The cognitive agent: Overcoming informational limits

Orlin Vakarelov

Abstract

This article provides an answer to the question: What is the function of cognition? By answering this question it becomes possible to investigate what are the simplest cognitive systems. It addresses the question by treating cognition as a solution to a design problem. It defines a nested sequence of design problems: (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? This provides a corresponding nested sequence of system classes: (1) autonomous systems, (2) (re)active autonomous systems, (3) informationally controlled autonomous systems (autonomous agents), and (4) cognitive systems.

This article provides the following characterization of cognition: The cognitive system is the set of mechanisms of an autonomous agent that (1) allow increase of the correlation and integration between the environment and the information system of the agent, so that (2) the agent can improve the selection of actions and thereby produce more successful behavior.

Finally, it shows that common cognitive capacities satisfy the characterization: learning, memory, representation, decision making, reasoning, attention, and communication.

1 The Thinning of Cognition

The concept of cognition is thinning. Once focusing on the human mind, cognitive science has gradually begun to investigate cognitively simpler organisms. Mammals (and birds) have been part of psychology since the time of Pavlov, but more recently much simpler organisms have attracted the attention of cognitive science. Investigations of insects, especially social insects, have demonstrated incredible sophistication of adaptive behavior, including complex pattern recognition, communication, and learning.[1] There have also been some provocative studies of bacteria and bacterial colonies suggesting that the notion of cognition may be used to describe the organization and behavior of microbial organisms.[2] Here is the problem: cognitive science has gotten away with a vague, prototype-driven concept of cognition, but stretching the concept to insects and single-cell organisms demands a more systematic discussion and ultimately a definition of a *theoretical* concept that outlines the subject matter of the discipline.

Without such a definition, debates about whether some (or all) bacteria possess some form of cognition are susceptible either to trivialization of cognition (on the pro side) or to cognitive chauvinism (on the con side). In this article I offer such a definition.

Surprisingly little has been said about what cognition is, that would allow us to address the problem of conceptual thinning. Available approaches to the nature of cognition can be grouped into one of four categories:

1. Within cognitive science, when offered, explicit definitions describe prototypical features of cognitive systems. A typical example of such definitions

Department of Philosophy, University of Arizona

Corresponding author:
Orlin Vakarelov, Department of Philosophy, Social Science Building Room 213, University of Arizona, Tucson, Arizona, 85721, USA
Email: okv@u.arizona.edu

is: "Cognition refers to the mechanisms by which animals acquire, process, store, and act on information from the environment. These include perception, learning, memory, and decision-making" (Shettleworth, 1998, p. 5). Here is another example, this time focusing on bacterial cognition: "The term cognitive refers to processes of acquiring and organizing sensory inputs so that they can serve as guides to successful action. The cognitive approach emphasizes the role of information gathering in regulating cellular function" (Shapiro, 2007, p. 812). Such definitions are suggestive of the nature of cognition, and play an important rhetorical function in the monographs where they appear, but they do not offer the precision needed to analyze the thinning problem. However, they express important heuristic ideas, such as the idea that cognition is related to processing of information from the environment to guide behavior, or the idea that capacities such as "perception, learning, memory, and decision-making" are central for cognition.

2. The artificial intelligence (AI) motivated foundational debates about cognition (or intelligence), which have been most influential for philosophical debates about cognition, have followed an *architectural* approach. They have focused on the general functional mechanisms of implementing cognition, and on the modeling tools needed to describe it. This category includes the symbolic computational (GOFIA) approach, but also the connectionist, dynamicist, and distributed approaches to cognition. It is probably a bit unfair to group all of these approaches together because only the computational approach has attempted an explicit characterization of cognition/intelligence—the physical symbol system hypothesis (Newell & Simon, 1981). However, if we must interpret each of these approaches as offering a necessary condition for cognition (sufficient condition will trivialize cognition), then it means that cognition is characterized (partially) by the nature of its architecture. This ought to be unsatisfactory because there is no reason to think that the kind of phenomena that cognitive science studies can be characterized by a common necessary architecture. A general architectural approach can at best offer simulatability; it cannot offer a definition.

3. There has been a renewed interest in connecting more closely the phenomenon of cognition with life. Unlike the more functionally based architectural approaches to cognition, which view life as but an implementation medium for cognition, this *biogenic* (Lyon, 2006) approach insists that "cognition is a biological phenomenon and can only be understood as such" (Maturana & Varela, 1980, p. 7). The thesis is a version of what Godfrey-Smith (1996) described as the *strong continuity thesis* about the relation between life and cognition. According to it, life is a necessary condition for cognition. A cognitive system must be living if it is to count as cognitive. In this tradition, the problem of defining cognition has been reduced partly to defining life/metabolism, or to identifying an aspect of life that supports cognition. Thus, Maturana and Varela define the notion of *autopoiesis* (see 2.1), which is supposed to imply an organizational definition of both life and cognition. Observing that the notion of autopoiesis is too weak to imply interesting cognition, Bitbol and Luisi (2004) argue that cognition is a special kind of adaptive metabolism. The problem with existing nontrivial definitions/characterizations of cognition within the biogenic program is that they do not connect properly to more advanced forms of cognition. This is not a criticism of the program itself (this article can be viewed as a part of the program) but only of the limited attempts of defining cognition. Note that weaker forms of the continuity thesis, which are more plausible, cannot use the nature of life as a defining condition for cognition. They demand an independent definition.

4. There are a very limited number of attempts to offer an explicit definition of cognition with a set of sufficient conditions—a mark of the cognitive—that are not limited to high cognition. I know of only one such attempt by Rowlands (2009, p. 8):

A process P is a cognitive process if and only if: (1) P involves information processing—the manipulation and transformation of information-bearing structures. (2) This information processing has the proper function of making available either to the subject or to subsequent processing operations information that was (or would have been) prior to (or without) this processing, unavailable. (3) This information is made available by way of the production, in the subject of P, of a representational state. (4) P is a process that belongs to the subject of that representational state.

Rowlands uses this definition to argue for an extended cognition thesis whereby it is shown that some extended systems (a human using some artifacts) qualify as cognitive systems. It is unlikely that Rowlands' definition suffices for the thinning problem because it is not targeted to the simplest cases of cognition. The problematic notion here is the notion of "representation," which for Rowlands has a strong intentional dimension requiring consciousness.

While none of the attempts at a definition are satisfactory for solving the thinning problem, they offer important insights to be preserved in a successful definition.

So, how should we proceed towards a definition of cognition? What would it mean to have offered a characterization of the natural phenomenon of cognition, appropriate to outline the domain of cognitive science? I will make the following assumptions: (1) modern cognitive science has a fairly good implicit grasp of the *domain* of higher cognition; (2) cognition, like most other biological categories, defines a *gradation*, not a precise boundary—thus, we can at best hope to define a direction of gradation of a capacity and a class of systems for which the capacity is relevant; (3) cognition is an *operational* capacity, that is, it is a condition on mechanisms of the system, not merely on the behavior of the system—to say that a system is cognitive is to say something general about *how* the system does something, not only *what* it does seen from the outside; (4) cognition is a phenomenon of *organized complexity* as modus operandi—it is a product of the gradual emergence of complex structure in the world through various processes of incremental self-organization of far-from-equilibrium systems; and because of this, a general theory of cognition must be sensitive to the inherent historicity and hierarchical organization of complex systems (Nicolis & Nicolis, 2007).

I assume the following model: There is a thermodynamically open system O, that is, a system exchanging matter and energy with the environment. The interaction is organized and modulated by some (functionally) internal subsystem C of O. The question of what makes the system cognitive is a question about what role or *function* C has for O. Here I use the term *function* in a sense akin to Cummins' sense, that is, the function is determined by the role C plays in the organization of O and the interaction of O with its environment (Cummins, 1975). If we call O the organism of the scenario, the question about what systems are cognitive can be analyzed via the question: *What is the general function of cognition in an organism?* The systems that have and use mechanisms that fulfill the function are the cognitive systems. Of course, as is usually the case with functions, the same mechanism may have other functions, but cognitive science investigates the *cognitive* function of the cognitive mechanisms of cognitive systems.

I approach the analysis of the function as an analysis of a *design specification* for a system. We outline a general problem that must be solved—the design problem—and we can ask whether a particular mechanism solves (or improves towards a solution, *approaches*) the problem. We can identify the function of cognition by first identifying a design problem faced by a class of systems. Identifying the class of solution strategies to the problem allows us to identify 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 class of strategies for solving the problem. Because *solution* is a success term, while some problems can only be approached, as a marginal improvement—the problem related to cognition will turn out to be of this class—in the discussion to follow the terms *solution* and *strategy* will also refer to approaches.

