1 Introduction

Within a biological organism it is possible to identify a number of general functions facilitated by internal mutually connected mechanisms. The organization responsible for material and energy exchange with the environment and the integration of those into the operation of the organism is broadly classified as the metabolic system. The organization responsible for identification of external invaders and the intervention and elimination of the invaders is classified as the immune system. The organization responsible for generating of a copy or a closely resembling organism is the reproductive system. The organization responsible for controlling the interactions of the organism with the environment, usually to facilitate other functions, is classified broadly as the cognitive system.\footnote{In philosophical discussions the term “cognition” is commonly used to refer to the high level capacity or activity of thought, and it is contrasted with perception, affection, instinct, and other “lower” capacities.} Other such general functions
may be identified as well. The distinction of biological functions need not correspond to an anatomical separation of the organization — the same organs (or organelles) may serve different functions for different systems. Nevertheless, different general functions warrant independent investigations and distinct general theories. Thus, it is desirable to develop a general theory of metabolism, a general theory of the immune system, a general theory of replication, or a general theory of cognition. In this paper I initiate a general theory of cognition that is both sensitive to the biological origins of cognition and sufficiently abstract to apply to non-biological systems as studied in AI. Particularly, I identify the proper internal function of cognition and demonstrate how traditional cognitive capacities satisfy this function.

Didn’t we identify the function of cognition: “controlling the interactions of the organism with the environment”? What more can be said about its function? It is useful to distinguish between the systemic function and the organizational function of a system. A systemic function of a system for an organism is a function identified by a system theoretic approach to the description of the organism, where the organism is modeled as a black box and its behavior is modeled by tracking relations with other systems in the environment.\(^2\) If \(O\) is an organism and \(S\) an alleged system of \(O\), \(S\) satisfies a systemic function for \(O\) if what counts as a satisfaction of the function can be determined by regarding \(O\) as a black box. In the case of the metabolic system, the function, as defined above, can be described purely by describing the matter/energy interactions of \(O\) and the environment: \(S\) is that system in \(O\) “responsible

Modern cognitive science, however, has realized that such traditional distinctions do not isolate distinct capacities that can be investigated in isolation. Instead, it is necessary to unite all such capacity into a broad notion of cognition; and, that cognition must be investigated as a phenomenon tightly constrained by the biology of organism. In contrast, a lot philosophical discussions of “cognition” are based on a Cartesian methodological intuition of the relationship between mind/cognition and the body/organism, in that it is assumed that cognition can be described and investigated in isolation form the role it plays in an actual cognitive agent. To a large degree, the idea that the concept of epistemic rationality is central for understating cognition is a consequence of the Cartesian program. See Wheeler (ref) for an analysis and criticism of the Cartesian program and a comparison to an alternative insisting that cognition must not be isolated from the organism.

\(^2\)Within system theory an organism can be (and is) modeled as a complex system with internal subsystems. However, always one assumes bottom level systems that are treated as black boxes. Maintaining such a level of abstraction in the description of a system is ultimately the strength of the system theoretic approach. (refs) The treatment of cognition in this paper is deployed within the broad conceptions of system theory.
for material and energy exchange with the environment and the integration of those into the operation of the organism”. One need not actually describe the operation of the metabolic system in specifying its function. Similarly with the systemic function of cognition. In fact, it is possible to specify a more precise systemic function for cognition, going beyond control of interactions. Cognition is often associated with systems involving more “sophisticated” control relationships described as adaptive systems and anticipatory systems. (refs) Adaptive systems are systems that are capable, through appropriate feedback mechanisms, to adjust their behavior to a broad set of environmental conditions. Anticipatory systems (a sub-class of adaptive systems) are systems that can anticipate the future state of the environment and modulate their behavior in light of the anticipation.\(^3\)

While the systematic investigations of such systemic functions is important for a general theory of cognition and cognitive science broadly construed, a general theory of cognition requires a characterization of the organizational function of cognition. An organizational function of a system \(S\) of \(O\) is a function determined by specifying a major role a system \(S\) plays in the organization of \(O\). To specify an organizational function one needs to consider the internal operation of \(O\). For example, the function of the heart to produce circulation is an organizational function. Of course, usually the internal operation of \(O\) is strongly constrained by demands on its behavior and interaction with the environment. Organizational functions usually subserve systemic functions. The function of the heart to pump blood subserves both the function of transportation of nutrients and the locomotion of immune agents (and other system functions as well). My goal is to identify a general organizational function of cognition. I will say more about organizational functions in the next section, but first let us identify the proposed function with an intuitive example.

Consider, by all standards, a highly complex cognitive task: the discovery of a proof of a mathematical theorem. A mathematician starts with a collection of simple to grasp mathematical facts. Let us think of these as the inputs to the mathematical system that

\(^3\)Both adaptive systems and the subclass of anticipatory systems can exit in contexts not directly related to cognition.
the mathematician is implementing. Then, with a large collection of formal transformations, embeddings in complex mathematical structures, utilization of various spatial, quantitative and other intuitions, etc. the mathematician is able to generate a complex proof (a proof sketch really) that provides support for the statement of the theorem. Why was this complex process necessary? Why didn’t the mathematician simply directly grasp the theorem the way she may grasp many of the simpler results? In an important sense, the proof was necessary because there is a difference between the status of the simple inputs and the status of the theorem. The inputs are easy to comprehend and justify, while the theorem is hard. The difference in status depends both on the complexity of the statements and on the capacities of the mathematician. If the mathematician had unlimited intellect — if she was god — than she would have been able to more or less directly comprehend the theorem, like a human (with sufficient education not to be underestimated) can comprehend that $1 + 1 = 2$. But the mathematician is not god, she is intellectually limited in her capacity to directly comprehend complex mathematical theorems. She needs to do significant cognitive work to overcome this limitation by a systematic, locally guided proof construction. The process of proving a complex theorem and the capacities needed to accomplish it, it seems, is needed exactly because the human mind is quite limited — they serve to reduce an informational limitation of the agent. I claim that this phenomenon is the root of not merely mathematical cognition but all cognition. I claim that the organizational function of cognition in general is reducing informational limitations that actual agents face in order to facilitate the systemic function identified above. The rest of the paper will serve to justify this claim.

