How "Authentic Intentionality" Can Be Enabled.
A Neurocomputational Hypothesis

Abstract According to John Haugeland, the capacity for "authentic intentionality" depends on a commitment to constitutive standards of objectivity. One of the consequences of Haugeland's view is that a neurocomputational explanation cannot be adequate to understand "authentic intentionality". This paper gives grounds to resist such a consequence. It provides the beginning of an account of authentic intentionality in terms of neurocomputational enabling conditions. It argues that the standards, which constitute the domain of objects that can be represented, reflect the statistical structure of the environments where brain sensory systems evolved and develop. The objection that I equivocate on what Haugeland means by "commitment to standards" is rebutted by introducing the notion of "florid, self-conscious representing". Were the hypothesis presented plausible, computational neuroscience would offer a promising framework for a better understanding of the conditions for meaningful representation.

Keywords Authentic Intentionality, John Haugeland, Neural Representation, Constitutive Standards of Objectivity, Visual Perception

Introduction

Beliefs, desires, intentions, perceptions, perhaps emotions, are all examples of mental representations. Representations are interesting objects because they bear a semantic relation to the world. They have certain *content*, or a certain *meaning*.[1] They represent something, are about things, properties, or states of affairs extrinsic to them. The topic of what follows is how mental representations "mean", namely: how they say something meaningful about their objects to their consumers. Call this the "problem of intentionality".

John Haugeland, one of the contemporary leading figures in the philosophy of psychology, has articulated a number of arguments to treat the problem (Haugeland 1998a, 2002a). One of my two aims here is to give a critical evaluation of some of the consequences of Haugeland's way of dealing with the problem. Although Haugeland is sympathetic with the aim of understanding the mind scientifically, he takes inspiration from Kant and Heidegger and argues for a "new existentialist" philosophy of mind which is concerned with people. He argues that people's cognitive capacities differ fundamentally from those of other animals and of (current) machines. With respect to the problem of intentionality, what Haugeland is looking for is how mental states have "authentic" content, or "authentic intentionality" (Haugeland 2002a). That is, how mental representations can feature in a process of understanding. For Haugeland, understanding, making sense of things, is the mark of the mental.

[1] Often 'content' is used solely to deal with linguistic-like mental representations. However, one common assumptions of computational neuroscience - the field I am going to explore, is that the nervous systems represent without the need of positing linguistic-like items. Hence, I shall adopt the term more broadly; 'content' and 'meaning' will be used interchangeably.

My second aim is to provide grounds for why computational neuroscience may help us make some steps towards a better understanding of how "authentic intentionality" is *enabled*.[2] I shall borrow ideas and concepts developed in computational cognitive neuroscience to show what computational properties of the neural realizers of mental representations may enable authentic intentionality. The hope is to suggest a new, fruitful way of looking at an old problem in order to make small steps towards a solution, or at least towards the specification of some constraints for an adequate solution.

The paper is structured thus. Firstly, I set the stage by making some conceptual distinctions useful to individuate my tack on the "problem of intentionality". I shall confine my examples to one species of mental representations: visual perceptions. The section concludes by clarifying how computational neuroscience conceives of representation. I then formulate and explain the main steps of Haugeland's argument for what the conditions for "authentic intentionality" are. One of the consequences of his argument is that an account of meaning couched in terms of computational neuroscience is insufficient. On the contrary, I argue that such an account might suffice to pin down conditions sufficient to make authentic intentionality possible. The section revolves around the notions of value, emulation, encoding-decoding, adaptation, and perceptual learning. It shows how these notions may be interpreted in order to identify the beginning of *one* possible neurocomputational account of authentic intentionality. Finally, I consider an important objection to my account, and argue that the objection is built on certain confusions, and, therefore, can be rebutted. The paper concludes with a summary of the points made and defended.

Setting the Stage

Computational neuroscience is the use of mathematical modeling and computer simulations to understand the brain. By itself, this does not entail a commitment to *computationalism*, namely the claim that brains are computing devices. In the explanatory framework of computational neuroscience, however, it is often assumed that brains are in fact *kinds* of computers (Churchland and Sejnowski 1992; Dayan 1994; Eliasmith 2003; Montague 2007; O'Really and Munakata 2000). One way to motivate this claim is to argue that brains are kinds of computers because "the defining function of nervous systems is *representational*" (Churchland and Grush, 1999, p.155; see also Shagrir 2006 for an argument along these lines). The argument is as follows. Computation presupposes representation - ("No computation without representation" is the Fodorian slogan (Fodor 1981)). The brain appears to be essentially a representational

[2] 'Enable' is here used as a "making-possible" relationship. 'Enabling conditions' are sufficient to make possible some personal-level fact such as authentic intentionality (see McDowell 1994; Hurley 2008).

device in that brain states represent states of something else. It follows that a prerequisite of both computation and cognition (given that cognition depends on brain activity) is representation. Assuming that representations are entities that possess meaning, both computation and cognition presupposes meaning. From this, it is inferred that the brain is a kind of computer because it carries out tasks by transforming meaningful representations, *and* computation is transformation of meaningful representations. Hence, brains are computational devices.

All the premises of the argument are not obvious. And the argument itself is not valid; it is, in fact, an instance of the fallacy of affirming the consequent (or inference to the best explanation). Moreover, it is unclear which notion of representation-transformation is in place. What matters, for present purposes, are not the independent reasons to subscribe to this view. Rather, by taking a neurocomputational perspective, I would like to formulate and defend a claim about the conditions under which a mental representation is meaningful. To address this problem, let's now make some bits of terminology straight.

Representations are, in a general sense, entities that bear information about something else. To facilitate the discussion that follows, this characterization can be further articulated thus. Compare the way metal detectors represent with the way a football player perceives an offside. Whereas metal detectors simply detect metallic object, the football player's perceptual mental state of an offside represents by uncovering the "semantic structure" of its object to the football player. In both cases there is a correlation between a representation and something that is represented. In both situations there is something that carries information about something else. But the way the information is carried and consumed seems to make a difference. Metal detectors *indicate* that their target is there, metal detectors consume their representations witlessly in that they don't know how to use the content of their representations. The football player's perception of an offside brings about information that makes sense to its consumer; the representation can feature in a process of understanding; and the consumer can act upon the content of her representation. Call the type of representations that fall in the class exemplified by the first case: *detection*-type representations. Call the type of representations that fall in the class exemplified by the second case: *authentic*-type representations. The latter type of representations presupposes understanding: they make sense to their consumers. The football player understands what an offside is, and can tell when a player is to be considered in offside during a game of football. The metal detector, instead, does not understand what metal is, and it cannot act upon the content of its representations.