With these considerations in mind, I will confront 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. Every design problem for a class of systems defines a set of solutions for the problem—those systems within the class that satisfy the design specification. Once such a class of solutions is defined, we may define a further design problem related to improvement strategies for the solution, and so on. I will apply this methodology until I reach a class of systems for which we can define a design problem that demands the kinds of strategies normally associated with cognitive mechanisms. Once we identify the function of cognition, we can isolate the class of cognitive systems.

The strategy can be described as follows: There is a distant place we want to go to and surround it with a fence in a natural way—we want to go to the uncontroversial cognitive mechanisms and outline an inclusive domain of cognition. We start from a place with secure foundations and no contamination of cognitive terminology. The place is a domain of natural systems. We categorize the domain and ask: which of the categories is most likely to lead us to the remote place—which subdomain is most likely to contain the target place? We proceed until we have reached the smallest neighborhood of the target place that can be isolated by natural, local concepts. In this way we avoid circularity in the definition.

This methodology of analysis is particularly useful for phenomena in organized complex systems, because the sequence of design problems and solutions can correspond to a sequence of steps of development of complexity of organization. Thus, the sequential narrowing of the class of systems offered by the process is not merely a constructive argumentative technique that can be discarded after it is performed. It is an essential part of the understanding of cognition. To use a slightly modified Wittgensteinian metaphor from the *Tractatus*, we are building a ladder to reach to our target concept. We start from a safe place—living systems—and reach to cognition. But unlike Wittgenstein who at the end kicked the ladder, we cannot. The ladder is part of the final product. We can at most reposition the ladder on an alternative footing. We can move from the realm of living systems to artificial cognition (how to do that will be the subject of further work), but the structure of the ladder will always be imprinted in the cognitive systems.

Let's clear the suspense and introduce the main idea about the function of cognition with an analogy. Imagine a mathematician proving a complex theorem. She starts with a collection of inputs—simple mathematical facts. With a large collection of formal and informal transformations, our mathematician generates a complex proof that supports the statement of the theorem. Why is this complex process necessary? Why does the solution require "cognitive" work? Why doesn't the mathematician simply _grasp_ the theorem directly the way she may grasp many simple theorems? In an important sense, the proof is 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, importantly, on the capacities of the mathematician. If our mathematician had unlimited intellect—if she were a god—then she would directly comprehend the theorem, just as we can comprehend that $1 + 1 = 2$. The mathematician is not a god, however. She is intellectually limited in her capacity to comprehend complex mathematical theorems. She needs to do significant cognitive work to overcome this limitation. To this end, she uses a systematic, logically guided proof construction. The process of proving a complex theorem and the capacities needed to do it are necessary precisely because the human mind is quite limited—the work of doing the proof compensates for the mathematician's informational limitation. Here is what I am up to. I argue that this phenomenon of _compensating for informational limitation by doing work_—as we see with the mathematician—is at the root of all cognition. I claim that the general function of cognition is to compensate for the informational limitations that actual agents embedded in a world face. The rest of the article will argue this.

The article is organized as follows: Section 2 develops the sequence of design problems, and simultaneously a nested sequence of system classes, leading to the ultimate problem whose solution requires cognition—the informational limitation problem. Section 3 offers an analysis of the strategies for approaching the problem, offers a precise definition of cognition, and argues for the correctness of the definition. Section 4 briefly suggests how the question of artificial cognition may be approached based on the definition offered here. Finally, Section 5 offers a few concluding remarks.

2 Autonomous Agents and Informational Limits

In this section, I will describe a nested sequence of design problems. I will associate each design problem with a class of systems whose organization can be regarded as a "good" solution to the problem. The nested sequence of 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), and (4) cognitive systems. The distinctions are not sharp: almost everywhere there is gradation among the system classes.

2.1 System Persistence and Autonomy

The most rudimentary design problem begins here: if there is cognition, there must be a system. 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 arise. The first design question that must be examined is: _What allows systems to persist as individual entities_? More specifically: _For which of those systems that persist is a capacity of cognition relevant_? This design question is targeted to naturally emerging systems. Later (in Section 4) this will be abstracted to allow for artificial/designed systems. However, a biogenic approach to cognition must be faithful to the biological origins of cognition and must include a story accounting for naturally emerging cognition and a further story about how artificial cognition is possible.

We can identify two broad strategies for system persistence: _robust_ and _dynamic_. For the purposes of this article, I will not define these notions, but will rely on intuitive examples. Thus, a rock is a robust persistent system. It 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 outside the rock—in fact, what is considered to be inside and outside the rock depends on the strength and topological connectedness of the bonds. Robust persistent systems are among the longest persistent systems in the universe. However, strong bonds cannot do much more than persist. There is no need for cognition. Rocks don't need to think any more than they need to eat or sleep. (But see Section 4.)

Dynamic stability is a more complex matter. This is the realm of _dissipative systems_ (Prigogine, 1961;

Prigogine & Nicolis, 1977). A dissipative system is an open nonequilibrium thermodynamical system that maintains stability of an *organizational parameter* by dissipating matter and energy from and to the environment. (i.e., there exists an appropriate parameterization of the system such that an important parameter has a stable dynamical orbit. See Pillay, 2008, for an introduction to stability theory.) We can distinguish two classes of dynamically stable dissipative systems: *heteropoietic* and *autopoietic*. In heteropoietic systems, stability (maintaining the organizational parameter) is determined by the boundary conditions of the system as well as a gradient of free energy that can drive the dynamics of the system. Standard examples of such heteropoietic systems are: Bénard cells and water eddies. Bénard cells form when oil in a container is heated from below sufficiently quickly 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. The eddies are driven by the energy gradient of the flowing water and the structure of the channel. Heteropoietic systems can be quite stable—the Great Red Spot on 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 systems themselves, not merely the boundary conditions, are responsible for maintaining stability. The term *autopoiesis* was coined by Maturana and Varela (1980) to describe a phenomenon where the conditions for maintaining the structure of a system are present within the system. They 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 of the theory is, I think, unsatisfactory, so we will ignore it.[3] One of the most interesting characteristics of autopoietic systems is that they support *process closure*. That is, 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. We can think of the closed system of product formation rules determined by the processes—the rules specifying how a compound is obtained from (or decomposed by) other compounds—as defining the system (given a fixed interval of variation of the external conditions).

Autopoietic systems are interesting for two important reasons. (1) Their dynamic self-maintenance allows them to persist within shallow energy wells—the bonds that hold them together can be extremely weak in comparison with static systems. Moreover, they are systems that genuinely *do* something about their persistence—the internal processes that generate the system perform *work*. (2) The process closure that defines them determines the essential compounds/mechanisms for maintaining the closure, as well as determining a fixed set of functional 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 are maintained by the system provide the *goal* of the processes (no intentions assumed). Autopoietic systems can be regarded as the simplest kinds of *autonomous* systems (Barandiaran & Moreno, 2006; Christensen & Hooker, 2000; Thompson, 2007). I will call a system autonomous if (1) it can be described as having a goal that it "tries" to achieve, and (2) the control mechanisms of the system that veer it towards that goal are part of the system.[4] In autopoiesis the goal is persistence and the control mechanism is derived from the processes in the closure set.

The definition of autopoiesis admits resistance to fluctuations in the external environment, 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 (2005) describes, autopoiesis is a structural condition that a system either satisfies or does not—either a system maintains process closure or does not. The notion of autonomy, however, is a graded notion: for a fixed goal, a system can be more or less autonomous depending on how sensitive the system is to the external conditions and to what conditions it can adapt. Autonomous systems are, therefore, not merely autopoietic systems, but autopoietic systems with further capacities. The capacity that tracks gradation of autonomy is often recognized as *adaptability* (Barandiaran & Moreno, 2006; Christensen & Hooker, 2000; Collier & Hooker, 1999; Di Paolo, 2005). The simplest autopoietic systems possess a minimal sense of adaptability in that they are capable of repairing damage, but the organization of the closed system need not be sensitive to larger variations in the environment. Autopoietic process closure is a purely internal condition of the system organization. A more adaptable system must be open not only to the transmission of matter and energy, which are the resources of the autopoietic process, but also the process itself must be sensitive to the state of the environment. The process need not be entirely internally closed (Barandiaran & Moreno, 2006; Christensen &

Hooker, 2000). The dimension of system organization related to adaptability of the organization-maintaining processes is commonly associated with cognition, either directly, or as I would prefer, as pointing in the direction of cognition (to use the fencing metaphor again).

