The Cognitive Agent

Orlin Vakarelov
Department of Philosophy
University of Arizona
Tucson, Arizona
okv@u.arizona.edu

January 15, 2009

Abstract

In this project I investigate what minimal conditions can allow a system to be regarded as a cognitive agent, i.e. what the subject of cognitive science is. An agent is a highly organized complex system which has *intrinsic goals* (in its minimal sense, persistence within viability boundary) and whose dynamical organization supports an *informational description*. Every agent, it follows, is *informationally deprived*. This deprivation raises a design problem: How can the organization begin to overcome this informational limitation? I claim that cognition is the general strategy for resolving this design problem. More precisely, I define cognition as follows:

Cognition is the set of the mechanisms/organizational constraints of an autonomous agent that allow lowering of the conditional information entropy of selected important informational sources in the environment on the control structure of the agent*, so that the agent can improve the selection of actions to produce successful behavior in light of its information gathering and carrying limitations.*

I suggest that standard cognitive capacities — learning, memory, feature detection, representation, reasoning, etc. — can be viewed as special cases of this general strategy.

The concept of *agent* has many uses in philosophical discussions. Epidemiologist assume different characteristic of agents from the ones moral philosophers assume; which are yet different from the ones metaphysicians assume in (e.g.) debates over free will. Such differences of desired characteristics are relatively innocent because the target of investigation is usually human agency, which possesses all the needed characteristics for each of the problems, and then more. In this context, the differences are merely difference of emphasis of characteristics. The luxury of human agenthood is not available, however, when one is interested in investigating general agenthood characteristics of simpler natural systems, the kind of systems that, in an evolutionary process, may have lead to the human kind of agency. In this paper I explore the problem of what general minimal characteristics warrant describing a system as being a *cognitive agent.*

Few words on what I mean my cognitive, and how I plan to approach the problem. It is useful to distinguish between capital 'C' *Cognition* and small 'c' *cognition*. By Cognition I mean cognition as the high level capacity of discursive thought, etc. — this is, the cognition that Descartes was talking about; the cognition that most philosophers talk about when using the term. By cognition I mean the notion that modern cognitive science studies, including when cognitive scientist talk about insect or even bacterial cognition. It is, roughly, the intricate capacities of organisms to navigate and interact with their environment. I am interested in characterizing the notion of small 'c' cognitive agent — the kind of agent that cognitive science studies. Not that Cognitive agents are not interesting, they are very interesting, they are the ultimate interest. Understating cognition, however, is a prerequisite for providing a theory (philosophical or scientific) of Cognition.

One approach to characterizing cognitive agenthood is to look are the various notions of Cognitive agenthood (of which the various philosophical notions of agenthood are arguably special cases) and extract common characteristics that must be associated with cognitive agenthood. Such a top-down approach may produce some interesting suggestions, however it suffers from one considerable difficulty. It is very difficult to isolate what characteristics of human like Cognitive agenthood must be preserved in simpler systems. Debates about the nature of intelligence have shown that such an enterprise is near hopeless. I take a bottom-up approach where my goal is to consider simple systems that possess interesting characteristics on the way towards an intuitive notion of cognition and to isolate important organizational/design landmarks. I characterize the notion of agent and cognitive agent as solutions to distinct design problems that nature provides. The characteristics appear somewhat stipulative, but this comes with the approach. The ultimate arbiter of the success of such a bottom-up, model building approach is the ability of the theory to capture the right high-level concepts once one gets to them — the proof of the pudding is the eating, as they say. It is beyond the scope of this paper to reach the ultimate goal, but there is enough room to lay out the main minimal distinctions and motivate their formulation. Let us begine!

