we need it.

Subjective experience

Subjective experience can be seen as the complete state of activation present in a person's brain at any moment. It is personal and unique in several fundamental aspects:

- it is different for every person. Indeed, every person has a unique neural network, developed by biological growth and psychological experiences. Therefore, experience can never be accurately communicated to another person. Moreover, since symbols only cover a very small part of the meaning that resides in our mind, it also cannot be expressed symbolically.
- it is different for any context, event or situation, because no two situations will be so similar that they exactly activate all the same neurons to the same degree.

- it is different from one instant to the next, because it immediately changes into another state depending on new perceptions and the internal dynamics of spreading activation and "fatiguing" of neurons.
- it is intrinsically "intentional" or "affective": it is not a neutral, objective registration of what happens in the environment, but an evaluation of aspects that are potentially positive or negative for the person having that experience. Experience prepares the person for goal-directed action, so as to be ready to counteract negative aspects (**disturbances**), and exploit positive ones (opportunities or **affordances**).

In conclusion, experience is fundamentally *subjective*, *transient*, *idiosyncratic* and *ineffable* (impossible to be expressed in language).

Feeling without action

But why do we have experience even when we are passive, when nothing happens, when no actions are taken, or when no goal is (consciously) aimed at? What function can it have if it does not have any observable consequences? This is Chalmers's objection against a functionalist account of consciousness (functionalism means explaining mental phenomena in terms of the use or function they have for the organism).

From our cybernetic perspective, in those seemingly non-functional cases the implicit goal of cognition is **anticipation**, i.e. being maximally prepared for anything that *might* happen. This implies letting activation spread from all presently perceived or conceived phenomena to whatever other phenomena that are associated with them (because they have to some degree co-occurred in the past and are likely to occur again). These other phenomena too will be evaluated in terms of associations and possible dangers or opportunities, albeit not as intensively as the phenomena in focus. We may assume that the associated phenomena will at least be **primed** for potential activation later.

As also argued by Jeff Hawkins [2004], it is the whole of all these implicitly anticipated, associated phenomena that constitutes our "understanding", experience, or feel of the phenomena in focus. Thus, even an at first sight passive state is characterized by plenty of active processes propagating and bootstrapping within the brain. Whenever something does change, externally or internally, the mind is primed for appropriate action. Without this subjective state of anticipation, the brain would have to start processing the new situation from scratch, thus being much less efficient in its reactions.

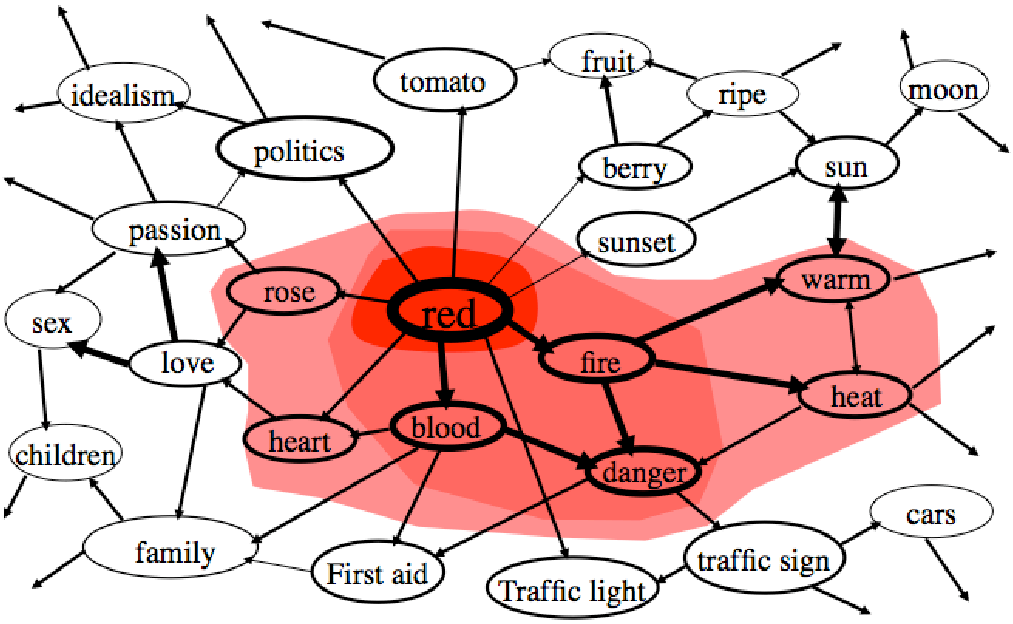
A zombie without such subjective experience would be much more clumsy than a human being in reacting to any new situation—about as clumsy as present-day robots or computers, who lack this complex state of anticipation. In other words, a true zombie would behave *distinguishably* different from a real human being. The zombie thought experiment is actually inspired by a mechanical notion of a cognitive system that is purely reactive, without a complex internal state based on learned connections.

What then is a "quale"?

Consider the **quale** of "redness", a classic example in the philosophy of mind.

Note that the pure experience of red is actually quite unrealistic: we always see red in a context of other phenomena, such as a red rose, a traffic sign or a red Ferrari. Perhaps we could experience pure red only in a laboratory under controlled conditions, where all we see is undifferentiated red light, while all other stimuli, such as sounds, are suppressed. But even then, the phenomenon of neural fatigue would ensure that the experience of redness would weaken and eventually disappear, while making place for varying thoughts and imaginations.

Even if we could remove it from any concrete context, the abstract quality of redness would still remind us of concrete phenomena that it tends to co-occur with, such as: blood, fire, roses, sunsets, a political party, danger signs, red traffic lights… All of these remembered phenomena would be to a smaller or larger extent primed by the view of red.



For example, the association of red with fire, blood and the sun makes us experience red as a "warm" color, in contrast with "cool" colors, such as white and blue. This means that we implicitly expect a red room to be warmer than a blue room. Therefore, the thermostat might be put one degree lower in such a room without us perceiving it as colder (this is another example of feedback from anticipations to perceptions). The association of red with blood, fire, and traffic signs and the fact that it is relatively rare in nature moreover make us experience red as something important, that signals a potential danger. Thus, red tends to activate, to demand attention, in contrast with "calm" colors such as grey, brown or green.

It is the whole of these—stronger or weaker, explicit or implicit—anticipations that together can be said to constitute our "feel" or "quale" of what redness means. The picture above proposes a first visual impression of how this feeling is generated by a process of spreading activation. The initial activation of the concept "red" spreads to associated concepts, proportionally to the strength of the association (stronger associations are depicted by thicker arrows). From these "neighbors" in the associative network, such as "blood", "rose" and "fire", the activation spreads further to concepts, such as "First aid" and "love", that are more indirectly associated, but which may still undergo some degree of priming by the conception of "red".

Emotions

Arousal

The fuzzy state of neural activation defines our general "feeling" about a situation. An **emotion** is something more distinct and intense, with stronger, more focused activation that tends to spread through the body as a whole. This bodily state of activation is called **arousal**. It is typically characterized by higher blood pressure, faster heart rate, deeper breathing, increased sweating, etc., and is triggered by hormones such as adrenalin (epinephrine). This focused activation is similar to access consciousness, except that it appears controlled by the older, "reptilian" parts of the brain, and not by a global workspace or symbolic cognition.

The function of emotion is, like the function of cognition, in the first place preparation. The difference is that cognition prepares for as yet hypothetical events that the mind tries to anticipate. Emotions prepare for actual physical action, not just for interpretation and reflection. That is why they activate the body as well as the brain. Forceful action is needed when a true **diversion** has taken place, i.e. an event that requires an active intervention and change of behavior, not merely a continuation of your present pattern of activity. Examples are an encounter with a tiger, but also a meeting with a person you would like to get to know better.

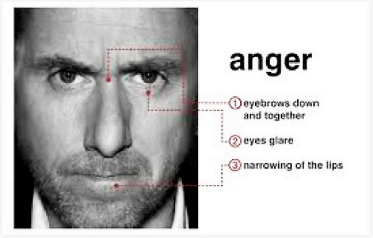
A good metaphor for an emotion is the concept of "force" in Newtonian mechanics. Without any forces working on it, a material body will continue moving in a straight course at a constant speed. Forces are what make the body accelerate, decelerate, stop, start, or change direction. Similarly, a normal pattern of activity moves in a given direction with a more or less constant speed. When a significant diversion occurs, this course of action needs to change so as to take into account this unplanned event. An emotion acts as the mental and physical "force" that produces this change of speed or direction. Similarly, an emotion can function like the driving force that gets you going starting from a state of relative rest or inactivity (e.g. when you are passionate about achieving some goal), or that suppresses your normal activity (e.g. when you are depressed or paralyzed by fear).

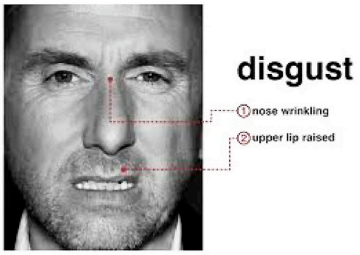
Emotions can be categorized in the first place by the amount of arousal that accompanies them. Most emotions are characterized by more arousal than normal, because the situation demands extra energy, e.g. joy, fear, anger, ... However, sometimes the situation requires less arousal than normal, e.g. in situations of depression, boredom, or contentment, where the change in behavior is towards saving energy or keeping a low profile.

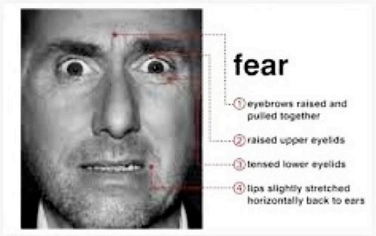
Emotions can be artificially induced by injecting the equivalent of adrenalin into the blood stream to create arousal. People who get such an injection (without knowing what it is) tend to interpret the resulting feeling cognitively as an emotion caused by some phenomenon they perceived. However, people witnessing the same phenomena (e.g. pictures shown by an experimenter) who got an injection with water do not feel any particular emotions about them. This is another example of the feedback between conception, perception and experience, or between mind (feeling) and body (arousal).

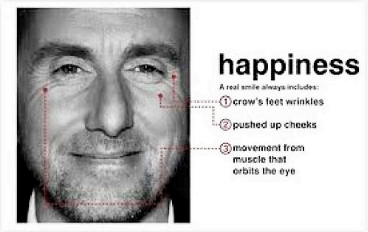
Types of emotion

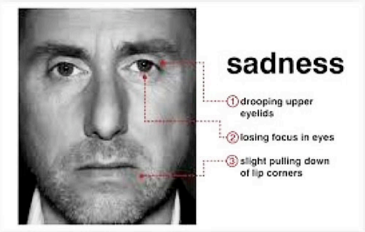
On the basis of facial expressions, psychologists distinguish 6 universal emotions, which are recognized in all cultures. They are: Joy, Sadness, Anger, Fear, Disgust, and Surprise. Sometimes a seventh one is added: Contempt, which is a feeling of moral superiority towards another person.

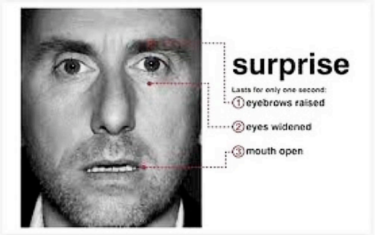


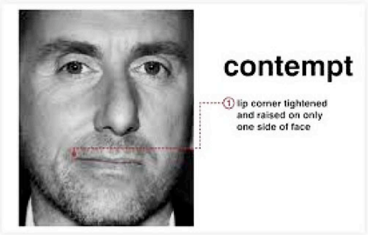



fear
① eyebrows raised and pulled together
② raised upper eyelids
③ tensed lower eyelids
④ lips slightly stretched horizontally back to ears


happiness
A real smile always includes:
① crow's feet wrinkles
② pushed up cheeks
③ movement from muscle that orbits the eye







But there are actually an infinite range of possible emotional states that differ between people, cultures and situations. Many psychologists have proposed models, dimensions and taxonomies for classifying these emotions, but none of them seem to capture all the subtle differences and similarities. On the basis of our cybernetic framework, we can suggest at least the following distinguishing properties.

Emotions are triggered typically by situations with an element of **surprise** (i.e. failure of anticipation) that requires a change of arousal or activation level so as to adjust activities, plans and expectations.

- When the diversions are *positive* (affordances), e.g. receiving a present or getting a promotion, they result in positive, pleasurable feelings such as pride, confidence or satisfaction. When the surprises are *negative* (disturbances), e.g. failing for an exam or being in an accident, they result in feelings of displeasure, sadness or pain. The degree of positivity or negativity of a feeling is called its **valence**.
- Emotions can be about diversions that are expected in the *future*, e.g. hope, fear, confidence, or that have been experienced in the *past*, e.g. guilt, sadness, satisfaction.
- Emotions can be *social*, i.e. concerning your relations with other people, e.g. pride, anger, shame and love, or individual, e.g. self-confidence, fear.
- Emotions can be triggered by something under the person's *control*, e.g. pride (if it went well) and guilt (if it went badly), or not under control, e.g. fear, depression, curiosity.

The properties of arousal, valence, control, time and social relations together allow us to classify and explain the most important emotions. However, like all subjective experience, the cognitive interpretation of a situation is unique for every individual, and dependent on earlier experiences, including cultural learning. This explains why even while there are universal emotions, some aspects are dependent on culture (e.g. shame is a much more important emotion in collectivist cultures, like the Japanese, than in individualist ones, such as the European culture).

Bounded Rationality and Cognitive Biases

Rationality and its limitations

In the science of economics, the fundamental assumption is that human beings are intrinsically **rational**: they try to make the *best possible decisions* whenever they are confronted with a problem in which they have to choose. This makes sense intuitively: why would anybody want to make a less than optimal choice? It also makes sense from the point of view of evolutionary epistemology or evolutionary psychology: if certain types of individuals would have a tendency to systematically make less good decisions, they would lose the competition with more rational individuals, and thus eventually be eliminated by natural selection. Therefore, we may assume that over the millions of years that evolution has had to shape our brain, it would have produced a very efficient decision-making mechanism, which carefully weighs the pros and cons in order to come to a near optimal decision.

On the other hand, we all know that people sometimes make stupid mistakes, are guided more by emotion than by reason, and often behave plainly irrationally, going so far as to engage in behaviors (such as being addicted to gambling, overeating, starting a war, or smoking) that everybody knows to be harmful. Such deviations from assumed rationality must be investigated and explained. Several hypotheses have been proposed.

Bounded rationality

This concept was proposed by Herbert Simon, who was not only one of the founding fathers of cognitive science and artificial intelligence, but a Nobel prize-winning economist. Inspired by his research in problem solving and information processing, Simon noted that most real-world problems are simply too complex for us to find optimal solutions.

Example: suppose you want to buy a car. There are dozens of different models on the market, with different features, qualities, prices and limitations. You can never know all the complexities that go into building a car, nor what will happen with the car in different circumstances, such as during a crash, in wintertime, in very hot weather, when you suddenly need to brake, when you are carrying a lot of luggage, etc. You also do not know in which precise circumstances you will use the car, and therefore you cannot say for sure whether comfort is more important than power, or air conditioning more important than size of the trunk. Even if you had all that information about all the potential car models, you would never be able to design a decision model that takes all these data and criteria into account and that allows you to calculate which car optimally satisfies all these different criteria.

Simon therefore proposed that when making decisions, people do not *optimize* (choose the best possible option), but **satisfice** (choose an option that is good enough to satisfy them, but is probably not the best). However, this still does not tell us how people really make a decision. It also does not explain clearly irrational decisions, such as starting to smoke, that are not only not optimal, but far worse than satisfactory.

Cognitive heuristics

Inspired by Simon's work on problem solving and cognitive psychology, Kahnemann and Tversky [1982; 1996; Kahnemann, 2003] studied human limitations on rationality in much more detail, coming up with a variety of systematic deviations from optimality that they called "**cognitive biases**". Their explanation is rooted in the concept of **heuristic**, which we defined as a rule of thumb that people use to simplify problem solving when the search space is too large to be systematically explored. Heuristics are not guaranteed to find optimal solutions, but because they make search much more efficient, they help you to find acceptable solutions ("satisficing") in a reasonable time.

