SPECIAL ISSUE

FUNDAMENTAL PRINCIPLES AND MECHANISMS OF THE CONSCIOUS SELF

Alexei V. Samsonovich[1] **and Lynn Nadel**[2]

([1]Krasnow Institute for Advanced Study, George Mason University, Fairfax, VA, USA; [2]Department of Psychology, University of Arizona, Tucson, AZ, USA)

ABSTRACT

We start by assuming that the self is implemented in the brain as a functional unit, with a definite set of properties. We deduce the fundamental properties of the self from an analysis of neurological disorders and from introspection. We formulate a functionalist concept of the self based on these properties reduced to constraints. We use the formalism of schemas in our functionalist analysis, i.e. – a symbolic level description of brain dynamics. We then reformulate the functionalist model at a connectionist level and address the emergent "context shifting" problem. We suggest how the model might be mapped onto the functional neuroanatomy of the brain, and how it could be used to give an account of a range of neurological disorders, including hippocampal amnesia, various forms of schizophrenia, multiple personality, autism, PTSD, hemineglect, and reversible anosognosia. Finally, we briefly discuss future perspectives and possible applications of computer implementations of the model.

Key words: episodic memory, hippocampus, contextual reinstatement

INTRODUCTION

Neuropsychologists have found a certain class of phenomena hard to account for within the traditional frameworks of cognitive psychology. These include episodic memory, self-awareness, dreams (day dreams and night dreams), personality, value systems, the subjective experience of emotions, Theory-of-Mind (the human ability to understand other minds), various forms of higher-order and meta-cognition, and finally, consciousness and free will. All these problems appear to have one thing in common, namely, their relation to the concept of the self.

In addition to problems related to understanding these normal mental states, there is a counterpart set of problems related to neurological disorders: how can one characterize the nature of schizophrenia and related delusions and hallucinations (e.g., delusion of passivity), multiple personality disorders, the sort of personality observed in a split-brain case, post-traumatic stress disorder (PTSD), autism, hemineglect, reversible anosognosia, and so on? It appears that these disorders also have something in common, namely, that they can be viewed as "agency disorders" (Frith, 1996; Frith and Gallagher, 2002). Thus, these disorders also have something to do with the notion of the self.

These considerations suggest that a clarification of the concept of "self" would be very useful. In what follows we will suggest that a relatively simple model based on multiple, simultaneously active brain representations of the "I" (e.g., I-Now, I-Previous, I-Next, I-Yesterday, I-Future, I-Imaginary, I-Goal, I-Metacognitive, I-Pretended, etc.) may better explain the entire complex of foregoing cognitive phenomena than models which either ignore the "I" or offer a rather primitive account of it, based on a single mental perspective of a subject. Therefore, the aim of this work is to present a model for how the brain generates these multiple instances of the "I", or the self. But, what exactly is the self?

There are many ways to understand the self, or the "I". Usually when scientists talk about the self in relation to a human subject, they mean either the identity of the individual, as opposed to others, or the internal image of the body, as opposed to the rest of the world (Damasio, 1999). These basic, root, core notions of the self play the grounding role with respect to the more elaborate concepts of personhood and the like (Damasio, 1999), thereby supporting an idea of automatic, perhaps "hardwired" mechanisms of linkage of our experience to those higher-level concepts. The self, however, is seldom studied in the psychological literature at the higher level *per se*, as the agent-author of cognitive acts, or as the target of attribution of first-person experiences, or as the subject of the first-person experience. Nevertheless, we believe the latter three aspects – the agent, the target of attribution and the subject – constitute the main essence of the self-concept. This is the sense in which we use the word "self" in the present work.

Since the introduction of the concept of a "soul" by Descartes (1637), philosophers from Hume (1739/1965) to Dennett (1991) have asked if the self understood as the subject of experience actually exists? Hume's (1739/1965) point about the self is that he thinks we do not find any self in addition to the phenomenological contents we assign to the self. Yet this observation could characterize an ability to see things rather than the things *per se*. As pointed out by Kant (1781/1929), our intuitive

knowledge in general, and self-knowledge in particular, may have inherent limitations. Among the modern interpretations of the self-concept, most notable is Dennett's discussion (1991, Chapter 13) which gives us (along with other related writings) the now popular notion of the self as an illusory "narrative center" (i.e., the "main hero") of all subjective experiences and memories in a given individual. Blackmore (2002) pushes this line to its extreme, claiming that our self plus all our conscious experiences are illusory, i.e., they simply do not exist. Also relevant is Thomas Metzinger's (2003) book, in which the author claims that "No such things as selves exist in the world." (p. 626). Moreover, this conclusion appears to the author "a rather trivial one" (p. 626). The book is interesting, though, in that it also develops extensive connections between the discussion of the self and a range of neuropsychological deficits. Among those who advocate the existence of the self "as a thing" is Galen Strawson (1997). David Rosenthal's (2003) adaptation of his higher-order-thought model of consciousness to the unity of consciousness also stresses the difficulties of describing the evolution of our mental states without the self as a real, functional unit, yet offers a substitute (unconscious higher-order thoughts). This framework, as we shall see below, appears to be very close and relevant to ours.

At present the ontology of the self, together with the ontology of consciousness, stand together as a controversial problem in the phenomenology of mind (Chalmers, 1995, 1996, 2003; Tye, 2003; however, we should point that Chalmers' "Hard Problem" is not directly related to Hume's observation about the self). This situation is due to the fact that the "I" and consciousness are known to us only subjectively, from our private, first-person perspective (Searle, 1998). Nevertheless, together with Searle we believe that objective facts about selves can be established and studied based on an empirical scientific approach. We do not accept Searle's (1980) view that only a brain-like substance can instantiate consciousness and the self. We prefer Chalmer's (1995, 1996) view on this issue, assuming that the functional organization of the brain viewed as an information-processing system is all that practically matters for the self and consciousness, according to the *principle of organizational invariance*. From this point of view, consciousness and the self could be implemented in a computer, in analogy with their implementation in the brain.

Our primary task here is to identify the fundamental principles that objectively allow for an implementation of this sort. Our secondary task will be to make a connection between these principles and the brain and to apply them to the analysis of the foregoing agency disorders, as well as of the normal state of consciousness. In addition, we discuss the advantages of our approach based on the self-concept, considering it as a potential general computational method in cognitive modeling and in artificial intelligence. Given this agenda, we eschew here any focus on the ontology of the self, delegating this topic to our related work (Samsonovich and Ascoli, in press). Instead, starting from an epistemological analysis, we focus on the cognitive-psychological, functionalist description of the self. Stated differently, we intend to sidestep any question of what kind of *thing*, if anything, the self is, and to concentrate on what features of a psychological system we have in mind when we describe that system as having a self. While making this choice, we are aware that the subjective features we take as constitutive of a psychological system that can be appropriately described as having a self may well be illusory. The system may seem to have these features, even though it does not actually have them. This sort of analysis, however, lies beyond the scope of the present work.