Let us recap the progress so far. In my attempt to characterize the function of cognition, I isolated the class of autopoietic systems as the only naturally emerging systems (not artifacts) for which cognition could be relevant. This was because these are the only systems whose conditions of *persistence* are determined and controlled from within. The key functional condition here is autonomy. Cognition, minimally, must be related to maintaining autonomy.

Not every form of autonomy should be regarded as a product of cognition however. Cognition, as I suggested in the introduction, should be an internal mechanism that is doing some specific work. The task, the design problem at hand, is adaptability. Adaptability, however, is a behavioral condition. Thus I cannot adopt it as a target notion of the theory. I want to focus on the internal mechanisms that facilitate it.

2.2 Active Systems

One mechanism for increasing the scope of possible viable environments is to maintain process closure that can switch between different modes of operation depending on the state of the environment. Bitbol and Luisi (2004) distinguish between different *kinds* of metabolism depending on whether the system can be in different *modes* of operation based on available nutrients or cell damage. For example, depending on whether lactose or glucose is present in the environment, *Escherichia coli* bacteria can activate different genes that can be expressed to produce enzymes appropriate for breaking down the corresponding sugar (Ben-Jacob et al., 2005). However, there are only so many modes that a system can adapt to and, most importantly, what modes the system needs to adapt to depends on accidents in the environment.

A more flexible strategy for coping with environmental variation is to have some control over the environment—get to where food is available, or make the food come to you, avoid places where you are food, and so forth. 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 systems 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 (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 waste. This can usually be modeled with thermodynamics and theories of diffusion. In such systems it is not especially interesting, even when possible, to model the relations with the notion of control. The notion of control becomes interesting when (1) the coupled dynamical interactions can be decomposed into isolated subprocesses 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, that is, the system can be modeled effectively with the machinery of *control theory* (a branch of dynamical systems theory; Levine, 1996; Hinrichsen & Pritchard, 2005). Whether this is desirable—that is, 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 paddling and "monitoring" nutrition sensors (Blair, 1995), 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 interactions of type (a) and (b) and triggering relations between them, the system can be effectively modeled as a *control* system.[5] When this is possible, we can term the environment-to-system interactions *control inputs,* and the system-to-environment interactions *control outputs*. I call such systems *active* autonomous systems. This allows us to state a second design problem for active systems: *How can an active system perform better control outputs in order to improve its chances of persistence?* The question shifts the focus from conditions on behavior to conditions on control outputs that affect behavior. A normative condition on the entire system is analyzed via a normative condition on the control outputs of the system.

The simplest strategy is to affect the environment in a uniform way regardless of its state. A system may release a chemical that may attract food or repel predators; or a system may move by paddling randomly. A more flexible strategy is to make the current state of the environment relevant for the control outputs. The simplest way of doing this is through implementing a triggering relation between control inputs and control outputs—a *fixed action pattern*. Autonomous systems that operate in this way can be described as *reactive* systems. There are many examples of fixed action pattern behavior in the animal kingdom that are discussed in the ethological literature. It is not clear, however, whether there are natural systems that are only reactive systems. Common wisdom has dictated that many simple animals are reactive systems, but considering

the literature on insect or bacterial cognition, there may not be too many purely reactive systems.

The path to cognition leads in the direction of more "sophisticated" strategies for making the environment relevant for effective control outputs. I focus on this next.

2.3 Agents

As a solution to the design problem of persistence, reactive systems have a weakness. They essentially define a more-or-less functional relation between the control inputs and control outputs—a response function. I describe the relation as *more-or-less* functional, because the systems are complex dynamical systems and in such systems stability of behavior is a difficult luxury. It cannot be expected that the same control inputs would produce exactly the same control outputs, but the outputs must be sufficiently close; otherwise it would be pointless to describe the systems with the machinery of control theory. Going back to the weakness, whether the functional relation in fact allows the system to approach its goal is contingent on the stability of the environment. If the environment changes sufficiently, the same fixed action patterns may have a detrimental effect. The problem is that the response function need not be sensitive to the success of its operation, so that it can be adjusted based on the relation of the organism to the goal.

Now, there may be response functions whose dynamical implementations are based on some feedback mechanisms that corrects the response because some control input is correlated with how close the system is to its goal. Such feedback mechanisms alone do not deserve systematic investigation in terms of cognitive machinery; simple cybernetics may suffice. In the effort to locate cognition we need to consider systems that have dedicated mechanisms for response function control and modulation based on the goal. Such dedicated mechanisms would certainly utilize dynamic feedback, homeostasis, or other such stability-inducing processes as the means to achieve appropriate goal-directed control, but they must be investigated at a level of abstraction that goes beyond dynamics and control.

We can consider systems that admit the following internal functional decomposition: The system possesses a subsystem M, (for a model or a "cognitive" map) that mediates the control relation between the inputs and outputs. The system possesses another system P (for a purposeful filter) that modulates M in light of the relation of the system to the goal. This system organization (and terminology) was introduced and investigated by Nauta (1970) following Ackoff (1958). I have expanded on his work to argue that systems with such an architecture can be regarded as semantic *information systems*, where it is insisted that information is a dynamical systems phenomenon (Vakarelov, 2010). In such systems, M can be regarded as having its own macrostates, and the states of M can be correlated with an external system or a collection of features of the environment, S, with its own macrostates. Here by a *macrostate* I mean the standard notion from dynamical systems theory of a collection of microscopic dynamical states (Hinrichsen & Pritchard, 2005; Katok & Hasselblatt, 1996). In this setup, the modulating second-order control role of P, which is sensitive to the relation of the whole organism to the goal, provides the basis for the *determination* of the macrostate structure of M and S.

The idea for this determination is as follows: The organism is engaged in highly organized coupled interaction with its environment. The patterns and invariances of this interaction can be described in terms of macroscopic relations on the systems based on macroscopic distinctions inherent in the viability conditions of the system and thus its goal states. In the case of autopoietic systems, as discussed above, the viability states are determined by the system states that maintain the appropriate organizational condition of autopoiesis—the appropriate process closure. When probability distributions of the likely future trajectories of the system based on environmental contingencies are available, that is, when one can provide a measure of how close the system is to the goal (or to the danger boundary) it becomes possible to assess how the behavior and the control functions of the system relate to these macroscopic distinctions. Thus, the entire system–environment complex can be analyzed with a higher level of abstraction (reducing the free parameters of the system). In this case, we can determine the appropriate distinctions in the environment relevant to the organism (the macrostructure of S) as well as the relevant distinctions witnessed in M (the macrostructure of M). In this case, the macrostates of M can be interpreted as informational states and the states of S as, to borrow Gibsonian language, something like the affordances in the environment related to S. There is a sense in which the states of M can be regarded as being about the states of S, but I urge caution when making this semantic connection. In Vakarelov (2010) I define the meaning of the states of M to be the *interface role* they play in the control system, which includes, but is not identical to the correlation between M and S. Nonetheless, if appropriate probability measures exist, relative information measures, as in Shannon's theory of communication (Shannon, 1948), can be defined between M and S. I will use such measures in Section 3. The nature of semantics in this account of information systems is not important for the current project.

Let us consider a simple example of an information system. Consider a bacterium that has a detector for two types of nutrients that may be present in the environment. The bacterium needs to be able to switch between three metabolic cycles depending on the availability of the nutrients (one default with no nutrients). Assume a mechanism, a control gene, that is triggered by the sensor and that initiates the different metabolic cycles depending on the state of the detector. Assume also that extracting energy from one of the nutrients has some negative side-effect for the bacterium, such that the nutrient should be digested only if the bacterium is in real need of food. Otherwise, the benefit is less than the possible harm. Imagine, then, another mechanism that is sensitive to the general health of the bacterium so that a compound is produced in proportion to the health that can bind to the control gene and modulate its expression. In essence, the second mechanism can modulate the role of the control gene and allow switching to the nasty food metabolism only if the bacterium is in danger of starving. In this system the control gene can act as M, the modulating mechanism can act as P, the goal is, naturally, maintaining good health, and the source of the information is the near environment divided according to three macrostates—presence of the preferred nutrient, presence of only the nasty nutrient, and neither. This constitutes a simple information system. The system is so simple that it may not be worth describing the control mechanism of the bacterium as an information system, but the option is available. Real bacteria, as the literature cited earlier indicates (see note 2), have more complex control systems.