The key abstract idea that needs to be included in the study of cognition is *autonomous agent*. The precise definition of autonomous agent is not agreed upon, and is still a subject of philosophical debates; however there exists a common thread in all accounts: an autonomous system is a self-governing system. Autonomous systems are contrasted with heteronomous systems, which are other-governed (if governed at all). Thus, an autonomous system[1] is a system for which at least two notions make sense: the system can be described as being *governed* (or *controlled*); and there exist a sense in which there is a locus of governance that belongs to the system. Governing further implies that there is a *minimal goal* that the governing is trying to achieve — in the simplest case the preservation of some quality, such as system integrity. The *minimal* goal is always an immediate goal — something whose success can be evaluated at any moment. The goal cannot be something that must happen in the future. Such a goal is not implied from a system being an autonomous agent.

Naturally, autonomy is a gradual notion in at least two dimensions: a system may be better or worse in terms its ability to satisfy the goal in various conditions; and, the system may be partially autonomous, partially heteronomous. The gradation of autonomy is especially important for the account of cognitive agency because the different degrees of autonomy may be characterized according to what "design" specifications the systems satisfy, which may allow us to characterize cognition as the solution of a type of design specification. In fact, it is possible to view autonomous systems as defining a sequence of more and more specific design problem, each emerging from what naturally is a good solution to the previous. The first such problem is the *preservation* of the organizational constraints that define the system as autonomous. Here it is assumed that the minimal goal of the governing process is the preservation of the organizational constraints. If the goal is different, than the most general design problem is the achievement of that different goal.

So far it is not obvious why there should be any natural systems that can be regarded as autonomous. There are, however, natural systems that exhibit the minimal conditions for autonomy. There are dynamical systems whose organization is such that the defining/identity conditions of the systems are de-

[1]Throughout it is, of course, assumed that the system is not in thermodynamic equilibrium, and it is in constant interaction with a dynamic environment. This is important because only such systems need governing.

termined by the very systems, i.e. by the very defining conditions. Such a circular inter-determination of a system is possible if the system, in the process of persistence, generates its own structure, which contains the conditions for the self-generation. Such a system is said to satisfy *process closure*. Systems that can generated themselves out of external resources from the environment and externally available energy, and that can release any waste products back to the environment are called *autopoietic* systems.[2] (Maturana & Varela, 1980) The theory of autopoietic systems was developed as an abstract structural theory of the organization of the living cell and was extended as a model to living system in general.

Autopoietic systems can be regarded as possessing natural teleology because they are self-determining[3], and the internal operations of the system that maintain it can be regarded as having an immanent goal within the organization. Namely, the processes have the function to maintain the organizational constrains of the systems, and thus the persistence of the very processes. The theory of autopoietic systems can be viewed as providing a model of autonomous systems, possibly a model of the easiest to emerge autonomous systems. This is because the defining processes of the system (1) maintains a dynamical separation between the system as a being, and the environment, (2) bestow the system with the natural goal of dynamical persistence, and (3) defines a locus of governance withing the system in the sense that the processes, which are an essential part of the system, govern the structural characteristics of the system and its interaction with the environment. In the simplest kinds of autopoiesis the notion of governance may be rather trivial, but it is real never the less — it is a feature of the dynamics of the system, not merely an imposed mode of description.

The design problem of preserving the organizational constraints of the system places only a few constraints on the nature of autonomy. However, there are strategies that improve the solution of this problem. Let me introduce some technical vocabulary. The history of dynamic interactions of a system with its

[2] A system is *autopoietic if:*

1. it has a semi-permeable boundary,
2. the boundary is produced from within the system, and
3. it encompasses reactions that regenerate the components of the system. (Varela 2000, via Luisi 2003, and Bourgine & Stewart 2004)