We start by making the following basic assumptions. (i) We assume that the self as a functional unit is implemented in the brain neurophysiologically. We shall use the term *substratum* to refer to the set of physiological properties (e.g., patterns of neuronal activity) that directly contribute to the functional organization associated with the self. (ii) We assume that multiple instances of this unit, the self, may be present and co-active in one and the same brain, at the same time. Although this assumption may be intuitively hard to accept, it is consistent with the neurophysiological and behavioral data, as will be explained below. (iii) The concept of the "I" (a token-self concept, not to be confused with the linguistic concept of the personal pronoun) can be introduced as a functional unit that is directly associated with conscious awareness, sensory perception and voluntary control in a given individual. In particular, in this view consciousness is a property of the current instance of the "I", and no conscious experience can exist without the "I".

So far attempts to find the substratum of the self in the brain have failed. Since Descartes (1637), there have been few serious proposals associating a particular part of the brain with the self. At present there is no precisely known physical subsystem or process in the brain corresponding to the self, as a functional unit. One reason for this is the lack of a clear scientific concept of the self. Historically cognitive psychology was essentially based on "elimination" of scientifically "inappropriate" mental concepts, such as free will, qualia, mind, thought, consciousness, subjective feeling, etc. This was accomplished either by neglecting them entirely or by "translating" them into concepts acceptable from a neuroscientific or contemporary cognitive-psychological point of view (Churchland, 1988). For comparison, the term "memory" only became accepted in science in the middle of the last

TABLE I
Normal Features of Consciousness

	Fundamental subjective feeling of a conscious subject revealed by introspection	Corresponding subjective feature of consciousness in a third-person view	Example(s) of a consequence of its absence in the subject's first-person perspective	Example in which it persists while being objectively, evidently false
1	"I exist as a conscious being.[1]"	Existence of 'I' as the subject of awareness	Zombie	Detailed study of brain dynamics by the subject
2	"My 'I' has no internal parts or mechanisms, nor can it be divided into separable parts."	Atomic nature of the conscious self (or the "I")	Dissociative states	Split-brain double-consciousness (Mark,1996)
3	"There is only one real 'me'. I am one and the same 'me' in all my diverse senses, feelings, thoughts, actions and memories; continuously throughout my entire life."	Uniqueness and integrity of the conscious self (or the "I")	Dissociative states, multiple personality	Hidden observer
4	"There is only one true awareness: my awareness. It always reflects my actual, genuine present experience, is unique and not transferable."	Privacy, authenticity, uniqueness, nontransferability, infallibility of awareness	Paranoid schizophrenia	False experience induced by hypnosis
5	"There is only one true memory: my memory.[2] It is unique and not transferable, including my general knowledge about the world and my episodic memory that always reflects my past experience."	Privacy, authenticity, uniqueness, nontransferability, infallibility of memory	Paranoid schizophrenia	False memories
6	"I possess a 'free will': when I imagine things, make decisions, or initiate voluntary actions, what drives my process of thinking is myself."	Free will of the conscious self (or the "I")	Hypnosis	Libet's (1985) experiment
7	"I can only exist at a definite location, at a particular time, under a definite set of circumstances. I can imagine myself being anywhere, however, even in my imagination I always find myself in one (sometimes implausible or impossible) situation, and never in a superposition of two different situations at once."	Self-localization of the conscious self (or the "I") in one particular context	Sleep dreams, hallucinatory states	Dissociation between visual and kinesthetic modalities in the spatial domain
8	"In a definite sense, I always think and behave consistently with myself, including continuity of my thoughts, their consistency with my present and past experience and behavior, consistency of awareness across my modalities, and consistency of my voluntary actions with my conscious intentions."	Consistency of awareness with personal construct, its self-consistency across modalities and over time, consistency of awareness with episodic memory, with sensory input and behavior	Bipolar and schizophrenic disorders	Anosognosia
9	"By analogy with the above feelings about myself, I can understand other conscious beings, but I can never take their awareness as a genuine awareness, and their memory as a genuine memory. Therefore (see also 3, 4, 5), I am the unique representative of the kind; I am the only one true 'I'."	Social autonomy and unique self-identity of the conscious individual; tacit substitution of other conscious beings with their limited models conceptually distinct from the subject's self-model	Co-dependence	Wrong social image

[1] Here "I" should be understood as the subject of awareness, as opposed to "I as a physical entity".

[2] Each reference to memory should be understood as a reference to explicit memory, wich by definition is available for conscious retrieval.

century. In this work we hope to show that other "inappropriate" mental concepts such as the self also can be given precise scientific meaning and prove useful in objective empirical studies.

Functionalist Self-Concept

Assuming the existence of the substratum of the self in the brain with a specific functional role assigned to it, we can start by considering its functional role from an objective point of view, for which we need a functionalist concept. As a functionalist minimum, one can think of the self as a "black box" with a certain set of properties that characterize its functions. In order to clearly formulate these properties, we need to take into account the following. First, there is the entire complex of unexplained phenomena associated with normal states of consciousness we wish to understand. These phenomena are captured by the fundamental feelings of a normal conscious mind

revealed by introspection (Table I, left column). Although the subject may not be continuously aware of these subjective feelings and may learn that some or all of them are in error (see below), they characterize persistent beliefs of the cognitive system about its own normal consciousness as it is experienced from moment to moment, day to day.

A second source for seeking a deeper understanding of the normal self concept is the wide range of psychological conditions associated with abnormal states of consciousness, each of which stands unexplained at present, because we lack insight into the normal condition, and each of which might help us understand that condition if looked at from the right perspective. Consider the fact that many of the normal features of consciousness (Table I, left column) can be violated, leading to abnormal states (Table I, 3rd column). For example, feature n. 1 (existence as a conscious being) may not be present in a zombie state; feature n. 3 (integrity of the individual) is absent in multiple personality disorder; features n. 2 (indivisibility), n. 3 (integrity), n. 4 (non-transferability) and n. 7 (space-time constraints) can dissolve in sleep dreaming; feature n. 6 (free will) disappears under hypnosis; feature n. 8 (consistency) is violated in bipolar personality and schizophrenia; feature n. 9 (social autonomy) is inoperative in co-dependence, and so on.

On the other hand, most of the above properties may persist as subjective feelings even when their fallacy is made objectively evident and explicit (Table I, right column). Thus, feature n. 2 was objectively false in a split-brain, double-consciousness case when the two co-existing in one brain subjects of experience, with different beliefs, did not realize what was wrong, being totally unaware of their multiplicity (Mark, 1996); features n. 4, n. 5 (infallibility, authenticity of awareness and memory) and n. 8 become objectively false in cases of false experience and false memories induced by hypnosis (Beahrs, 1983); features n. 7 and n. 8 become objectively false in virtual reality experiments that induce dissociation between visual and kinesthetic modalities in the spatial domain, while the feeling of unity persists; feature n. 8 becomes objectively false in transient anosognosia (e.g., Ramachandran, 1995) and in phantom limb phenomena (e.g., McGonigle, 2002), and feature n. 6 in some experiments conducted by Libet (1985), when a patient explained that he pushed the button after being deceived by the experimenter's trick.