The paper is organized as follows: In section 2, I discuss the methodological steps needed to support my claim. In section 3, I discuss how autonomous agents naturally begin to demand cognitive capacities and how this is related to informational limitations that they all have. Section 4 is the culmination of the paper, where I identify the organizational function of cognition. Moreover, I demonstrate that canonical examples of cognitive capacities — learning, memory, feature detection, representation, and reasoning — represent instances of
solutions to the informational limitation problem.

2 Methodological Remarks

The notion of “function” is both intuitive and philosophically tricky. Its intuitiveness is exemplified in the ubiquitous functional ascriptions that we make when explaining how things work. There is something comfortable when describing the heart as having the function to pump blood, when saying that the function of the hammer is to drive nails, etc. Its trickiness is connected, as Millikan correctly points out (ref), with the tendency to associate two potentially orthogonal ideas with functional ascriptions. On one hand the notion of function is connected with the notion of a role in the operation of a system. According to this conception an object, say a carburetor, has a function for a system, say a car, if the object fulfills a fixed role, say oxidizing the fuel, which role is determined by the global organization of the system. This use of the notion of function is connected to the set of doctrines known collectively as functionalism. On the other hand the notion of function has a teleological twist — it is an intentional notion. An object has a function if it is intended to satisfy a particular purpose, which it, as a matter of fact, may not satisfy. A broken or poorly designed carburetor may still have the function of oxidizing the fuel, even though it never actually achieves its function. Note that according to the role conception of function a broken carburetor is not really a carburetor, because to be a carburetor an object must successfully fulfill its role. I have nothing especially insightful to say about the notion of function here, other than to conclude that when I set on a task to identify the (organizational) function of cognition, I must be very careful about how I achieve that, negotiating between the role and the intentional aspect of function. I cannot rely on a stable, agreed upon notion of function. In a sense, the method that I follow to elucidate the function of cognition is what determines the notion of function that I use.

What do we want the identification of the general organizational function of cognition
to achieve? (1) The primary use of a functional characterization of a class of systems is, of course, categorization. (Millikan) A goal is to group all those candidate systems that satisfy the functional specifications into the general category of cognitive systems. The categorization should not be arbitrary. It should identify the systems because there is something important shared among them to justify scientific focus. The cognitive systems should be a proper domain of scientific investigation — there should be a benefit in isolating the class of cognitive systems and distinguishing them from other systems, or using a different taxonomy of systems all together. The function of cognition that is identified must highlight what is interesting about the categorization. (2) Unlike the heart, which can be identified as an anatomical organ, the cognitive system cannot be easily identified in an independent way. (While the brain of many creatures is a good start, it does not provide a sufficiently general class of systems, partly because the anatomical characteristics of the brain do not map directly onto its functional role.) Therefore, the problem of identifying the function of cognition is not a problem of identifying what a particular object or a system does in an organism, but rather identifying what the system is in the first place. This is why identifying a systemic function for cognition is easier. (3) Cognitive science is interested primarily with the mechanisms underlying cognition. A general theory of cognition should be a general theory about the nature of cognitive mechanisms. This is why we cannot stop with the systemic function of cognition, but must characterize its organizational function. We must recognize one difficulty however — there may be many possible cognitive architectures, it is not clear that they share anything more in common than the systemic function. Providing a completely satisfactory general account of organizational function may be difficult. It is my hope that the informational limitation account that I propose is both sufficiently general and sufficiently connected to actual cognitive mechanisms to be a good account of the organizational function of cognition. But, here lies the biggest contention of the project.

A proper characterization of the function of cognition as a categorization of cognitive systems must be faithful to the practice of cognitive science, in the sense that it must not
omit anything that the practice shows to be important. Simultaneously, the categorization must be naturally emerging within the larger domain of systems of which the cognitive systems are a subclass. By saying that a distinction is naturally emerging I mean that the distinction can in principle be characterized from below without consideration of any higher level distinctions. In principle, it should be interesting to characterize the collection of systems that are "cognitive" even if we have no independent notion of cognition. In other words, if we investigate the various systems in a world, looking at interesting regularities and taxonomies, we should expect to stumble upon the class of systems that are cognitive — we should be motivated to say: "Now, this is an interesting class of systems! We should develop a branch of science to study them systematically!" By specifying the function of cognition, ideally, we should be able to characterize the most restricted naturally definable class of systems that is faithful to the practice of cognitive science.

How do we achieve this goal? I suggest that we don’t fight the intentional dimension of the notion of function. We should not expect that cognition has a function because someone intended it so, but we could not fight the perspectival character of function. Indeed, I suggest that we embrace it by approaching function from a design prospective. We can identify the function of cognition by identifying first a design problem that a class of systems faces the solution of which determines the function of cognition. We must, therefore, simultaneously identify the class of systems for which a design problem can be defined, and characterize the general strategy for solution of the problem. In doing so we must operate simultaneously at two levels — at the level of the system out there in the world, and at the level of the model of the system. Particularly, we can use differential characteristics of different modeling approaches to identify classes of systems.