As we noted, most heuristics are very domain-specific, and therefore it is difficult to derive a general theory of biases or limitations on rationality from the notion of heuristic. However, some heuristics are more general in scope, and thus may explain common mistakes that people make. For example, Tversky and Kahnemann proposed the *representativeness heuristic*, which posits that you can infer the properties of a category of things from the properties of a representative sample of that category.

<u>Example</u>: suppose you are choosing a car, and considering buying a Honda. You ask your neighbor, who owns a Honda, about his experience with that car. The neighbor tells you he had a serious problem with the brakes last year. Following the representativeness heuristic, you may infer that Hondas are not safe, and therefore that you should buy a different brand. However, it is clear that this particular problem with your neighbor's car may be pure bad luck, and not representative for Hondas in general. However, given that you can never get to know all experiences of all Honda owners, you must simplify the problem, and at a certain moment make a decision based on the limited sample that you know.

The problem with heuristics and the notion of bounded rationality is that they are still rooted in the symbolic view of cognition, which tends to see all decision making as based on conscious reflection and manipulation of various explicit concepts or symbols. As we know from connectionist models, most cognitive processes are subsymbolic and subconscious. This means that they do not use general and explicit procedures as implied by the notion of heuristic, but are based on activation spreading along learned associations. Let us further examine this connectionist mechanism to see which biases it implies.

Towards a connectionist theory of cognitive biases

The basis of connectionist models is the mechanism of Hebbian learning or its more sophisticated implementation of delta learning, which strengthen connections that made good predictions, while weakening those that made wrong predictions. It can be shown that through delta learning the strength of a link A → B will eventually, after many experiences of correct or incorrect predictions, converge to the conditional probability P(B|A), i.e. to the proportion of cases in which A was effectively followed by B.

Example: assume there were 150 cases in which A was followed by B, and 50 cases in which B did not follow, then the link strength of A→B becomes equal to P(B|A) = 150/(150 + 50) = 0.75

If this is correct, the learning mechanism appears to be accurate or unbiased: its results correspond to the objective probabilities as experienced by the agent. This means that, although the neural network will of course make mistakes in its predictions, on average these mistakes will cancel each other out: it will make as many mistakes in the one direction (expecting B, when B does not happen) as in the other direction (expecting that B will happen, when it does not).

Definition: a **cognitive bias** is a systematic deviation from the correct average. This means that the average estimate of what is expected to happen will converge to a value that is either significantly higher or significantly lower than the average of what truly happens over an extended period.

Example: in the above case, there would be a bias if link strength (A→B) = 0.85 (overestimate of the probability of B following A), or if link strength (A→B) = 0.5 (underestimate of the probability of B following A).

Variability of the environment

A first source of bias is that Hebbian/delta learning needs time to converge to the correct average. If the agent only experienced a single episode of A followed by B, then the average probability as yet is 1. However, immediately setting the link strength to 1 seems overconfident, since it is dangerous to deduce such certain prediction from a single episode: it is unlikely that because something happened once, it will always happen in the same circumstances. That is why Hebbian or delta learning only makes a relatively small adjustment to the strength of a link with each learning experience, as measured by the learning constant c.

The learning constant moreover takes into account the fact that averages may change over time, e.g. that the probability of B following A may increase because of changed circumstances. For that reason, recent reinforcements of a link contribute more weight than older reinforcements, but not enough weight to immediately erase the effect of older experiences (which include experiencing no association at all between A and B). This principle may be illustrated by the *recency effect*: the last items on a list are remembered better than the preceding ones.

On the other hand, for a single new experience the adjustment (say 5%) may be too low: perhaps this experience is just the first occurrence of many, and if this occurrence is important (e.g. a great danger or opportunity) failing to anticipate the next one because of a too weak adjustment may be fatal. That is why the cognitive system tends to pay special attention to new or unexpected events, which may signal a drastic change in the situation. This may be illustrated by memory biases such as the *primacy effect*, which—as a complement to the *recency effect*—notes that people remember the first items from a list better than the following ones, and the *von Restorff effect*, which notes that an item that stands out (e.g. because of being marked in a different color) is also remembered better.

Hyperbolic discounting

The variability of the environment can be used to explain one of the most studied biases: *hyperbolic discounting*. This means that when people have to choose between two rewards they tend to prefer a smaller one that comes sooner (for example, receive $100 now) to a larger one that comes later (for example, receive $200 in six months). This makes sense: the longer you wait, the larger the chance that some unknown environmental variation or disturbance would occur that prevents you from getting the expected reward (for example, the person who promised you the $200 changes his mind, dies, or loses all his money).

What is less obvious is that this "discounting" (gradual reduction) of the value of a reward over time happens more quickly in the immediate future than in the farther-away future. For example, when offered the choice between $100 in a year, and $200 in a year and half, most people would prefer the $200 option, even though the interval between the options is just as long as in the first situation, and therefore the probability of some disturbance intervening in between would seem to be just as large. But this reasoning assumes that the probability of unpredictable disturbances (uncertainty) would remain the same. In a truly variable environment, not only situations but probabilities can change in unexpected ways. The safest assumption is that whatever uncertainty you have about the near future, this uncertainty (i.e. the probability of something going wrong) will only get bigger in the farther-away future. Therefore, opportunities in the immediate future are more reliable, while opportunities in a farther-away future are intrinsically much more difficult to estimate. Therefore you might as well prefer the very uncertain $200 in 18 months over the almost as uncertain $100 in 12 months.

Utility more important than probability

Another important source of biases is that learning mechanisms are not there just to estimate the conditional probabilities of events (Hebbian learning), but the average reward or punishment received (**reinforcement learning**). For an agent trying to achieve its goals of survival and proliferation in an uncertain environment, the primary function of the cognitive system is to maximally achieve those goals, not to accurately predict what will happen. From the evolutionary and cybernetic perspectives the most important objective is **control**. Prediction is useful only insofar as it helps achieve control. Therefore, the cognitive system can afford to make systematic errors in prediction as long as it minimizes its errors in control. Maximizing control means maximizing the expected reward, value, or utility achieved by actions, while minimizing the risk of serious accidents. This means that we can expect a cognitive bias towards the outcome with the *highest average utility*. This is normally the outcome that would receive the strongest accumulated reward and the weakest accumulated punishment.

We can divide this bias in two subsections:

1. avoiding the danger of strong punishment. This leads to the biases of **paranoia**: people tend to overestimate the probability of serious external threats or dangers being realized.
2. seeking the opportunity of large reward. This leads to the biases of **optimism**: people tend to overestimate their own ability to achieve positive outcomes.

Together, they explain why people tend to behave like "paranoid optimists" [Haselton & Nettle, 2006].

Examples of paranoid biases:

- *Phobias* are exaggerated fears for potentially (but usually not so) dangerous things, such as spiders, snakes, and heights. The probability of falling from the grassy border of a high cliff is not larger than the probability of falling from your lawn into your flowerbeds—both of which are very small. Yet, people will not hesitate walking just next to the flowerbeds, while they may start to panic when coming anywhere near the edge of the cliff. The reason is simply that the punishment in case of falling from a cliff is immensely greater than the punishment in case of falling into a flowerbed. Avoiding the very small risk of this happening is worth the relatively small effort that is needed to stay far away from the edge. This fear of heights makes perfect rational sense, but the same exaggerated reaction tends to appear in situations where the objective risks are so small as to be negligible, like in fear of flying, or fear of speaking in public.
- The *negativity bias* is a tendency for people to pay much more attention to bad news or to potential signs of danger, than to good news or signs of opportunity. For example, the news that 100 people died in a fire will attract much more attention than the news that life expectancy increased last year with a month, even though the latter implies that many more than 100 people who would otherwise have died now survived. This makes sense because of the utility bias: ignoring the danger of dying in a fire may cost you much more dearly than ignoring the fact that people live longer now. In the first case, there is something you may potentially do to avoid the danger. In the second case, there is not much that you yourself need to do.
- *Loss aversion* refers to the observation that people would rather miss the opportunity to gain $200 than run the risk to lose $100, even when both events have the same probability.

> Losing something is experienced as more negative than gaining the same thing is considered positive. In another experiment, people may not want to pay more than $1 in order to get a coffee mug, but once they have it, they will not sell it for anything less than $2: the cup becomes more valuable for them simply because they now own it. This may be explained by the fact that a loss, because it is negative, is considered *a priori* more important than an equivalent gain, which is positive.

<u>Examples of overconfidence biases:</u>

The negativity bias in emotions is compensated by something called the "*positivity offset*": generally people tend to feel good about their situation, and to be optimistic about the future. When people are asked to estimate the probability of getting good outcomes for themselves (e.g. succeeding in an exam, or answering a question correctly), their estimate tends to be higher than the real outcome. Moreover, a large majority of people think that they are smarter, nicer, or happier than the average person, while objectively less than 50% should fall into that category.

While these misjudgments seem rather innocent, the bias for overoptimism can lead to dangerous behaviors, as exhibited by the gambling addict who is convinced that next time he will win big, the skier who takes exaggerated risks, or the smoker who believes that it is only the others who will get lung cancer.

A plausible explanation for such overconfidence is that it motivates people to actively seek opportunities and overcome problems, even when objectively their chances of success are only small. If the cost of being active is less than the cost of potentially missing out on a great opportunity because you thought that it was anyway not worth the effort, then the bias goes towards assuming that activity will be successful. It is sufficient that on average being confident, and therefore active, gets bigger rewards than being pessimistic, and therefore passive—even when the pessimists tend to be more accurate in their predictions.

An example that combines paranoia and overconfidence is the observation that most people feel much safer driving a car than sitting in a plane, even though innumerably more people die in car accidents than in plane crashes. The reason is that the driver of the car feels in control and therefore tends to assume he will be able to deal with any problems. Passengers in a car feel somewhat less safe, because they have no control, but still safer than plane passengers. Planes are particularly frightening because plane crashes often appear in the news because of the negativity bias—unlike car crashes, which have typically fewer victims, and are intrinsically less newsworthy because they are so common. Moreover, fear of flying may be inspired by fear of heights, which is a general instinct.

Immediate context more important than overall statistics

Another important source of biases is the fact that cognitive processes function on the basis of **spreading activation**: concepts, memories and experiences become activated depending on how closely connected they are to the immediate **context**, i.e. the cognitive items that are being activated (by thought or perception) at this moment, together with the ones that were recently activated so that a low level of activation or priming remains, making it easier to activate them again. This means that what comes most easily to mind are the things that are most directly associated with the present situation: things said, observed, remembered or considered now or in the recent past. These are also the things that seem subjectively most important or most probable.

On the other hand, probability is calculated formally as the total number of times some phenomenon occurred divided by the total number of times that this phenomenon could have occurred. For example, the probability of a coin falling with heads up is 50%, because if you had tossed a coin 100 times, it would have produced "heads" about 50 times. Calculating such a "base rate" would require an exhaustive search through long-term memory of all known occurrences of that event (e.g. remembering all the times you witnessed a coin being tossed, and the outcomes of all these events). Given the limitations of long-term memory, this is obviously not possible. Moreover, as we noted when discussing the learning constant, there is no guarantee that because something happened with a certain frequency in the past, it would happen with the same frequency in the future.

Therefore, the cognitive system implicitly relies on a "context-dependency heuristic", which assumes that recent or easily remembered occurrences are more important than the long-term "base rate". It does this by first activating those immediate memories, and using them as a starting point for further propagation of activation. At the same time, it tends to ignore or neglect the base rate (this is called the *base rate fallacy*).

Examples of contextual biases:

- The *anchoring bias* can be demonstrated by a little experiment in which people are asked to randomly choose a number between 0 and 1000, and write it down. They are then asked to estimate the year in which the Mongols first invaded Europe. It turns out that the answer they give is strongly correlated with the random number: people who chose a smaller number tend to guess an earlier date for the Mongol invasion. Since the two numbers are by definition independent, this is clearly irrational. The explanation is that the random number remains in short-term memory as an "anchor" from which activation is propagated in order to build an estimate.
- The *conjunction fallacy* may be illustrated by an experiment in which people are given a description of an independent, intellectual and politically engaged woman. They are then asked to estimate the probability that this woman would be: (a) a bank teller; (b) a bank teller and active in the feminist movement. They typically estimate a higher probability for case (b), even though all (b) cases are also (a) cases, and therefore the probability of (b) cannot be larger than the one of (a)! The larger base rate (frequency of occurrence) of (a) is simply ignored. The reason is that the description of the woman has strong associations with the

profile of a typical feminist, but not with the profile of a typical bank teller. Therefore, more activation propagates to (b) than to (a).

- The *confirmation bias* is the tendency to pay special attention to information that confirms your preconceptions, while neglecting information that contradicts them. For example, a racist will especially notice all news reports in which a crime was perpetrated by someone from another race, considering it as evidence that confirms his prejudices, and ignore most of the crimes performed by someone from the same race. The reason is that the preconception primes the cognitive system to quickly recognize situations that fit in with the preconception. In the racist example, there exist strong associative links between the concepts of "other race" and "crime": whenever the one gets activated, a little activation will spread to the other, priming it to immediately become fully activated if further evidence appears. The ensuing co-activation of the concepts will further strengthen their association because of Hebbian learning. No such priming occurs with the concept of "own race", making the effect of additional evidence here much weaker.

Conclusion

Given the intrinsic uncertainty and complexity of the world, it is clear that cognition cannot in general make optimal decisions or perfect predictions. Therefore, rationality is bounded. However, we may assume that evolution would at least have avoided a cognitive system that makes systematic mistakes. Yet, many dozens of cognitive biases have been observed in a variety of psychological experiments and social or economic situations. (For an extensive list, see http://en.wikipedia.org/wiki/List_of_cognitive_biases). We have proposed a number of plausible, general explanations for such systematic deviations from rationality.

Most generally, cognition uses heuristics to simplify complex problems, thus systematically ignoring factors that are difficult to establish or process, such as base rates, in favor of more easily available data, such as contextual cues. Moreover, what seems to be most rational, such as calculating probabilities on the basis of formal statistics, may in practice not be the most useful for survival. When long-term fitness is taken into account, paying special attention to serious dangers or remaining confident in spite of low probabilities of success may be the best general strategy, even when this results in a "wrong" estimate of probabilities.

As a practical guideline, it seems best to be maximally aware both of your natural cognitive biases and of the formal rules of statistics. Then you can use either the one or the other (or a combination of both), depending on what you know about the situation, in order to make decisions. Sometimes, the situation is too ill defined or variable for statistical rules to apply, and then you can better rely on your intuition. In other cases, the laws of probability apply unambiguously, and you can decide without hesitation, e.g., that being a bank teller is more probable than being a feminist bank teller. Even when statistical rules are not readily applicable, however, it is worth being aware of your cognitive limitations, and e.g. look systematically for contradicting evidence in order to counter the confirmation bias, or check the statistics about plane crashes before worrying about boarding that plane…

Individual Differences

Differences in cognitive competence

It is well known that some people achieve much more on the cognitive level than others: they are able to tackle very complex problems, or exhibit unusual feats of creativity. There are plenty of examples of people that are considered to be a scientific, artistic or political genius: da Vinci, Einstein, Edison, Mozart, Caesar, Gandhi... Apparently their cognition functions better than the one of average people. On the other hand, some people have great difficulty with a normal level of cognitive functioning, e.g. mentally retarded people or people with the Down syndrome. Given that we all have essentially the same brain organization and basic cognitive abilities, what causes these differences in cognitive performance? Let us review the main components of cognitive ability.