The fact that objectively false or "wrong" feelings can be *subjectively* experienced as correct or "right" feelings, raises a question: are these cases fundamentally different? Or, could it be that in these special cases and in general, under "normal conditions" our basic subjective feelings about self do not reflect some ineluctable fact of nature, but rather are fundamental errors, consequences of specific psychological mechanisms that are hidden from our awareness? If this is the case, then our task is to find out exactly what those mechanisms are.

As we argue below, these mechanisms consist in certain constraints that work within the brain. Borrowing a term from Golosovker (1936/1987), we will call these constraints *Error Fundamentalis* (in Golosovker's work this term refers to pre-assigned outcomes of acts of imagination involved in a process of myth creation; however, in our usage of this term, it is applicable to conscious cognition in general). Based on Table I, we formulate the Error Fundamentalis as follows below. This list constitutes our functionalist concept of the self.

It should be clearly understood that the Error Fundamentalis listed below are not ontological postulates about the self: they do not describe the self as it exists in nature! On the contrary, they characterize the self epistemologically, in that *they are the fundamental beliefs of the "I" about self* (these beliefs always remain in effect, even during times when the "I" is not conscious of them, or when the "I" theoretically speculates, as we do now, that some or all of these beliefs are errors). Stated differently, these constraints apply to brain representations of the self and its properties rather than to the self or the "I" per se. The reader should make no mistake here: they tell us what the system represents rather than what the system is! In this, and only in this sense, they tell us how this functional unit, the "I", or the self in general, is built into our mind and should be implemented in a cognitive-psychological model, or a computer. (Compare the following list to Table I):

Error Fundamentalis of the "I"

- EF n. 1: Any given experience instantiated in the brain has its subject, an instance of the "I" (the self of the given individual). This instance is characterized by its perspective and attitude (see definitions of these terms below).
- EF n. 2: The "I" is an autonomous, atomic entity, in that it is indivisible, with no apparent internal structure or discernable intrinsic mechanisms, independent of any external agent in its controlled actions.
- EF n. 3: The "I" possesses a unique identity. There is only one instance of the "I" at each moment of time. The "I" is consistently one and the same entity in all its diverse experiences through its entire life. Its existence is continuous in time.
- EF n. 4: All current experiences attributed to the current instance of the "I" are immediately and directly available to that instance. There is no conscious experience in the given individual that is not attributed to the current instance of the "I".
- EF n. 5: The past experiences associated with past instances of the "I" are potentially available via retrieval of these instances together with their

original, previously experienced mental perspectives (episodic memory retrieval).

• EF n. 6: The cause of all currently happening behavioral and cognitive actions of the system that are consistent with the current working scenario (see below) is associated with the currently active instance of the "I". In other words, these events are interpreted as voluntary actions of the "I".

• EF n. 7: Each instance of the "I" is characterized by a proper subjective time and is associated with one particular context, typically including a definite location in space.

• EF n. 8: The content of the current experience of the "I" is internally consistent, in terms of the semantic knowledge of the given individual about the world and about self, including logic and common sense. Accordingly, the content of the stream of consciousness and episodic memory of the "I" (which consists of a sequence of causally related mental perspectives attributed to the "I") is also internally consistent in the same way.

• EF n. 9: The "I" is aware of its presence in its current perspective and at other moments of time. The "I" is aware of its identity, in that it is distinct from others. Generally, the "I" is aware (not necessarily consciously aware) of all the above properties 1-9 taken for granted by the "I" as ineluctable facts of nature.

How can we understand these properties within a functionalist framework applied to the brain? As mentioned above, one approach is to note that they can be implemented in a model as constraints. That is, they imply certain semantical constraints imposed on all possible narratives about the self, and the system of representations of all present and past experiences instantiated in the brain must be constrained in a similar way. Insofar as the brain can be considered as a dynamical system, the Error Fundamentalis can be implemented as dynamical constraints. They are "errors" in the sense that they cannot be taken as literally true facts of the physical world, from an objective point of view (for example, from an objective point of view, the number of subjects of experience and their identities may be fundamentally indeterminate: Parfit, 1984; see also our related work: Samsonovich and Ascoli, in press). We nevertheless view the Error Fundamentalis as true constraints on the intrinsic sementics of representations, understood in terms of a cognitive-psychological model applied to the brain.

For example, here is how we interpret EF n. 9, which says, in particular, that the self is aware of EF n. 1-EF n. 9. This means that, objectively, the self represented in the system must behave as if it had *strong beliefs* EF n. 1-EF n. 9 about itself (again, meaning "my beliefs about the world" rather than "my beliefs about constraints that predetermine all my possible thoughts and actions"). This in turn implies, e.g., that acts of reasoning about self attributed to the "I" (including self-analysis, self-awareness, introspection) must conform to EF n. 1-EF n. 9 in order to be instantiated in the brain. Ultimately these constraints reflect hard-wired properties of the normal brain, though they of course undergo epigenetic development and are hence susceptible to disruption by abnormal conditions of either a genetic or experiential nature.

Finally, here we explain the meaning of several terms as they are understood in this text. The term "perspective" refers to the instance of the self: e.g., a first-person perspective of the subject, as opposed to a third-person perspective. Generally speaking, the *perspective* of a subject can be uniquely specified by a short list of determining parameters. These parameters include the personal identity (e.g., I vs. John vs. Mary) and the proper time of the subject (e.g., me now, at 11:31 AM, vs. me yesterday, at 2:45 PM), as well as the status of the given instance of the subject (e.g., me actual vs. me imagined vs. me as I remember myself from yesterday's dream vs. me as I pretend to be, etc.) and the position in a Theory-of-Mind hierarchy (e.g., me vs. my view of John's belief about me), etc. According to our point of view, multiple perspectives of these kinds together with their associated contents may be co-active in the same brain at any moment of time.

The term "attitude" here refers to the kind of mental state, i.e., to the kind of mental position of the subject with respect to the content of experience, and it can be different for each element of awareness (e.g., Panzarasa et al., 2002); however, the notion of the current overall attitude of the subject also makes sense, via the notion of the focus of attention. For example, if I am conscious of my present condition and at the same time I am focused on thinking about the past, then my attitude is "past", although my perspective is "now". If, on the other hand, I re-experience my episodic memory, then I put my self into the perspective of my "I-Past". Within that perspective, my attitude with respect to the retrieved experience could be "now", although the experience could belong to yesterday.

Attitudes and perspectives have a lot in common and may even be considered as elements selected from one and the same set of possibilities. The difference between these notions becomes clear, when one uses a spatial metaphor. In this case, the perspective is given by the location of the subject, and the attitude is given by the location of the target of attention of the subject. Another notion that is easy to capture with this spatial metaphor is the distinction between allocentric and egocentric representations that we will need below. Traditionally defined within the spatial domain only, this notion is easy to extend to the temporal domain (Samsonovich and Ascoli, 2002), as well as to other cognitive dimensions (including the Theory-of-Mind mental perspective hierarchy: Langdon and Coltheart, 2001; Vogeley and Fink, 2003). In this case, "allocentric" means "in coordinates or notations not related to the subject",

whereas “egocentric” means “in a coordinate system centered at the subject's perspective”.