Why are information systems relevant in the attempt to define cognition? Clearly, there is a tradition in cognitive science to view cognition as involving utilization of information from the environment (see Section 1, discussion of category 1). This consideration, while suggestive, is not sufficient. The important consideration is *what* information systems offer. First, information systems allow the goal of the system to enter explicitly in the control mechanism P, as a dedicated subsystem. The performance of the system in light of the normative significance of the goal need not be described only through the global behavior of the system, but may be described through the performance of the dedicated mechanism. Second, the mediating subsystem M focuses the environmental significance on action control so that the control relation can be modulated by the purposeful system.

Information systems are important for cognitive science not simply because they offer a new level of complexity of behavior and adaptability lacking in reactive systems; although they ultimately do that. They are important because they contain an independent, functionally localized goal-directed environmentally sensitive control system—a system needing independent investigation, needing its own science. An organism whose internal organization is so deluded as not to allow a useful separation between different functions, does not require a separate science of cognition. The behavioral complexity of such an organism alone is not a reason for calling it a cognitive system. Cognitive science is not the science of complex behavior; it is the science of the dedicated mechanisms that generate the complex behavior.

No claim is made that information systems are sufficient for cognition; however, they offer an important stepping stone towards cognition. I suggest that we call an organism that is partly controlled by an information system an *agent*. This is not offered as a conceptual analysis of *agent*. It is offered as a theoretical definition motivated by the intuition that agents are distinguished by their capacity for systematic goal-directed behavior. It will be useful, and suggestive, to call control inputs to an agent's information system *percepts*, and the corresponding control outputs *actions*.

The idea is to view cognitive systems as a type of agent. Within the informational system framework I formulate the third design problem for the class of agents: *How can an agent use better information to control its actions?* The question shifts the normative focus from the actions to the quality of the utilizable information from the environment.

Before I investigate this design problem in the following section, let me eliminate some possible misunderstandings and confusions related to the suggested notion of information system and the suggestion that it is necessary for cognition.

(1) I regard the concept of information system to be more primitive than the notions of information and information state. Information systems are special kinds of highly organized open dynamical systems. The notion of information is viewed as the currency of the information system, analogous to the way money is the currency of an economic system. The notion of information state, which has the role of the information vehicle (the data), is determined by the macrostructure of the coupled dynamical system. This is a pragmatic conception of semantic information where notions such as data or meaning are ultimately determined by looking at the mode of interaction of an organism with an environment, and the control role the informational mechanisms play in the interaction.

Thus, (2) in light of using information systems to move towards a notion of cognition, I must note that this notion of information more closely resembles Gibson's (1979) notion of information than the communication or processing notion of information more commonly used in discussions of cognitive science.

The notion of information system does not determine what happens to the information inside the system, how it is processed or whether some "computation" takes place. The notions of *information system* and *informational processing system* are distinct and orthogonal. An information system may be implemented with an information processing (even symbolic processing) mechanism inside, but it doesn't have to be. One key property of information processing systems is their functional separation from other systems. Information processing systems can be described completely by specifying inputs, processing/computation operations, and outputs. An information system need not admit such a functional separation. In the case of biological systems especially, such a functional separation may be impossible, except in limited cases such as explicit symbolic manipulation in humans. In naturally emergent information systems, embodiment and close-coupled relations between the organism and the environment are ineliminable. In fact, they structure both the informational states of the system and the external source of the information.

This takes us to, (3) that information systems need not be representational systems. I do not claim that *M* represents *S*. The philosophical treatment of the concept of representation is messy. I do not wish at this stage to enter into debates about what constitutes a representation. Information systems should not be viewed as nonrepresentational systems either. Some information systems may legitimately be described as operating with representations. Human cognition certainly relies heavily on representations. Any position that denies this is based on ideology, not on science. (In light of this, I regard the division of the approaches to cognition between representationalist and anti-representationalist as a red herring.)

Finally, (4) the importance of the notion of information for cognition has been explicitly criticized[6] by proponents of the dynamicist program (Chemero, 2009; Thelen & Smith, 1994; van Gelder, 1998) who insist that cognition is a dynamical system phenomenon (an ontological claim), and it should best be investigated with the machinery of dynamical systems theory (a methodological prescription). What is the connection between my approach to cognition and the dynamicist program? I clearly deny the methodological prescription. In fact, I deviated from the methodological prescription in Section 2.2 already. Dynamical system description of even fairly simple systems—systems for which cognition is relevant—is not plausible in practice. The complexity of even a simple bacterium is so great that an explicit description with differential equations is outlandish. Much more effective descriptions are available. In some cases it may be possible to describe aspects of the behavior of the bacterium as an information system in a more manageable way. I however do not deviate from the ontological claim. Indeed, the very concept of information system is a concept of dynamics. In Vakarelov (2010) I sketch an in-principle way of describing information systems using dynamical system theory, including more exotic developments such as synergetics (Haken, 1993, 2000). Thus, the language of dynamical systems is probably indispensable for a mature science of cognition. Indeed, some of the examples offered by dynamicists, such as walking or finger tapping (Kelso, 1995), bouncing a ping-pong ball, or even performance on the A-not-B task (Thelen & Smith, 1994), may be best described as dynamical systems. Conclusions of this can be, however, that some of the examples were incorrectly regarded as cognitive phenomena (some aspects of walking or finder tapping may be like that, even if a brain is involved), or that information-based cognitive phenomena only modulate the dynamical system (in the case of bouncing a ping-pong ball), or the informational mechanisms create a platform with intrinsic dynamics that is locally best described with dynamical system models (in the case of the A-not-B task). The list is not exhaustive.

2.4 Informational Limits

Any natural system would be severely limited in terms of what information from the environment reaches it at any moment, and how it can respond based on the information. It is important to understand this claim in the context of information systems. In an information system, the notion of information contains several dimensions:

1. The first dimension is related to the possible differences in the environment that may be relevant for the operation and well-being of the system. In the case of autopoietic systems, any difference in the environment that has an effect on the state of its structural organization is a relevant difference. For example, fluctuation in the distribution of viable matter is relevant, so is the existence of remote meteors that may potentially strike the system. Some differences in the environment may be irrelevant for the system. At the extreme end are differences such as whether a neutrino is passing through the system; other more macroscopic irrelevant differences may include minor fluctuations in nutrients, or aspects of the internal organization of other organisms whose external influences are filtered by their structure.
2. The second dimension is related to whether the differences can propagate physically to the system. Many things can get in the way: broken causal

links, too much noise in the environment that washes away the correlation, or insignificant effects on the system. It is not sufficient for there to be a correlation with external differences in principle. It may be the case that every event in the universe is reflected in every subsystem of it—imagine some quantum coupling does that. Even if every system can serve as a measuring apparatus for every difference in the universe, it does not mean that the information system is sensitive to the difference. Electromagnetic radiation (not light) from my laptop has a differential effect on the state of my brain as a physical system, but it does not follow that my brain, qua cognitive control system, is sensitive to the radiation.

3. The third dimension is related to whether the differences can have a control significance for the system. The human eye has more than 100 million light receptors, each capable of a large number of possible responses (say N). Thus, each eye is capable of making more than N^{10^8} distinctions. That many distinctions can never be relevant for a natural agent, for it cannot perform a compatible number of distinct actions. Only a small number of (equivalence classes of) received distinctions ever obtain a systematic control role for outputs of the system.

4. The fourth dimension is related to whether those distinctions which may have a control role can be modulated by the goal-tracking mechanisms. The distinctions generated by the sucking motion of a baby's mouth on a mother's nipple can propagate to the mothers body initiating the secretion of milk. The process, however, is not modulated by a purposeful system on the side of the mother (while, it is on the side of the baby; and of course, the mother can override the causal link by pulling away the baby).