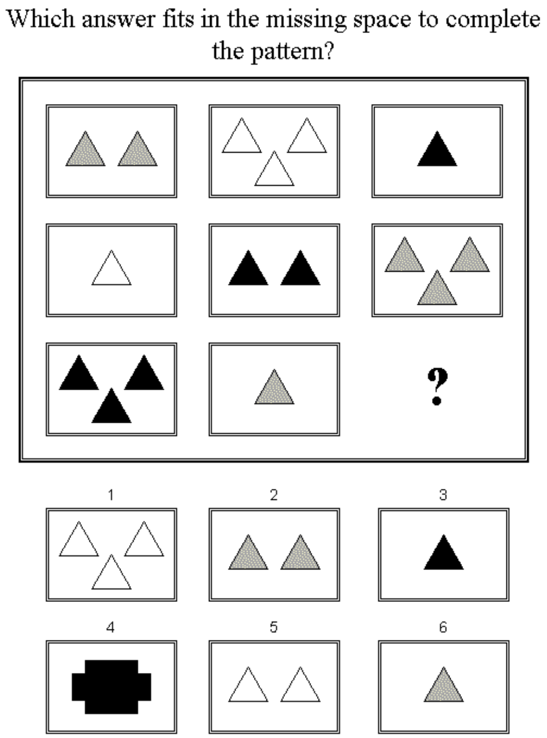
Expertise

A first important component is the amount of specialized or advanced knowledge, or what may be called **expertise**. As we saw when discussing AI, intelligence is useless without plenty of knowledge. Learning takes a lot of time, effort, and undergoing a variety of experiences. Not only the quantity, but also the quality of the experience is important: the information input must be good. This implies good education and training, the reading of good books, etc. It is estimated that to become an expert in any advanced field (e.g. painting, physics, politics, chess…) typically takes at least ten years of hard work. This requires at least sufficient motivation, discipline and patience to achieve it, and an environment that provides the right challenges and feedbacks.

Intelligence

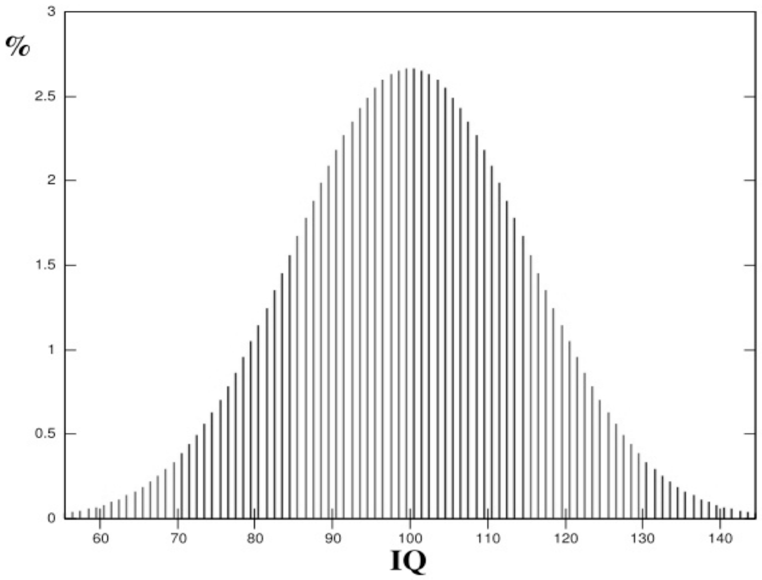
A more general component of cognitive competence, in the sense that it does not require specific knowledge, is what is conventionally known as **intelligence**, i.e. the general ability to solve problems. A well-known way to conceptualize the difference between knowledge-dependent and knowledge-independent factors is Cattell's [1963] distinction between "fluid" and "crystallized" intelligence. Crystallized intelligence is the result of the accumulated knowledge and experience that we bring to tackle common, concrete problems. It typically increases unrestrictedly with age. Fluid intelligence is the quickness and versatility of thinking that is needed to solve the most abstract, novel problems. Fluid intelligence increases during childhood, but reaches a plateau by the end of puberty (around 16 years) and tends to decrease with older age.

IQ scores


Which answer fits in the missing space to complete the pattern?
?
1
2
3
4
5
6

Intelligence is normally measured by **IQ** tests. These combine a variety of multiple choice questions on different subjects and tasks domains. While neither of the questions on its own gives a reliable estimate of intelligence, in the sense that they all require some specific expertise, the more questions there are in the test, and the more diverse they are, the better the total score distinguishes more from less intelligent individuals. This is because highly intelligent individuals typically have gathered a very broad experience, and tend to be better than average in practically any cognitive domain. The most reliable tests of fluid intelligence require anticipation of abstract patterns. A classic example of such a test is called “Raven's Progressive Matrices”, in which a pattern of abstract pictures is shown and the subject needs to choose which other picture must be added (see picture).

To get a reliable IQ score, multiple IQ tests must be taken, since results may vary depending on the type of test, the degree of concentration or tiredness, etc. The average of several tests, usually at least three, taken within one year, may be considered that person's "true" IQ score.



IQ tests are normally calibrated so that the average intelligence level of the population is 100 and the standard deviation 15. The distribution of scores around the average has the bell-shape of the well-known Gauss curve, as depicted below. This means that about 68% of people fall within one standard deviation from the average, i.e. in the range 85-115, about 95% in the range 70-130, and about 99.7% in the range 55-145. In other words, less than 0.15% of the people have an IQ higher than 145.

IQ and success in life

The scores on IQ tests are obviously not perfect, but they are in general good indicators of future performance, including success in education, success in professional life, eventual level of income, and even life expectancy [Gottfredson, 1997, 2004]. The latter can be explained by the fact that more intelligent people know more about health, understand better what they should do to remain healthy, and manage better to foresee and deal with the type of dangerous situations that lead to accidents or illness. In fact, IQ is a better indicator of future life chances than any other psychological measure. This fits in with our general philosophy that cognitive ability is what allows an agent to cope with the problems it encounters, and thus better survive and thrive in its environment.

Socially, IQ is very important in the job market: the most difficult, important and best-paid professions, such as manager, doctor, lawyer, engineer, researcher, etc. typically require a high minimum intelligence level, and the higher the IQ, the higher the level a person is likely to achieve within the profession. As technologies become more complex and the demands for productivity increase, even less well-paid jobs demand increasingly advanced cognitive skills. This may explain why in advanced societies there is a tendency towards the creation of a class of long-term unemployed, who simply cannot find a job whose intellectual demands are in line with their (limited) abilities. In addition to unemployment, people with a low IQ are also more likely to fall prey to a variety of social problems, such as poverty, drug addiction, criminal behavior, and becoming a single mother.

It has further been observed that the economic development of a country or region is strongly correlated with its average IQ level [Lynn & Vanhanen, 2002]: the national IQ appears to be highest in the states that are wealthiest (mostly Western) and/or fastest growing (such as China and South-East Asia). As a rough approximation, an increase of 10 points in mean IQ corresponds to a doubling of the per capita GDP.

Origin of IQ differences

Differences in IQ are dependent to an important extent on genes or inheritance [Jensen, 1998]. This was measured by comparing the IQ of identical twins (who are genetically identical) that were raised separately (e.g. adopted into different families). Their difference in IQ is on average smaller than the difference between non-identical twins (who are genetically different) that were raised together. However, IQ is also strongly influenced by the socio-economic environment. This is spectacularly illustrated by the **Flynn effect**: the on-going increase in IQ of the population.

The Flynn effect

In the developed countries (for which enough data are available), average IQ has been increasing with about 3 points per decade over the last century [Flynn, 1987; Neisser et al., 1996]. This means that people now are on average some 20 IQ points smarter than people in 1940. People with a perfectly normal IQ of 90 then would according to present norms merely score 70, i.e. as having a mild form of mental retardation! Since people now are genetically almost identical to the people in 1940, this means that the (important) changes in environment must be responsible for this increase.

The most probable explanation is that higher intelligence results from the presently better life circumstances:

- better health care (less serious illnesses that can delay or damage brain development)
- better nutrition (more proteins, fats, vitamins, minerals such as iron and iodine, etc. to build and support the brain)
- higher levels of education (although the direct effect of schooling on IQ seems relatively small).
- higher cognitive stimulation by an increasingly complex environment

This last factor may be particularly important. Indeed, our everyday world offers ever more information to be processed ever more quickly—in the form of advertisements, news items, magazine articles, movies, television, computer games, Internet, etc.—and this requires ever more activity from the brain, thus “training” it to become more intelligent.

Another plausible factor contributing to intelligence increases is that families have become smaller: with fewer children, parents have more attention and resources to invest in each child. The effect on intelligence is confirmed by the observation that first-born or single-born children are on average some 2 to 3 IQ points smarter than second or third-born children, who had to compete with their siblings for parental attention.

Finally, it is worth mentioning Flynn’s own, more skeptical interpretation of his observations. In his view, what has increased is not so much our general intelligence but our ability to reason on a more abstract, formal level. He illustrates this with observations of people from non-industrial societies (such as Siberian hunter-gatherers) who are obviously very knowledgeable, adaptive and skilled in their own domain, but who cannot seem to make the kind of “logical” classifications and inferences that we take for granted, because they always start from the concrete context they know. For example, they may classify rabbits together with chickens because both can be eaten, ignoring the abstract categories of “mammal” or “bird”. Or if told that all bears in the Arctic are white, and that Peter shot a bear in the Arctic, and then asked what the color of the bear was, they would answer: “Well, brown, of course. All the bears I have ever seen were brown!”

Flynn concludes that the counterfactual reasoning and abstract classification that we expect in such problems is a highly advanced skill that needs to be learned. Obviously, our technological society puts increasingly strong emphasis on teaching such skills to its members. Since IQ tests typically make use of such skills, people will score higher on them because they have mastered these skills better, not because they have become intrinsically more intelligent…

Creativity

Creativity is another general aspect of cognitive competence. However, it is more difficult to measure than IQ. Indeed, who is judge that some solution is more creative than another one? Sometimes creativity is measured by divergent thinking skills. Convergent thinking refers to systematic problem solving that converges on the one correct solution. This is typically measured by IQ tests. Divergent thinking, on the other hand, is seen in something like brainstorming, where many different possible approaches are proposed. Divergent thinking can be measured by the number of different "solutions" generated in a given time span to an open-ended question such as: how many uses can you think of for a brick?

It turns out that creativity is correlated with IQ, but not exactly the same. A minimum IQ of about 140 seems necessary for exceptional achievement, such as the one exhibit by creative geniuses like Einstein or da Vinci. For higher IQs, however, there is no clear correlation with creativity.

Giftedness

"Giftedness" is defined as a person's potential for exceptional mental achievement. It can be seen as a combination of unusually high intelligence, creativity, and motivation to make cognitive advances (in domains such as art, science, literature, politics, etc.). Unfortunately, this potential is often not realized. This is in part because gifted people need support from their social environment, which they don't always get because they are not being recognized as gifted. Failure to realize the potential can also happen because gifted people can be emotionally insecure, and therefore doubt so much about their goals and abilities that they never manage to realize what they are capable of.

A gifted person exhibits a typical personality profile, characterized by the following traits:

- Cognition: excellent problem-solver, makes unusual associations, original ideas, vivid imagination, fast learner, good memory
- Feelings: sensitive, intense, passionate, "overexcitable", good sense of humor
- Motivation: ambitious, risk-taking, very curious, broad interests, persistent, perfectionist
- Social relations: non-conformist, autonomous, feels different from other people, feels empathy and compassion for others, has strong sense of ethics and justice

The g-factor

The g-factor ("g" stands for "general") is the most general factor underlying intelligence, IQ and all forms of cognitive ability [Jensen, 1998]. It seems in practice equivalent to what we have called *fluid intelligence*, and is probably the basis for giftedness.

The g-factor is derived statistically, as the factor that all tests of mental ability have in common. Such a factor exists because different test scores are all positively correlated, including tests of apparently very different abilities such as spatial insight, verbal fluency, general knowledge, extent of vocabulary, technical understanding, mathematical ability, abstract reasoning, pattern recognition, creativity (divergent thinking), and even so-called "emotional intelligence". In practice, this means that someone who has a more (or less) than average creativity is also likely to have a more (or less) than average vocabulary, spatial insight, verbal fluency, etc. The reason is that all these abilities, while measurably different, make use of a common core, the general cognitive ability or g-factor. If that factor is stronger (weaker) than average, then most of these abilities will also tend to be stronger (weaker) than average.

The g-factor can be interpreted as a measure of information processing efficiency: how well does the brain handle incoming information? It is positively correlated with a number of physical-psychological characteristics:

- brain volume
- energy efficiency of the brain (more intelligent people require less energy to perform a given task once they have had a little training with it [Haier, 1993]).
- size of working memory
- reaction speed (more intelligent people react more quickly on very simple tasks, such as pushing a button if a word they are shown represents an animal)

These observations have led to a number of neural hypotheses to explain the differences in g between individuals, looking for physical characteristics of the brain that may lead to increased efficiency:

- glia: these support cells that bring energy to the neurons have been found to be more numerous in Einstein's brain
- myelin: this fatty substance that provides electrical insulation around neural connections (axons) may explain why electrical signals propagate better in some brains. Together with the glia, myelin occupies a sizeable part of the brain, possibly explaining why more efficient brains also tend to be larger.
- neural plasticity: easier formation of connections (synapses and axons) between neurons; mice genetically engineered to develop neural connections more quickly appear to be behave more intelligently.
- metabolism: more efficient energy production, e.g. because of more small blood vessels in the brain; fluid intelligence decreases in old age when atherosclerosis makes blood circulation more difficult

The hypothesis of neural propagation depth

For my own hypothesis about the origin of g, I start from the general connectionist interpretation of cognitive processes, which is based on activation propagating from neuron to neuron across synapses. This spreading activation process dissipates much energy and is prone to making errors. Therefore, we may assume that activation weakens with each crossing of a synapse. We have seen that propagation stops when activation drops below the threshold for activation of the neuron. This leads us to the following.

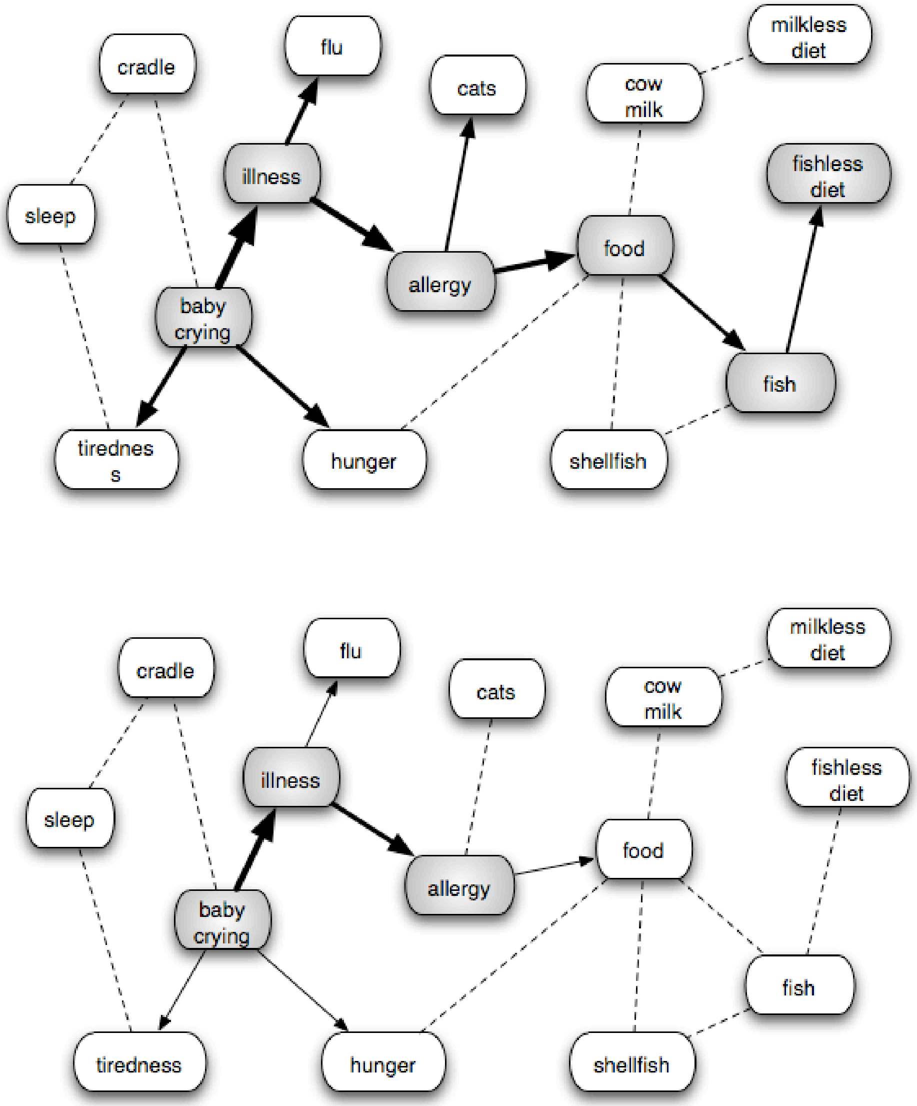
Definition: **Propagation depth** is the average number of steps a coherent propagation process (train of thought) takes before it stops

Let us apply this to the process of problem solving, which we have introduced as the essence of intelligence. A problem is represented by an initial combination of concepts, its solution by a final combination of concepts that satisfies certain criteria. Problem solving is then the propagation of activation along associations between concepts so as to reach the final state from the initial state. For example, the problem that a baby cries repeatedly might be solved via the following, 6-step association path:

> baby cries (initial state) → illness → allergy → food allergy → fish allergy → fishless diet (solution)

Assume now a brain characterized by a lower propagation depth of 3 steps. This would result in the following train of thought:

> baby cries → illness → allergy? (process stops, but no actual solution found)



The figure illustrates the activation spreading over a network of associated concepts. Highlighted nodes represent activated concepts, dotted lines represent associations, and solid arrows represent the amount of activation propagated over an association (thicker arrows = more activation). Because of decay, activation decreases with subsequent steps in the propagation process. The diagram above represents a network with relatively high propagation depth, where six concepts in sequence get activated. In the diagram below, there is much more decay so that activation stops propagating after three steps.