The notion of time, as it is used here in most cases, is understood as *subjective time*, i.e., the time subjectively perceived by the self, which from a functionalist point of view means nothing but the time stamp of the personal experience. Generally, subjective time of the current experience may differ from the physical timing of the related physiological processes in the substratum at a short time scale, up to 500 milliseconds (Libet et al., 1979).

The last essential notion that we need to explain is that of a *schema*. The term “schema” (plural “schemata” or “schemas”) was introduced by Kant (1781/1929), and currently has an extremely broad usage in science, with different semantics in different fields: from history, philosophy and linguistics to cognitive psychology (e.g., Cheng and Holyoak, 1985; Iran-Nejad and Winsler, 2000) and neuroscience (e.g., Arbib et al., 1998). Within computer science alone the word has perhaps a dozen different senses. The most advanced up-to-date mathematical theory of schemas exists in the field of evolutionary computation (Langdon and Poli, 2002; De Jong, 2005). The approach taken in the present work does not rely on any of the foregoing specific interpretations of the notion of a schema, or their related schema theories. While common points and conceptual overlap between our and the foregoing frameworks are inevitable, here the understanding of the term “schema” should not be confused with, or biased by the above frameworks.

In the present work, the notion of a schema is understood in a very general sense applicable to cognition. We view schemas as any-level abstractions that may have neurophysiological implementations in the brain. Specifically, *we define a schema as an abstract model or a template that can be used to instantiate and process a certain class of mental categories*. These categories include all possible types of elements of the subjective world: concepts (e.g., objects, properties, events, relations), feelings, sensations (qualia), intentions, cognitive and behavioral actions, etc. Any complex or product of schemas can be viewed as a schema on its own. We will generally say that a schema has a *state*, when its instance is bound to some external content. For example, perception of a red circle can be described via a state of the schema of red bound to other content, e.g., to a state of the schema of a circle. Logical reasoning can be described in terms of states of the schemas of inference, etc. In our terminology, a state is considered *mental* (not necessarily conscious), if it is attributed to a subject of experience. In addition, this attribution modifies the semantics of the word “state” in this case, which now refers to a state of the subject, in addition to a state of the schema. In our framework, schemas are dynamical objects: they can be created and modified. The entire set of schemas in a given individual constitute that individual’s *semantic memory*. In a simple version of the model we can assume that all schemas are kept together in one pool and are available for usage at all times. We assume that schemas, at a level of their definition, take into account (and thereby implement) the fundamental constraints noted earlier: the Error Fundamentalis. In particular, the dynamics of a system of mental states, as determined by schemas, generally will conform to Error Fundamentalis.

Conceptual Framework in Action

How might this work? Assume that one is looking at a tree, and one’s brain has a representation of this experience: “I see a tree”. In fact, the actual content of the experienced state could just be “there is a tree”, without the self-awareness component (it is only when one is conscious of oneself as seeing a tree that one comes to have some mental state with content about oneself and one’s seeing). In both cases, the “I” must be attached to this raw experience as a label, in order to signify that it is an experience consciously or automatically attributed to “myself”. Alternatively, if I think of somebody else seeing this tree, the “labeling” would be different: it should signify “he” or “she” instead of “I”. Thus, the entire complex involves two components: one representing the content of the experience of which the subject is aware, and the other labeling the subject who is experiencing this content (even though the subject may not be aware of self at the moment). Again, we call this complex a *mental state*.

We can imagine complex structures composed of mental states in the following way: Assume one has a set of experience at a given moment of time (Figure 1, green): “ I see a cherry tree”, “I walk”, “I am entering a garden”, “I recall yesterday’s party: cherries in cocktail”. This could be an approximate snapshot (mental simulation) of my current state of awareness, which quickly changes to another snapshot: “ I hear birds”, “I am in a garden”, “I am thinking about yesterdays’s episode”. This condition in turn may evolve in its time frame, causing my re-experience of the retrieved episode from its “native perspective” I-Past, which in turn my force me-present (I-Now) to think about my future plans and then to focus on them at the next moment of time (I-Next), and so on. This picture (Figure 1) is completed by linking together all my experiences and related instance of my self labeled in a self-explanatory manner, as follow: I-Previous, unifying all that I was aware of just a moment ago; I-Next, unifying my expected state of awareness at the next moment of time; I-Past, pointing to a set of mental states that once happened together to constitute a first-person experience for me in the past; I-Future;