An information system has some information, that is, M is in an informational state, only if all four dimensions are active: that is, (1) if there exists a relevant difference in the environment, (2) the difference can be reflected in the system, (3) the reflected difference can, in principle, have a control significance, and (4) significance can be modulated by the purposeful system.[7]

If we view an agent as facing the problem of performing the best (or good) action in light of the state of the environment, it would seem that a successful strategy would demand that the distinctions relevant in the environment would propagate to the control system of the agent so that they can be used for steering the goal-directed behavior. Let us call this a know-it-all strategy. Looking for such a strategy would make sense only if there is a reasonable possibility of the relevant distinctions entering the agent as information, that is,

satisfying all four dimensions. The described dimensions, however, are susceptible to several information bottlenecks—in particular, between dimensions 1 and 2, 2 and 3, and 3 and 4. Thus we can ask (a) can the distinctions in the environment be reflected in the system? (b) can the distinctions reflected in the system acquire control significance? and (c) can the distinctions that have control significance be modulated by the purposeful mechanism? Questions a and b are especially vulnerable. In complex environments, the kind of environments where natural agents may emerge, it can be assumed as a natural fact that the number of distinctions in the environment relevant for an organism is astronomically larger than any distinctions reflected in the system. Think of this as a poverty-of-stimulus argument on steroids. Moreover, it can also be assumed that the number of distinctions reflected in the system is considerably larger than the distinctions that may have a control significance. (But I will not assume that the number of distinctions relevant for purpose-guided behavior is considerably smaller.) For the purpose of this article, I consider these assumptions to be logically contingent, but empirically and conceptually sound.[8]

All natural agents therefore, including the mathematician that I discussed in the introduction, are severely informationally limited, in the sense that the structure of the environment is vastly too complex to be internalized in the control system. 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 information limitation.*

Let us call this the *information limitation principle*. I regard this as a naturalistic constraint on any theory of cognition. Any theory of cognition that ignores or idealizes away ILP is either not a theory of cognition or not a naturalistic theory. A consequence of this principle is that the know-it-all strategy is not an available option for agent architecture.

By adopting ILP we can formulate a fourth design problem: *How can the internal organization of the control mechanisms of the agent be improved to reduce the informational limitations?*

Note that the design problem does not demand that the information limitation is eliminated. This would be impossible. The problem cannot be solved by only getting more information. Rather, the design problem calls for strategies that reduce the limitation. I claim that this is the class of the cognitive systems, and the mechanisms that reduce the limitation are the cognitive mechanisms. To be cognitive is to be limited and to be able to do something about it.

3 The Function of Cognition

The central proposal, the theoretical hypothesis of this article, is that the most general conception of cognition is that cognition refers to the set of mechanisms in an agent that address the fourth design problem—the information limitation (IL) problem. The work in Section 2 served to identify the problems for which cognition is relevant and to describe the classes of systems for which the problems arise. While an informal conception of cognition was used in the process that guided the theory construction that led to the information limitation problem, the concept of cognition did not enter the actual theoretical definitions. In this section I will argue that the IL problem gives a good general conception of cognition that may be useful in addressing the thinning problem. To this end I will accomplish three tasks: (1) I will analyze possible strategies that address the IL problem; (2) I will offer a precise characterization of the function of cognition; and (3) I will demonstrate how prototypical capacities associated with cognition are captured naturally by the characterization. Task 3 will serve as the primary inductive support for the claim that the characterization offers a good theoretical definition of a concept of cognition.

3.1 Overcoming Informational Limitations in Agents

What does it mean to have a strategy for overcoming information limitation in an agent? It is useful to take the statement apart. According to the above definition (see Section 2.3), an agent is a system with at least one goal and whose behavior is partly controlled by an information system sensitive to the goal. The source of the limitation is the huge discrepancy between the differences in the environment relevant to the agent's goal and the ability of the control system of the agent to act selectively based on the differences. Thus, a measure of the limitation would be related to the connection between the differences in the environment and the witness of the environment for the information system, namely the subsystem M. Note, however, that the definition of an information system does not include the entire environment as a potential source of information. Rather, it specifies a subsystem of the environment, S, as the source, whose macrostate structure is determined by its dynamic interactions with the agent (this includes its intrinsic organization). Thus, at any moment the relevant connection is between M and S. However, what S is, and what its macrostates are, depends partly on the agent's behavior; thus, S itself can be modified as a result of changes in the agent's organization and control outputs. In an agent, therefore, a strategy for overcoming the information

limitation would be an activity (performance of work) that allows a better coordination of the different macrostates of S and M so that the purposeful system can modulate the control system towards more accurate and adapted actions in light of the goal.

ILP eliminates some strategies; in particular, it eliminates the brute force solution to the IL problem. It cannot be assumed that all the agent needs to do is get more raw data. A natural agent would get overwhelmed quickly. Figuratively speaking, and quite suggestively so for the idea of cognition, the solution needs to be "smarter."

We must also disregard external intervention strategies. That is, a strategy cannot depend on an external system modifying the agent so that the information limitation is reduced. In the case of evolvable systems, the strategy cannot be for a new, better agent evolved by some mechanism of random variation and selection. Evolution can produce more adaptive systems, including better cognitive systems, but itself is not a mechanism of cognition. We are interested in strategies that involve modifications of the agent by the agent itself. Put generally, when the problem of information limitation is investigated for our purpose, we must not compare different agents, but a single agent across time. (Exotic forms of Lamarckian-like evolution, where acquired traits can be passed to the next generation, can be included. It is unlikely that such mechanisms exist in natural evolution, but they can be imagined in artificial evolution.)

For the purpose of investigating the strategies for overcoming information limitations we can make the following assumptions: (1) Some aspects of the agent's organization are fixed for the information system. Such aspects, however, can play a central role for the operation of the system. The constraints on the sensing system(s) of an organism (the size, position, and makeup of the eyes, for example) or its body (the length and arrangement of the bones, or the elasticity of tissues) are essential for the functioning of the control system of the organism. Nonetheless, these fixed constraints cannot be modified by the organism. They affect the extent of the information limitation of the organism, but they cannot be a part of an improvement strategy. (2) A strategy must involve mechanisms that modify the internal organization of the information system, the relation of the agent to the environment, and the environment itself (and potentially the very mechanisms). The mechanisms may be parts of the information system, or they may be additional systems that do not directly participate in control or modulation of control. They may also be implemented by actions—control outputs—as when the system changes its position to see better, or when the system modifies the source of the information, as in cutting a fruit to

look inside. The mechanisms may also be implemented by internal changes of organization that are not outputs of the information system (or are outputs of other parallel information systems).

It should be obvious by now that we cannot hope to characterize every possible strategy for reducing the information limitation of the agent. However, it is possible to outline several classes of strategies. These strategy classes demonstrate why the function of cognition can be connected to reduction of the information limitation.

One way of reducing the discrepancy between the source of information S and the medium M is to build in more structure to M (and its control dispositions) than can be facilitated by the immediate informational connection to S. This may be done in various ways, but the most immediate is to take advantage of the historical dynamics of M, and its interaction with S and the rest of the information system. It is always the case that the current state of M (and its control dispositions) depend on both history and current control inputs. We get this simply by the fact that M is a dynamical system. However, if its dynamics is right, the history of its interactions may "collect" temporally extended information that allows patterns of S's behavior to be reflected in the way M controls the outputs of the system. If the patterns of historic interaction with a system "contain information" about the dispositions of the system *now*, and if the "information propagates" to the correct dispositional state of M, the agent may be able to "anticipate" the future behavior of S. Thus, the agent's actions may be targeted to more effective satisfaction of its goals. The idea of *anticipation*, and the corresponding notion of *anticipatory system*, has been suggested to be central for cognition (Collier, 1999; Collier & Hooker, 1999; Davidsson, Astor, & Ekdahl, 1994; Dennett, 1991; Rosen, 1985). In my account it emerges naturally as a consequence of a strategy class to the IL problem. The notion of anticipation, however, is behavioral (except in the original systematic treatment of Rosen, which uses the idea of an internal model of the environment). One advantage of my account is that anticipation emerges from an investigation of internal mechanisms in the agent.

A more complex way of building more structure in M is, in addition to temporal dynamics, to provide internal mechanisms that modify M concurrently in a way that offers a better match between S and M. This method is favored by representationalist models of cognition. The cognitive system is assumed to receive a decoupleable input, the input is processed and analyzed, and then an output is generated. The focus of this account of cognition is the processing part. If information is processed symbolically, the approach reduces to GOFAI. The goal of information processing is usually extraction of more (or more salient) information from the input, as well as other related purposes, such as selective storage of information for future use. Traditional representationalist accounts of cognition differ from my approach, or more generally from the enactive/situated tradition to which I belong, primarily in that for the enactive tradition the physical structure of the agent and its interactions with the environment play an indispensable role. However, the proposed cognitive mechanisms investigated by the different schools of representationalism fit naturally into my framework, when representational machinery is properly integrated in an information system, as mechanisms that offer a reduction of information limitations.