This illustrates how a higher propagation depth, as could be expected in more efficient brains, would lead to better problem-solving abilities and better cognitive performance. Assume that giftedness is proportional to the g-factor, and that g is proportional to propagation depth D.

Higher D implies:

- more problems can be solved because of wider propagation of activation
- larger working memory, because more concepts can remain simultaneously activated

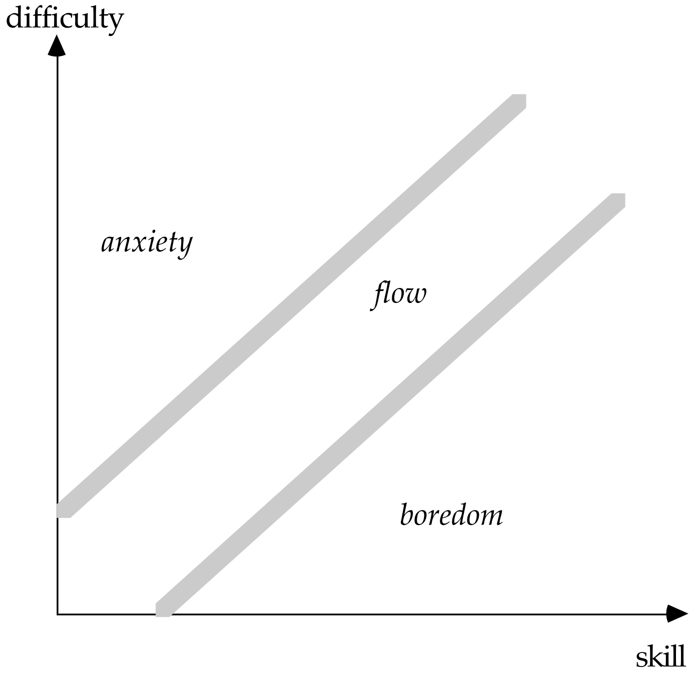
- more “far-fetched” associations (concepts are combined that have no direct associative connections)
- higher ability of abstraction (ability to classify apparently very different phenomena, such as cars and motorboats, into the same category by noticing the weak or indirect associations that they have in common)
- stronger co-activation of remote concepts and therefore better Hebbian learning of new association betweens these concepts, and as a result:
- better long-term memory for these associations

These implications of higher propagation depth can also explain other traits of gifted people in the domain of imagination and feeling:

- Perceptual sensitivity: activation propagates more efficiently from sensory stimuli (perceptions) to conscious conceptions
- Emotional intensity: perceptions and conceptions elicit more intense and detailed feelings, via the same mechanism of enhanced propagation of activation
- Vivid imagination: conceptions more easily activate detailed perceptual memories which may appear in more unusual combinations
- Empathy: imagining oneself in someone else’s place, and feeling the accompanying emotions, becomes easier as well, thus leading to a stronger sense of compassion, and therefore of justice.

Interaction between intelligence and motivation

The psychologist Csikszentmihalyi [1990, 2004] observed that people seek “flow” in their activities. Flow is a pleasurable, happy, focused state, characterized by a feeling of being in control, and of advancing smoothly towards one's goals. It requires a balance of challenges (problems, task difficulty) with skills or competences. When the challenges are higher than the skills (i.e. the task is too difficult), the person will feel stress and anxiety. When the challenges are lower than the skills (i.e. the task is too easy), the person will feel boredom.



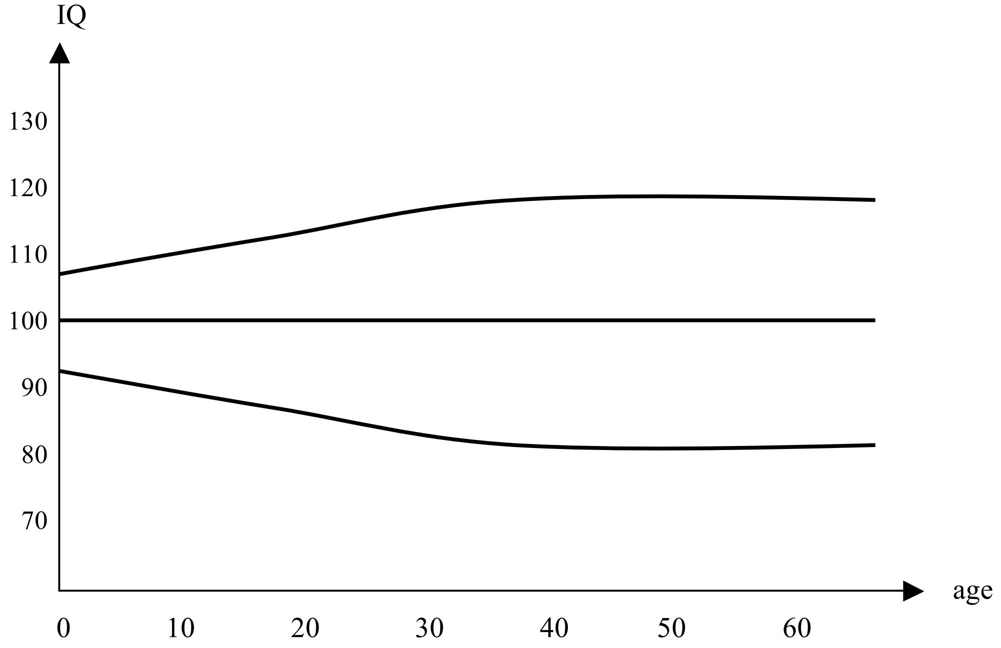
From this perspective, to be happy in your activities, you should find the right challenge level: not too easy, not too difficult. However, as people become more skilled because of learning and experience, they need to raise their challenges to remain satisfied. Therefore, people will gradually take on more ambitious tasks (e.g. more advanced studies, jobs with more responsibility, more complex works of art, more advanced chess levels, etc.).

Gifted people are particularly skilled at information processing and problem solving. According to the flow logic, this implies that they will be ambitious, perfectionist, and seek difficult challenges. They are also skilled at learning. This explains their intense curiosity and wide range of interests. As such they will progress quickly in any domain that they choose to explore. The more progress they make, the more they will be motivated to go even further, and tackle even more complex problems. Thus, their brain will be ever more stimulated to develop itself.

Unfortunately, the same positive feedback cycle applies to people with a low intelligence level, but now working in the opposite direction. When they notice that they are not doing as well as the others in common challenges (e.g. when studying at school), they will seek easier challenges (e.g. drop out of school, or choose an easier study domain). This will in turn reduce their cognitive stimulation. Lower stimulation (less or easier problems to solve or information to process) means a less active brain, and therefore a reduced training of the neurons, synapses, glia, blood vessels and other active components of the brain. The result is likely to be a reduction of their propagation depth, g factor, or IQ relative to people of average IQ.

This mechanism has been proposed by Dickens and Flynn [2001] to explain why there is at the same time such a strong correlation between IQ and genetic predisposition (as illustrated by twin studies) and such a strong effect of environmental stimulation (as illustrated by the Flynn effect).

According to this hypothesis, small differences in IQ determined by the genes are amplified by the positive feedback sketched above: more intelligent people tend to seek more cognitively stimulating environments, thus becoming even more intelligent with respect to the average, while less intelligent people do the opposite, becoming even less intelligent with respect to the average. The resulting divergence is depicted in the picture below, which shows how a hypothetical IQ difference at birth (0 years) increases during the period of major cognitive development (until about 35 years) and then stabilizes.



Amplifying intelligence

The conclusion must be that even intellectually disadvantaged people should be maximally stimulated to develop themselves cognitively, albeit in such a way that they remain in “flow”, i.e. don’t become anxious because the tasks are intrinsically too difficult for them. The cybernetic paradigm suggests an efficient method to promote flow: concrete and immediate feedback, so that neural mechanisms that produce good decisions are reinforced, and those that produce bad decisions are suppressed. The immediacy of such feedback means that there is no need to sustain activation over prolonged periods in working memory, which is typically a heavy burden on the brain because of the mechanisms of decay and fatigue.

An example of a very efficient environment to produce such flow is a computer game: every action by the player produces immediate and clear feedback, keeping the attention focused, and helping the player to improve his or her performance. Moreover, games normally have many different “levels” of difficulty, so that both novice and expert can find the right level of challenge. It is precisely the continuing flow experience that makes computer games so addictive, keeping people motivated to play it again and again. However, if the game would be designed so as to promote core cognitive skills, it could be turned into a very powerful educational tool that would stimulate even low-IQ people to maximally develop their intellectual capacity.

The nutrition hypothesis for the Flynn effect suggests another approach to increase intelligence: nutritional supplements that help the brain to develop neuronal connections and to produce enough energy to sustain high levels of activation. A number of supplements have been investigated as potential "smart drugs" that support cognitive functioning. These include the herb Ginkgo Biloba, caffeine, Omega-3 fatty acids, Acetyl-L-Carnitine, vinpocetine, and the synthetic drug piracetam. Their effect tends to be most pronounced in people who have (temporary or accumulated) cognitive difficulties, e.g. due to fatigue or old age. There is as yet no evidence for long-term augmentation of IQ, although the Flynn effect suggests that this may be possible, especially if the supplements would be taken from an early age by the people who are most disadvantaged.

Problems of the gifted

Being exceptionally intelligent does not only have benefits. In their relations with others, highly gifted individuals tend to question authority and accepted wisdom, as they prefer to think for themselves. This may obviously get them in problems in society, as illustrated by Galileo's conflict with the Catholic Church. They also often feel alienated or out of step with other people. This follows from the flow model: assuming that the cognitive skills of a gifted person are much larger than the skills of a "normal" person (i.e. of average intelligence), a typical interaction or conversation between the two will result in anxiety for the normal person, who cannot follow half of what the other says, and in boredom for the gifted person, who already has anticipated most of what the other has to say.

Another common problem is that the gifted may not be recognized as such: they have a too broad range of interests, and thus do not fit the "expert" stereotype that people have of smart people as being exceptionally good in specialized technical domains such as mathematics, chess, engineering, ... Because of these stereotypes they rarely recognize themselves as gifted, and therefore usually do not understand why they seem to be so different from others. They often feel lonely or misfit, and tend to accept dismissive views of others. Other common problems are that because of their very wide range of interests they do not know what to focus on, that they can be so perfectionist that they never finish their work, that they have unrealistic expectations of others, and that they can be emotionally and physically too sensitive. While psychologists have found no real evidence for the widespread notion that genius and madness are related, these common problems of the gifted may explain why they are sometimes perceived as such…

Collective cognition

Collective intelligence

Cognitive superagents

We have defined cognition as a systems phenomenon. This means that it does not need to remain limited to individual agents. An organized collective of agents, such as a society, an organization or a group, defines a **supersystem**. If we assume that the members of this group share goals or values and act cooperatively to reach these goals, then we can interpret this supersystem as an agent in its own right: a superagent. Our general conception of how an agent functions implies that this superagent too must exhibit cognition: it must process its perceptual information in order to decide about the most appropriate action to achieve its goals. The question now is in how far such group cognition is different from the cognition of its individual members.

Collective action is in general physically more powerful than individual action: a hundred individuals pushing together can move a much heavier weight than a single individual. However, if we assume that all individual agents would react in exactly the same way to the same perceived conditions, then the collective would react in the same way as well. The collective decision as to what action to perform would therefore be identical to the individual decision. In this case, the collective would be just as intelligent as an individual agent: it would lack any group advantage.

However, different individuals normally have different forms of knowledge or expertise, because they underwent different experiences. If we could add all that knowledge together, then the collective would know more than any of its members, and be able to make wiser decisions. Moreover, the collective might distribute the information processing to be done across different members or subgroups (division of labor), so that the system as a whole would be able to tackle more complex problems that demand more information processing than a single individual can perform. (For example, when a large search space needs to be explored, the space might be divided into segments, and each individual would be responsible for searching a particular segment.) In that case, we may say that the system exhibits **collective intelligence**.

The first problem, however, is that there is no simple and obvious way to “add up” the knowledge that is implemented as different neural networks residing in independent brains or the solutions that have been found by different individuals searching in different places. This is the problem of **aggregation** of knowledge.

Reducing errors through aggregation

The simplest and most common aggregation method for making collective decisions is voting: when a choice needs to be made between alternatives, every individual expresses his or her preference. The preferences are added up, determining the overall preference for each alternative as a percentage of the number of votes. Typically, the option with the highest number of votes is then chosen. This method has shortcomings, though, since it can lead to a choice that a majority of participants dislike, and since it ignores the contributions of all those whose preference failed to get enough votes.

A more accurate aggregation method can be used if the decision to be made is quantifiable, i.e. when the preferred alternative can be expressed as a number. Examples of such decision problems are: how many beans does that jar contain, or how many pounds does that cow weigh? In this situation, all individuals propose a number as their best guess. The *average* of all the numbers is then calculated. Perhaps surprisingly, in many cases it turns out that this average is much more accurate than any of the individual guesses!

The explanation is simple: because of limited experience each individual has a certain bias. This bias results in an error, i.e. a deviation from the correct number. We will assume, however, that these errors are *independent*, i.e. they are randomly distributed around the correct number. This means that the probability of a deviation in one direction (too large) is the same as a deviation in the other direction (too small). The different deviations therefore tend to cancel each other out when added up to calculate the average. Because of the statistical “law of large numbers”, the more people participate in the guessing, the more “canceling” occurs, and therefore the closer the average guess will approximate the correct number.

Example: This principle can be illustrated by a simulation made by Norman Johnson [1998]: different software agents explore a maze (labyrinth) until they find the exit. They are programmed to remember the overall shortest path they found, i.e. which option to choose at each intersection of routes (excluding the detours they made that only brought them back to the initial path). The chosen option gets a preference of 1; the other options get a preference of 0. All preferences of all agents for each intersection are then added up or averaged, defining a collective preference for each option at each intersection. A new path is constructed that takes the option with the highest collective preference at each intersection. It turns out that this “collective” path is shorter than any of the individually found paths. This is because for individual paths remaining deviations from the shortest possible paths are as likely to be in the one direction as in the other. Therefore, these deviations tend to cancel each other out when the preferences are averaged.

Limitations on collective intelligence

This aggregation mechanism will not eliminate errors or biases that go in the same direction, because then the errors accumulate instead of canceling each other out. This can have several possible reasons.

1) agents have similar biases

Example: all farmers at a fair tend to underestimate the weight of a cow because they only have experience with sheep.

In this case the collective (aggregated) guess is as good as the average individual guess, but worse than the best individual guesses. The collective is neither more stupid nor more intelligent than a randomly chosen individual.