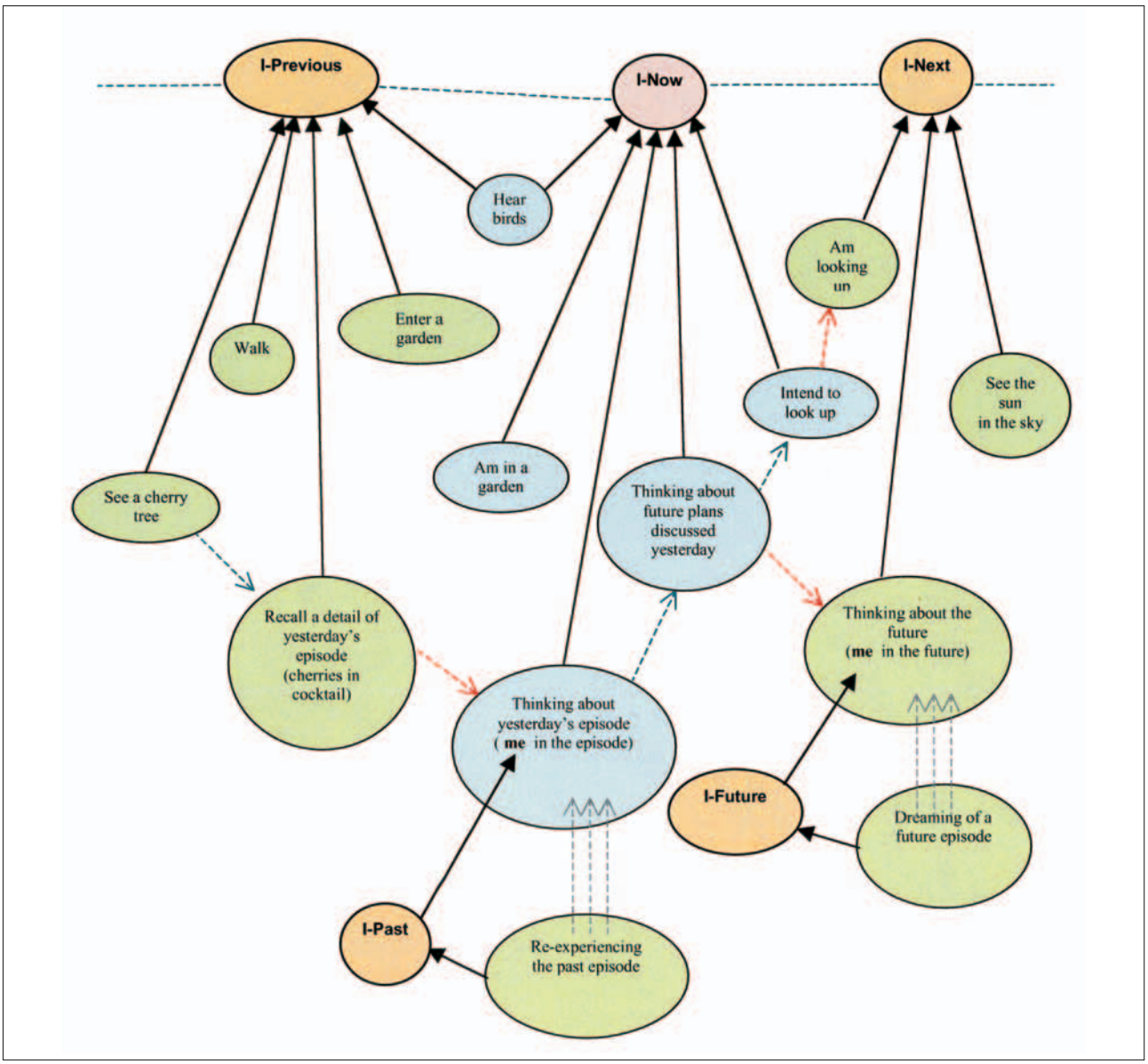

Fig. 1 – *A possible fragment of a snapshot of a state of working memory, viewed as consisting of mental states. The labels "I-Now", "I-Previous", "I-Next", "I-Past", "I-Future" taken together with their links refer to the substrata of the subject (the "I") of the corresponding mental perspectives: like the perspective of my (first person) now, or the perspective that I had a moment ago, or that I am going to take next. Other labels and links instantiate states of schemas, as described in the text. Different colors mark "snapshots of awareness" (mental simulations) of different instances of the "I".*

etc. Although these states may refer to various moments of time, we consider them as being simultaneously active in the brain. The time attributed to them is really a time stamp, or, subjective time, rather than the physical time of their manifestation in the brain (see the previous section). If we choose to call all mental state instantiations in the brain "memories", then we should have clear notions of a memory of a recent or remote past, of a memory of the present, and of a "memory of the future". In addition, there would be "memories" of things imagined, etc. All these "memories" can be realized as multiple instances of the subject, each with its own perspective, that are simultaneously active in the brain and interacting with each other, thus together constituting the content of "working memory". We are now in a position to ask how consciousness fits into this picture.

Because consciousness is, generally speaking, an experience attributed to the subject, the self, it must be defined in terms of the self and it's associated experiences. The self can be associated with experiences (i) as the subject-possessor of the experience and (ii) as a part of the experiential content. In our framework, (i) corresponds to the notion of consciousness, while (ii) corresponds to the notion of self-awareness. Therefore, these notions do not imply each other. Specifically, we propose that all active mental states in the brain, that directly include the label "I-Now", constitute consciousness (Figure 1, blue). Other active mental states (Figure 1, other colors) may influence the contents of consciousness, resulting in subjective

feelings of recall, intention, etc., but they do not constitute consciousness per se. Because thus defined conscious mental states are immediately available for introspection in the first-person perspective of the subject (I-Now), this proposal is consistent with most operational definitions of consciousness (see our discussion of essential indexicals below). One interesting possible alternative to this view will be discussed later.

This functional interpretation of the notion of consciousness is different from others found in the mainstream philosophical literature, probably the closest of which involves higher-order-thought theories (HOT) (Armstrong, 1980; Rosenthal, 1986, 2003; Lycan, 1996). In view of this analogy, we see the position of Rosenthal (2003) as the closest to ours, with our "self" playing the HOT role in Rosenthal's (2003) framework. Is the difference between the two points of view reducible to a difference in terminology? Probably not, because we do not interpret the self as a thought, but rather as a *thing* possessing definite characteristics (cf. Strawson, 1997; discussion of this issue is beyond the scope of the present work: see our related work, Samsonovich and Ascoli, 2003, in press). Again, here we eschew any focus on the ontology of this thing, merely assuming its indirect representations; constraints that we call Error Fundamentalis.

Indeed in this picture (Figure 1) the foregoing Error Fundamentalis can be considered as constraints that underlie certain relations among the representations attributed to different instances of the self (some of these relations are symbolically represented by dashed lines in the figure): they are states of relational schemas. The constraints are in effect because they are built into schemas that give rise to mental states and their relations. For example, consider the pair of mental states attributed to two consecutive instances of the self: I-Previous, "recall a detail of yesterday's episode", and I-Now, "think about yesterday's episode". When I recall a detail of yesterday's episode, then the properties of the psychological system of episodic memory make it likely that in the next moment I will be thinking about the episode itself – a detail serves as a retrieval cue to complete the neural activity pattern representing the entire episode.

This psychological property, which is instantiated in the neural dynamics of the memory system, embodies the causal schema of recall, which in turn conforms to the constraint that the self (and therefore mental actions attributed to it) must be consistent over time. Relations like these taken at any level of description (from the activity of single neurons, through patterns of activity of neural ensembles and their connections, through schemas and mental states, up to the Error Fundamentalis) are sensitive to the labels specifying the instances of the self, because these labels determine the subjective temporal relations among mental states and in addition signify that all these experiences belong to one and the same self. If these would be two unrelated selves, then there would be no explicit logical connection between the two cognitive acts (although the two acts still could be connected implicitly, by priming and interference mechanisms that work at the neurophysiological level).

Here the reader can ask, e.g., the following question. Consider a person named George; for that person, the label 'I' functions differently from the label 'George'. That person might, after all, not know that his name is 'George' (or that the current date is April 1st), and so he might well represent George's April 1st states as belonging to somebody other then himself (or to some other day than today). In virtue of what does "I-Now" function as the "current-self" label, as opposed to functioning as a label that simply happens to apply to some individual, taken at some moment of time (e.g., "George-April-1st")? In other words, why does this I-Now label, in contrast with other similar labels, define consciousness rather than another kind of memory of the past or a mental simulation? Neurophysiologically, the answer can be found in the unique status of the states labeled "I-Now", based on the role this mental perspective plays in all sensory and control functions of the brain of the given individual (I-Now has direct access to all of them, while other mental perspectives do not). Semantically, the mechanism uses what Perry (1979) calls the *essential indexical*: the label "I-Now" refers to the self in this essentially indexical way.

In summary, the laws of evolution of a dynamical system of mental states defined by schemas can be inferred from the Error Fundamentalis in conjunction with the laws of causality, common sense and logic. And vice versa: Error Fundamentalis together with the general laws of causality, common sense and logic limit possible schemas, which apply to each individual subjective world (i.e., each individual's semantic memory). The set of schemas in turn can be viewed together as a higher-level objective description of the laws of human brain dynamics.