[3] This idea of natural teleology in this context goes back to Kant's *Critic of Judgment* where he claims that a system possesses natural teleology if it is simultaneously its own cause and effect.(Weber & Varela, 2002; Thompson, 2007)

environment I call the *behavior* of the system.[4] This is a very weak sense of behavior — according to this definition even a rock can have a behavior. In this terminology the original design problem is the problem of dynamically maintaining the existence of behavior. Because autonomous systems are in constant interaction with the environment, it is useful to isolate the interactions with the environment that is produced or modulated by the system. I call such interactions *proto-actions*.[5] Some proto-actions can be well defined isolatable causal events, such as a movement, release of a chemical, etc. Probably most interactions that autonomous systems have with the environment are best described as close-coupled processes, in which case a proto-action may be a subtle modulation of a close-coupled process. It is likely impossible to specify a precise boundary for distinguishing proto-actions from generic dynamic interactions of the system with its environment, but some interactions are clearly identifiable as proto-actions in virtue of the dynamics of the system. Proto-actions are interesting because they provide a mechanism for the system to have a targeted effect on the environment, including on the relation between the system and the environment. Now, the behavior of the system depends crucially on the state of the environment — the environment must be within the limits of viability of the system. Therefore, proto-actions can help (or hurt) the future state of the system. This observation induces a further design problem for an autonomous system: How can an autonomous system (proto) act in a way to improve its chances of survival?

Some proto-actions, like releasing a chemical that may attract food, or moving randomly, may on average improve one's changes of survival regardless of the state of the environment. The success of other proto-actions is contingent on the state of the environment. Swimming in a straight line is helpful only if the direction of motion is along an increasing gradient of food. (Of course, it may be helpful in other instances as well.) A proto-action may be more successful if its performance can be made contingent on the state of the environment in a way that the proto-action is effective in that state. In other words, an autonomous system may be more successful in changing its environment beneficially if the environment itself plays a control role for its proto-actions. It is, thus, useful

[4] The word "behavior" has been used in many different ways in philosophical discussions. Here it is used in a purely technical sense. It is not to be confused with other technical or informal, intuitive uses of the term. This being said, my use is consistent with the use of the term "behavior" in system theory.

[5] Because the autonomous system is a part of the environment, in some cases it may be useful to include the system as being effected by the proto-action.

to also isolate interaction between the autonomous system and the environment whose effect is internal to the systems, and than to consider those such interactions that play a control role for proto-actions. I will call such interactions, *proto-percepts*.

How do proto-percepts, as modes of interaction, differ from mere dynamical interactions of the system with the environment? Both a photon reaching a photo cell, and a bullet entering the skull are interactions whose effects are internal. However, the photon is such that it affects the system by changing its dispositional characteristics in a subtle way, so that the system can engage in a particular proto-action as a result. The bullet, which also changes the dispositional characteristics, does so in an indiscriminate way, causing arbitrary change in the system's behavior. The difference is that between control and mere effect. The difference does not lie in the interaction itself but in the internal organization of the system and its dispositional characteristic with respect to the interaction — the disposition to change the system's dispositions; a kind of second order disposition. The photon can have a control function because the internal organization of the system is such that it is sensitive in very particular ways to the reception of the photon, and differences made by the photon are propagated to immediate (or future) differences in the proto-actions of the system. If the system is organized in such a way that an incoming bullet can produce a subtly sensitive change in the system, the bullet can be regarded as having a control role. This being said, like in the case of proto-actions, the difference is not always sharp, but it is sufficiently sharp in many cases. The vagueness does not come from our inability to sharpen the difference theoretically, but from the gradual nature of the kinds of internal organizations that can exist in a system.

Whenever the system's interactions with its environment can be described in terms of internal organization that mediates and controls the interactions, whenever the internal organization is sensitive to particular aspects of the environment and uses those aspects to selectively modulate the dispositional characteristics of the system, especially the disposition to selectively affect the environment, it becomes useful to describe the system in informational terms. The notion of information is a complicated and often misunderstood notion (or family of notions). There are no completely satisfactory and all encompassing treatments of it.[6] And, this is not the place to provide such a treatment. I can

[6]There are, however, quite a few sophisticate philosophical treatments of different aspects of information and related notions. See (Floridi, 2003) especially the bibliography and the

provide few pointers, however, to what features of a system may warrant an informational description, and how it differs from other kinds of description.