With these considerations in mind, I will approach the problem of identifying the general function of cognition by identifying a nested sequence of design problems that co-determine a nested sequence of system classes. A design problem for a class of systems defines a set of solutions for the problem, which defines a set of systems that implement the solutions, on
which set we may define a further design problem related to optimization strategies for the solutions, etc. I will apply this methodology until I reach the function of cognition. Which will leave us with one final problem: arguing that the class of systems that I identify is indeed the class of cognitive systems, and the design problem that I identify indeed specifies the organizational function of cognition — and this is a toughy. The best I can do about arguing that the function I identify is the function of cognition is to demonstrate that many central general cognitive skills can be viewed naturally as fulfilling the identified function — i.e. they are strategies for solutions of the design problem. This may not provide a completely satisfactory argument that I have identified the most restricted naturally definable class of systems that is faithful to the practice of cognitive science, but it is the best I’ve got.

3 Autonomous Agents and Informational Limits

In this section I will describe a nested sequence of design problems and with each design problem I will associate a class of systems whose organization can be regarded as a “good” solution to the design problem. The nested sequence of design problems will be the following (as new terminology is introduced the problems will be reworded): (1) How can a system persist? (2) How can a system affect its environment to improve its persistence? (3) How can a system utilize “better” information from the environment to select better actions? and (4) How can a system reduce its inherent informational limitations to achieve more successful behavior? The corresponding nested sequence of systems will be: (1) autonomous systems, (2) (re)active autonomous systems, (3) informationally controlled autonomous systems — autonomous agents, (4) cognitive systems. None of these distinctions are sharp; almost everywhere there is gradation among the system classes.
3.1 System Persistence and Autonomy

The most rudimentary design problem that must be solved for there to be even a systemic function of cognition is for there to be a system on the first place. Without a condition allowing a system to exist as an entity discernible from its environment and persisting sufficiently long as that same entity to allow qualification of its dynamical behavior, the question of cognition does not even arise. The first design question that must be examined is What allows systems to persist as individual entities?, and moreover, For which of those systems that persist is a capacity of cognition relevant?

We can identify two broad strategies for system persistence: static and dynamic. An example of statically persistent system is a rock. A rock is held together by strong chemical bonds. The stability of the system is derived from the stability of the bonds. The separation of the system from the environment depends on the sharp difference between the bonds of atoms within the rock and the bonds with atoms not in the rock — in fact, what is considered to be in and out of the rock depends on the strength and connectedness of the bonds. Statically persistent systems are among the most persistent systems in the universe — when it gets to persistence one should go static. However, strong bonds cannot do much more than be persistent. There is no need for a function of cognition. Rocks don’t need to think any more than they need to eat.

Dynamic stability is a more complex matter. This is the realm of dissipative systems. A dissipative system is an open non-equilibrium thermodynamical system that maintains stability of an organizational parameter\(^4\) by dissipating matter and energy from and to the environment. We can distinguish two classes of dynamically stable dissipative systems: heteropoietic and autopoietic.\(^5\) In heteropoietic systems stability (maintaining the organis-

\(^4\)That is, there exist an appropriate parametrization of the system such that an important parameter has a stable dynamical orbit. All of these notions are made precise in stability theory. See (refs) for an introduction.

\(^5\)The words derive from the Greek poiesis (ποιησις) meaning creation, with the auto- and hetero- corresponding to the prefixes for self- and other-. Thus, a heteropoietic system is an other-created system, while autopoietic system is a self-created system. My use of the term “heteropoietic” is somewhat non-standard. In cybernetics heteropoietic systems are those that are artifacts of design (refs). I use the term in a way
zational parameter) is determined by the boundary conditions of the system as well as a gradient of free energy that can drive dynamics of the system. Classic example of such heteropoietic systems are Bénard cells and water eddies. Bénard cells form when oil in a container is heated from below sufficiently quickly so that a temperature gradient exists. The system self-organizes into a collection of convection currents that settle one next to another, appearing like a collection of cells. In this configuration of the fluid dynamics the system dissipates the heat energy more efficiently. Water eddies are stable structures that emerge in a water current when the river bed has appropriate irregularities. All that drives the eddies is the energy gradient of the flowing water and the structure of the channel. Heteropoietic systems can be quite stable — the spot of Jupiter has existed for more than 400 years — but like rocks, heteropoietic systems don’t “do” much from within. Autopoietic systems are dynamical systems where the system themselves, not merely the boundary conditions, are “responsible” for maintaining stability. The term autopoiesis was coined by Maturana and Varela (refs) to describe a phenomenon were the conditions for maintaining the structure of a system are present within the system. More precisely, they define autopoiesis as follows:

"A system is autopoietic if:

1. it has a semi-permeable boundary,
2. the boundary is produced from within the system, and
3. it encompasses reactions that regenerate the components of the system."

(Varela 2000, via Luisi 2003 and Bourgine and Stewart 2004)

Maturana and Varela introduced the notion in an attempt to provide a general characterization of living systems, where the paradigm example of an autopoietic system is the biological cell. They also claimed that autopoietic systems possess cognition, but this part that the “other” is not a designer but conditions outside of the system, like boundary conditions.

6Within the text I will use quotes to identify words that have a strong intentional connotation, yet I use them without assuming intentions. Unfortunately, most of our experiences are with intentional beings, and many of our words about action, etc. are infused with intentions. As philosophers, we must control our intuitive inclinations to embed intentions onto such words when used in context where intentions are not present.
of the theory I think is unsatisfactory so we will ignore it. (see Di Paolo ref for a systematic criticism). One of the most interesting characteristics for autopoietic systems is that they support process closure. All the machinery needed to regenerate and maintain the system is included within the system (or is readily available in the environment in the form of matter and free energy) and is itself a product of the system.\(^7\) We can think of the closed system of product equations determined by the processes as defining the system (given a fixed scope of external conditions).