As a next step, instances of the self together with their perspectives and with the attributed contents can be represented as separate *charts* with mental simulations developing on them in parallel (Figure 2). If one can think about the charts represented in Figure 2 as structures, then mental states are the elements of these structures. Each mental state is an instance of a schema bound to a particular chart and to the related content (formally speaking, all mental states that belong to one chart together can be viewed as one, *complete* mental state of the given instance of the self). As one can see, this framework comes from a generalization of the Theory-of-Mind framework in its simulationist version (Nichols and Stich, 2000, 2003): there are

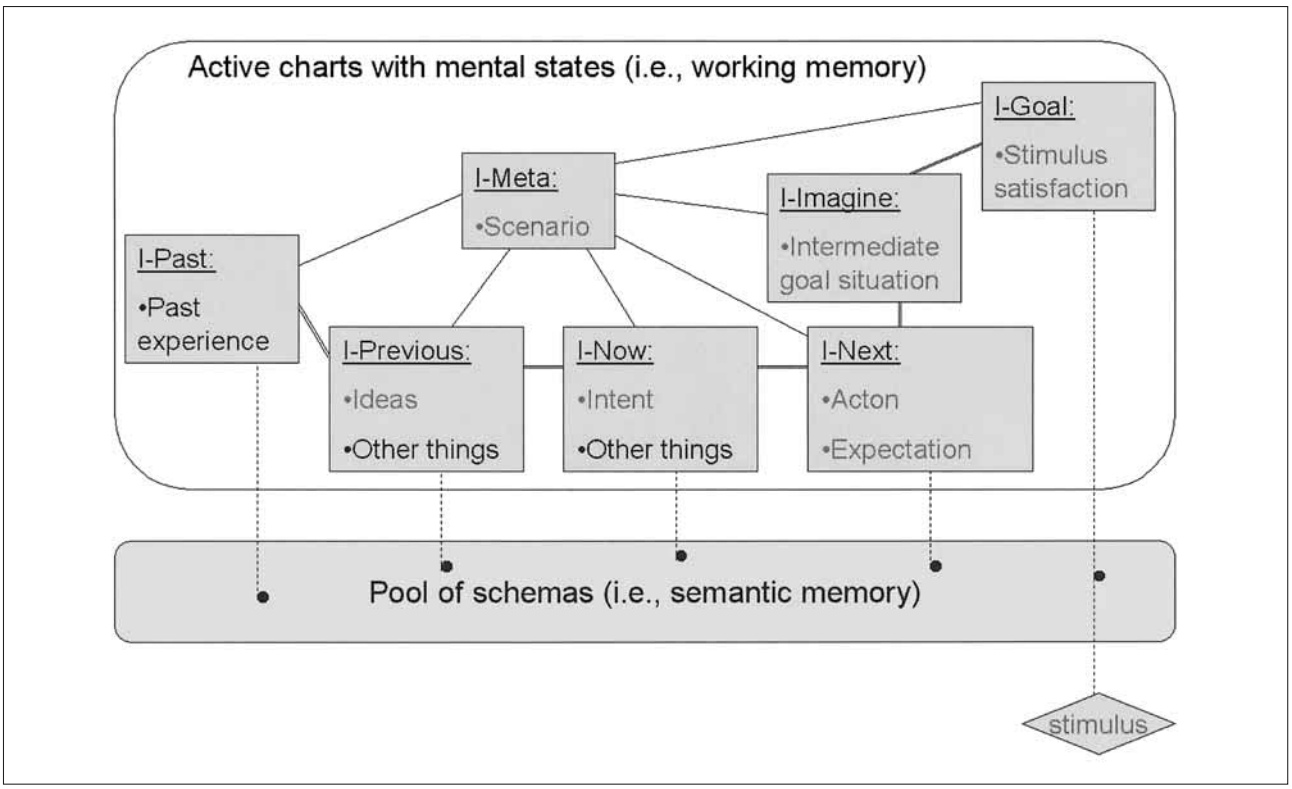

Fig. 2 – *The nature of voluntary action. The diagram represents a system of active charts that constitute the working scenario (connected by the double line, from I-Past through I-Now to I-Goal) and the underlying pool of schemas. An example of schema processing is shown by the rightmost vertical dashed line: the stimulus (a driver) activates a certain schema (e.g., the schema of eating food) and creates its state in I-Goal (if necessary, I-Goal is created too during this process). The chart I-Meta instantiates a metacognitive perspective of reasoning at a large time scale.*

hierarchical relations among perspectives (charts), and the contents of these charts are mental states. At a larger scale there is some sort of organization among charts: e.g., I-Now and I-Previous are consistent with each other. There is a line that can be traced through the sequence of charts leading to some sort of a goal mental perspective: I-Goal, if this is a goal-directed behavior. There is also an extension of this line into the past, which makes it consistent with my previous states: I-Past. The "mainstream" sequence of charts bound by this line (the double solid line in Figure 2) constitute a *working scenario*. In addition to the working scenario, there may be a number of other charts present in working memory (not shown in Figure 2). The structure shown in Figure 2 is fluent, as well as in Figure 1. Therefore, this is not an architecture in the ordinary sense, but a possible snapshot of a dynamical state of the system. A snapshot taken at the next moment of time may show different components.

This high-level cognitive-psychological model can also be viewed as a general framework and an architecture underlying a possible computational approach to the study of human mind (Samsonovich and DeJong, 2003, 2004). As a computational framework, it will be presented elsewhere; however, one detail of the computational version of the model is worth mentioning here, as it may have counterparts in the brain implementation too. In addition to mental states, there may be states that do not belong to any chart and therefore are not mental. One kind of schema that gives rise to states of this sort we call "drivers" (Samsonovich and DeJong, 2003). They mediate automatic processing and interactions of mental states across charts. More generally, a *driver* in this computational framework is an active unit that processes schemas and their states. One example of a driver is shown in Figure 2. It is a simple kind of a driver: we call it a *stimulus* (such as hunger, for instance). This stimulus is bound to a schema (of eating food, in our current example), and its action is: (a) to create a mental state of the schema on some chart where the state would fit; if necessary, creating a new chart for this purpose; and (b) if this chart is not I-Now and not I-Next, then, to give this chart a goal status. The process could be more complicated in a human brain: e.g., the goal must be consistent with the system of values, current priorities, etc.

Now we can see how voluntary action emerges in this picture. First, we assume that the dynamics of mental states results in generating a coherent sequence of charts connecting I-Now to I-Goal, a working scenario. Given this, a voluntary action paradigm starts from the generation of a set of ideas that are mental states representing plausible actions in the current situation. Among them one is spontaneously selected as the intent (one of those that fits best into the working scenario) and is