It may be argued that the distinction is spurious since it relies on a difference in complexity of the mechanisms that enable the metal detection, in one case, and the offside perception, in the other. “Add a color detector, a shape detector, a motion detector, a language module, … to the basic metal detector – it may be argued - and the metal detector will be able to perceive, or detect, as it were, an offside, and tell when a player is to be considered in offside”. Hence, the difference between detection and authentic representation would be one of computational power and complexity. It would not be a difference in kind. A natural reaction to this line of argument is to raise some Chinese-Room kind of worry (Searle 1980). Whatever the cogency of Searle’s argument, this type of worry would at least flag a distinction between representations that enter a process that merely leads to appropriate response given an input, and representations that enter a process of understanding. Only the latter, not the former, would bear information *for* their consumer (Haugeland 1992; Clark 1997b, sec. 2). Accordingly, the representation of an offside of the modified metal detector wouldn’t really make sense to its consumer since its representations do not enter any process of understanding since metal detectors understand nothing. Instead, when the football player perceives an offside her representation makes sense to her. This distinction will not be defended and articulated further. It simply serves me as background to introduce Haugeland’s position on the conditions for authentic, meaningful representation.

Before unpacking Haugeland’s proposal, let’s ask how representations can be conceived of in computational cognitive neuroscience - which is the playground of my argument. The information theoretic account of codes (Shannon 1948) offers *one* way to understand what neural representations are (Eliasmith 2003, sec. 3.1). On this view, neural representations are identified by encoding *and* decoding (Dayan and Abbot 2001 Ch. 1-3). Neurons code physical properties with their activity in response to stimuli: Neuronal activity systematically co-varies with and selectively responds to properties such as light intensity, colour, orientation, and so on. Neural encoding provides a mapping from stimulus to response. Given a stimulus, neural encoding determines how neural activation in a certain brain area transduces the stimulus as function of some non-neural parameter – for example, determining how a neuron in the primary visual cortex responds as function of the orientation of a bar of light. The activation *given* stimulus relation is typically captured by the neuron tuning curves (e.g. Gaussian-like tuning to the angle of the bar in the receptive field). The tuning curve to a feature of a stimulus such the orientation of a bar is therefore the curve describing the average response of a neuron as a function of the values of the features. Neural decoding provides a mapping from response to stimulus. It extracts an estimate of what physical property is coded in a particular neural activation, thereby determining the relevance of the encoding for the system. Given a certain

neural activation, the study of neural decoding amounts to guess how likely it is that a certain type of stimulus is in the environment – for example, determining the orientation of a bar of light given a certain activation in the primary visual cortex.

It can be concluded that, *prima facie*, neural representations so conceived are detection-type representations. The meaning of a neural activation, like that of a metal detector activation, is grounded in the environmental conditions which that particular activation has the function to carry information about. "A neuron's representation – O'Reilly and Munakata (2000, p. 25) claim - is simply that which it detects". Thus, neural spike trains may bear information about something but they do not seem to represent "authentically". That is, even if neurons code information about certain objects, they do not really represent, or stand in for objects. This poses an immediate concern: Since neurons don't seem to carry authentic-content, how can we hope to advance towards an understanding of the conditions for "authentic-intentionality" working in the framework of computational neuroscience? Understanding and meaning belong in the personal-level, not in the sub-personal level (Dennett 1969). Meaning is for people, not for their brains. How can we hope to bridge the two levels working at the level of brains instead of that of people? Before addressing these concerns, the next step is to consider Haugeland's tack on the problem of intentionality.

Haugeland on Authentic Intentionality

Perceptions exhibit intentionality. 'Intentionality', following Haugeland, is a synonymous of semantics (or meaning). The difference between authentic-type representation and detection-type representation, accordingly, amounts to the difference between authentic intentionality and *ersatz* intentionality (Haugeland 1992). Consider the chess-playing computer Deep Blue and a human chess-player. Haugeland thinks that whereas the human chess-player perceives pieces, positions, and moves that make sense to her, Deep Blue may detect certain positions of different types of pieces on the chessboard, but its perceptions don't make any sense to it. Whereas the human chess player's perceptions exhibit authentic intentionality, the intentionality of Deep Blue is *ersatz*, it's "as-if intentionality". What grounds this difference in chess perception? Why can the human chess player have as object of her perception a knight-fork, and Deep Blue cannot even if we say that it plays as-if it had the same perception?

Haugeland argues that the difference is grounded in the capacity for commitment. His argument is transcendental. The transcendental method applied to the study of the intentionality consists in an investigation of the conditions necessary for intentionality. The first step in Haugeland's argument is that genuine intentionality requires the "capacity for objectivity". At least some human perceptions are objective in a way that animal and robot perception is not.

Objective perception is perception that is true of objects *as* such, *as* they are (Haugeland 1996). In this respect the capacity for objective perception is different from the capacity to detect something reliably. 'Object', in the way Haugeland uses the terms, is not (only) what is perceived. "Objectivity – he writes (Haugeland 1998a, p. 241) – in perception is a kind of structure that involves the perceiving, that which is perceived and the relation between them". The relation between perceiving and that which is perceived is crucial.

Let's return to the chess example to put that relation into focus. Deep Blue doesn't have authentic representations because it doesn't have the capacity for objective perception. It cannot perceive a knight-fork *as* such. The reason why this is so is that it lacks the capacity that would enable its perceiving (or detecting) to be in the right relation with that which is perceived (or detected). This relation does not depend either on proper functioning or on consensus. Instead it depends on the capacity to *care* about the rules of the game. But Deep Blue doesn't have this capacity. Notice that, according to Haugeland, caring is independent from both consciousness and emotion. To clarify the point Haugeland argues that a perception of a knight-fork is objective only in the context of a game of chess. A game of chess to be a game of chess, and not another, requires the existence of certain rules, or standards, that not only govern the game, but also *constitute* it. Without those rules, the game of chess wouldn't exist. These standards are not only to be followed - in the same way drivers follow some driving rules on the road, but they call for a commitment. When one plays a game of chess she is not merely subject to the rules of the game; it is constitutive to that activity she is engaged that it is governed by those rules. Driving through a certain town in a foreign country that happens to have certain rules, one is subject to the rules of that country authority. But it is not constitutive to the activity of driving through that town that it is governed by those rules: that area, for example, could have come under the control of a different country with different rules. Instead, it is impossible that you play a game of chess, which is constituted by certain rules, and you are not subject to those rules.