Autopoietic systems are interesting for two important reasons. (1) Their dynamic self-maintenance allows the systems to persist within shallow energy wells — the bonds that hold them together can be extremely weak in comparison to static systems. They can maintain significant complexity without dispersing in chaos (as is usually the case with heteropoietic stable systems). Moreover, they are systems that genuinely “do” something about their persistence. (2) The process closure that defines them determines what are the compounds/mechanisms that are essential for the closure to be maintained, and determines a fixed set of roles for them. This all depends on the system itself, not on any particular external interpretation of how the system operates. Autopoietic systems, therefore, can be described in functional terms where the structure of the process closure defines the participants and their functional roles, and the stable organizational parameters that is maintained by the system provides the (natural) teleology, as the “goal” of the processes (no intentions assumed).

Autopoietic systems can be regarded as the simplest kinds of autonomous systems. We will call a system autonomous if the system can be described as having a goal that it “tries” to achieve, and the control mechanisms of the system that veer it towards that goal are part of the system.\(^8\) In autopoiesis the goal is persistence (always an immediate goal) and the

\(^7\)In the theory of chemical process dynamics, the phenomenon of process closure is called autocatalysis. The notion of autopoiesis adds the further requirement that the system maintains its own boundary. For an analysis of the connection between autopoiesis and autocatalysis see (ref).

\(^8\)In philosophical discussions one usually uses more specific notions of autonomy that assume intentional beings, like human beings, with explicit goals and consciously controlled actions and decisions. Obviously, such high level cognitive skills cannot be assumed in the current discussion.
control mechanism is derived by the processes in the closure. The definition of autopoiesis admits resistance to fluctuations in the external environments, but it does not imply that an autopoietic system can adapt to more complex changes in the environment. One reason for this is that, as Di Paolo describes it, autopoiesis is a structural condition that a system either satisfies or it does not — either a system maintains process closure or it does not. The notion of autonomy, however is a gradual notion: for a fixed goal, a system can be more autonomous depending on how sensitive the system is to the external conditions and to what conditions it can adapt. Autonomus systems are, therefore, not merely autopoietic systems, but autopoietic systems with further capacities.

In light of the systemic function of cognition, cognition is related to the ability of the system to maintain its autonomy in diverse circumstances. To move in the direction of identifying the class of cognitive systems we must investigate the gradation that exist among autonomous systems, particularly, the gradation that allows better adaptability.

3.2 Active Systems

One mechanism for increasing the scope of possible viable environments is to maintain process closure that can switch to different modes (branching sub-processes) depending on the state of the environment. Bitbol and Luisi (ref) distinguish between different kinds of metabolism depending on whether the system can operate in different modes based on what kind of nutrients are available in the environment (and presumably based on what damage must be repaired in the system). Cellular organisms can operate in different modes for such reasons. (refs) Still, there can be only that many modes that a system can adjust to, and most importantly, what modes the system needs to adapt to depends on the flukes of the environment.

A more effective strategy for coping with the variation of the environment is to have (in addition) some control on the environment — get to where food is available, or make the

---

9 Autonomy is a gradual notion in other dimensions as well: how many goals a system can pursue simultaneously, how much of its behavior is control internally and how much by external systems, etc.
food come to you, avoid places where you are food, etc. We need to make sense of this idea.

A system is in a constant dynamical interaction with its environment, the state of the system always affects the state of the environment. In the language of dynamical system theory, the two systems are coupled. How do we isolate those interactions that can be interpreted as the autonomous system controlling the environment. For simple autopoietic systems all the interactions that exist (as far as the process closure is concerned) reduce to absorption of matter and energy and release of lower grade energy (e.g. heat) and useless byproducts. This can usually be modeled with thermodynamics and theories of diffusion.\(^{10}\) In such systems it is not especially interesting to model the relations with the notion of control. (Although it is possible to force such a model.) The notion of control becomes interesting when (1) the coupled dynamical interactions can be decomposed into isolated sub-processes either (a) from the environment to the system, or (b) from the system to the environment. And, (2) the processes can be given appropriate functional roles in terms of control relations, i.e. the system can be modeled effectively with machinery of control theory.\(^{10}\) (in addition to general dynamical systems theory). Whether this is possible, i.e. whether one kind of model is more effective than another, ultimately depends on the organization of the system and the nature of its interactions with the environment. For example, when a bacterium moves in the direction of increased nutrient gradient by moving paddles and “monitoring” nutrition sensors (ref), the interaction can be modeled more efficiently in terms of control relations than with the dynamics of diffusion.

When it is possible to decompose the coupled interaction between the system and the environment into the interaction of kind (a) or (b) that can be modeled with the language of control effectively, we can describe such interactions as proto-percepts and proto-actions respectively, and such systems as active autonomous systems.\(^{11}\) This allows to state a second

\(^{10}\) Control theory, of course, is a branch of dynamic system theory. What control theory allows is focusing on special parameters of the system, leading to a large reduction of the state space of the system.

\(^{11}\) Describing the behavior of systems in terms of control relations, feedback and the emergent dynamics in the domain of classical cybernetics (to be distinguish form general cybernetics which is a more general philosophical theory (refs)). Thus, it is also appropriate to call such systems cybernetic systems.
design problem for active systems: *How can an active system proto-act better in order to improve its chances of persistence?*

The simplest strategy is to affect the environment in a uniform way regardless of its state (or depending only on the metabolic mode). A system may release a chemical that may attract food or pollinators, or repel predators. Or, the system may move a paddle randomly to cause it to be mobile, etc. A bit more flexible strategy is to make the current state of the environment relevant for the proto-action. The simplest way of doing this is through implementing a triggering relation between a proto-percept and a proto-action (or a fixed action pattern). Autonomous systems that operate in this way can be described as *reactive* systems. As we move towards cognition it is interesting to consider yet more “sophisticated” strategies for making the environment relevant for effective proto-actions.