2) the group is influenced by the same agents

Example: some agents are allowed to argue why a particular alternative is the best one. The others listen, and base their choice on what they heard the speakers say. If the speakers are randomly chosen, and the listeners are equally affected by each speaker, the collective decision will represent the opinions only of the speakers. This is better than the average individual decision, but worse than the decision where all agents vote independently, because the group whose opinion is taken into account is smaller.

If the speakers are experts on the issue, the result is not necessarily better, because experts are more likely to have similar biases rather than independent ones. If the experts' biases are independent of each other, the errors they make are likely to be smaller. In that case, the result is likely to be more accurate than when the speakers are not experts, but not generally more accurate than when everybody can vote independently. Even though this includes bigger individual errors, this also produces more error canceling. For example, in the Johnson [1998] simulation, averaging the votes of only the best performing agents ("experts") does not improve the overall quality of the decision.

Collective stupidity

In some cases, the collective decision is worse than the majority decision or even the worst of individual decisions. This phenomenon is called **groupthink** [Janis, 1971, 1972]: the group converges on a poor solution, without anyone daring to criticize it, because everyone thinks the others all agree with it.

Possible mechanism: opinions are expressed in a certain order, e.g. alphabetically, randomly, or the most "expert" group members first. Since every listener is influenced by the previous speakers, new speakers will tend to confirm what those before them said. This can be understood from two mechanisms, one cognitive (priming) and one social (conformity). If you haven't yet developed an independent idea about a topic, hearing others' thoughts will prime your mind to think along the same lines, thus making very different ideas less likely to come up. Moreover, if you feel part of a group, then you will want to maintain and strengthen your ties with that group by reinforcing rather than contradicting the group opinion.

Because of this dynamic, if several speakers have all more or less said the same, further speakers become increasingly unlikely to express an opinion deviating from the previous ones. Therefore, the opinion expressed most often in the beginning tends to become dominant, especially if the first speakers were "experts" or authorities. If this opinion was very inaccurate, the collective decision will also be inaccurate.

An even worse situation occurs with **polarization**. Assume that the group needs to choose an option ranging between two extremes (poles), e.g. "let's bomb country X", "let's make peace with country X". Individually, each member may prefer some intermediate solution because he feels rather uncertain about what to do, e.g. "let's embargo country X while negotiating with them". Assume now that the first speakers tend more to the one extreme (e.g. "let's bomb"). This will influence further speakers to express arguments supporting this extreme. As more and more speakers all add arguments for the same type of approach, everyone feels more primed and encouraged to further develop that approach, and thus potentially receive recognition from the others. Thus, the group may take a final decision much more extreme or risky than any member would have chosen individually. The reason is that they all think: "well, I may have been uncertain, but everybody else seems to support this option, so I can safely go along with them". This phenomenon may explain some catastrophic decisions made by committees, such as the invasion of Iraq by the US.

Conclusion: requirements for collective intelligence

According to Surowiecki [2004], the requirements for a group of people to exhibit collective intelligence (or *wisdom of crowds*, as he calls it) are the following:

- **Diversity**: the more diverse the knowledge and experience possessed by the different members of the group, the less the group is likely to overlook certain aspects, or to fall prey to a similar bias, and therefore the better their collective judgment can be.
- **Independence**: individuals should express their opinion independently from other members of the group; otherwise, when the opinion of the one is influenced by the opinion of the other, biases can get reinforced, resulting in groupthink and collective stupidity
- **Decentralization**: people should as much as possible be able to gather their information independently, that is, exhibit a division of cognitive labor, so that they can together cover an as wide range of situations as possible.
- **Aggregation**: collective intelligence requires an effective mechanism, such as voting, averaging or discussion, for aggregating a diversity of individual opinions into a single, collective decision.

The last requirement, aggregation, is the least obvious, since there exist many conceivable ways to integrate information from different sources into a single conclusion. The apparently best method, averaging, only works for quantitative decisions. Some of the most common methods, such as discussion in committee meetings, do not obey the criteria of independence and decentralization, and therefore may lead to poor results.

The result can be improved if the different members express their opinions independently and anonymously (e.g. on a computer-supported discussion system) before they start responding to the opinions of others, and if the discussion is guided by a neutral moderator, who ensures that everybody duly answers to all the important questions. The anonymity makes sure that everybody's ideas are given equal attention (instead of the discussion being dominated by the more authoritative people). This is the basis of the so-called *Delphi procedure* that aggregates the ideas of a panel of experts, via several rounds of anonymous discussion.

Without aggregation there is no collective decision, and therefore a cognitive superagent may need to use less than optimal methods to reach a decision if no simple alternative is available. The next sections will discuss some aggregation mechanisms that exist in the real world of interacting agents—rather than in the ideal situation where a computer calculates the average of many independent estimates.

Meme propagation

Knowledge may also be constructed by a collective without aggregation of opinions that are expressed in parallel. Aggregation can happen *sequentially*: each individual in a chain or sequence adds something to a piece of information that is being transmitted from the one to the next. This is an aggregation mechanism that obviously ignores the requirement of independence, and therefore its results will exhibit specific biases. We will call such a piece of information that is being transmitted from individual to individual a **meme**.

Suppose that person A has some knowledge (e.g. a new idea or something s/he observed or read) that others don't have. A tells the information to B, B tells it to C, C to D, etc. Such sequential transmission will tend to affect the information that is being passed on. Each new version will be slightly different from the previous one, because communication is not perfect, as people forget things, misunderstand, exaggerate, or express themselves poorly. This variation is not random, though, because of biases intrinsic to the individuals and to the way they communicate. Some elements are more likely to be transmitted than others, e.g. elements that are more interesting, or that confirm the hearer's prejudices or own observations. Other elements are more likely to be left out, e.g. elements that are difficult to understand or formulate, or that seem contradictory with experience. This constitutes a natural selection of the information propagating across a chain of individuals. As a result of such evolutionary dynamics, after many transmission processes the information may well have changed beyond recognition. This change is biased in particular directions, however, and as such to some degree predictable.

The most obvious case of bias is when the meme becomes a **myth** or legend, i.e. something that appears like knowledge but that has no practical use or foundation in the real world. An example is the “Mozart effect” [Bangerter & Heath, 2004]: the belief that babies become smarter if they listen to classical music. There is no scientific basis for this assertion. The origin is a newspaper report of an experiment in which adults scored temporarily better on some psychological tests after listening to classical music (perhaps just because it is relaxing). After several retellings in which one journalist quoted or paraphrased another one, temporary improvement in scores has become permanently increased intelligence, while adults have become children and finally babies. Other examples are the different “urban legends” [Heath, Bell & Sternberg, 2001]: spectacular stories that are being told and retold as if they really happened, but for which no evidence can be found. Examples are the man who found a dead rat in the cola bottle he has just been drinking from, or gangs that specialize in drugging their victims, after which they steal a kidney to sell on the black market for organs [Heath & Heath, 2006].

Such biased construction processes are not always negative, though. In some cases, a collective decision needs to be made for which there is no objective basis, because no option if objectively “better” than any other. This is the case for **conventions**, where everyone needs to agree about doing things in a certain way in order to facilitate communication or coordination, but the different ways are essentially equivalent. A meme may thus evolve into a standard that is accepted by the whole group, e.g. a symbol for a particular shared concept, such as the word “dog”, or the convention that you should drive on the right (or left) side of the road. This mechanism may explain the process of **social construction** of knowledge.

While common, such strongly biased developments are not the general rule. Memes can also evolve without losing their basis in reality. This happens typically when the observation that inspired the meme is easy to repeat. In that case, the culturally transmitted version is regularly confronted with the original evidence, so that it cannot deviate too much from that original. In that case, the meme may become a piece of “common sense” knowledge. An example is the idea that lions are dangerous. Most people have not been attacked by lions, and therefore know this only through hearsay. However, from time to time, someone witnesses an attack by a lion, thus being able to compare the meme with reality, and correcting the meme if necessary.

What type of information propagates well (is a good meme) depends on a variety of criteria, some increasing the quality of the information, some likely to decrease it. We will here list the most important properties of a meme that make it more likely to propagate well [Heylighen, 1997; Heylighen & Chielens, 2008; Heath & Heath, 2006]. To illustrate their concrete meaning, we will use an example of an extremely successful meme, i.e. a piece of knowledge that everybody knows, namely the rule stating that pregnancy lasts 9 months.

Objective criteria

These are criteria that depend on the outside world to which the meme is supposed to refer, independently of the people who transmit the meme.

- *Evidence*: information that is supported by many independent observations is more reliable. Each time a meme is confirmed by a new piece of evidence, it becomes more plausible.
  —For example, each time you hear about another person's pregnancy lasting about nine months, the "9 month" rule will become more strongly established in your mind.

- *Distinctiveness*: information that refers to something precise, distinct or detailed can be confirmed (or contradicted) more easily by observation than vague or fuzzy statements. It therefore appears a priori more reliable.
  —For example, the observation that a pregnancy lasts 9 months will be taken more seriously than the observation that a pregnancy lasts somewhat less than a year, even though this latter version is less likely to be incorrect in any particular case.

- *Invariance*: information that remains valid over a wide range of contexts or situations is more stable and broadly applicable. It will therefore be remembered more easily.
  —For example, the duration of 9 months is applicable to practically all human pregnancies. The observation that first pregnancies tend to last slightly longer applies to a much smaller number of cases, and is therefore a priori a less universally valid piece of knowledge.

Subjective criteria

These are the criteria that depend on the use or meaning of a meme for the individual subject.

- *Utility*: information that is valuable or useful to its carrier is more likely to be remembered, applied and passed on.
  —Knowing how long a pregnancy lasts is extremely important for all expecting mothers, fathers, and family members. Therefore, they will be motivated to seek out, remember and communicate any information in this regard.

- *Affectivity*: information that provokes strong emotions is more likely to be remembered and passed on. Emotion here refers to instinctive states of arousal, such as fear, desire or disgust [Heath, Bell & Sternberg, 2001], without need for reflection like in the case of utility.
  —Pregnancy is biologically an extremely important state, which is accompanied by very strong emotions including hope and fear about the outcome. Therefore, all information concerning this state will have an immediate impact on the mind.

- *Coherence*: the better information fits in with the knowledge that individuals already have, the more easily they will understand and accept it [Thagard, 2002; Heylighen, 2001].
  —The 9-month rule for pregnancy fits in with our general expectation that difficult, important processes require a lot of time. Moreover, the measurement in months fits in with our tendency to associate fertility processes with months (approximately the duration of a menstrual cycle) rather than with weeks or days.

- *Simplicity*: short, simple messages are easier to understand, remember and transmit.

  —It is easier to remember an estimate of 9 months than the slightly more accurate estimate of 266 days.

- *Novelty*: information that is unexpected will attract more attention.

  —The 9-month rule is so well-known that you will hardly find anybody for whom it is unexpected, but you could imagine how impressive it would sound if you would hear for the first time that this very complex and individual process always has the exact same duration.

- *Repetition*: repeated exposure to the same message helps it to be assimilated and retained.

  —Family members of pregnant women will hear again and again when the baby is expected to be born, so that the 9-month rule is constantly reinforced.

Intersubjective criteria

These are the criteria dependent on the social transmission, from subject to subject.

- *Publicity*: the more effort an individual puts into spreading a message, the more people will receive it.

  —Family members of pregnant women will not only repeatedly hear but also speak about when the baby is expected.

- *Formality*: messages formulated explicitly and unambiguously are less likely to be misinterpreted. Informal, context-dependent formulations, on the other hand, only make sense within the given context.

  —Compare "the duration of a pregnancy is 9 months" (formal) with "she will give birth 9 months from now" (context-dependent). The first sentence is always clear and unambiguous. The latter sentence loses its meaning when the circumstances change, e.g. at a later time or when speaking about a different person.

- *Expressivity*: information must be easy to express in a given language or medium.

  —Thanks to the existence of common words like "pregnancy", "duration" and "month", the rule is very easy to formulate in such a way that everybody understands it. Imagine instead having to say "the gestation period is equivalent to 9 menstrual cycles"…

- *Authority:* an authoritative, trustworthy source of the information makes it more likely to be accepted.

  —The 9-month rule is confirmed by medical specialists, handbooks and encyclopedias, as well as by mothers who have actually experienced it themselves.

- *Conformity*: information confirmed by more people is more easily accepted [Boyd & Richerson, 1985]. The larger the majority of people that agree with a meme, the more difficult it will be for anybody to hold on to a "dissident" view.

—Everybody around will constantly remind you that a pregnancy lasts 9 months, and you will not find any truly divergent opinions.

- *Collective utility*: information, if adopted by a group, may help the group to function better, independently of its individual utility. Examples of such collectively useful memes are standards, linguistic conventions, and traffic rules: these only become useful if many people apply them.
  —The 9-month rule uses the implicit convention that pregnancy starts with conception (and e.g. not with the first missed menstruation). This makes it easier for pregnant women, doctors and hospitals to coordinate the follow-up of the pregnancy, since they will a priori all agree on what e.g. "the sixth month" refers to.

A cognitive interpretation of memetic criteria

Most of these criteria can be derived directly from our connectionist model of individual cognition. The principle that knowledge is learned through repeated **reinforcement** implies that:

- *useful* information will be rewarded and therefore reinforced each time it is successfully used, and such a reward will be anticipated even when it has not been used yet.
- information that triggers *emotions* will elicit stronger activation and therefore be registered more strongly in memory
- *repeated* information will be reinforced with each repetition and therefore be better remembered
- similarly, information confirmed by many individuals (*conformity*) or repeatedly by the same individual (*publicity*) is better remembered
- information with independent *evidence* will not only be reinforced each time a bit of evidence is encountered, but different types of evidence will reinforce different associations in the neural network, thus "grounding" the information more broadly and deeply in the cognitive system. This is confirmed by the experiment of Ebbesen & Bowers [1974], which showed that people become more convinced of an idea as they hear more different arguments for it. The results of the experiment can be accurately reproduced by a connectionist simulation [Van Overwalle & Heylighen, 2006].

The mechanism of **spreading activation** implies that:

- complex messages activate many different nodes, and require several steps of processing via propagation across related nodes. This activation will therefore diffuse or dissipate more easily than the one created by *simple* messages that activate just a few nodes and links, and therefore leave a stronger memory trace.

- *distinct*, detailed perceptions create strong patterns of activation that more easily lead to clear, focused conceptions

The importance of **anticipation** explains that:

- *invariant* information allows anticipation in a much wider range of situations. It is therefore intrinsically more useful and more likely to be confirmed by evidence.
- *novel*, unexpected messages elicit strong attention (activation, consciousness) and a quick registration in memory, because they contradict implicit anticipations, and therefore indicate potentially important shortcomings of the cognitive system
- *incoherent* messages do not fit in with the existing cognitive structures. They will receive inhibitory signals from the knowledge they contradict, so that their overall activation is suppressed, while failing to receive reinforcing signals from knowledge they confirm. They are therefore difficult to understand, to associate with existing knowledge, to learn and to remember
- when people repeatedly express true messages, we expect their next message to be true as well. Therefore we start to trust them, giving them an aura of *authority* or trustworthiness, so that the next time they say something we anticipate that this information will be correct.

The mechanism of the **social construction** of knowledge, by which groups get to agree on common concepts, symbols and rules, explains that:

- Knowledge that is difficult to represent in the form of agreed-upon symbols will be difficult to *express* by the sender, to understand by the recipient, and therefore to communicate
- Expressions that use ambiguous, context-dependent symbols, such as "he saw it there", may be easy to understand in the right context, since that context primes the mind for grasping what the symbols refer to, but that understanding gets lost as soon as the same message is expressed in a different context. Therefore, an expression using more formal symbols, such as "John Smith saw the Eiffel tower in Paris", will maintain its meaning across many more transmissions.
- Concepts and rules that have become consensual and that are useful for coordination within the group will be used frequently, and therefore will be continuously reinforced. Moreover, the success of groups that use them will incite others to imitate these conventions so that the meme spreads to other groups.