According to Haugeland, chess players must be *committed* to those rules. They must *care* about the rules. This distinguishes human chess players from Deep Blue. Deep Blue may "know" the rules of the game, and may "follow" those rules, it may also "want" to win the game, and it may be extremely good at winning games. But Deep Blue doesn't "give a damn" (Haugeland 1998a, p. 47) about the rules. This is so because it lacks the *capacity to care*. He cannot commit itself to the standards essential to games of chess. Deep Blue lacks the capacity of commitment. "Existential commitment" to the constitutive standards that govern an activity is condition of possibility for objective perception. Existential commitment is the capacity to recognize and take responsibility for the standards and skills with which an agent copes with

things in a certain domain. Deep Blue lacks the capacity for objective perception. Therefore its representations are not authentic. Only humans can be consumers of authentic representations because only humans have the capacity to care and to be existentially committed. The notion of existential commitment is important and requires clarification.

When Haugeland argues that objectivity is constituted by a commitment to certain (constitutive) standards, he is after the articulation of a kind of resilient skill, or know-how which is essential to understanding. A resilient skill, here, is the ability to recognize and cope with the phenomena in a domain. In chess, for example, it amounts to the ability to tell whether a certain piece is a bishop or that a certain move is a castling and is permissible at that point in the game. Haugeland argues that the mark of this resilient skill is the capacity of recognizing an impossible event *as* impossible. The possible, here, is the realm of "legal moves" in a given constituted domain of objects. Phenomena which fall out of this realm are impossible since they are excluded by the constitutive standards of the domain. Only people are in the position of possessing the capacity for resilient commitment. To understand why, let's move from chess perception to everyday perception. Haugeland (1996) asks us to imagine a situation where the sensible properties of different things in our environment start mixing up. What looks like a table smells like an orange juice; what smells like an orange juice, looks like rock; what moves like a football, sounds like a thunder, and so on and so forth. Also their parts start mixing up; parts of tables with a hemisphere of the football, bits of rocks attached to a quarter of ham, and so on and so forth. Now, since some combination of properties doesn't *seem* to be permissible in one and the same thing, how would you react to this situation? For Haugeland you would reject what you seem to perceive. "That's impossible!" you would say. Recognizing this impossibility *as* impossibility means that what you seem to perceive goes against the rules – in the same way an illegal move in chess goes against the rules. Those are the rules that constitute the domain of things which you perceive, or, as it were, the standards in virtue of which that which you perceive makes sense to you. If something doesn't seem to accord with the standards, you (as a human being) are in the position to double-check for a mistake, and then, once recognized something as impossible, to reject it since it *cannot* be that way. If the anomaly persists in spite of your check, you (as a human being) can realize that the standards themselves are wrong, and hence you can revise them. The way we should understand "realize" here is problematic. For the moment, I assume that "realize" does *not* imply that we need consciousness. I shall expand on the point in the fourth section.

Robots and animals are not in such a position for Haugeland. Animals in the situation just described will be confused, disoriented, but they will not be able to reject what they seem to perceive on the grounds that that cannot be possible – let alone revise their standards if the

anomaly persists. For animals and (current) robots don't hold things to some constitutive standard about what is and isn't possible. Because of these reasons, Haugeland concludes that only human beings are in a position such that their perceptions can be authentic representations. Only human beings can understand – at least currently.

How Neurons Can "Count on Standards"

I don't intend to criticize Haugeland's argument directly. My target is one consequence of his argument: Neurons cannot care, cannot commit themselves to constitutive standards. Consequently an account of the conditions for authentic intentionality in terms of neurocomputational functions is doomed to failure. I would like to give some grounds to resist this conclusion. The upshot of my argument is neither reductionism nor a naturalization of meaning. For, on the one hand, I don't know how to derive high-level properties such as "being-responsible" from low-level neural properties, and on the other, I don't have a theory of meaning to defend. What I'm after is to unlock the possibility to make some steps toward a better understanding of the conditions for authentic representations from a neurocomputational perspective. In so doing I wholeheartedly endorse a *co-evolutionary* approach to understand cognition (Churchland 1986). I assume that any account of cognition needs to be informed and constrained by our knowledge about the neural structures where cognition is realized. Personal-level theories of representation must co-evolve with sub-personal theories of how the brain works.

What computational functions may neurons have in order to enable the consummation of authentic representations? In light of Haugeland's argument it is sufficient that neurons possess such computational functions that can make possible the ability to "take responsibility" for standards that govern a certain domain of objects. This ability can be further analyzed in three abilities:

i) The ability to realize when a representation "goes wrong", that is when it falls outside the realm of the legal moves in a given domain.

ii) The ability to reject a representation *as* wrong, or *impossible* since it falls in a zone excluded by the constitutive standards of that domain.

iii) The ability to revise the standards of possibility if the anomaly persists.

The remaining of this section is devoted to articulate *one* possible account for how neurons can *count* on standards that constitute a domain of object. It will serve as the beginning of an explanation - grounded in neural computation - of the enabling conditions for what is required for

authentic intentionality. I will have something to say also about abilities (i – iii), in particular on how constitutive standards can be revised, but my focus will be on "counting on constitutive standards of objectivity".

Some words of *caveat* are in order now. In discussing some computational functions of neurons, at times I shall stretch the intentional vocabulary, thereby blurring the distinction between sub-personal and personal level. I don't mean to suggest that neurons themselves can understand, or can be aware of anything in the same sense people can. In doing so, my aim is to ascribe specific computational functions to neurons, circuits of neurons, and brains. I shall also pay little attention at the structural level (either single neurons, or circuits, or systems, or the entire brain) where the relevant computational functions have to be ascribed since I am not aiming at providing a full-blooded story, but only the coarse-grained beginning of a plausible explanation.

Neurons that Care

There are many ways in which representations can "go wrong". "Going wrong" presupposes the existence of standards. Realizing that something goes wrong presupposes a kind of know-how. This know-how is not necessarily conscious. In virtue of this know-how representation-consumer systems can coordinate with their environment by acting upon their representations. Neurons have two functional properties that might underlie this kind of know-how. The first is valuation. The second is emulation. Let's examine each in turn.