Some proponents of dynamical systems description of cognition (and intelligence) view informational (and information processing) descriptions as being fundamentally different from the dynamical. It is suggested that a dynamical description excludes an informational description. This contrast depends on an overly restricted conception of information, and on a failure to appreciate the difficulty of describing the dynamics of highly complex and organized dynamical systems. An informational process is not categorically different from a dynamical process. Every informational system is a dynamical system. The question is, is every detail of the dynamics, every microscopic feature of the system, relevant for the understanding of general behavior of the system. If the dynamics of the system is highly constrained by complex internal organization, the organization can filter out microscopic details, making the system selectively sensitive to only certain fluctuations. A purely dynamical description of such a system may not be explicitly sensitive to the redundancies that are imposed by the organization. The question, then, is not whether a system can be described with dynamical or informational language — it is always possible to do both — but weather the informational *vs.* the dynamical description is desirable for capturing the structure of the system. While this may appear as a pragmatic problem, and to certain extent it is, what description is more appropriate for a given system ultimately depends on the system. Ultimately, it depends on how organized the system is.

A systems organization is a precondition for possibility of the system to engage in informational interaction with other systems; however it alone in not sufficient to count as information. The kinds of dynamical processes where informational descriptions are interesting consist of (at least) two interacting dynamical systems, (one of which can be the environment) where the organizational structure of the systems becomes somewhat correlated, and moreover, where the correlation allows the behavior of one of the system to be partially controlled by the other. In this context one can describe one of the systems as containing information about the other, and some of the interactions between the systems as conveying the information. The process of informational transmission and utilization emerging between the systems is a feature of the coupled dynamics of the systems; it is not contrasted to it. Moreover, the fact that there

other entries in the volume.

may exist a dynamical correlation between two systems that can be described more effectively in informational terms does not preclude there being other, closely coupled interactions between the systems best described within dynamical systems theory (using coupled differential equations). The informational process may actually control the close-coupled process — think of the close-coupled process of bouncing a pig-pong ball, that can be molded and controlled by somebody issuing commands of moving to the left or to the right.

From this extremely limited discussion of information we should concentrate on two things: (1) the central question about informational systems is not whether there exist a quality (or quantity) of information that the system manipulates, but whether there exist a level of abstraction[7] on top of the dynamical description of the system that can be captured with informational concepts, and that provides a better, more economical description of the system. (2) Whether such a level of abstraction exists depends on the system and its interaction with other systems, including the environment — particularly, it depends on the organization and correlations among organized structures in the systems, and the control relations in which they can enter in virtue of the organization.

Going back to the problem of (proto) acting better, the realization that the actions must somehow be contingent on the environment can be rephrased by saying that information from the environment must control the proto-actions (hopefully in a way that is beneficial for the overarching goal of persistence). In other words, proto-actions become a part of an informational control system. Some more terminology is in order: I call a proto-action that participates in such information control system *an action.* The corresponding proto-percepts I call *percepts*, a foreword looking stretch of terminology. And, finally an autonomous system for which an informational description is beneficial I call an *agent.* Thus an *agent* is an autonomous system whose organization is sufficiently complicated to admit informational interaction with the environment.

Immediately a further design problem emerges: How can we have a system, an autonomous agent, that utilizes more and better (more relevant) information about the environment? This question is not only about the information channels between the environment and agent, but also about the ability of the agent to utilize the information — good record keeping alone will not do.

[7]For a discussion of the connection of the notion of information to levels of abstraction, with a precise theory of he later, but without explicit connection to dynamics, see (Floridi & Sanders, 2004)

What are possible design strategies for such a problem? Here is architecture of an informational über-agent: The agent contains a gigantic table listing all possible states of the environment and the corresponding commands to execute. Together with the table the agent has the abilities to identify the particular state of the environment, and thereby it can check the table and decide on an action. Let us bracket the problem of whether such a system, as a subsystem of the world, is metaphysically possible. Clearly such a system is outrageously implausible. I hope the intuition points to a generally valid observation about autonomous agents: all agents are severely informationally deprived, both because their interactions with the environment is limited with respect to the possible relevant states of the environment, i.e. the information channels that they can use are extremely limited, and the systems themselves are substantially simpler[8] than the whole environment. I take this observation to be a requirement for any naturalistic account of agents. The proper design strategy for the need for more and better information is not to throw in the information that is needed, but to find strategies for overcoming and reducing the informational limitation that all agents face. Thus, we have another further design problem: How can an agent cope and overcome its intrinsic information limitations to be able act more successfully. We need to go back to the notion of information again.