### 3.3 Agents

A more sophisticated active system would be sensitive to more complex patterns in its environment and allow particular patterns to affect its action in a context sensitive way. For example, the chemical signature of a piece of food may produce differential response based on whether a light sensor detects the time in the daily cycle. When this happens we can brake the direct triggering link between proto-percepts and proto-actions, and describe the system in terms of its dispositional characteristics. A detection of a pattern does not merely produce a corresponding action but affects the dispositional characteristic of the control system to affect actions based on other (including future) patterns of proto-percepts. This is especially interesting when the changes of disposition improve the ability of the system to proto-act in a way favorable to its goal (most generally persistence).

In the most general account of informational content that I am aware of, Donald MacKay defines the meaning of a message as:

“[T]he meaning of a message can be defined very simply as its selective function on the range of the recipient’s states of conditional readiness for goal-directed
While I am quite sympathetic to such an account of content, I will not endorse this as a theory of content here. Instead, I will use this as a suggestion that when a system can operate in a way that a proto-percept from the environment can play a selective function on the systems dispositions (the “conditional state of readiness”, in MacKay’s terminology) in a way that systematically contributes to better proto-actions, it is useful to describe the system in informational terms. Proto-percepts can be viewed as signals that have very targeted significance for the system. The system can be sensitive to different states of the signal in a way that a particular state has a systematic effect on the systems dispositions, that is sensitive to the role the source of the signal has on the system. For example, if the system can detect a class of predators based on the chemical signature of each predator, the system can respond differentially based on the signal in a way that optimizes the response to the particular predator. In more abstract terms, the context sensitive response to the signal (including whether to ignore the signal) is attuned to the source of the signal.

Informational models, like control models, and dynamical system models are very general — they can be forced onto many systems. As explained in section 2, the goal is not to isolate the class of systems to which the model can apply, but the isolate a (vague) class of systems to which one model is more effective than another. The features just described make it likely that an informational model is more effective than, say a control system model, or a low level dynamical model. As the complexity and dynamical constraints on the interactions increase an informational model will be more effective. We should avoid insisting that an informational model should not apply to some especially simple systems. Some of the classic examples are thermostats and sunflowers following the sun. It is claimed, if a theory of information makes a thermostat an information system, then there is something wrong with the theory of information. (ref) This is a mistake! A thermostat, after all, can easily be a part of a more complex information system, where its behavior plays an information processing role. The problem with the thermostat is not that it is not an information system.
The problem is that there is no particular advantage to model it as an information system over, say, a control system or a thermodynamical system. Some systems, in some aspects of their behavior, are described more efficiently in informational terms. The informational description highlights patterns in their organization and interaction with the environment that other descriptions miss or make less salient, and therefore less predictively or explanatorily or instrumentally significant. The boundary between such systems and those where there is not significant advantage in an informational description is not sharp. There is a gradual transition between the effectiveness of the model, and in the intermediary cases which model is more desirable may be user dependent. Now, with the help of powerful computers and efficient numeric techniques for solving differential equations, a model with a system of thousand non-linear differential equations may be quite acceptable. Still, as systems become more complicated (in the right way), one cannot avoid informational models.\textsuperscript{12}

A system for which an informational model for important aspects of its behavior is more effective than a control theoretic model I call an autonomous agent. The proto-percepts and proto-actions that participate in informationally described behavior of the agent I call percepts and actions respectively. Not all interactions of the agent with the environment must be modeled with informational terms. Many human interactions with the environment are best modeled with dynamics as close-coupled interactions — walking, bouncing a ping-pong ball, even performing an A-not-B task (refs). Such tasks, however, can be modulated by informational processes — somebody can instruct you to walk to the left, or stop bouncing the ball, etc.

We can now formulate the third design problem for the class of agents: \textit{How can an agent use “better” information to control its actions?}

The term “better” is left as term of art. It is not equivalent to “more” information,

\textsuperscript{12}Similar points have been made by Dennett in his argument for which systems are best describable taking the intentional stance. While an intentional description can be given to a lecture podium, such a description does not depict any \textit{real pattern} over and above pattern depicted by a design or a physical description. For a human being, however, an intentional description is much more effective than a physical description in than it captures patterns of behavior with predictive significance that are made salient in the intentional description, but that may be missed completely in other descriptions. (refs)
although in many cases more information can be better. Also, it is important that the information be used (in principle) to control actions. Information is not the same as data. Often it is claimed that the retina can transmit many gigabits of information per second — it has about 100 million receptors sensitive to large variation of light intensity. This is a claim about how much data can be transmitted. This is not part of the information of the system because the control mechanisms of a human are not sensitive to all the distinctions that the retina can make. Ultimately, what “better” means depends on the particular design solutions proposed. There are some general things to be said about this design problem, and we will move to this next.

3.4 Informational Limits

Let us consider a brute force solution to the third design problem: We can attempt to build an agent that for every relevant for the agent state of the environment (passing neutrinos don’t count as relevant) the agent has a well determined efficient act. Can one build such an agent? In the early days of AI, when researchers were following the micro-world approach to AI, devising systems that operate in worlds with small sets of states, it was assumed that the agent has a complete knowledge of the state of the world (and a possible representation for all other states). The goal was to devise an algorithm that will allow the agent to act towards some goal for every possible state of the world. In essence, we what something similar from our agent, only without the micro-world restriction, and without the need of complex, long-running algorithms. We want to build a real über-agent. Is this a feasible strategy to try to solve the design problem? Can we hope to build an über-agent? I would argue that it is realistically impossible.