A key function of brains is evaluation (Montague and Berns 2002; Montague 2007). By evaluating internal states, external stimuli, and behavioural outputs, brains have found a way to be efficient systems. Efficiency, in this context, is a measurement of the way a system uses its energy. One way a computational system like the neural system can operate efficiently in an uncertain world is by means of goals and rewards. If the system has some goals, that is, some state to pursue, then it is endowed with knowledge about how to invest its energy among the possible courses of actions it may take given a certain state. The system must also be endowed with some guidance that enables it to distinguish what goals are worth pursuing. *Reward*-values can play the role of guidance signals. What counts as reward can be defined operationally as the positive value that a system places on the attainment of a certain state. Accordingly, the nervous system might assign to each of its representations a certain reward value whose functional role is to "motivate" the system to work in order to get and consume high-value representations. For the computational neuroscientist Read Montague, neural representations literally carry their own "value-tag" (Montague 2007). The value-tag is a measure of the worth of the prospective goals that can be reached by pursuing a certain action. Values are measures of

the degree to which the nervous system should *care* about certain representations (Montague 2007, p. 19). Also the neuroscientist Edmund Rolls (2001, pp. 164-166) hints at the same point. He asks how brain representations have "content", how they are "grounded in the world". And he suggests that "one type of meaning of representations in the brain is provided by their reward (or punishment) value: activation of these representations is the goal of actions" (p.166).

Although it's not completely clear what Montague and Rolls have in mind when they talk about "meaning" or "content", the point that we can drive home by relying on this line of research is that there are evidential grounds that (some) neurons possess such computational functions as evaluating and integrating value to that which they represent. The value integrated to e.g. a perceptual representation serves as a "motive" to act (or abstain from acting) upon that representation to obtain a reward (Niv, Joel, and Dayan 2006). The same representation might motivate, in function of its value, different courses of action that will lead to different representations thereby stamping value-links between representations. All the value-links contribute to develop a know-how underlying an agent's capacity to navigate the world.

Representation-*cum*-value is not enough, however, to enable the capacity to find out when something goes wrong. Values might set goals that the nervous system can represent. This by itself does not imply that neurons dedicated to evaluation can play their role in the *absence* of incoming external stimuli. But without this kind of de-coupling, goals would be stored representations that the system can only compare on-line to the current representation. The nervous system wouldn't have the capacity to act resiliently upon its representations since it wouldn't be in the position to use its know-how *off-line* thereby figuring out what may go wrong in counterfactual futures. What seems to be sufficient for such decoupling capacity is emulation.

An emulator simulates the functions of a system using another system. It takes as input information about the states of the target system at a time and about the transformations which the information would undergo. The emulator yields as output a prediction of the state of the system. The computational function of an emulator is to model the processes of another system. At least some brain circuits seem to act as models of other, usually extra-neural, systems (Grush 2004). Emulator-circuits, thus, possess a forward-modeling capacity. They run in parallel with other systems, often off-line. As a consequence, they produce "expectations" about e.g. sensory feedback, and predictions of the outcome of a possible action. Emulation bestows the capacity for consuming "full-blooded" representations on the system on which they run (Clark and Grush 1999). Courtesy of emulation – convincingly argue Clark and Grush (1999, p.8) - it would be possible to identify within the central nervous system circuits of neurons whose functional role is to *stand-in*, as surrogates, for other specifiable states – either neural or extra-

neural - without the need for continual on-line interaction. These are what Clark and Grush "full-blooded representations". For present purposes, the important consequence of emulation is that it would bestow the capacity to represent "what could have been" on neurons.

Once we grant that emulation and valuation are computational functions of (some) neurons and that they can be integrated, some grounds are available to explain how neurons may enable us to realize that a representation has gone wrong. The marriage between emulation and valuation can be couched in terms of one tenet of reinforcement learning (Sutton and Barto 1998): temporal prediction errors. A prediction-error signal carries information about how well the actual reward attained after a choice tallies with the reward that was expected. When the actual reward doesn't match the predicted one, the stored value of the representation that predicted it is updated. Courtesy of emulation, brains can simulate counterfactual futures. Through valuation they can endow representations with value. Integrating these two computational functions brains are in a position to draw upon their current and past representations to predict what future representations may hold, and to consume the most valuable of these possible representations. Temporal difference prediction errors encoded in neural activity would enable us to "realize" that something has gone wrong by quantifying the difference between what was expected and what is actually consumed.

Summing up, valuation and emulation may be sufficient to enable a system to realize when something has gone wrong. Available evidence indicates that (some) neurons posses these properties. Valuation enables the system to care about certain representation; emulation enables the system to consume "full-blooded representations". At least some neurons seem to be in a position to enable a system to realize when something goes wrong because they enable it to care about certain valuable representations, and they enable it to realize that it was wrong about its expectations courtesy of prediction errors.

Presumably, however, the notion of standards implicit here is different from Haugeland's. The objection is that – even if we grant that the interpretations given to valuation and emulation are apt to describe some (non-empty) class of neural activities - the account sketched above confounds constitutive with regulative standards. If we consider evaluation and emulation alone, we are implicating "regulative standards" rather than constitutive standards. This objection seems to have bite. Given a domain of objects, in fact, "regulative standards" govern the choice among the set of legitimate moves in that domain. Constitutive standards are essential to the domain itself. But the rules underlying the "know-how" enabled by valuation and emulation are *not* essential to the representational domain itself. If, for example, the valuation or emulation systems were impaired, the brain (or parts thereof) would carry on representing internal states,

external stimuli and behavioural outputs. It is not essential to the activity of representing that one representation carries a certain value-tag or is linked to another representation courtesy of prediction-reward errors.

Our brain may be such that it does not enable us to make “good choices” among the set of legitimate moves in our representational domain because of impairment to the valuation or emulation systems. But emulation and valuation are not enough to enable us to identify standards *constitutive* of the representational domain itself. The task of the next sub-sections is to identify one possible neurocomputational foundation of the ability to count on constitutive standards. I shall first say something on the enabling conditions for the capacity to realize and reject a representation *as* “wrong”, that is for the capacity to realize that a representation falls outside the realm of possibilities of a given domain and reject it as “impossible”.

Rejecting an impossibility *as* impossible

How may neurons enable us to deal with perceptual “impossibility”? To make sense of this question and suggest an answer, "impossibility" needs qualification. We need to be clear on the standards which define what is possible and what is impossible in a given domain. To reiterate the point made above, possibility/impossibility has not to do with the standards implicitly defined by the guidance system based on valuation and emulation. The guidance system may ground representations-in-use, or “full-blooded” representations. But in that context, “going wrong” doesn’t really seem to capture what Haugeland has in mind. Emulation *plus* valuation does not tear apart a zone of impossibility (and of possibility) in a given domain. Therefore, we should consider other computational functions of neurons as plausible candidates to make sense of the question of how neurons may enable us to deal with perceptual “impossibility”.