In its simplest form, the classical mathematical theory of communication (Shannon, 1948) deals with how information (messages) can be transmitted between a source and a receiver through an information channel.[9] The messages are sent through a channel that has a certain *capacity*. Capacity is a measure of the maximum complexity of the message that the channel can transmit in a unit of time. It is, in a sense, the generalization of the notion of bandwidth that we see in, e.g. Internet connections. A modem has relatively low bandwidth in comparison to a DSL channel, which has a relatively low bandwidth to a fiber-optic cable. A higher capacity channel can transmit more information in a given unit of time than a lower capacity. The goal of information transfer, however is not merely getting messages though the channel, but getting information from the source to the receiver. If the receiver has no relation to the source, all the information that the receiver can obtain is limited by what can pass through

[8] Although, they may have high "concentration of complexity" in relation to the rest of the world. It is possible that the brain is the most complex, organized system in the universe, still the universe is vastly more complex then the brain, and not merely because it contains it.

[9] See for example(Cover & Thomas, 2006) for a more accessible modern treatment.

the channel. However, if the state of the receiver is already correlated to some extent with the source, i.e. the receiver already has some information about the source, then one only needs to transmit the "novel" information. In other words, the informational relation between the source and the receiver may be much richer than what can pass through the channel. For example, the single bit message "I do" can convey the information that the person next to you will be your spouse, a state with fairly rich content, provided you know that you are in the context of a marriage ceremony.

Most applications of the theory of information have to do with the structure of the communication channel and various coding procedures for optimizing the transmission of information. This is appropriate because one usually takes an external perspective towards the system, whereby all elements of the communication system can be manipulated. When dealing with the issue of agent architecture, and the ability of the agent to enter in informational relations with its environment, the manipulable constraints of the problem are different. When designing a communication system that faces the problem of informational limitation, an engineer has the option to use a higher bandwidth channel, or eliminate the noise in the channel; or, she may be able to reduce the equivocation of the source, making it, in a sense, to "speak more clearly". An agent designer has only the option of modifying the receiver of the information.[10] But, how may the system at the end of the information flow be modified in order to minimize the informational limitation? According to the model of communication channel assumed by communication theory, the only solution is to make the receiver need less bandwidth by increasing the correlation between the structure of the receiver and the structure of the informational source. In the mathematical theory of communication this relation is expressed with the notion of *conditional information entropy* of the source on the receiver. The conditional information entropy can be interpreted as measuring the uncertainty of the source from the point of view of the receiver. If the conditional entropy is maximal, then the receiver "does not know" anything about the source and

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

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

the elimination of the uncertainty requires the transmission of a lot of information. If the conditional entropy is low, the state of the source is only a little bit uncertain, and only a little bit of information is necessary to be transmitted in order to eliminate the uncertainty. Thus, the strategy for overcoming the informational limitation is for the receiver to somehow lower the conditional information entropy of the source with respect to itself.