Let us imagine what it would take to be an über-agent in the real world. How difficult would it be to lay in bed and to watch PBS on the TV in the neighbor’s house, while reading 756 book simultaneously directly from the shelf of the library 20 miles from the house, while knowing the position and mental states of every of the 2 million people in the city, while
tracking the positions and velocities of the many million asteroids in the asteroid belt? Does this sound challenging? Well, this is probably an insignificant fraction of the relevant states that an über-agent must be able to track and have a ready response.

No naturalistic account of agents can hope that a real physical system can do even a fraction of all that. Any natural system would be severely limited in terms of what information from the environment reaches to it at any moment, and how it can respond based on this information. All natural agents, including the mathematician that I discussed in the introduction, are severely informationally deprived. I suggest that we adopt this as a fundamental principle about real systems — a principle that should not be idealized away.

**ILP:** All agents operate under the condition of severe informational limitation.

Let us call this the *information limitation principle.* Adopting this principle we can formulate a fourth design problem: *How can the internal organization of the control mechanisms of the agent be improved to reduce some of the informational limitations?*

Note that this problem targets specifically the internal organization of the system. It therefore, defines an organizational function. Note also that the design problem does not demand that the informational limitation is eliminated, this would be impossible. Rather, the design problem calls for strategies that reduce the limitation. A solution class would not consist of über-agents, but of agents that have some interesting organization that can be described as reducing the limitation. I claim that this is the class of the cognitive systems, and the mechanisms that reduce the limitation are the cognitive mechanisms.

4 The Function of Cognition

4.1 Lessons from MTC

In its simplest form, the classical mathematical theory of communication (MTC) deals with how information (messages) can be transmitted between a source and a receiver through an
information channel. The messages are sent through the channel, which is assumed to have certain **capacity**. Capacity is a measure of the maximum complexity of the message that a channel can transmit in a unit of time. It is, in a sense, the generalization of the notion of bandwidth that we see in, e.g. Internet connections. A modem has relatively low bandwidth in comparison to a DSL channel, which has a relatively low bandwidth to a fiber-optic cable. A higher capacity channel can transmit more information in a given unit of time than a lower capacity. The goal of information transfer, however, is not merely getting messages through the channel, but getting information from the source to the receiver. If the receiver has no relation to the source, all the information that the receiver can obtain is limited by what can pass through the channel. However, if the state of the receiver is already correlated to some extent with the source, i.e. the receiver already has some information about the source, then one only needs to pass the "novel" information. In other words, the informational relation between the source and the receiver may be much richer than what can pass through the channel. For example, the single bit message "I do" can convey the information that the person next to you will be your spouse, a state with fairly rich content, provided you know that you are in the context of a marriage ceremony.

Most applications of the theory of information have to do with the structure of the communication channel and various coding procedures for optimizing the transmission of information. This is appropriate because one usually takes an external perspective towards the system, whereby all elements of the communication system can be manipulated. When dealing with the issue of agent architecture, and the ability of the agent to enter in informational relations with its environment, the manipulable constraints of the problem are different. When designing a communication system that faces the problem of informational limitation, an engineer has the option to use a higher bandwidth channel, or eliminate the noise in the channel; or, she may be able to reduce the equivocation of the source, making it, in a sense, to "speak more clearly". An agent designer has only the option of modifying
the receiver of the information. But, how may the system at the end of the information flow be modified in order to minimize the informational limitation? According to the model of communication channel assumed by communication theory, the only solution is to make the receiver need less bandwidth by increasing the correlation between the structure of the receiver and the structure of the informational source. In the mathematical theory of communication this relation is expressed with the notion of conditional information entropy of the source on the receiver. The conditional information entropy can be interpreted as measuring the uncertainty of the source from the point of view of the receiver — it is the informational deficit that the receiver has about the source. If the conditional entropy is maximal, then the receiver "does not know" anything about the source and the elimination of the uncertainty requires the transmission of a lot of information. If the conditional entropy is low, the state of the source is only a little bit uncertain, and only a little bit of information is necessary to be transmitted in order to eliminate the uncertainty. Thus, the strategy for overcoming the informational limitation is for the receiver to somehow lower the conditional information entropy of the source with respect to itself.

The communication model of the mathematical theory of communication makes certain assumptions about the source of the information that a model taking as primitive the interaction of an agent with the environment does not. Particularly, the model assumes that the source is well defined, and that its informationally relevant states are defined. For example, it may be assumed that the source produces messages in a fixed alphabet of symbols with a fixed frequency distribution of the symbols. The formal analysis begins after this assump-

---

13 The situation may be a bit more complicated. One may count improvements of the sensory capacities as improvements of the communication channel. It may still be possible, however to isolate the parts of the communication channel that depend only on the environment. Also, in many cases it is reasonable to hold the sensory capacities fixed because they may be more difficult to modify, while concentrating on adaptable internal structures in the agent. This is especially true when looking at cognitive systems.

A further source of complication emerges from the possibility that the agent may modify the environment so that it can improve the capacity of the channel or lower the equivocation of the source. It is reasonable to assume that this will happen only in sophisticated agents, so it need not be assumed in the general case.

14 In MTC the conditional information entropy of a system $X$ on a system $Y$ is given by the expression: $H(X|Y) = -\sum_{x \in X, y \in Y} P(x \& y) \log P(x|y)$ where $P(x \& y)$ is the joint probability and $P(x|y)$ the conditional probability of the states $x$ and $y$. 