The coding-decoding combination, which is one way to characterize neural representations, is the obvious place to look at to approach this problem. Realizing that a representation is impossible presupposes the existence of a mapping between an “alphabet” used by a representational system and an “alphabet” used by the represented system (Eliasmith 2003). Realizing that something is impossible means that there is a mismatch between the two alphabets, given the possible mappings. In neurocomputational terms, the encoded alphabet is given by neural responses, and the decoded alphabet is given by physical properties. The encoding-decoding combination defines the mapping between the two, thereby defining possible and impossible representations.

By pursuing the characterization of neural representation assumed here, everyday perception is enabled by an encoder-decoder cascade (for a similar modeling assumption see

e.g. Seriès, Stocker and Simoncelli 2009). According to this view, “the encoder is defined by the probabilistic response of a population of neurons, and decoder transforms the population activity into a perceptual estimate” (Ibid.).

There are different ways to model decoding (see Dayan and Abbot 2001, Ch. 3). Perhaps the most common is maximum-likelihood decoding, which is an optimal method to reconstruct a stimulus from the firing rate of a neural population. Given a neural response, the decoder yields the most likely stimulus that produced it. If *s* is a specific physical feature of an external object, **r** is a spike train, and we know the encoding model P(**r**|*s*), then the maximum-likelihood estimate $\hat{s}$ of the stimulus *s* is the stimulus that has maximal probability of having caused the response **r**, that is $\hat{s} = \text{argmax}_s P(\mathbf{r}|s)$.

Notice two points about this characterization. First, even though it isn’t known at the moment whether neurons in fact perform maximum-likelihood estimation, there is evidence that neurobiologically plausible architectures can perform this kind of computation (Jazayeri and Movshon 2006; Deneve, Latham and Pouget 2001). Second, this characterization of the decoder underwrites the fact that neurons have graded responses to stimuli, thereby suggesting that neurons are not detectors that determine that either there is a stimulus *s* of that there is not *s*. P(*s*) is the prior probability of the stimulus: it is meant to reflect the distribution of values of a perceptual stimulus *s* in the world. Everything that is perceived in the world depends on prior assumptions about the world. That is, if P(*s*) = 0, then whatever the neurons say, for any *s*, P(*s*|**r**) = 0. This may be a case of impossible representation, where the standards essential to the neural representational domain are *statistical*. Brains, therefore, have to start with a statistical model of the world which defines the standards of possibility for our representational domains. The model can be updated on the basis of experience by interacting with the world. This hypothesis has independent supported from the evidence that small children and young animals possess a rough, hard-wired model of the statistical structure of their world (Eliasmith 2005, sec. 5; Spelke and Kinzler 2007; Spelke and Van de Walle 1993; McCloskey 1983). This innate model needs not be very detailed; it needs to be good enough to establish a domain of representational legal moves.

Now, in light of the kind of constitutive standards just considered, let’s re-examine the question which opened this section. What are the neurocomputational principles which may enable the capacity to realize that something has gone wrong *according* to the neural statistical model of the world (and hence has to be rejected)? A promising way to try to gesture at an answer is to consider perceptual illusions. This is not because illusions are here taken to be a kind of impossible representations; rather, because illusions have been convincingly suggested

to be diagnostic as to the computational principles underlying the neural coding of information (Clifford et al. 2007).

Some perceptual illusions have been observed to be linked to sensory adaptation, which is a fundamental computational function of sensory neurons. Indeed, it seems that neural computation is sensitive also to the *temporal* dimension of the statistics of the environment. The timescale over which neural computations - and the behaviour they enable - and environmental statistics may affect each other can be millennia, months, or minutes and second. *Neural adaptation* is the process by which sensory systems alter their operating properties in function of changes in the environment where they are embedded. It can generate weaker or stronger neural responses in function of the kind of environmental changes. One consequence of neural adaptation is that unchanging properties of a scene tend to fade from view (Martinez-Conde, Macknik, and Hubel 2004). The temporal context of a stimulus is given by what has been perceived close in time to the stimulus. Temporal dependencies can result in adaptation which, in turn, may lead to perceptual biases such as the "tilt after-effect" (Gibson 1937).

The tilt "after-effect" depends on the temporal context where a stimulus is perceived. A visual stimulus of a particular orientation, say lines tilted 15 degrees clockwise from vertical, is presented before a target, say vertical lines. The two stimuli overlap in space but one is presented before the other; thus their "temporal context" is different. After having gazed for at least thirty seconds at the first stimulus, the vertical lines of the target stimulus will appear repulsed away from the "temporal context orientation", say tilted counter clockwise (see Schwartz, Hsu, and Dayan 2007, p.523, Fig. 1). The neural changes underlying adaptation usually occur over relatively short time scales – tens of milliseconds to few minutes; and, given a constant stimulus, they lead to a decrease in neurons' responsivity and to modifications in tuning curves properties (Ibid., p. 525, Fig. 3).

During sensory adaptation a change in the encoder - as defined above - is usually observed. Given this change, and given the assumption that a brain representation is identified by an encoding-decoding cascade, we should examine whether the decoder changes accordingly, or remains "unaware" of the changes in tuning. In the former case, in order to maintain the optimality the decoder would dynamically adjust to match the changes in the encoder. In the latter case, the decoder would be unaffected by adaptive changes in the encoder, thereby becoming sub-optimally mismatched to the altered encoding population. The 'optimality' involved here is statistical and has to do with the constitutive standards discussed above. The goal of visual perception, it is assumed here, is to yield an "efficient" representation of the world. This claim can be spelled out in many different ways: given the conception of 'neural

representation' adopted here, neuronal representations are optimal when they encode "as much information as possible in order to most effectively utilize the available computing resources" (Simoncelli and Olshausen 2001, p.1195).

Electrophysiological and psychophysical data are usually linked by assuming the latter case, namely: the decoder receives the adapted input without being aware of those changes thereby "going wrong" - for example in yielding absolute orientation. This possibility is referred to as "coding catastrophe" (Schwartz, Hsu, and Dayan 2007), and has been recently suggested that the perceptual illusions brought about by adaptation arise as a result of it (Seriès, Stocker, and Simoncelli, 2009). From the normative viewpoint suggested above, the performance of perceptual systems shouldn't be disrupted so easily: the decoder should "know" that changes have occurred due to adaptation and adjust accordingly. Although, at present, it is problematic how to understand the computational and mechanistic properties of coding catastrophe, this notion might suggest that sensory adaptation may be sufficient to ground a "resilient capacity" to realize that a representation is "wrong" – according to some constitutive statistical standards - and has to be rejected.