Distributed cognition

Communication networks

Information can travel both sequentially (from person to person) and in parallel (several messages reaching the same person more or less simultaneously or being sent by the same person). Information can travel between agents, but also between agents and objects (e.g. books, computers), and even between objects without human intervention (e.g. from computer to computer via the network). Agents and objects that contain information can be seen as **nodes** in a network. Objects are passive: they store the information, but don't change it. Agents are active: they to some degree change the information (if only by deciding to pass it on or not).

Nodes are **connected** or linked if information frequently passes from the one to the other. The strength of the link is the degree to which the information transmitted across this link will tend to be accepted. It represents the trust that the receiver has in the sender. Each time a message is successfully sent across a link (i.e. the receiver accepts it), the link is strengthened, because the receiver gets more trust in the sender. This mechanism is similar to Hebbian learning in a connectionist network.

The network of trust links between agents is equivalent to a **social network**. Trusted senders are “friends” of the receiver. Close friends are trusted more than superficial acquaintances. Such a network is analogous to a neural network, and therefore in principle able to process complex information in an adaptive way. Information or "memes" transmitted along the network follow a pattern of spreading activation: each agent that assimilates a meme is "activated". The strength of this activation will depend on:

- the strength (fitness) of the incoming meme (how well it satisfies the selection criteria),
- the number of incoming links along which it arrives (each additional confirmation of the information from another source increases its activation).
- the strength of these links (how much the receiver trusts the senders)

Stronger activation of an agent means that this agent becomes more likely (or more active) to transmit this meme to his acquaintances. Received activation can also be negative, when the message transmitted contradicts a meme received earlier, thus inhibiting its further transmission. Different memes that mutually confirm or support each other, when reaching the same agent, may be aggregated by that agent into a "synthetic" meme that is transmitted as a whole to further agents.

This "connectionist" interpretation of meme propagation across a social network appears like a complex but inspiring model for distributed cognition. It deserves to be explored in much more depth. Applying these ideas, my colleague Frank Van Overwalle has already implemented a simulation of agents exchanging information in this way [Van Overwalle & Heylighen, 2006].

While sticking to relatively simple situations, this simulation managed to accurately reproduce the results of various classic experiments in social psychology, which investigated how people process information collectively.

Division of Labor

In society, each individual specializes in a particular domain: the domain in which that person has most experience (typically because the person has been most active in that domain). This implies a self-reinforcing dynamic: an agent more active in the domain gathers more experience and therefore becomes better at solving problems in that domain; therefore, the agent will tend to take on more problems in that domain, and thus gather further experience [Gaines, 1994]. By becoming an expert, the individual may come to possess some unique (or nearly unique) knowledge. Therefore, there will be little competition for that individual's services, and the individual can earn more money from his expertise.

The more specialized individuals are, the more diverse the knowledge available in the society [Martens, 2005]. Imagine that everybody in the group has the same knowledge or experience, of the amount K. For example, they may be all farmers, like in an African village, who know how to tend their specific crops and animals, and who are essentially interchangeable: when one farmer is ill, another one may look after his goats and fields. In that case, the group as a whole would have the amount of knowledge K. Now imagine a group of thousand people who have completely different, non-overlapping knowledge, each of an amount K. For example, one is a lawyer, one is a doctor, another is an engineer, an artist, a farmer, etc. The total knowledge available in the group is now 1000 K, much larger than in the previous case. Such advanced specialization explains in part why there is much more knowledge in a highly developed, postindustrial society (such as ours) than in an agricultural society (such as the African village).

However, individuals still need to be able to communicate. This implies that at least part of their knowledge (language, conventions, general culture, common sense, ...) should overlap or be shared by all. Otherwise they cannot understand each other, and no cooperation is possible. Moreover, it is good to have some overlap in expertise for the sake of redundancy: if one expert is not available, another should be able to replace him. Also, different experts on the same subject are likely to have somewhat different biases and experiences, so that the one can correct for the errors of the other, and their collective decisions are more reliable.

In an ideal society, knowledge would be distributed in the most efficient way, so that as much as possible useful expertise is present, easy to locate, and reliable or robust. Yet, agents can still easily communicate and exchange their knowledge so that anyone can get any information he may need.

Stigmergy

In addition to social networks, there is another mechanism to easily get the right information to the right person: **stigmergy**. With stigmergy, information is exteriorized and stored in a commonly accessible medium. When another agent browses through that medium, s/he will recognize the information that is relevant to him/her, and act on it, e.g. using a condition-action rule. The stored information forms a stimulus ("stigma") that entices the agent to perform work ("ergon") on it. This changes the externally stored information so that it satisfies new conditions. These conditions may incite another agent to recognize it and act upon it.

In this way several "specialists", each having their own condition-action rules, can collaborate efficiently on a shared external memory:

- without need to communicate about who is going to do what
- without need for the agents to know each other or their specific expertise
- without need for a particular planning, order or sequence in which the agents consider the information

<u>Example:</u> Wikipedia

Stigmergy explains why Wikipedia, the freely editable web encyclopedia, is so successful in stimulating its contributors to produce consensual knowledge. People normally read pages on subjects they are interested in, e.g. Belgian football, cognition, butterflies, ... Because they regularly read about the same subjects, they tend to develop some expertise in these topics. When they notice that something in the page is missing or incorrect, they are inclined to edit the page in order to add the missing information, so that the page becomes better. Since many people read these pages, many people with diverse expertise will add to the knowledge in them. As a result, a typical Wikipedia page will contain a wealth of relevant information with little or no errors.

This appears like a quite effective method for **aggregating** the knowledge of diverse individuals into a single description. It must be noted, however, that this aggregation does not fulfill the strict criterion of **independence**, since later contributors will be influenced to some degree by what previous contributors have written. On the other hand, the interaction between contributors is much more indirect than in a face-to-face discussion or meeting, and the **diversity** of contributors (who are self-selected from the global population of Internet users) appears to be much larger. Therefore, the risk for **groupthink** seems to be much smaller.

<u>Example:</u> termites collaborate on building a termite hill.

Initially, individual termites drop some mud randomly. When another termite perceives a heap of mud (condition), it tends to add its own mud to it (action). In this way initially small heaps grow into towering columns that touch each other.

<u>Example:</u> ants collaborate on mapping their surroundings.

When an ant finds foods it leaves pheromone marks on its way back to the nest. When another ant comes out of the nest, it tends to follow an existing pheromone trail. The stronger the trail, the stronger the probability that it is followed. Yet, there is also a small probability for the ant to deviate and leave the trail, thus possibly finding a *new* source of food. When the ant comes back with food, it adds pheromone to the existing trail thus making it stronger. This increases the probability that subsequent ants would follow this same trail. If the food source is exhausted, the pheromone trail is no longer reinforced and gradually evaporates. After a while, a complex network of trails is formed, connecting the nest to the different sources of food in the most efficient way. This network functions like a collective, external memory for the ant colony, telling them how and where to find their food.

Note that this example also does not follow the independence criterion, since later ants are directly influenced by the decisions of earlier ants, but this is to some degree compensated by the fact that ants can deviate from the trail and thus increase the diversity of the paths they explore [Heylighen, 1999].

Towards a global brain

The following methods of distributed information processing should ideally all be integrated:

- **aggregation** of diverse opinions into a collective estimate
- communication via **social networks** based on trust or acquaintance
- **meme** propagation based on the degree of "interestingness" of the message
- **division of labor** and coordination between the differently skilled specialists
- external memories that facilitate collaborative knowledge improvement via **stigmergy**

Together they would form a complex, self-organizing information network that delivers information where it is needed, that processes information so as to improve it, and that creates new information based on the recombination of existing ideas. Such a network could be implemented using the Internet. For example, social networks are already implemented in community systems such as Facebook and LinkedIn. Stigmergy is already implemented in Wikipedia and similar websites that stimulate users to add to the work. Memetic propagation already happens via email forwarding. These different systems should further be integrated. For example, interesting emails could automatically be forwarded to all friends in the social network, depending on their degree of interest in the subject and trust in the sender. External repositories of information could be linked depending on their mutual relevance, so that a person interested in the one is automatically directed to others that are relevant. Collective decisions could be made by aggregating people's preferences, taking into account their position in the social network (e.g. giving more authority to people that are trusted by many).

Such developments would lead to a distributed information processing system similar to the brain: a “Global Brain” for humanity. There is little doubt that the most important technological, economic and social development of the past decades is the emergence of a global, computer-based communication network. This network has been growing at an explosive rate, affecting—directly or indirectly—ever more aspects of the daily lives of the people on this planet. A general trend is that the information network becomes ever more global, more encompassing, more tightly linked to the individuals and groups that use it, and more intelligent in the way it supports them. The web doesn't just passively provide information, it now also actively alerts and guides people to the best options for them personally, while stimulating them to share their experience. To support this, the web increasingly builds on the knowledge and intelligence of all its users collectively, thanks to technologies such as blogs, wikis, ontologies, collaborative filtering, software agents, and online markets. It appears as though the net is turning into a nervous system for humanity.

The “Global Brain” is a metaphor for this emerging, collectively intelligent network that is formed by the people of this planet together with the computers, knowledge bases, and communication links that connect them together. This network is an immensely complex distributed cognitive system. It not only processes information, but increasingly can be seen to play the role of a brain: making decisions, solving problems, learning new connections, and discovering new ideas. No individual, organization or computer is in control of this system: its knowledge and intelligence are distributed over all its components. They emerge from the collective interactions between all human and machine subsystems. Such a system may be able to tackle current and emerging global problems that have eluded more traditional approaches. Yet, at the same time it will create new technological and social challenges that are still difficult to imagine.

Conclusion: the new science of the mind

The origin of cognitive science

Cognitive science [e.g. Luger, 1994; Bechtel, 1999; Thagard, 2005] studies the functioning of the mind. Its history can be seen as a prolonged struggle against **dualism** and the **homunculus** view of the mind, i.e. against the idea that the mind is a separate, unanalyzable entity, which somehow resides inside our brain from where it perceives the world and controls our actions. While this view is simple and intuitive, it does not explain *anything* about how the mind functions: it merely attributes mental properties, such as knowledge, intelligence and consciousness, to this unspecified, mysterious entity, while leaving us in the dark about the components, processes or mechanisms that constitute this entity.

Philosophy of mind started with the debate between rationalists and empiricists: does our knowledge come from internal thoughts (**rationalism**) or from external observations (**empiricism**)? After several centuries of discussion, the conclusion was that both thinking and observing are necessary, and that they interact in complex ways. But philosophy lacked a methodology for testing its hypotheses, and therefore its theories remained disparate, subjective and vague.

Psychology turned the study of the mind into a science, by introducing experiments that can be used to confirm or refute theoretical hypotheses on the basis of objective data. Initially, this experimental approach led to the paradigm of **behaviorism**, which insisted that theories of cognition should deal *exclusively* with externally observable phenomena: **stimuli** (triggers of perceptions) and **responses** (ensuing actions). Mental activity was then reduced to the learning and use of associations between stimuli and responses.

The introduction of computers with their internal memories, programs, and processors, however, inspired the **information-processing** paradigm: after being perceived, stimuli are internally processed using various complex schemas stored in long-term memory. This paradigm can be seen as the proper start of **cognitive science**: the mind has now become a cognitive *system*, i.e. an analyzable whole consisting of connected components that processes informational input and transforms it into observable output. The task now was to discover precisely what these components are, how they are connected, and how together they process information.

The symbolic paradigm for cognition

Initially, the guiding metaphor for analyzing cognitive processes was the manipulation of **symbols** according to a complex program or algorithm. This led to the symbolic paradigm for cognition. Its basic assumption is that knowledge is an abstract, internal representation of the external environment. This representation consists of symbols, each representing a particular component or aspect of the environmental situation. The task of cognition is to solve **problems**, in the sense of answering questions about that environment. This is done by manipulating the symbols according to given inference rules in order to find the combination that best answers the question. The symbolic paradigm thus sees abstract reasoning as the essence of cognition.

The symbolic paradigm was implemented in **artificial intelligence** (AI), a general approach to the simulation of cognitive processes by means of computer programs. However, symbolic AI has been much less successful than expected—in particular in terms of reproducing actual human performance. In contrast to the logical reasoning of AI programs, people's reactions are based on intuition, which is rooted in their subjective experience of the situation. This makes them much more flexible in dealing with complex and unforeseen circumstances. In part as a result of these failures, the symbolic view of cognition has come under harsh criticism over the past two decades [e.g. Bickhard & Terveen, 1996; Clancey, 1997; Suchman, 1990]. It has now been largely overtaken by a "new" cognitive science, which is inspired more by the concrete functioning of the human mind (biologically, neurologically, psychologically, socially) than by abstract theories of logic and computation.

One fundamental criticism of symbolic theories is that if you try to represent all the relevant aspects of the real world with symbols, your representation becomes much too complex to be systematically explored by a computer, and *a fortiori* by the human brain. Indeed, the brain is limited by the famous "**magical number**" (Miller, 1956): not more than about seven items can be held simultaneously in **working memory**. A sufficiently detailed description of a real-world situation will typically include hundreds of symbols (words, concepts, features) that can be combined in millions of different ways, making it essentially impossible to manipulate these symbols in order to systematically explore all their potentially relevant combinations.

Instead, the brain relies on its **long-term memory**, which can store millions of facts, to quickly *recognize patterns* in the incoming information. This long-term memory is implemented as a network of variable-strength connections between nodes or neurons. Recognized patterns function as stimuli that trigger appropriate responses or actions. Unlike a computer program, the **neural network** structure of the brain is very good at fitting fuzzy and ambiguous perceptions into known patterns, at learning to recognize recurrent patterns in incomplete and inconsistent data, and at associating perceived patterns with appropriate actions. However, it is very poor at simultaneously keeping several such patterns in mind while reasoning, because the corresponding patterns of neural **activation** tend to interfere with each other. Moreover, activation quickly decays because of diffusion and neuronal fatigue.

Finally, while long-term memory is very effective at **recognition**, it is rather poor at **recall**, i.e. reviving memory patterns without perceptual stimulation. In that sense, human memory is much less reliable than a computer memory for retrieving a fact outside of the concrete context that reminds you of that fact.

The extended mind

One of the key insights of the new cognitive science is that cognition is necessarily **situated** and **embodied** (Clark, 1997, 1999; Clancey, 1997; Anderson & Michael, 2003). This means that a cognitive system, such as the human mind, is always interacting with its environmental situation via its **sensors** that **perceive**, and **effectors** that produce **actions**. The complexity of the real world is dealt with not by manipulating an abstract internal representation, but by manipulating the world itself, i.e. by performing actions and monitoring their results via perceptions. This interaction is controlled via **sensory-motor feedback**:

- perceptions trigger actions;
- actions produce changes in the environmental situation;
- these changes are again perceived,
- these perceptions trigger new actions to—if necessary—correct or extend the effects of the previous actions.

Different situations will produce different perceptions, and therefore trigger different actions. Both cognition and action therefore are **situated**: they are determined much more by the concrete external situation than by internal reasoning or planning (Suchman, 1990; Susi & Ziemke, 2001; Clancey, 1997). This shifts most of the burden of memory and reasoning from the brain to the **environment**: instead of having to conceive, predict and remember the potential results of an action, the action is simply executed so that its results can be read off from the environmental situation.

Actions leave their mark on the environment. Insofar that this mark is made in a stable medium, such as stone, paper or silicon, it functions like an objective registration of what has happened, storing the information for later review by the brain. In that way, the brain can "offload" information and store it in an external memory that is more reliable and less energy consuming than its own working memory. In this case, we may say that the mind *extends* into the physical environment (Clark & Chalmers, 1998), or that cognition is **distributed** across the brain and various material supports (Hollan, Hutchins & Kirsh, 2000; Hutchins, 1995). A simple example is taking notes. The markings on the paper change as the results of our actions (writing). On the other hand, they remain safely stored while we do not interact with the paper. When perceived (read), they trigger thoughts and corresponding new actions, such as adding a related item to the list of already registered items.