20
tion is made; prior to this no model exists. In the situation of the agent interacting with an environment the source of the information and the states of the source that are informationally relevant are not determined upfront. This indeterminacy of the source opens another possible approach for reducing the informational limitation. The agent may be able to "select" the appropriately limited source, with appropriately limited informational states so that it does not need to transmit and operate with too much information. The selection mechanism in the most general case depends on the dynamical interactions of the agent with the environment — it depends on what subsystems of the environment can be isolated in terms of the interactions with the agent, and what aspects of those systems are relevant for the agent’s goals. It also depends on what structures in the agent can be correlated with the systems of the environment so that they facilitate informational channels — appropriately organized sensory sub-systems can play an important role in allowing targeted correlation with the systems in the environment.

The structure of the agent is central for its ability to correlate its dispositional characteristics to act with important systems or general characteristics of the environment. It is this ability to correlate selectively with target sources of information that allows the agent to begin reducing the informational limitations. It is the internal organization that not merely controls action, but controls action in a "smarter" way that deserves independent investigation. Indeed I claim that, be it in disguise, cognitive science studies (and it should study) those structures that can control the agent better than a generic response system does. This leads me to suggest the following characterization of the organizational function of cognition (FC):

\textbf{The cognitive system is the set of organizational constraints (mechanisms) of an autonomous agent that: (1) allow increase of the correlation and integration between the environment and the control structure of the agent, i.e. allow low-}

\footnote{Of course, there are some generic sources of information. For example, the entire environment may be regarded as the source, and all of its possible dynamical states as the informational states. It is exactly this source, however, that is outrageously complex for the agent to communicate with.}
ering of the conditional information entropy of selected important informational sources in the environment on the control structure of the agent, (2) so that the agent can improve the selection of actions to produce successful behavior in light of its informational limitations.

4.2 Why is this an Account of Cognition

I have made a claim that the identified organizational function is cognition. While I have demonstrated why identifying and investigating this function is interesting, I have not argued why we should connected with cognition, as the subject of the now well developed discipline of cognitive science. I need to show that the class of cognitive systems according to FC, is fateful to the practice of cognitive science. My strategy is to show that some of the central cognitive capacities can be regarded as strategies for solving the information limitation design problem, from which follows that they have the function identified in FC.

In what follows, I discuss how the cognitive capacities — learning, memory, feature detection, representation, and reasoning — can be captured naturally within my conception of cognition. While such capacities are uncontroversially cognitive, their precise definitions are not settled. There are debates in the various fields of cognitive science about what constitutes learning, memory, or feature detection; about what counts as representation, and what is required for reasoning. Each of the notions admits of a classification into a number of sub-capacities — there are different kinds of learning mechanism, there are different ways a system can have memory, there are different strategies for representation, etc. It is predictable that a more mature cognitive science will dispose of such distinctions, in everything but a very heuristic sense, and replace them with a more fine-grained collection of concepts. It is unreasonable, then, to expect that I can, at this point, provide a precise characterization of these capacities. I will only resort to general ideas about what kind of capacities these are. Being able to absorb the concepts into my account even without precise definitions is an advantage of my account, demonstrating its generality.
Learning: Let us turn to one of the most general concepts of cognition — learning. In its most general form, learning can be described as the capacity of a system to adapt its behavior to external conditions. This characterization however is too general for cognition. Too many systems are capable of adapting to external conditions. A thermostat is an adaptive system. More generally, every dynamical system that, through a negative feedback loop, can settle into a stable state (or an n-stable state) can be regarded as adapting to its environment. Intuitively, learning involves the further capacity of adapting the system’s dispositional characteristics in a way that past interactions of the system with the environment have the effect of modulating future interactions when the environment is sufficiently similar. The classical Pavlovian stimulus response associative learning mechanism is an example of this phenomenon, where the presence of correlated stimuli (bell and food) affects the dispositions of the system to react to the similar environmental conditions (bell) in a way the can be interpreted as anticipation of another condition (food). This model, of course, can be generalized. Conditions where environmental factors affect the dispositional characteristics of a system are exactly those that can be conveniently described in informational terms (refs MacKay). Thus, learning can be described as the capacity of the system to incorporate past information in a way that it aids the system’s behavior when novel, but sufficiently similar conditions hold. If one concentrates on the internal structures that modulate the systems disposition, we observe a process whereby past information gets integrated with the current information to enhance the information that the system contains above what is available at the moment from the current interactions with the environment. The result is that the conditional information entropy of the system is lower than what the current channel admits. Learning, by allowing past information to be maintained in the system in terms of the modulation of its dispositional characteristics, is a prototypical example of a capacity for overcoming the informational limitations of the agent.
Memory: Memory, in its essence, is a special kind of learning mechanism. As such, it is naturally captured by the proposal; however, it presents a specifically interesting example of information entropy lowering. Often the term memory is used to depict any past information that may be present in a system. In this case, every instance of learning can be described as using memory; however, such a general conception of memory does not provide a good basis for studying memory systems. Usually, when cognitive scientists talk about memory systems they describe systems that can, in one sense or another, reproduce an aspect of a past state of the system, often a state that is similar to or closely related to an input state of the system. Thus, memory can be seen as a special case of past interaction with the environment affecting current dispositions of the system — the systems of memory affect dispositions by regenerating (but not exactly) a missing internal state. Such capacity can provide more general utilisability of past information, because the information has the potential to being partially decoupled, or at least have more control implications, than more task specific learning mechanisms. Memory mechanisms, then, provide a clearer example of information entropy lowering; example reminiscent of the more restricted conception of information connected to the notion of coding.