How does neural adaptation help us to understand how that "resilient capacity" is enabled? Let's assume that the response properties of sensory neurons are matched, in function of some statistical optimization criterion, to the properties of the physical features to which they are exposed. Now, sensory adaptation is consequence of spatial and temporal contextual modulation of a stimulus. Encoding populations after a certain period of being excited by a constant stimulus "get tired" and their underlying tuning curves undergo a change. The important point is that one of the consequences of adaptation is that the encoder model P(**r**|s) changes. On a short time scale, a mismatch between the adaptive changes in tuning curves and the unaltered downstream processing may be observed. This mismatch typically leads to perceptual illusions which arise as a result of the temporary sub-optimality of the "unaware" decoder. Seriès, Simoncelli, and Stocker (2009)'s results underwrite such a conclusion. By investigating the effects of adaptation on representation of motion direction and contrast, they found that an adapted encoder coupled with an "unaware" decoder leads to predictions consistent with the psychophysical data.

We may speculate that adaptation leads to changes in the statistical structure of the representational domain of the encoding neuronal population. To such changes the decoder might be initially "unaware". The decoder will not accept as "possible" something which it deems to be impossible given its prior "knowledge" of the statistical structure of the domain of the encoder. This mismatching in the coding-decoding cascade is what I take to ground the ability to

realize that a representation is impossible and hence has to be rejected. Since sensory adaptation may be sufficient to produce such a mismatch, sensory adaptation may be sufficient for that ability. Impossible representations, therefore, are representations which do not follow from the normal statistical mapping between encoded “alphabet” and decoded “alphabet”. They fall outside the realm of legal moves for a certain domain. In this sense, the statistical structure of the world is criterial for the correct exercise of what Haugeland calls a resilient skill.

Learning to perceive

The argument above provides us with *one* obvious possible suggestion regarding the conditions under which neurons would modify the standards that govern their representational domain when a mistaken representation is suspected.

After the perceptual system has “tried” to resolve the coding-catastrophe unsuccessfully, on a longer time-scale the decoder may become “aware” of the changes in the encoder, and may adjust its standards accordingly. With the re-matching between encoder and decoder, new standards – and new “legal moves” - for possible representations would ensue. The computational functions that would enable this re-matching are described by the term *perceptual learning*. Those neurocomputational functions ground the capacity to improve in performance in perceptual tasks after training or extensive sensory experience (Fahle and Poggio 2002). Perceptual learning involves changes at different structural and functional levels in primary sensory cortices. Structurally, perceptual learning involves changes in early sensory representations. Functionally, it may lead to new ways in which these representations are used in a task. For example, orientation discrimination, after perceptual learning, can significantly improve: after training smaller orientation differences can be reliably detected. Thus, neurons, after perceptual learning, perceive “new” objects. Such changes in performance are underlain by changes in the brain. After perceptual learning, under certain environmental conditions, a decrease in neurons’ responsivity might be recorded analogously to the observed decrease in activity in the neurons tuned at and around the adapted perceptual target. This fact suggests that perceptual learning might be viewed as a phenomenon analogous to sensory adaptation, only operating on a longer time-scale (Teich and Qian 2003). In fact, whereas, the changes induced by adaptation are transient, the structural and functional effects of perceptual learning are long-lasting. Both adaptation and learning would be underlain by tuning curve changes that would, thereby, explain the psychophysical effects of both phenomena. It is possible that on longer time scale the perceptual system involves changes in the perceptual system whereby neurons ability to respond accurately to the physical features relevant to the task at hand improves. With learning, the decoder would dynamically adjust to the changes in the encoder.

And new standards of objectivity would thereby become available. In light of the discussion thus far it may now be helpful to make clear the notion of constitutive standards underlying my argument.

The natural environment has crucially contributed to the evolution of the areas of the brain responsible for the processing of perceptual information. The discussion on representation as encoding-decoding cascade suggests that the job of brain sensory systems is to extract information about the spatio-temporal environment encoded as patterns of activity across populations of neurons. The information extracted by the activation of such populations is about the significant contents of the represented spatio-temporal environment. Whatever standards brain sensory systems may happen to have to carry out their job, they must be dependent on the environment where the sensory systems evolved and develop. This is reflected in neurocomputational modeling where most models "start from the premise that the way the cortex decomposes images is influenced by the regularities inherent in the statistics of [the spatio-temporal input up to time t] associated with the input scenes" (Schwartz, Hsu, and Dayan, 2007 p. 527, see also Simoncelli and Olshausen 2001). The standards reflect a matching between the response properties of neurons and the properties of physical signals to which brains are exposed. The matching *constitutes* the domain of objects which can be represented by neurons. Given the statistical nature of the encoding-decoding cascade which may identify a neural representation, it seems, therefore, that ultimately the standards that constitute neural representations reflect the statistical structure of the world where brains are embedded. It is the world itself that imposes the standards. Before elaborating on this last claim in the next section, let's take stock by highlighting the main claims made and defended so far.

First, courtesy of emulation and valuation brains (or parts thereof) are in a position to enable the consumption "full-blooded representations."

Second, encoding and decoding may identify what a neural representation is.

Third, courtesy of adaptation brains (or parts thereof) may be in a position to enable the capacity to realize and reject a representation as impossible.

Fourth, perceptual learning enables brains (or parts thereof) to modify their standards.

Fifth, the representational standards which define the domain of objects brains (or parts thereof) can represent reflect the statistical structure of the environments where brain sensory systems evolved and develop.

If my argument is sound, then it would follow that one of the conclusions from Haugeland's argument has to be revised. Namely, ideas and evidence from computational neuroscience might offer a promising framework for a better understanding of the conditions of possibility of

“authentic intentionality”. Besides, it would turn out that animals and robots which are endowed with the kind of computational resources discussed above might have authentic representations.

A Fundamental Objection and a Tentative Reply

There is a fundamental objection to the argument I have just been articulating that I want to address. So goes the objection: “The conclusion you draw about commitment and standards miss the target. You trade on an equivocal sense of commitment to standards. Haugeland is after censoriousness and taking responsibility for constitutive standards of objectivity, which cannot be accounted for at the neural level.”