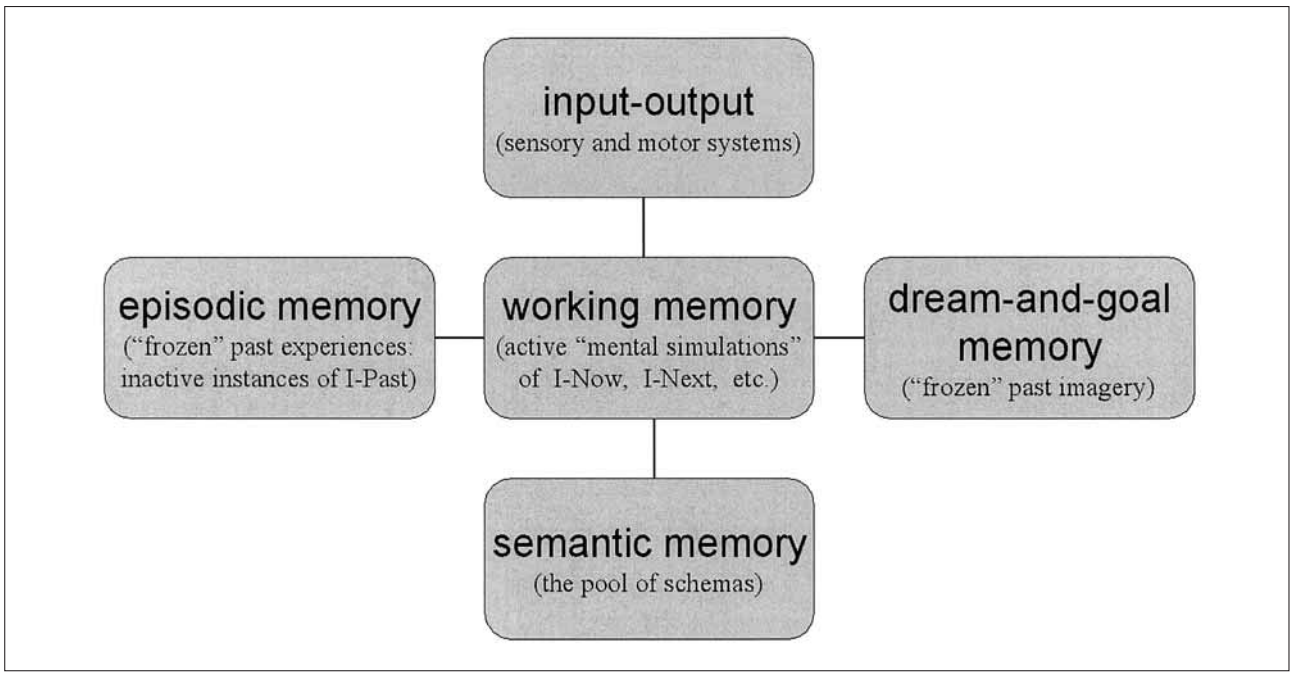

Fig. 3 – *The macro-architecture of the mind. The main components of the model are: working memory, episodic memory, dream-and-goal memory, semantic memory (the pool of schemas), and the interface module, which is responsible for sensory input and for behavioral output.*

scheduled for execution. It is executed at the appropriate moment of time, before which a prediction of the expected result (what will happen next and how it fits into the general strategy) is generated in I-Next. This is a general outline of the voluntary action paradigm. The mental states in Figure 2 signify that the author-subject who initiated the action is I-Now, and the agent who is going to complete this action is now I-Next (and will become I-Now at the next moment of time).

When expanding this view of the mind (Figure 2) to a larger scale (Figure 3), one would see a global picture consisting of 5 basic components: (1) working memory that corresponds to ongoing "mental simulations" (Figures 1 and 2); (2) episodic memory, which naturally is formed from states of working memory when some of them survive de-activation and become "frozen" for a long period of time (they also change their neurophysiological basis from neuronal activity to synaptic weights); (3) an interface component, including sensory input and behavioral output, dependent on a collection of brain systems; (4) a counterpart of episodic memory representing a system of values, or the memory of dreams, plans and goals (in principle, this kind of memory could be included in a generalized notion of episodic memory viewed as a memory of personal experience of any kind: e.g., a dreamed experience, with an appropriate perspective label); and (5) semantic memory, as explained above, consisting of schemas. Brain implementation of schemas could be conceived as stored spatiotemporal patterns of neuronal activity, resulting from learning in neuronal assemblies.

Before concluding this section, we need to emphasize the following. First, we propose that our Error Fundamentalis, traditionally viewed as the putatively "emergent" properties of the self, in fact are imposed a priori as constraints on schemas underlying brain dynamics. The idea is that one might be able to make progress in understanding brain dynamics by basing a theory on the Error Fundamentalis postulated as constraints and then trying to infer other properties of the conscious self, starting from these principles. We take this approach in the present work. In doing this, we do not *stipulate* the emergence of these fundamental properties of the self from the complexity of brain dynamics; on the contrary, we *postulate* them as initial constraints rather than as emergent properties. The question of how these constraints could arise through brain evolution (phylogeny) and its epigenetic development (ontogeny) is beyond the scope of the present work. The question of how they are given in this model has a simple answer. It should be clear from the above that Error Fundamentalis, together with the general laws of causality, common sense and logic, must be represented by schemas. Therefore, by associating each mental state with an instance of a self (i.e., placing it into a chart) and by allowing only those configurations of mental states that conform to the schemas, one automatically satisfies Error Fundamentalis. This is exactly what one should expect to happen in model dynamics, because these dynamics are governed by the schemas. How this mechanism could be implemented in a connectionist framework is our next topic of discussion. Finally, we emphasize

again that our "trick" with a functionalist "reduction" of the self to the Error Fundamentalis is not intended to address the ontological problem of the subjective existence of the self, which stands separate from our present agenda. Nevertheless, from the above analysis we can derive one conclusion relative to this topic: at least in the kind of cognitive system that we considered so far, *implementation of Error Fundamentalis via schemas is a precondition for the emergence of a conscious self together with its first-person subjective experience*, regardless of the nature of these things: real, virtual, illusory, etc. In support of this claim, in the section "Application to Neurological Syndromes…" we consider several cases when this precondition is violated.

Connectionist and Brain Implementation

Can this functionalist model be implemented in a connectionist network? Generally speaking, attributing representations to particular instances of the self can be implemented by linking certain nodes in a network. If there are nodes representing instances of the self, then they could be linked to (associated with) patterns representing experiences. So far so good, but a serious problem emerges with this approach that we call the "context shifting" problem. As time flows, what used to be I-Now becomes I-Previous, what used to be I-Previous, becomes I-Past, and so on. It is not easy within a connectionist network to switch the links formed previously, let's say, by Hebbian associative learning, and to transform them into links among different subsets of nodes (Fuster, 2003). How can this critical step be accomplished?

The problem can be explicated intuitively from a different perspective. How can something we recall be attributed to the past, given that it is experienced in the present at the time of the recall (and, presumably, also at the time of memory creation)? In other words, how can we experience something (specifically, a sense of a past experience) that matches neither what we actually experienced and memorized, nor what is happening at the moment?

One of the main features that discriminates between episodic and semantic memory, as noted by Tulving (1985, 2002), is whether or not memory is attributed to a particular source: the subject-agent of experience. In the case of episodic memory we attribute the memory to our previous self (Wheeler et al., 1997). In the case of semantic memory, like a memory that "$2 \times 2 = 4$", we remember facts without any attribution. This is related to the problem of context shifting. When we recall semantic knowledge and apply it to a current content, it is retrieved into working memory, and the mental state of awareness is formed by associating it with I-Now directly and through other content currently associated with I-Now. There is no problem in this case. When, however, we recall an episodic memory, we are retrieving a previously active mental state. Upon reactivation, if it is to be experienced as a memory of a previously experienced event, rather than a current event, then it must be referenced in my consciousness (the I-Now perspective) as a system of mental states bound to I-Past (see Figure 1). But the problem is that this previous experience was not associated with the label I-Past. It was associated with the label I-Now, which of course has been associated with many other memories in the interim. How can this problem be solved?