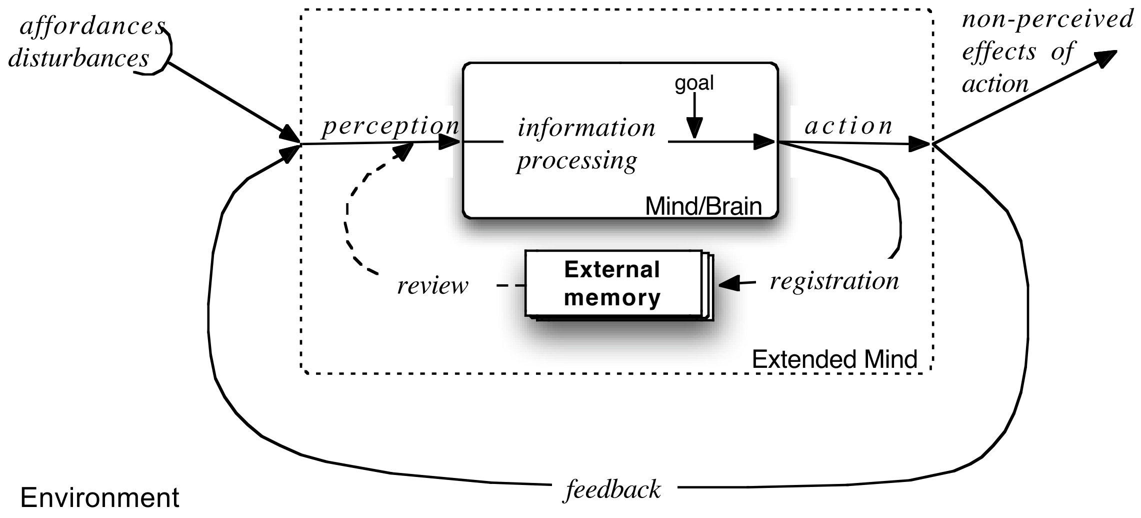
A useful paradigm to conceptualize the dynamics of such environmentally mediated activity is the concept of **stigmergy** (Parunak, 2006; Susi & Ziemke, 2001; Heylighen, 1999, 2007). An activity is stigmergic if the action by an agent leaves a mark (*stigma* in Greek) in the environment that stimulates an agent (the same or another one) to perform further work (*ergon* in Greek). This subsequent action will leave another mark which in turn will stimulate yet another action. Thus, different actions indirectly trigger each other, via the traces they leave in the environment. This allows a flexible coordination of the different **condition-action rules** that govern the agent's behavior. Thanks to stigmergy, even the extremely simple **reactive agents**, which lack a working memory, become able to execute complex activities (such as building a nest).

The mind as a control system

The environment not only provides a passive medium that registers the effect of actions: it actively intervenes in the agent's activity, producing opportunities to perform new actions or **disturbances** that make the actions' result deviate from what was intended. In situated cognition, opportunities for action created by the presence of specific objects or situations are called **affordances** (Norman, 1999). Because our brain has evolved to quickly adapt to its environmental situation, our perception is especially tuned to the recognition of both disturbances, that create problems that need to be addressed, and affordances, that may help us to solve problems and achieve our goals (Gibson, 1986).

A fundamental paradigm to understand this agent-environment interaction is the cybernetic notion of **feedback control** (Powers, 1973; Heylighen & Joslyn, 2001), which is also known as error-controlled regulation. A goal-directed agent, such as an ant or a human, tries to achieve its goals by eliminating any difference between its present situation (**perception**) and its desired situation (**goal**). A goal here should not be understood as a completely specified objective or end-state, but merely as an (explicit or implicit) preference for certain situations over others. For every perceived difference between the present situation and the goal, an **action** is performed to reduce that **deviation**, i.e. bring the situation closer to the preferred one. If the result as perceived is not sufficient, a next action is performed to again bring the situation closer to the goal, and so on, until the agent is satisfied.

Although some actions may be counterproductive (in that they increase the distance to the goal), the overall process tends to zoom in efficiently on the goal because of **negative feedback**: every new action tends to correct the errors created or left over by the previous action. External disturbances are dealt with in the same way: whatever caused the deviation or error, the system's reaction is to try to maximally reduce it, until there is no deviation left. In that way, the system remains in control of the situation, by efficiently counteracting any movement away from its desired course of action. In feedback control, there is no need for planning or for complex reasoning. This makes the mechanism very robust, and able to deal with the most complex circumstances (Gershenson & Heylighen, 2004).



The different components of this mind-environment interaction are summarized in the figure. We can distinguish two nested levels of mind:

1) the traditional idea of mind as inherent in the brain;
2) the "**extended mind**" which encompasses the brain together with any external memories that are used to support information processing.

In the traditional perspective, external memory is part of the environment. In the **cybernetic** or **distributed cognition** perspective, however, it is part of the **agent**, since it is completely controlled by the agent. The part of the environment that is not under control—i.e. which does not perform merely as the agent expects—intervenes in the agent's activity via what we have called affordances and disturbances. These, together with the feedback received via the environment about previous actions and the reminders stored in the external memory, determine the situation as perceived by the agent, and therefore the agent's further actions.

Anticipation and consciousness

The main shortcoming of feedback as a method to tackle environmental problems is that the reaction to a perceived problem may come too late, when the problem has grown too large or too complex to still be corrected by a well-chosen action. Therefore, an essential function of the brain is to **anticipate**, i.e. to conceive potential problems and consequences before the situation has become difficult to control, and thus prepare the agent to intervene in the most efficient way.

In a **neural network**, anticipation happens automatically thanks to the mechanisms of **Hebbian learning** and **spreading activation**. Hebbian learning creates or strengthens a connection A → B whenever B is experienced in close succession to A. Therefore, neural connections in the brain increasingly reflect causal connections in the outside world, so that the strength of the A → B connection becomes an ever more reliable estimate of the probability that A would indeed be followed by B. The activation of one or more nodes, representing perceived or conceived conditions, automatically propagates along these connections, thus activating or **priming** the conditions that are likely follow. This produces a mental and physical preparedness that helps the agent to deal with the situation.

If such anticipated conditions turn out to be wrong, the cognitive system immediately tries to correct its error by focusing the attention on the unexpected phenomenon, so as to gather more information. Furthermore, it will weaken to some degree the connections that produced the wrong prediction, so that making the same mistake becomes less likely. The focusing of attention implies a higher degree of **consciousness** and arousal. This is typically accompanied by **emotions**, such as joy, fear or surprise, and the creation of a stronger memory trace, e.g. in episodic memory. The other major cause of focused attention is a problem that is considered so important that a lot of cognitive resources must be devoted to it. This happens by inhibiting the flow of activation to phenomena that seem irrelevant to the focus, potentially "blinding" the mind to things that otherwise would certainly be noticed.

When nothing particularly important or unexpected happens, consciousness is in general more diffuse, taking the form of what we have called "**subjective experience**". This can be seen as the entire state of activation and priming created by a specific situation and the different perceptions and conceptions that accompany it. It functions as a general form of mental preparedness. This state is unique, personal, essentially context-dependent, and in constant flux. As such it cannot be expressed in any formal or symbolic way, e.g. by means of language or logic. This does not mean that it must remain forever outside of the domain of science, as some philosophers have argued, but only that it cannot be captured by traditional symbolic models of cognition.

Intelligence and its amplification

Different people differ in the degree to which their brain is good at anticipating and at solving problems. This depends partly on the amount of accumulated knowledge and experience (**crystallized intelligence**), partly on the intrinsic efficiency of the processes of spreading activation and learning (**fluid intelligence**). I have hypothesized that the most basic measure of fluid intelligence, the **g-factor**, reflects the number of steps that activation can propagate in the brain without losing its initial focus. More steps in a coherent train of thought means anticipation of phenomena that are farther in the future, less obviously related to the initial situation, and more uncertain. This enables the tackling of more complex, abstract and long-term problems, and especially the discovery of more creative or unexpected solutions.

While fluid intelligence partly depends on the genes, it is also strongly influenced by the support and stimulation received from the environment, as illustrated by the ongoing increase in average

IQ. This allows us to envisage methods to augment internal intelligence—as a complement to the methods discussed earlier that make information processing more effective by extending the mind into the environment. One promising approach is based on the concept of flow: a focused activity where the level of difficulty equals the level of skill, and where there is constant feedback. Being engaged in flow-producing activities stimulates people to maximally use their brain, and thus to train and improve their cognitive abilities. When the challenges are higher than the skills, on the other hand, they are likely to give up, thus losing an opportunity to sharpen their brain.

One practical application of cognitive science is to make us aware of the in-built limitations of our intelligence, or **cognitive biases**, thus helping us to overcome them. One fundamental reason for the existence of such biases is that cognition has evolved to maximize survival and growth, thus making us both **paranoid** (tending to exaggerate external dangers) and **optimist** (tending to exaggerate our own abilities to achieve progress). Another reason is that neural networks function through **spreading activation**. This tends to exaggerate the importance of recently activated conceptions and their associations (**context**) relative to the long-term "base rate" frequency of events.

A complementary method to increase intelligence and combat individual biases is by **aggregating** the experience of many different individuals. If this is done in a balanced way, e.g. by averaging the individual guesses, the result may be called **collective intelligence**. However, collective decisions are often subjected to various social biasing mechanisms, such as a tendency to conform to what others say. This may lead to **groupthink**, where the collective behaves more stupidly than the individuals it consists of. Such collective stupidity can be avoided by fostering **diversity**, **independence** and **distribution**, so that group members can collectively explore a maximum of different approaches.

Information from different individuals is rarely aggregated in a central place, though. It more commonly propagates from individual to individual, in the form of communicable ideas or "**memes**". Good memes should be easy to assimilate, remember, express and transmit, implying that they should fit in well both with individual cognitive systems and with social, linguistic and technological means of communication. Information will be propagated most reliably along the strongest links in a **social network**, i.e. between the people that best know and trust each other. Information can also propagate indirectly, by being stored in a medium, such as libraries or databases, that many can access. This has the advantage that the medium functions like a shared memory that can be improved independently by anyone. This is a **stigmergic** form of collaboration.

These different mechanisms for the propagation of information in society, such as networks and stigmergy, seem remarkably similar to the mechanisms used by the individual brain. This analogy becomes stronger with the growing use of electronic networks, which become ever more efficient in distributing and processing information. This opens up the perspective of the emergence of an unimaginably intelligent distributed cognitive system at the planetary level, i.e. a **Global Brain**.

Index

Recommended Reading

The following is a bibliography containing the most important works by the authors referred to in the text, complemented by some other important publications on the topics discussed.

Anderson, D. & Michael, L. 2003: Embodied cognition: A field guide, Artificial Intelligence 149, p. 91—130.

Arbib, M. A. (1995). The handbook of brain theory and neural networks. Cambridge, MA: MIT Press.

Ashby, W. R. (1947). The nervous system as a physical machine: With special reference to the origin of adaptive behavior. Mind, 56(221), 44–59.

Ashby, W. R. 1958. Requisite variety and its implications for the control of complex systems. *Cybernetica* 1, no. 2: 83-99.

Ashby, W. R. 1964. *An introduction to cybernetics*. Methuen London.

Ashby, W. R. Design for a Brain - The Origin of Adaptive Behaviour. Chapman and Hall, London, 1960.

Ashby, W. R., and R. Conant. 1981. *Mechanisms of Intelligence*. Intersystems Publications.

Aunger R. (ed.) (2001): Darwinizing Culture: The Status of Memetics as a Science (Oxford University Press)

Baars, Bernard J. (1988), A Cognitive Theory of Consciousness (Cambridge, MA: Cambridge University Press)

Baars, Bernard J. (2002) The conscious access hypothesis: Origins and recent evidence. Trends in Cognitive Sciences, 8 (1), 47-52.

Baars, Bernard J.(1997), In the Theater of Consciousness (New York, NY: Oxford University Press)

Bangerter, A., and C. Heath. 2004. The Mozart effect: Tracking the evolution of a scientific legend. *British* Journal of Social Psychology 43, no. 4: 605–623.

Bateson, G., 2000. *Steps to an Ecology of Mind*, University Of Chicago Press.

Bechtel, W. et al. Ed. (1999). A Companion to Cognitive Science. Blackwell Companions to Philosophy. Malden, MA: Blackwell Publishers.

Beer, R. D. (2000). Dynamical approaches in cognitive science. Trends in Cognitive Neuroscience, 4(3), 91–99.

Beer, R. D. 1995. A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* 72, no. 1-2: 173–215.

Bekoff, M., Allen, C., & Burghardt, G. M. (Eds.). (2002). The cognitive animal: Empirical and theoretical perspectives on animal cognition. Cambridge, MA: MIT Press.

Berger P. L., T. Luckmann: (1967) The Social Construction of Reality: A Treatise in the Sociology of Knowledge, Anchor.

Bickhard MH, L Terveen (1996) Foundational issues in artificial intelligence and cognitive science: impasse and solution, Elsevier Science Publishers.

Block, N., 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–287.

Bonabeau, E. Dorigo, M. Theraulaz, G. (1999) Swarm Intelligence, Oxford University Press.

Boyd, R. & Richerson, P.J., 2001. Norms and bounded rationality. *Bounded rationality: The adaptive toolbox*, 281–296.

Braitenberg, V. (1984). Vehicles: Experiments in synthetic psychology. Cambridge, MA: MIT Press.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation, RA-2(April), 14–23.

Brooks, R. A. (1991). Intelligence Without Reason. In Proceedings of the 12th international joint conference on artificial intelligence. Los Altos, CA: Morgan Kauffman.

Campbell, Donald T. (1974), "Evolutionary Epistemology." In The philosophy of Karl R. Popper, edited by P. A. Schilpp, LaSalle, IL: Open Court, pp. 412-463. .

Cattell, R.B., 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1), 1–22.

Chabris, C. F. (2006). Cognitive and Neurobiological Mechanisms of the Law of General Intelligence, in: Roberts, M.J. (Ed.). Integrating the mind. Hove, UK: Psychology Press.

Chalmers DJ 1995 Facing up to the problem of consciousness, Journal of Consciousness Studies,

Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory.* Oxford, England: Oxford University Press.

Changeux, Jean-Pierre. 1985. Neuronal man: the biology of mind. New York: Oxford University Press.

Chomsky, Noam (1965). Aspects of the Theory of Syntax. The MIT Press.

Churchland, Patricia Smith (1986). Neurophilosophy: Toward a Unified Science of the Mind-Brain. The MIT Press.

Clancey W.J. (1997), Situated Cognition, Cambridge University Press, Cambridge.

Clancey, W.J., 1997. Situated cognition: On human knowledge and computer representations, Cambridge Univ Pr.

Clark A. (1999): An embodied cognitive science, Trends in Cognitive Science 3:9, p. 45—351.

Clark A. and Chalmers D. (1998): "The Extended Mind," Analysis 58, p. 7-19.

Clark, A. (1997). Being there: Putting brain, body and world together again. Cambridge, MA: MIT Press.

Clark, A., & Toribio, J. (1995). Doing without representing? Synthese, 101, 401–431.

Cleeremans, A. & McClelland, J.L., 1991. Learning the structure of event sequences. *Journal of Experimental Psychology: General*, 120(3), 235–253.

Cowan, N., 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(01), 87–114.

Csikszentmihalyi, M. & Nakamura, J. (2002). The concept of flow. In: Snyder, C. R., & Lopez, S. J. (Eds.). The handbook of positive psychology. Oxford University Press, p. 89 – 105

Csikszentmihalyi, M. (1990). Flow: The Psychology of Optimal Experience, Harper Perennial.

Damasio, Antonio R. 1996 (1994). Descartes' error: emotion, reason and the human brain. London: Papermac [First published by New York: Grosset/Putnam].

Damasio, Antonio R. 1999. The feeling of what happens: body and emotion in the making of consciousness. New York: Harcourt Brace & Company.

Dehaene, S. & Changeux, J.P., 2004. Neural mechanisms for access to consciousness. *The cognitive neurosciences*, 3, 1145–58.

Dehaene, S., 2008. Conscious and nonconscious processes: Distinct forms of evidence accumulation. In *Decision making, the human mind, and implications for institutions. Strüngmann forum reports*. pp. 21–49.

Deikman, A.J. (1966): Deautomatization and mystic experience. Psychiatry 29 : pp. 324-338.

Deikman, A.J. (2000): A Functional Approach to Mysticism, Journal of Consciousness Studies 7, nr. 11-12, p. 75—91.

Dennett, D.C, 1993. *Consciousness explained*, Penguin London.