A broader characterization of this class of IL reduction strategies is: utilization of internal mechanisms that modify in-agent information vehicles in order to extract more information from the available inputs. This idea has been suggested by others as well. For example, Ben-Jacob et al. (2005) and Ben-Jacob (2009) describe this as extracting *latent information* from the environment, where "by *latent information* [Ben-Jacob et al.] refer to data embedded in the environment that, once processed cognitively, initiates change in the organism's function or behavior" (Ben-Jacob, 2009, p. 79). Extraction of latent information is seen as a central characteristic of cognition. A characteristic that, Ben-Jacob argues, is found in some bacteria and bacterial colonies.

A second way of reducing the information limitation is when the organism carefully controls its interaction with the environment, where the limited channel of interaction between S and M is systematically monitored to assure that the most relevant information for current actions is available. This idea historically has been emphasized by the ecological approach to cognition pioneered in psychology by Gibson, and in AI by Brooks. The idea is that the organism need not internalize the world completely. Rather, it suffices for the organism to maintain the right invariance in its sensory array (its control inputs) and only react to a small number of distinctions (Gibson, 1979). A metaphorical description is that the organism offloads its informational problem to its connection with the environment: "The world is its own best model" (Brooks, 1991, p 417).

A third way of reducing the information limitation is by changing the source of the information, S. This strategy class can be viewed as an extension of the second strategy class. I need to clarify first what I mean by "changing." I do not mean the situation where the agent physically modifies the system S by its actions. I would regard this kind of change as part of the second class. By change of S I mean a change produced by modifications of the information system of the agent. Remember that the information source of an

information system depends on the system—it depends on which subsystem of the environment interacts with the information system and which macrostates of the subsystem are relevant for the goal-directed behavioral patterns. Therefore S, as a component of the information system, depends on the entire information system. The key is not to think of S as an independent object in the environment with independent properties that are reflected in the information system. Instead, S is a system partially constructed in the dynamical interaction of the organism with its environment. In many cases, no doubt, there may be independent ways of identifying S and some of its macrostate structure; after all, lions are real independent objects significant for the well-being of an antelope. Still, S can be changed by changing the dynamics of the information system, and the dynamics can be changed by parts of the information system that are in or near the agent. Let us consider an example. The information source can be a patch of the night sky. The physical system, the system of stars, is way back in the past light cone of the organism, but it interacts (in one direction) with the organism by fluctuations in the electromagnetic field mediating between the systems. With the naked eye few stars are visible in the patch. However, if a telescope is placed on the eye, the macrostates of the source change completely—now many more dots are visible. The source, qua source, has changed, even if it, qua physical object, has not changed.

Not every modification of the source provides a reduction of the information limitation; however, some do. This can happen by reducing selectively the number of states connected to M—*focusing*. The system uses its limited informational resources by extracting information from only a part of the original source (the new source is a subsystem of the original source), but it can extract more accurate or detailed information from the part. A cheetah can switch from exploring the savanna to focusing on a particular gazelle and using minute changes in the gazelle's behavior to skilfully modulate its chase. The distinctions and systematic dynamics of the whole savanna, including all the gazelles, trees, and bugs are too complex to effectively guide the cheetah's behavior, but by focusing on the gazelle and a few other aspects of its environment, the cheetah can accurately modulate its behavior at 90 km hr^{-1}. For the short minutes of the chase, the world of the cheetah is effectively smaller, but she is more accurately attuned to it. The key is not simply reducing the source, but reducing the source and selecting more important macrostates for the control of the goal-directed behavior—selecting a few important differences.

Other strategy classes of reduction of the information limitation exist as well. For example, proponents of extended cognition have insisted that some organisms may utilize external artifacts, such as paper and pen, to extend their cognitive systems, effectively making them more powerful (Clark, 2003, 2008). The same strategy can be described for an arbitrary agent, without assuming that it is a cognitive strategy (although it will turn out to be according to my analysis). Such strategies, I suspect, will be found only in fairly complex organisms, that is, humans, whose cognitive capacities are not in doubt. Thus, for the purposes of this article, I will not discuss these strategies further.

3.2 Defining Cognition

It's time for the punch line: I propose that the term *cognition* be used to describe the various mechanisms in an agent that implement the strategies that reduce the information limitation. I want to make this idea a bit more mathematically precise. For this purpose I will resort to a measure defined in Shannon's mathematical theory of communication (MTC) (Shannon, 1948; Weaver & Shannon, 1963). The measure that captures the idea of one system being informationally correlated to another is *conditional information entropy*. Conditional information entropy of a system X on a system Y is defined 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. For those unfamiliar with MTC I recommend Cover and Thomas (2006) for a more modern technical introduction. In the context of an information system, $X = S$ and $Y = M$. The states are, naturally, the corresponding macrostates of S and M. We assume that the probabilities are determined (somehow) by the global dynamics. Since both S and M are subsystems, probabilities may be defined even if the total dynamics are deterministic. This is because specification of macrostates of S and M underdetermines the state of the global system, thus probabilistic relations may depend on the existence of random latent variables in the system.

The conditional information entropy, $H(S|M)$, is usually interpreted as the amount of information deficit in M about S. Thus, if $H(S|M) = 1$, M and S are statistically independent—the agent has no information about the source. If $H(S|M) = 0$, then there is a perfect correlation between M and S—the agent has perfect information about S, and if S is the entire environment, the agent is an epistemic god. Transformation of the agent from M_1 to M_2 such that $H(S|M_1) > H(S|M_2)$ is a conditional information entropy lowering transformation. Similarly, a transformation of the source from S_1 to S_2 such that $H(S_1|M) > H(S_2|M)$ is also a conditional information entropy lowering transformation. Both cases have the effect of making the agent more

attuned to its source of information. The ILP for an agent is the observation that for most potential sources S, $H(S|M) \approx 1$. A strategy for reducing the IL problem is a transformation of the agent by some internal mechanism such that $H(S_1|M_1) > H(S_2|M_2)$, that is, a conditional information entropy lowering operation.

It is important to keep in mind that the measure H is completely general. It can apply to any two systems (or variables). The measure has the desired interpretation only in the context of an information system, where M and S, and their corresponding macrostates have a special significance for the agent's goal-directed behavior. The notion of "information" in MTC is much broader than the pragmatic/semantic notion of information in an agent. Under no circumstances should it be assumed that because I take advantage of a mathematical measure from MTC, I have therefore switched to "Shannon information." Thus, when the notion of conditional information entropy is used below in the characterization of cognition, one must keep in mind the demanding context of information systems.

With this discussion in mind, I propose the following characterization of cognition:

- **The cognitive system** is the set of mechanisms of an autonomous agent that (1) allow increase of the correlation and integration between the environment and the information system of the agent, that is, allow lowering of the conditional information entropy of selected important informational sources in the environment on the information medium in the agent, so that (2) the agent can improve the selection of actions and thereby produce more successful behavior in light of its goal(s).

3.3 Why Is This an Account of Cognition?

My strategy for arguing that the definition of cognition just presented is indeed the right one is to demonstrate that we can view the accepted instruments of cognition as strategies for reduction of informational limitations. Going back to the metaphor of fencing cognition, I must demonstrate (1) that we have fenced the right things, and (2) that the fence is sufficiently constrained to define a self-contained scientific discipline. The fencing metaphor has a weakness, however. As I insisted in Section 1, we should expect cognition to be a graded concept. The definition provides a graded concept because it relies on mechanisms that produce a marginal effect—reduction (or lowering). Thus, we should not regard the "fence" as a strict border. Rather, we should regard it as a vague outline of a region of system space, where an independent science of cognition (as opposed to only biology) becomes important for

modeling organism structure and behavior. Thus, we should think of the outline of the space of cognition in the same way we think of the outline of, for example, the tropics.