One way to identify the mark of authentic intentionality, for Haugeland, revolves around the difference between realizing that something is strange, puzzling, and realizing that something that seems like it is in a certain way *cannot* be possible, and therefore has to be rejected. For my argument to have some validity, some neurocomputational property has to be sufficient to enable that kind of “taking responsibility for standards”. In a sense it’s reality itself that imposes those standards through evolution. Because of efficiency, or evolutionary fitness, in fact, neural computation is *regulated* by standards such that neurons try to consume “valuable” representations. Courtesy of computational mechanisms such as temporal difference prediction errors, on the one hand, brains have found a way to realize when a representation goes wrong, and to re-adjust the values that guide them to consume certain representations rather than others. Courtesy of adaptation, on the other, brains have found a way to alter their representational operating-properties in response to the space of possibilities defined by the environment in which they are embedded.

Haugeland may acknowledge all *that* (Haugeland 2002b, pp.142-143). He may acknowledge that brain representations follow certain standards because of their evolutionary causal history, that neurons sometimes “misfire”, that they can adapt to perceptual conditions, and that they can learn. But Haugeland is not really concerned with all this. He is after an appearance-reality distinction. In the case some “illusion” persisted, people, but only people, have the resources to reject it. He asks us to consider the case of a magic show. In front of the illusion produced by the magician, people, but only people, can refuse to believe that that which seems in a certain way cannot really be that way. If after careful consideration, the illusion does persist, people, but only people, can realize that the appearance-reality standards they have, are likely to be wrong, and thereby they can revise them.

“As creatures who won’t accept as real anything they deem impossible, and who can realize that their convictions about possibility are wrong, we are also creatures who can realize not merely that their actual representations of things are defective but that their very

representational resources are defective – and they can fix them. This capacity is certainly a cultural achievement, perhaps the highest so far." (Haugeland, 2002b, p.144)

The problem with this conclusion is with "realize" and with "constitutive standards". Let's focus on standards first. While Haugeland will admit that (either biological or cultural) evolution is important to explain how people came to commit to constitutive standards, he will reject that commitment is intrinsically an evolutionary process. The reason is that we can hope to come to grips to the kind of normativity involved in this commitment only by considering what is *ontologically* (not simply causally) required for the possibility of non-accidental objective truth. Then, for such a task, an appeal to evolution is ultimately unsatisfactory (Haugeland 1998b).

My argument however does not appeal (only) to evolution. My argument on constitutive standards has two sides. One side does involve evolution. Trivially, brains are the way they are because of evolution. Evolution might provide us with constraints to understanding the causal origin of neural capacities involved in representing. The second side, however, points out that it is the world, with its statistical structure, in conjunction with specific principles of neurocomputation that discipline the representing capacities of neurons. This is what *constitutes*, not only regulate, a possible domain of objective representations. To want to know about this constitutive relation amounts to want to know about the "structural" relations between mathematical/statistical properties of electrical signals in the brain and the mathematical/statistical properties of the represented object in the world (Shagrir 2006).

This issue does not depend on evolution. It is onto-logical, in that in order to understand how neural representational standards constitute perceptual content, we have to understand the statistical structure of the world where brains are embedded, and the computational constraints underlying the neural coding of information. Hence, objective perception actually requires that brains "count on", and "insist on" constitutive standards. My use of intentional vocabulary notwithstanding, I didn't mean to suggest that neurons are conscious or aware of anything like people are. My argument assumed that neurons "commit" themselves to constitutive standards, and enable the capacity to realize when some representation is defective in an a-conscious, deterministic way. I didn't assume that Haugeland's argument presupposes consciousness either. But if this is really so, then it becomes difficult to understand what marks the real difference between people, who can realize that their convictions about appearance and reality are mistaken, and dogs, which cannot.

Haugeland (2002b, p. 143) argues that the essential difference is "whether convictions about what is and isn't possible can figure among the considerations upon which [the conclusion that a representation has to be rejected on the grounds that it would be impossible for things to be the

way we represent them to be]". But this difference seems to rely on a difference in *complexity*, and not in *kind*, in the representational resources available to a consumer of representations. Specifically, it seems to depend on a capacity for a particular kind of use consumers of representations can make of their representations. In short, the difference seems to be between thinking, and thinking about thinking. *This* appears to be "a good candidate for a distinctively human capacity" (Clark 1997a, pp. 208-209).

Haugeland doesn't deny that dogs can be wrong about things, and realize that they are wrong about things and perhaps reason and entertain hypotheses. What Haugeland denies is that dogs, robots (at present), and neurons can distinguish between possible and impossible. But this capacity of distinguishing between possible and impossible, I believe, is not dissociable from a capacity for thinking about thinking. Specifically, it is not dissociable from a capacity for "florid representing", which requires the ability to "thinking about thinking" (Dennett 1998). Thinking about thinking is the ability to self-consciously, deliberately use of a token of thought for further thought. This is a "florid" way of using representations. But this seems exactly the way in which people can realize that their convictions about possibility and impossibility are wrong, whereas animals cannot. For if we suppose that the ability that people have to take responsibility for standards of (im)possibility is dissociable from their ability to use their representations in a florid, "knowing" way, then we would have independent grounds to argue that there is no real difference between dogs and people rejecting some of their perceptions as impossible. Both dogs and people may be mere believers, both – if we assume that dogs have a "theory of mind" - may entertain beliefs about beliefs, and therefore they can *unwittingly* reject something as impossible on the basis of their beliefs, and their responsivity and discrimination. Haugeland would probably reject this argument. But then, he should clarify whether resilient commitment is dissociable from florid representing or not. If it is, then the argument just made would re-apply. If it isn't, then we must conclude that self-consciousness is presupposed for commitment since florid representing involves self-consciousness.

If this is correct, two consequences follow. The first is that in order to make his argument about authentic intentionality *while* maintaining the irrelevance of neural computation, Haugeland should explain, without appealing to commitment (on pain of circularity), under what conditions self-consciousness arises. For authentic intentionality requires commitment, and commitment requires florid-representing which involves self-consciousness. In default of such independent argument, the conclusion that an appeal to neurons' computational properties cannot be the right way to account for authentic representation has to be resisted. This is the overall conclusion of my argument. The second consequence, with which I conclude this section, is that it is not clear to what extent an account of florid-representing can be successful by adopting the

explanatory framework of computational cognitive neuroscience, since florid-representing seems to require skills that would extend cognition beyond the boundaries of the neural realm (Clark 1997a).