Di Paolo, E. A., Noble, J., & Bullock, S. (2000). Simulation models as opaque thought experiments. Artificial Life, VII, 1–6.

Dickens, W.T. and Flynn, J.R. (2001) Heritability estimates versus large environmental effects: The IQ paradox resolved, Psychological Review 108, pp. 346—369.

Dreyfus, H.L., 1992. *What computers still can't do: a critique of artificial reason*, The MIT Press.

Ebbesen, E. B., & Bowers, R. J. (1974). Proportion of risky to conservative arguments in a group discussion and choice shift. Journal of Personality and Social Psychology, 29, 316–327.

Fauconnier, Gilles; and Turner, Mark. 2002. The way we think: conceptual blending and the mind's hidden complexities. New York: Basic Books.

Flynn J.R.: Massive IQ gains in 14 nations, Psychological Bulletin 101:2, 171-191 (1987).

Flynn, J.R., 2007. *What is intelligence?: beyond the Flynn effect*, Cambridge University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. In S. Pinker & J. Mehler (Eds.), Connections and symbols. Cambridge, MA: MIT Press.

Fodor, Jerry A. (1983). The Modularity of Mind. The MIT Press.

Gaines B.R. (1994), "The Collective Stance in Modeling Expertise in Individuals and Organizations", *International Journal of Expert Systems* 71, 22-51.

Gärdenfors, P. (1994). How logic emerges from the dynamics of information. In J. van Eijck & A. Visser (Eds.), Logic and information flow (pp. 49–77). Cambridge, MA: MIT Press. http://www.lucs.lu.se/Abstracts/LUCS_Studies/ LUCS15.html.

Gärdenfors, P. (2000). Conceptual spaces. Cambridge, MA: MIT Press.

Gazzaniga (ed.), Michael S. (2004). The Cognitive Neurosciences III. The MIT Press.

Gershenson C. & F. Heylighen (2004). How can we think the complex? in: Richardson, Kurt (ed.) Managing the Complex Vol. 1: Philosophy, Theory and Application.(Institute for the Study of Coherence and Emergence/Information Age Publishing)

Gershenson C. (2004) Cognitive paradigms: which one is the best? Cognitive Systems Research 5, 135–156

Gibson, J.J. (1986). The Ecological Approach to Visual Perception. Hillsdale (NJ): Lawrence Erlbaum.

Glimcher, Paul W. (2003). Decision, Uncertainty, and the Brain. The MIT Press.

Gottfredson, L.S. & Deary, I.J., 2004. Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science*, 13(1), 1.

Gottfredson, L.S., 1997. Why g matters: The complexity of everyday life. *Intelligence*, 24(1), 79–132.

Grassé P.-P. (1959). La Reconstruction du nid et les Coordinations Inter-Individuelles chez *Bellicositermes Natalensis et Cubitermes sp.* La théorie de la Stigmergie. *Insectes Sociaux*, 6:41-84.

Haier, R. J. (1993). Cerebral glucose metabolism and intelligence. In P. A. Vernon (Ed.), Biological approaches to the study of human intelligence (pp. 317–373). Norwood, NJ: Ablex.

Harnad, S., 1990. The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3), 335–346.

Harnad, S., 2002. Symbol grounding and the origin of language. *Computationalism: new directions*, 143–158.

Haselton, M. G., and D. Nettle. 2006. The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases. *Personality and Social Psychology Review* 10, no. 1: 47.

Haugeland, John. (1985). Artificial Intelligence: The Very Idea. Cambridge, MA: Bradford Books, MIT Press.

Hawkins, J. (2004) On intelligence (Times books)

Heath, C. & Heath, D., 2007. *Made to Stick: Why Some Ideas Survive and Others Die*, Random House.

Heath, C., Bell, C. & Sternberg, E., 2001. Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology*, 81(6), 1028–1041.

Hebb, D. O. (2000). The organization of behavior. Minds, brains, and computers: the foundations of cognitive science: an anthology.

Hershberg, U., & Efroni, S. (2001). The immune system and other cognitive systems. Complexity, 6(5), 14–21.

Heylighen F. & Chielens K. (2008): Evolution of Culture, Memetics, in: Encyclopedia of Complexity and Systems Science, ed. B. Meyers (Springer).

Heylighen F. & Joslyn C. (2001): "Cybernetics and Second Order Cybernetics", in: R.A. Meyers (ed.), *Encyclopedia of Physical Science & Technology* (3rd ed.), Vol. 4 , (Academic Press, New York), p. 155-170.

Heylighen F. (1991): "Cognitive Levels of Evolution: pre-rational to meta-rational", in: The Cybernetics of Complex Systems - Self-organization, Evolution and Social Change, F. Geyer (ed.), (Intersystems, Salinas, California), p. 75-91.

Heylighen F. (1997): "Objective, subjective and intersubjective selectors of knowledge",Evolution and Cognition 3:1, p. 63-67.

Heylighen F. (1999): "Collective Intelligence and its Implementation on the Web", *Computational and Mathematical Theory of Organizations* 5(3), p. 253-280.

Heylighen F. (2001): "Bootstrapping knowledge representations: from entailment meshes via semantic nets to learning webs", Kybernetes 30 (5/6), p. 691-722.

Heylighen F. (2007). Why is Open Source Development so Successful? Stigmergic organization and the economics of information, in: B. Lutterbeck, M. Baerwolff & R. A. Gehring (eds.), Open Source Jahrbuch 2007, Lehmanns Media, 2007, p. 165-180.

Heylighen F., Heath M., F. Van Overwalle (2004): The Emergence of Distributed Cognition: a conceptual framework, Proceedings of Collective Intentionality IV, Siena (Italy)

Hilton, D.J. & Slugoski, B.R., 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1), 75–88.

Hollan, J., Hutchins, E. and Kirsh, D. (2000): Distributed cognition, ACM Transactions on Computer-Human Interaction 7:2, p. 74—196.

Holland, J. (1992). Adaptation in natural and artificial systems (2nd ed.). Cambridge, MA: MIT Press.

Hutchins E (1995): Cognition in the Wild (MIT Press).

Jakobi, N., Husbands, P., & Harvey, I. (1995). Noise and the reality gap: The use of simulation in evolutionary robotics. In Lecture notes in artificial intelligence: Vol. 929. Advances in artificial life: Proceedings of the 3rd European conference on artificial life (pp. 704–720). Berlin: Springer.

Janis IL (1971), *Groupthink*. Psychology Today, pp. 43–46.

Janis, I. L. (1972) *Victims of groupthink*. (Boston: Houghton Mifflin).

Jensen, A. R. (1998). The g factor: The science of mental ability, (Praeger, Westport, Connecticut)

Johnson, N.L., 1998. Collective problem solving: Functionality beyond the individual. *Los Alamos National Laboratory Working Paper LA-UR-98-2227*.

Kahneman, D. & Tversky, A., 1996. On the Reality of Cognitive Illusions. *Psychological Review* 103, 582-591.

Kahneman, D., 2003. Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review*, 93(5), 1449-1475.

Kahneman, D., Slovic, P. & Tversky, A., 1982. *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press.

Kelso. J. A. Scott (1995) Dynamic Patterns: The Self-organization of Brain and Behavior (MIT Press; Cambridge, MA).

Kirsh, D. (1996) Adapting the Environment Instead of Oneself. Adaptive Behavior, Vol 4, No. 3/4, 415-452.

Kirsh, D. (2000): A few thoughts on cognitive overload, Intellectica 1:30, p. 19—51.

Kosslyn, S.M., Ganis, G. & Thompson, W.L., 2001. Neural foundations of imagery. *Nature Reviews Neuroscience*, 2(9), 635–642.

Lakoff, George & Johnson, Mark. 1999. Philosophy in the flesh: the embodied mind and its challenge to western thought. New York: Basic Books.

Lakoff, George & Nuñez, Rafael. (2001). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. New York: Basic Books

Langton, C. G. (Ed.). (1989). Artificial life, Santa Fe institute studies in the sciences of complexity. Reading, MA: Addison Wesley.

Lenat, D.B., 1995. CYC: A large-scale investment in knowledge infrastructure, Communications of the ACM 38 , Issue 11, p. 33 - 38

Lévy P. (1997): Collective Intelligence, Plenum.

Libet, B. 1999. Do we have free will? *Journal of Consciousness Studies, 6* 8, no. 9: 47–57.

Loewenstein, G., 1994. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116, 75–75.

Loftus, E.F. & Ketcham, K., 1996. *The myth of repressed memory: False memories and allegations of sexual abuse*, St Martins Pr.

Loftus, E.F., 1974. Reconstructing memory: The incredible eyewitness. Jurimetrics J., 15, 188.

Luger, G. (1994). Cognitive science: the science of intelligent systems. San Diego: Academic Press.

Lynn, R. & Vanhanen, T. (2002). *IQ and the wealth of nations*. Westport, CT: Praeger.

MacLeod, C.M., 1991. Half a century of research on the Stroop effect: An integrative review. *Psychological bulletin*, 109(2), 163–203.

Maes, P. (1990). Situated agents can have goals. Journal of Robotics and Autonomous Systems, 6(1–2).

Maes, P. (1994). Modelling adaptive autonomous agents. Journal of Artificial Life, 1(1–2).

Martens B. (2005): The cognitive mechanics of economic development and institutional change, (Routledge)

Maturana, H. R., & Varela, F. J. (1987). The tree of knowledge: The biological roots of human understanding. Shambhala.

Maturana, H.R., 1974. Cognitive strategies. Cybernetics of cybernetics, 457–469.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group (Eds.). (1986). Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2: Psychological and biological models. Cambridge, MA: MIT Press.

McCrone J. (1999). Going Inside: A Tour Round a Single Moment of Consciousness. Faber and Faber.

McCulloch, W.S. & Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 5(4), 115–133.

McFarland, D. (1981). The Oxford companion to animal behavior. Oxford: Oxford University Press.

McLeod, P., Plunkett, K. & Rolls, E. T. (1998). Introduction to connectionist modeling of cognitive processes. Oxford, UK: Oxford University Press.

Miller, G.A. (1956): The magical number seven, plus or minus two, Psychological Review 63:2, p. 81-97.

Minsky, Marvin L.; Seymour A. Papert (1969). Perceptrons: An Introduction to Computational Geometry.

Neisser U. (1976): Cognition and Reality, San Francisco: Freeman.

Neisser, U. Boodoo, G., Bouchard, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. American Psychologist, 51, 77-101.

Newell, A. (1980). Physical symbol systems. Cognitive Science, 4, 135–183.

Newell, A. (1990). Unified theories of cognition. Washington, DC: Harvard University Press.

Newell, A., & Simon, H. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.

Noe, Alva (2005). Action in Perception. The MIT Press.

Norman, D.A. (1999). Affordances, conventions and design, Interactions 6 (3) pp. 38 – 43.

O'Regan JK , A Noe (2001) A sensorimotor account of vision and visual consciousness, Behavioral and Brain Sciences 24, 939–1031

Ochse, R (1990). Before the Gates of Excellence: The Determinants of Creative Genius, Cambridge University Press

Parunak, H. V. D. (2006) A survey of environments and mechanisms for human-human stigmergy, in: *Environments for Multi-Agent Systems II*, (Lecture Notes in Computer Science, Vol. 3830, Springer Berlin), p. 163-186.

Pavlov, I. P., & Anrep, G. V. (2003). Conditioned reflexes. Courier Dover Publications.

Piaget, J. (1968). Genetic epistemology. New York: Columbia University Press.

Port, Robert and Tim van Gelder (1995) Mind as Motion: Explorations in the Dynamics of Cognition. (MIT Press, Cambridge, MA.).

Powers, W. T. (1973) Behavior: the Control of Perception. Aldine, Chicago.

Putnam, Hilary (1988). Representation and Reality. The MIT Press.

Pylyshyn, Z. W. (1984). Computation and cognition: Toward a Foundation for Cognitive Science. Cambridge, MA: MIT Press.

Quiroga, R.Q. et al., 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045), 1102–1107.

Reber A.S. (1993) Implicit learning and tacit knowledge, Oxford University Press

Rescher, N., 1973. The coherence theory of truth, Clarendon Press.

Riegler, A. (2002). When is a cognitive system embodied? Cognitive Systems Research 3, 339–348.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.). (1986). Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations. Cambridge, MA: MIT Press.

Russell, B., Whitehead, A. N. (1910–13). Principia mathematica. Cambridge: Cambridge University Press.

Ryle, Gilbert [1949] (1984). The Concept of Mind. Chicago: University of Chicago Press

Searle JR 1980 Minds, brains, and programs, Behavioral and Brain Sciences 3, p. 417–24.

Searle, John (1992). The Rediscovery of the Mind. The MIT Press.

Shortliffe, E. (1976). Computer based medical consultations: MYCIN. Amsterdam: Elsevier.

Simon, Herbert A. (1970). The Sciences of the Artificial. The MIT Press.

Simons, D.J. & Chabris, C.F., 1999. Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, 28(9), 1059.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century-Crofts, New York.

Smolensky, P. (1988). On the proper treatment of connectionism. Behavioural and Brain Sciences, 11, 1–23.

Steels L. & Brooks R. (eds.) (1995): The Artificial Life Route to Artificial Intelligence: Building Embodied Situated Agents (Erlbaum).

Steels, L., 2005. The emergence and evolution of linguistic structure: from lexical to grammatical communication systems. *Connection Science*, 17(3), 213–230.

Stewart, J. (1996). Cognition = life: Implications for higher-level cognition. Behavioural Processes, 35(1–3), 311–326.

Suchman L. A. (1990): Plans and situated action, Cambridge University Press

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Susi, T. & Ziemke, T. (2001). Social Cognition, Artefacts, and Stigmergy: A Comparative Analysis of Theoretical Frameworks for the Understanding of Artefact-mediated Collaborative Activity. Cognitive Systems Research, 2(4), 273-290.

Thagard P. (2005) Mind, Introduction to Cognitive Science. (2nd edition) (Bradford Books)

Thagard, P. 2002. Coherence in Thought and Action. MIT Press.

Thelen, Esther and Linda Smith (1994) A Dynamical Systems Approach to the Development of Cognition and Action (MIT Press; Cambridge MA).

Theraulaz, G. & Bonabeau, E., (1999). A Brief History of Stigmergy. *Artificial Life*, 5(2), 97-116.

Thornton, C. (2000). Truth from trash: how learning makes sense. Cambridge, MA: MIT Press.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungs problem. In Proceedings of the London Mathematical Society, 42(2), 230–265. http:// www.abelard.org/turpap2/turpap2.htm.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, LIX(236), 433–460. http://www.abelard.org/turpap/ turpap.htm.

Van Gelder, T., 1998. The dynamical hypothesis in cognitive science. Behavioral and brain sciences, 21(05), 615–628.

Varela, F., Thompson, E., & Rosch, E. (1991). The embodied mind: Cognitive science and human experience. Cambridge, MA: MIT Press.

Vauclair, J. (1996). Animal cognition. Cambridge, MA: Harvard University Press.

von Bertalanffy, L. (1968). General system theory: Foundations, development, applications. George Braziller.

von Foerster, H. (1981) On Constructing a Reality. In: H. von Foerster, *Observing Systems*. Intersystems Publications, pp. 288–309.

von Glasersfeld, E., 1984. An introduction to radical constructivism. *The invented reality*, 17–40.

Walter, W. G. (1950). An imitation of life. Scientific American, 182(5), 42–45.

Walter, W. G. (1951). A machine that learns. Scientific American, 185(2), 60–63.

Watson, J. B. (1998). Behaviorism. Transaction Pub.

Wiener, N. (1948). Cybernetics; or, control and communication in the animal and the machine. Cambridge, MA: MIT Press.

Wilson, Robert A.; Frank C. Keil (eds.) (2001). The MIT Encyclopedia of the Cognitive Sciences (MITECS). The MIT Press.

Winograd, T. & Flores, F., 1986. Understanding computers and cognition: A new foundation for design, Ablex Publishing Corporation.

Ziemke, T., 2003. What's that thing called embodiment. In *Proceedings of the 25th Annual meeting of the Cognitive Science Society*. pp. 1305–1310.