I will address (2) first, because there isn't all that much that could be said in support. The question of (2) is whether it is theoretically useful to make the notion of cognition more restricted. Arguing for (2), then, is arguing for a claim of the following form: for every further restriction of the class of cognitive systems, there exists a cognitive system that is omitted. Arguing for such universal claims is very difficult in an empirical domain. It is more productive instead to issue a challenge: Can anybody offer a more restricted definition of cognition, based on a further design problem, that can address the thinning problem in a non-arbitrary way? If it is possible, then we have made further theoretical progress. I am skeptical, however, because I suspect any further narrowing will restrict the *mode* of reduction of the informational limitation. Such a restriction, I suspect, would be either too chauvinistic or too ad hoc. Of course, it is possible that a more restricted class is defined in a completely different way. If the challenge is viable, I think, it is in this way. I admit, this is not a deductive argument. It is, however, a sufficient reason to regard the claim that the definition is the narrowest systematic definition of cognition that can be provided as the "null hypothesis."

How do we know that we have fenced the right systems? We can examine the basic cognitive capacities investigated by cognitive science and be convinced that they indeed serve the function of reducing the agent's informational limitation. In this case, we see that the common and overarching characteristic of all basic cognitive capacities is the reduction of the informational limitation. We must focus on basic capacities, because derived capacities may have all sorts of functions. The ability to produce poetry has no interesting connection to reducing informational limitations.

I suspect that the discussion of the various classes of strategy for the IL problem in Section 3.1 suggested how familiar cognitive capacities are captured, but I was careful not to introduce cognitive language to avoid circularity. Some of the most obvious capacities that emerge from the discussion are capacities such as learning and memory. *Learning* clearly is a mechanism for reduction of information limitation (conditional entropy lowering) because it is a capacity that allows temporal patterns of interaction to modify the response function of an organism so that limited control inputs can produce behavior that is sensitive to larger dynamical patterns in the environment. Learning an association means that information about only one of the associates can be received, yet the control mechanism can function as if information about both were

received. *Memory*, which can be regarded as a special learning mechanism where information is stored in a way that it can be recalled, that is, aspects of the control input can be regenerated and integrated back into the control system, is also clearly a mechanism of integration of information over time, where the information can be selectively focused. Thus memory also incorporates the focusing strategy to the IL problem.

Maps, *modeling*, and *representations*, which are required for some systems of memory, also allow focusing of information. They also allow internalization of aspects of the environment that may be decoupled from the source. Thus, the system can use maps even if no information channel exists between the source of the map and the organism. Representations can also be "analyzed" by the system to extract information that is not discernible as an input. Such mechanisms for extraction of information are sometimes described as *reasoning*, and when they lead to differential action as *decision making*.

Mechanisms of *attention*, and selective decomposition of inputs based on *feature detectors*, are primarily mechanisms of focusing of the source. Such mechanisms can be implemented purely after the percept, but often they depend on external action control loops, as when attention is guided by the movement of the body, eyes, ears, nose, or even flagella.

Some of the most important parts of the environment of an organism are other organisms. When organisms coordinate not only their behavior, but also their information systems directly, we can regard the organisms as *signaling* or *communicating*. When communication serves as a mechanism where one organism can obtain information about the other organism's action dispositions, it functions to reduce the information limitation of one organism about the state and behavior of the other, including the important case of information about the percepts of the other. Note that as defined, not all instances of "communication" between organisms should be regarded as supported by cognition. All one needs is information systems. (If one insists that communication is a mark of the cognitive, then not all cases of signaling between organisms should be regarded as communication.)

This is not an exhaustive list of cognitive capacities, but only a sample of the way one may be convinced that cognition ultimately is about reducing informational limitation in an agent. I have not offered a specific architecture of cognition, however. It should not be expected that we should learn anything deeper about the vast array of cognitive capacities by realizing that they are all strategies of conditional information entropy lowering—no more than we can learn anything deeper about tango, waltz, or twist by realizing that they are all forms of dancing. The benefit of grouping

and studying tango, waltz, or twist as species of dancing, as opposed to species of having-a-good-time, is that one can discover more systematic relations and contrasts between them. Similarly, by viewing cognitive capacities as species of information limitation lowering strategies, we have a more compact science of them.

How can this definition help with the thinning problem? Its main benefit is that it uses only concepts that are not derived from high cognition. Let us say we have a target system—a bacterium. To determine whether it is a cognitive system, first we must identify mechanisms that make it an information system. Second, we must identify dedicated mechanisms that can reasonably be described as having conditional information lowering function. If this is possible, there is a theoretical value in grouping the bacterium with other cognitive systems. If such an analysis is impossible or if it appears as an arbitrary and unnecessary theoretical imposition, then it is best not to regard the bacterium as a cognitive system, but as a proto-cognitive agent, or an active autonomous system. Ultimately the verdict depends on the careful analysis of the internal operation of the systems. Complexity of behavior, or similarity to behavior observed in cognitive systems may serve as initial evidence that the target system is cognitive, but ultimately cognition depends on what is under the hood.

4 Moving the Ladder

So far the discussion of cognition presupposed that the candidate systems are biological organisms (or at least autopoietic organisms, if further conditions are demanded of biological systems). What sense can we make of the idea of artificial cognition? One possible way of addressing artificial cognition is simply to deny it. One can say: cognition is a biological phenomenon! Thus, no artificial, especially computation based cognition can exist. This, however, is throwing the baby out with the bathwater. Such a view would regard any artificial system that exhibits sophisticated behavior, even more sophisticated than human behavior, as a noncognitive system. This would be as chauvinistic a view of cognition as are views that insist on symbolic processing or intentionality.

Another possible way of addressing the problem is to observe that autopoiesis is itself a functional property, and thus it can be instantiated in alternative environments (or in molecular but constructed environments). Alternative forms of cognition are possible, but only through *alife*. In other words, if one wants to create an artificial cognitive system, one must first create an artificial living system. This view is more plausible, but it is still too restrictive.

By examining my characterization of cognition, we can see precisely which parameters can be relaxed to

allow alternative, including artificial, forms of cognition. It is useful to look down from the condition of cognition, and unraveling the necessary notions, identify what is indispensable and what is not. According to the characterization, cognition demands two things: an agent and a conditional information entropy lowering mechanism. There are no real constraints on the nature of the mechanism, so we must focus on the agent. An agent is a system that implements an information system; thus we must consider the constraints on an information system. An information system has three constraints: (1) there has to be an embodied system in a tightly constrained interaction with an environment that hosts a source, (2) the system must have a well-defined goal, and (3) the system must possess some minimal organizational complexity to support the sub-systems M and P. Condition (3) is functionally definable, so it places no constraints on implementation. Conditions (1) and (2), however, do place constraints. The dynamic bidirectional interaction with an environment condition and sensitivity-to-a-goal condition are necessary for the determination of the informational states and their semantics. Condition (1) thus demands embodiment and situatedness of the system. There is nothing special that autopoiesis offers here. A metal can with wheels will do. Autopoiesis offers a solution for the presence of internally controlled goal-driven behavior—autonomy. Autopoiesis, however, offers only a sufficient condition of autonomy. It is a mode of material organization that gives us simultaneously the mode of persistent operation and the conditions determining the goal of the system. The goal comes for free.

There is no reason that the persistent operations of the system and the goals must come from the same place. In naturally emerging systems they probably must, this is why autopoiesis may be necessary for the origin of cognition, but in artificial systems the goal may come from a distinct source. In other words, the goal can be decoupled from metabolism, in which case metabolism may be decoupled from cognition and may be replaced by tin, copper, and silicon. The ladder of cognition may be moved from autopoiesis to another kind of embodied, goal-directed system.

5 Conclusion

Let us recap the achievements of this article. I started with a question: *What does cognition do in a general system?* By answering this question we simultaneously answer the question *what systems are cognitive?*, and outline the domain of cognitive science. The strategy was not to describe a set of cognitive capacities, identified by empirical observation, but to identify a general *problem* that systems of particular kind need solving.

By identifying such a problem we can identify what, in general, cognition does—what the function of cognition is for a system. The problem that I identified is *reducing information limitation* to engage in successful behavior. I identified the problem by considering a nested sequence of more general design problems starting from a very generic problem that can be asked about any cohesive system: *How is it that it persists?* From there I considered systems that are responsible for their own persistence—the autonomous systems. Then a second design problem emerged: *How can an active system perform better control outputs in order to improve its chances of persistence?* I noted that systems whose outputs are sensitive to the environment in which they act offer a more adaptive solution to this problem. A particularly important class of systems that involve the environment in the determination of their control outputs is the class of information systems. I called such systems agents. For agents we could formulate a further design problem focusing on what information can be used from the environment. Finally, by observing that all naturalistically possible agents operate under severe information limitation, a new design problem focusing on reducing the limitation appears, thus leading us to cognition.