Conclusion

The problem of content is in general the problem of accounting for how mental representations mean. John Haugeland is one of the most influential explorers of this problem. I am sympathetic with Haugeland's general diagnosis about the conditions for authentic representation. However, if I understand it correctly, I disagree with one of the consequences of Haugeland's argument. Namely, the consequence that an account of the conditions for authentic intentionality in neurocomputational terms is doomed to fail. This paper has argued that this claim remains unsupported. It has provided the beginning of *one* possible account of how neural computation might enable authentic intentionality, and has suggested one way to understand the nature of the standards that re constitutive of objective representation. The study of brain computations might be a pervious path towards the quest for meaning.

Acknowledgments

I am sincerely grateful to Peggy Serie`s and Andy Clark for their encouragement and for their generous comments and criticisms on earlier versions of this paper. This research was partly funded by an Engineering and Physical Sciences Research Council (EPSRC) Studentship, awarded by the School of Informatics of the University of Edinburgh. The usual disclaimers about any error or mistake in the paper apply.

References

Churchland, P., & Grush, R. (1999). Computation and the brain. In F. Heil and R. Wilson (Eds.),*The MIT Encyclopedia of Cognitive Sciences*, (pp. 155-158). Cambridge, MA: MIT Press.

Churchland, P., & Sejnowski, T. (1992). *The Computational Brain.* Cambridge, MA: MIT Press.

Clark, A. (1997a). *Being There. Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Clark, A. (1997b). The Dynamical Challenge. *Cognitive Science,* 21(4), 461-481.

Clark, A., & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior*, 7(1), 5-16.

Clifford, C., Webster, M., Stanley, G., Stocker, A., Kohn, A., Sharpee, T., Schwartz, O. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research*, 47, 3125-3131.

Dayan, P (1994). Computational modelling. *Current Opinion in Neurobiology*, 4, 212-217.

Deneve, S., Latham, P.E. and Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*. *4*(8), 826-831.

Dennett, D. (1998). Making tools for thinking. In Dan Sperber (Ed.) (2000). *Metarepresentation*. New York: Oxford University Press.

Eliasmith, C. (2003). Moving beyond metaphors: Understanding the mind for what it is. *Journal of Philosophy*, C(10), 493-520.

Fahle, M., & Poggio, T. (Eds.) (2002). *Perceptual Learning*. Cambridge, MA: MIT Press.
Fodor, J. (1981). *Representations.* Cambridge, MA: MIT Press
Gibson, J.J. (1937). Adaptation, after-effect, and contrast in the perception of tilted lines. II. Simultaneous contrast and the areal restriction of the after-effect. *Journal of Experimental Psychology,* 20, 553-569.
Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences* 27, 377-442.
Haugeland, J. (1992). Understanding Dennett and Searle. In A. Revonsuo and M. Kamppinen, (Eds.), *Consciousness in Philosophy and Cognitive Neuroscience* (pp. 115-128). Hillsdale, NJ: Lawrence Erlbaum.
Haugeland, J. (1996). Objective Perception. In K. Akins, (Ed.), *Perception: Vancouver Studies in Cognitive Science*, Vol. V. (pp. 268-289). New York: Oxford University Press.
Haugeland, J. (1998a). *Having Thought: Essays in the Metaphysics of Mind*, Cambridge, MA: Harvard University Press.
Haugeland, J. (1998b). Truth and Rule Following. In *Having Thought* (pp. 305-361).
Haugeland, J. (2002a). Authentic Intentionality. In Matthias Scheutz, (Ed.), *Computationalism: New Directions* (pp. 159-174). MIT Press.
Haugeland, J. (2002b). Reply to Cummins on Representation and Intentionality", In Hugh caplin (Ed.). *Philosophy of Mental Representation* (pp. 138-144). Oxford University Press.
Hurley, S. (2008). The Shared Circuits Model: How Control, Mirroring and Simulation Can Enable Imitation, Deliberation, and Mindreading. *Behavioral and the Brain Sciences*, 31(1), 1-21.
Jazayeri, M., & Movshon, J.A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience 9*(5), 690-696.
Martinez-Conde, S., Macknik, S.L., and Hubel D.H. (2004). The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience 5*, 229-240.
McCloskey ,M. (1983). Intuitive Physics. *Scientific American*, 284,122-129.
McDowell, J. (1994). The content of perceptual experience. *Philosophical Quarterly* 44, 190-205.
Montague, R. (2007). *Your Brain is Almost Perfect: How we make Decisions*. New York: Plume.
Montague, R., Dayan, P, Person, C, and Sejnowski, T.J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. *Nature* 377, 725-728.
Montague, R., & Berns, G. (2002). Neural economics and the biological substrates of valuation. *Neuron*, 36, 265-284.
Niv, Y., Joel, D., and Dayan P. (2006). A normative perspective on motivation. *Trends in Cognitive Sciences*, 10(8), 375-381.
O'Reilly, R.C. & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.
Quartz, S.R., & Sejnowski, T.J. (2002). *Liars, Lovers, and Heroes: What the New Brain Science Reveals About How We Become Who We Are.* New York: Harper Collins Publishers Inc.
Rolls, E.T. (2001). Representations in the brain. *Synthèse*: 129, 153-171
Schwartz, O., Hsu, A., and Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience*, 8(7), 522–535.
Searle, J.R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*,3(3), 417-457.
Seriès, P., Stocker, A., and Simoncelli, E. (2009). Is the Homunculus "aware" of sensory adaptation? *Neural Computation 21*(12), 1-33.
Shagrir, O. (2006). Why We View the Brain as A Computer. *Synthese,* 153, 393- 416.
Shannon C. (1948). The mathematical theory of communication. *Bell System Technical Journal* 27,379–423.
Simoncelli, E.P., & Olshausen, B.A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193-1216.
Spelke, E.S., & Kinzler, K.D. (2007). Core knowledge. *Developmental Science,* 10, 89-96.
Spelke, E. S. & G. Van de Walle (1993). Perceiving and reasoning about objects: Insights from infants. In N. Eilan, B. Brewer and R. McCarthy (Eds.), *Spatial representation* (pp. 132-161). London: Blackwell.

Sutton, R., & Barto, A.. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.
Teich, A., & Qian, N. (2003). Learning and adaptation in a recurrent model of v1 orientation selectivity. *Journal of Neurophysiology*, 89, 2086–2100.