The communication model of the mathematical theory of communication makes curtain assumptions about the source of the information that a model taking as primitive the interaction of an agent with the environment does not. Particularly, the model assumes that the source is well defined, and that its informationally relevant states are defined. For example, it may be assumed that the source produces messages in a fixed alphabet of symbols with a fixed frequency distribution of the symbols. The formal analysis begins after this assumption is made; prior to this no model exists. In the situation of the agent interacting with an environment, the source of the information and the states of the source that are informationally relevant are not determined upfront.[11] This indeterminacy of the source opens another possible approach to overcoming the informational limitation. The agent may be able to "select" the appropriately limited source, with appropriately limited informational states so that it does not need to channel and operate with too much information. The selection mechanism, of course, in the most general case depends on the dynamical interactions of the agent with the environment — it depends on what subsystems of the environment can be isolated in terms of the interaction with the agent, and what aspects of those systems are relevant for the agent's goals. It also depends on what structures in the agent can be correlated with the systems of the environment so that they facilitate informational channels — appropriately organized sensory sub-systems can play an important role in allowing targeted correlation with the systems in the environment.

The structure of the agent is central for its ability to correlate its dispositional characteristics to act with the important systems or general characteristics of the environment. It is this ability to *correlate selectively* with target sources of information that allows the agent to begin overcoming the informational limitations. It is the internal organization that not merely controls action,

[11] Of course, there are some generic sources of information. For example, the entire environment may be regarded as the source, and all of its possible dynamical states as the informational states. It is exactly this source, however, that is outrageously complex for the agent to communicate with.

but controls action in a "smarter" way that deserves independent investigation. Indeed I claim that, be it in disguise, cognitive science studies (and it should study) those structures that can control the agent better than a generic response system does. This leads me to suggest the following characterization of the subject of cognitive science, and thus of cognitive agent:

> ***Cognition*** *is the set of organizational constraints (mechanisms) of an autonomous agent that allow increase of the correlation and integration between the environment and the control structure of the agent, so that the agent can improve the selection of actions to produce successful behavior in light of its informational limitations.*

The characterization can be restated in the language of information theory as follows:

> ***Cognition*** *is the set of organizational constraints (mechanisms) of an autonomous agent that allow* lowering of the conditional information entropy of selected important informational sources in the environment on the control structure of the agent*, so that the agent can improve the selection of actions to produce successful behavior in light of its informational limitations.*

Showing that this characterization of cognition is the right one for cognitive science, and that it leads to insights into high-level cognition is a demanding and complex task that it impossible to accomplish here. The strategy, however is to demonstrate that important capacities associated with cognitive systems can be viewed as special strategies for conditional entropy lowering — capacities such as: learning, memory, feature detection, representation, and reasoning, etc. Such capacities allow: accumulation and integration of information over time; targeting specific useful feature of the environment; building internal structures that encode information about the environment and its dynamics, using them to anticipate the future state of the environment based on limited information from perception; going beyond the immediate information thought informational transformations (with reasoning capacities), etc. It is not difficult to see that all these skills are modes of implementing the general strategy of making the internal organization of the agent better attuned to the structure of the environment and the available interaction of the agent with it, to minimize the burden of the need for too much immediate information coming through perception.

References

Bourgine, Paul, & Stewart, John. 2004. Autopoiesis and Cognition. *ArtiÞcial Life*, **10**, 327Ð345.

Cover, T. M., & Thomas, J. A. 2006. *Elements of information theory*. 2 edn. Wiley-Interscience.

Floridi, L. 2003. Information. *In:* Floridi, L. (ed), *The Blackwell Guide to the Philosophy of Computing and Information*. Oxford - New York: Blackwell.

Floridi, L., & Sanders, J. 2004. *Levellism and the Method of Abstraction*. IEG Research Report IEG-RR-4. Oxford University.

Luisi, P.L. 2003. Autopoiesis: a review and a reappraisal. *Naturwissenschaften*, **90**(2), 49–59.

Maturana, Humberto R., & Varela, Francisco J. 1980. *Autopoiesis and Cognition: The Realization of the Living*. Springer.

Shannon, C. 1948. The Mathematical Theory of Communication. *Bell Systems Technical Journal.*

Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. 1 edn. Belknap Press.

Varela, Francisco. 2000. *El fenomeno de la vida*. Santiago de Chile: Dolmen Ediciones.

Weber, Andreas, & Varela, Francisco. 2002. Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, **1**(2), 97–125.