The four problems were not selected arbitrarily. They took us from the question of *being* to the question of *acting*, to the question of *perceiving,* to the question of *thinking*. In a sense, although the discussion was based on a lower level system theoretic analysis, the concepts recovered are quite familiar. They came up, however, in a different order from when introduced by top-down approaches. In my order we see that *acting* is a way of *being*, *perceiving* is a way of *acting*, and *thinking* is a way of building complexity and order in the connection between *perceiving* and *acting*.

Funding

Notes

1. See for example: Alloway (1972), Gould (1986), Greenspan and Van Swinderen (2004), and Papaj and Prokopy (1989).
2. See Lyon (2007) for philosophical considerations, and Ben-Jacob, Shapira, and Tauber (2005) and Ben-Jacob (2009) for some physical/information theoretic arguments for bacterial cognition. See Shapiro (2007) for experimental arguments.
3. See Di Paolo (2005) for a systematic criticism. However, the matter is in no way settled. For a further defense of the thesis that autopoiesis implies cognition see Lyon (2004). Claims that autopoiesis is insufficient for cognition can

4. For a more detailed discussion of autonomy in broadly similar form see Christensen and Hooker (2000).

5. Bourgine and Stewart (2004) claim that this condition is sufficient to regard a system as cognitive.

6. It has also been criticized by Maturana and Varela, and some of their followers.

7. Some advanced systems, such as humans, can be said to have information without the system being able to do anything with it, except to retransmit it. This notion of "having information" is somewhat different from the one discussed here. It demands (cognitive) capacities not assumed in an information system. Discussion of this notion of information is beyond the scope of this article.

8. It may be possible to offer a stronger a priori conceptual argument as well. Consider two agents, Gad and Doity, that interact and that have complete relevant information. Because the actions of Doity are relevant to Gad, the state of the information system of Doity must be internalized in Gad. However, the same is true in reverse. Thus, Gad must internalize how Doity has internalized the state of Gad, ad infinitum. Either this is an incoherent situation, or it requires odd metaphysical assumptions, such as the possibility of infinite information.

References

Ackoff, R. L. (1958). Towards a behavioral theory of communication. *Management Science*, 4(3), 218–234.

Alloway, T. (1972). Learning and memory in insects. *Annual Review of Entomology*, 17, 43–56.

Barandiaran, X., & Moreno, A. (2006). On what makes certain dynamical systems cognitive: A minimally cognitive organization program. *Adaptive Behavior*, 14(2), 171–185.

Ben-Jacob, E. (2009). Learning from bacteria about natural information processing. *Annals of the New York Academy of Sciences*, 1178, 78–90.

Ben-Jacob, E., Shapira, Y., & Tauber, A. I. (2005). Seeking the foundations of cognition in bacteria: From schrödinger's negative entropy to latent information. *Physica A: Statistical Mechanics and its Applications*, 359, 495–524.

Bitbol, M., & Luisi, P. L. (2004). Autopoiesis with or without cognition: defining life at its edge. *Journal of the Royal Society Interface*, 1, 99–107.

Blair, D. F.. (1995). How bacteria sense and swim. *Annual Review of Microbiology*, 49, 489–522.

Bourgine, P., & Stewart, J. (2004). Autopoiesis and cognition. *Artificial Life*, 10, 327–345.

Brooks, A. (1991). Intelligence without representation. In J. Haugland (ed.), *Mind Design II*, 365–420, Cambridge, MA: The MIT Press (1997). (Originally in *Artificial Intelligence Journal, 47*).

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

Christensen, W., & Hooker, C. (2000). Autonomy and the emergence of intelligence: Organised interactive construction. *Communication and Cognition – Artificial Intelligence*, 17, 133–157.

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

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. New York: Oxford University Press.

Collier, J. (1999). Autonomy in anticipatory systems: significance for functionality, intentionality and meaning. In D. M. Dubois (Ed.), *Proceedings of CASYS'98, The 2nd International Conference on Computing Anticipatory Systems*. New York: Springer-Verlag.

Collier, J. D., & Hooker, C. A. (1999). Complexly organised dynamical systems. *Open Systems & Information Dynamics*, 6, 241–302.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). New York, NY: Wiley-Interscience.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy*, 72, 741–764.

Davidsson, P., Astor, E., & Ekdahl, B. (1994). A framework for autonomous agents based on the concept of anticipatory systems. *Cybernetics and Systems '94* (pp. 1427–1434). Singapore: World Scientific.

Dennett, D. (1991). *Consciousness explained*. Boston, MA: Little Brown and Co.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, 4, 429–452.

Gibson, J. (1979), *The Ecological Approach to Visual Perception*. Boston, MA: Houghton-Mifihin.

Godfrey-Smith, P. (1996). *Complexity and the function of mind in nature*. Cambridge, UK: Cambridge University Press.

Gould, J. L. (1986). The biology of learning. *Annual Review of Psychology*, 37, 163–192.

Greenspan, R. J., & Van Swinderen, B. (2004). Cognitive consonance: complex brain functions in the fruit fly and its relatives. *Trends in Neurosciences*, 27, 707–711.

Haken, H. (1993). *Synergetics: An introduction* (3rd ed.). Berlin: Springer.

Haken, H. (2000). *Information and self-organization: A macroscopic approach to complex systems* (2nd ed.). Berlin: Springer.

Hinrichsen, D., & Pritchard, A. J. (2005). *Mathematical systems theory I – Modelling, state space analysis, stability and robustness*. Berlin: Springer.

Katok, A., & Hasselblatt, B. (1996). *Introduction to the modern theory of dynamical systems*. New York: Cambridge University Press.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Levine, W. S. (1996). *The control handbook*. New York: CRC Press.

Lyon, P. (2004). Autopoiesis and knowing: Reflections on Maturana's biogenic explanation of cognition. *Cybernetics And Human Knowing*, 11(4), 21–46.

Lyon, P. (2006). The biogenic approach to cognition. *Cognitive Processing*, 7, 11–29.

Lyon, P. (2007). From quorum to cooperation: Lessons from bacterial sociality. *Studies in History and Philosophy of*

Science: Series C, Biological and Biomedical Sciences, *38*(4), 820–833.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Berlin: Springer.

Nauta, D. (1970). *The meaning of information*. The Hague: Mouton.

Newell, A., & Simon, H. (1981). Computer science as empirical enquiry. In J. Hougeland (Ed.), *Mind design II* (pp. 81–110). Cambridge MA: MIT Press.

Nicolis, G., & Nicolis, C. (2007). *Foundations of complex systems: nonlinear dynamics statistical physics and prediction*. Singapore: World Scientific.

Papaj, D. R., & Prokopy, R. J. (1989). Ecological and evolutionary aspects of learning in phytophagous insects. *Annual Review of Entomology*, *34*, 315–350.

Pillay, A. (2008). *An introduction to stability theory*. New York: Dover Publications.

Prigogine, I. (1961). *Thermodynamics of irreversible processes* (2nd ed.). New York: Interscience.

Prigogine, I., & Nicolis, G. (1977). *Self-organization in non-equilibrium systems*. New York: Wiley.

Rosen, R. (1985). *Anticipatory systems: Philosophical, mathematical, and methodological foundations* (Vol. 436). New York: Pergamon Press.

Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology*, *22*, 1–19.

Shannon, C. (1948). The mathematical theory of communication. *Bell Systems Technical Journal*, 27, 379–423, 623–656.

Shapiro, J. (2007). Bacteria are small but not stupid: cognition, natural genetic engineering and socio-bacteriology. *Studies in History & Philosophy of Biology & Biomedical Sciences*, *38*, 807–819.

Shettleworth, S. J. (1998). *Cognition, evolution, and behavior*. Oxford: Oxford University Press.

Thelen, E., & Smith, L. (1994). *Dynamics system approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Belknap Press.

Vakarelov, O. (2010). Pre-cognitive semantic information. *Knowledge, Technology & Policy*, *23*(2), 193–226.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*, 615–628.

Weaver, W., & Shannon, C. E. (1963). *The mathematical theory of communication*. Champaign, IL: University of Illinois Press.

About the Author

Orlin Vakarelov is a doctoral candidate at the Department of Philosophy and the Cognitive Science Program at the University of Arizona. He has an M.S. degree in logic and computation from Carnegie Mellon University. His main interests are in cognitive science, the philosophy of information, and the philosophy of science, with special emphasis on the emergence of informational phenomena in embodied and enactive models of cognition.