Feature detection: The mechanisms for feature detection demonstrate the second strategy for entropy lowering: the capacity of the system to select the appropriate source of information within the environment — particularly, to organize or be sensitive to useful restricted aspects of systems in the environment. Intuitively, a feature detector is a system (within the agent) that can correlate with particular features of systems of the environment. For example, in the visual system there are groups of neurons that are sensitive to the direction of patterns in the part of the visual field that they monitor (their receptive field); other groups of neurons are sensitive to whether the patterns are convex or concave, etc. Such groups of neurons are often termed feature detectors for the corresponding feature. Similar examples of systems that correlate features of the external environment, or the relation of
the external environment and the agent, are abundant in descriptions of brain organization. All too often neuroscientists, and cognitive scientists in general, ascribe the property of being a feature detector to subsystems solely on the observation that there exists a correlation between external features and the subsystems. A neuroscientist may implant an electrode near the subsystem and, observing the correlation in the data — usually the measurements of excitation in the neurons — concludes that there may be a feature detector. The question that is often not asked is for which system is the alleged feature detector detecting the feature, the system under investigation, i.e. the animal to whose brain the electrode is attached, or the larger system including the measuring apparatus and the scientist. The subsystem correlated to the feature can count as a feature detector for the system only if it has an effect on the rest of the agent, and ultimately on its behavior, that is dynamically sensitive to the correlation. The sweat glands on my forehead are correlated with the external temperature, but they hardly can be regarded as feature detectors for temperature for my cognitive system. A feature detector, then, must both correlate with a feature and interact with the rest of the cognitive system as to make the feature relevant for its operation. A feature detector is a kind of information filter. It is sensitive only to specific variations of the environment and the relation of the environment and the agent, and it is not sensitive to others. When appropriately integrated in an informational system in the agent, including other feature detectors and mechanisms to coordinate them, a feature detector, in virtue of its information filtering capacities can help isolate useful systems in the environment, making them effective sources of information with lower conditional information entropy to the agent than that of a larger, less selected super-system in the environment. In other words, the structural constraints that feature detectors impose allow the lower capacity channel to be sufficient for the control purposes of the agent by selectively targeting important features of the environment, and ignoring less important features.
Representation: Representation is a concept that has received special attention by philosophers. There isn’t however significant consensus about what kinds of systems must be regarded as representing. The lack of consensus remains in both the structural (syntactic) characteristics of representational systems, the nature of the relational (semantic) characteristics of the systems, and the relevance of pragmatic characteristics of representational systems. In other words, the topic is difficult, and therefore, I cannot hope to provide a theory of the nature of representation in the boundary of this paper; nor is it necessary to be convinced that the ability of a cognitive system to utilize representations, as narrowly or as widely one may interpret the notion, provides a mechanism for overcoming the informational limitations of the agent. Representation, which in its minimal rendition captures the idea that the agent possesses a mechanism that can correlate with actual or possible states of the environment (including internal to the agent states) or the relations between the agent and the environment, and that can do so in internal systematic ways utilized by the agent to take advantage of the correlations, facilitates both strategies for conditional information entropy lowering. Like in the case of memory, representations can allow past information to affect the present. Representations have the potential to move beyond memory in that they can be generated in other ways (e.g. reasoning, to be discussed below) that can enhance the informational content, i.e. the structure of the dispositional characteristics of the control system of the agent, beyond what is available through the occurent channels of information. Like feature detectors, representations may depict and organize in the representational system targeted features of the environment, focusing the source of the information. Representations move beyond the role of feature detectors, in that they need not be related to sensing systems that organize the current input of information, and in that representations can be more systematically organized. Representational systems are powerful informational devices exactly because they allow the integration of those strategies by

\[16\] Stronger conceptions of representation demand conditions such as, decouplability, i.e. the ability of the agent to use representations without the presence of external representant (Haugeland), homuncularity, i.e. the presence of internal isolated informational systems that communicate through the code of the representational system (Wheeler), etc.
simultaneously integrating and filtering information, increasing both the scope and accuracy of the informational control mechanisms of the agent.

**Reasoning:** The last general cognitive capacity that I discuss is reasoning — another philosopher’s favorite. In philosophical discussions the word ‘reasoning’ is used to refer to systems that manipulate propositional systems with the help of some inference rules. Cognitive scientist often use the term in a broader form — talking about e.g. spatial reasoning. I will call reasoning any general process that manipulates representational structures in a way that the structures are transformed to obtain alternative representational structures, with the goal of extracting more information, obtaining more accurate information, transforming the states in a way that it is easily utilized by another system, and/or constructing a representational state that must satisfy some specific condition (for example through search). Both discursive reasoning and spatial reasoning are instances of this more general characterization. But so can be the process that a composer engages in when generating and manipulating a piece of music in her head, or the process a chef uses to select the ingredients and procedures for creating a meal with a particular flavor. Reasoning is clearly a mechanism of conditional information entropy lowering because it is a mechanism for getting more out of whatever information is already there so that the novel state can match, hopefully, the environment, past, present and future better, and allow better action.

I hope it is clear that all these cognitive capacities — learning, memory, feature detection, representation, and reasoning — are mechanisms for overcoming the informational limitations of the agent. They become, of course, most interesting when working together; when the cognitive system can utilize several such capacities in accord. As a result, other cognitive capacities also become possible, such as planning, productive imagination, communication, etc. Saying that these capacities are mechanisms for overcoming informational limitation is not saying that they are merely such. The possession of selective sets of such capacities opens the possibility for systematic taxonomization of cognitive systems. They offer possibilities
where new goals may emerge, going beyond (and sometimes against) system survival — such as ones related to social organization of cognitive systems, or ones related to systems of ideas (memes). The characterization of the function of cognition that I provided is, in this sense, only a characterization of the beginning for cognitive science, not of cognitive science as a whole. The interesting work that cognitive science has done, and has yet to do is about what happens with the particular mechanisms for overcoming the informational limitations, and also with what novelties those mechanisms produce beyond the minimal characterization stated here.