The problem can be solved, we suggest, by introducing two sorts of memory indexing maps: allocentric and egocentric (see above). In all considerations up to this point, the notions of now, past and future were supposed to be represented by the same nodes over time. The same labels (I-Now, I-Past, etc.) used in our above description could be understood as the same nodes used in a connectionist model, or the same substrata used in the brain. This sort of labeling can be called *egocentric*, analogous to the use of this term in the spatial domain: it is dependent on the current sense of now (see above). On the other hand, memory indexing could be independent of the current sense of now, e.g., "I-2004-December 8", "I-2005-January-2nd", and so on. We call this kind of label allocentric (Figure 4). We need egocentric indexing, because the functioning of the agent is critically dependent on the essential indexicals (see above) via the egocentric map. The "I" must understand what is happening now, what has just happened before, and what is going to happen in the next moment, in order to be able to plan and to execute behavioral actions. On the other hand, as we saw, the allocentric map is necessary in order to keep track of all experiences and to be able to use rapidly created associations for retrieval in the future.

A solution to the context shifting problem based on the two maps can be imagined as follows (Figure 4). An experience I have is associated with two pointers (they constitute a substratum of my "I": one could think of them as patterns of active neuronal units) activated on the two maps: an egocentric pointer I-Now and an allocentric pointer I-2005-May31-1:10-PM. At the next moment of time, I have another experience, which is also associated with a couple of pointers (another substratum). The active pointer on the allocentric map is now different from what it was just a moment ago: e.g., it will be I-2005-May31-1:11-PM, while the pointer on the egocentric map, I-Now, is the same (Figure 4). Recalling what was being thought a moment ago involves inhibiting sensory perception and the sense of now, and simultaneously activating the sense of the previous moment of time: I-Previous. Assuming a "synchronization mechanism" for the two maps,

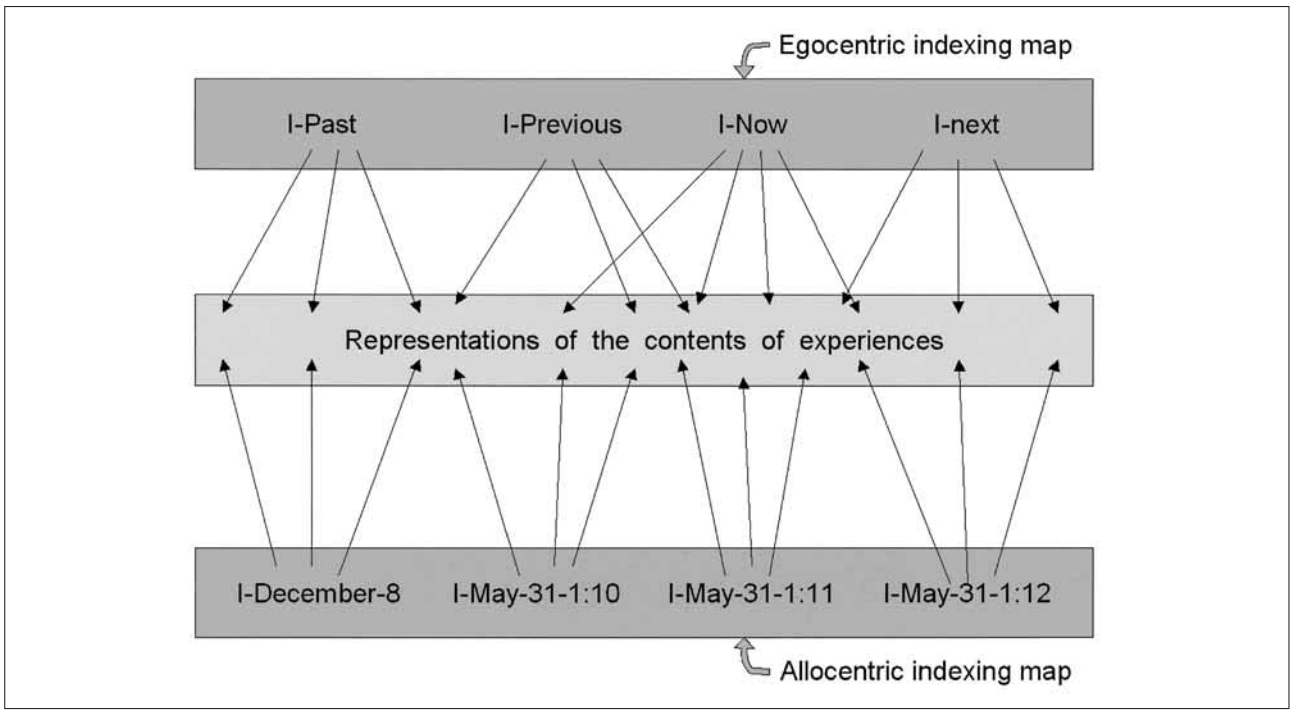

Fig. 4 – *Allocentric and egocentric memory indexing maps. The two maps "slide" against each other, as the time flows. I-Past, I-Previous, I-Now, I-Next are egocentric pointers, while I-December-8, I-May-31-1:10, I-May-31-1:11, I-May-31-1:12 are allocentric pointers to the representations of the contents of experiences.*

which we will not discuss here (e.g., in the analogous case of rodent spatial navigation it amounts to the so-called "path integration": McNaughton et al., 1996), this process will result in reactivation of the pointer I-2005-May31-1:10-PM on the allocentric map that was previously active. This reactivated allocentric pointer will in turn re-activate the ensemble of mental states previously associated with it. Now they will be co-active with I-Previous, and thus the necessary context shift will occur. This mechanism viewed at an abstract level is consistent with the multiple trace theory of memory consolidation (Nadel and Moscovitch, 1997; Moscovitch and Nadel, 1998; Nadel et al., 2000) and provides further insights into the latter.

How could this model be implemented in the brain? One possibility (Figure 5) is that the egocentric map is instantiated in some part of the prefrontal cortex (PFC: perhaps the dorsolateral and/or medial PFC). For comparison, rodent prefrontal cortex is known to be critically involved in spatially egocentric tasks (Kesner et al., 1989; Raggozino and Kesner, 2001; Ethier et al., 2001). And the allocentric map could be instantiated in the hippocampus. Indeed, there are representations in the rodent hippocampus known to be spatially allocentric (O'Keefe and Nadel, 1978). Consistently with our model, it has been shown by experiments with rodents that the two memory indexing maps, PFC and HC, work in parallel and can even substitute each other during a short time window of several seconds (Lee and Kesner, 2003). From this point of view, the notion of charts as it is understood here could be related to the notion of multiple charts involved in spatial representations observed in the rodent hippocampus (Samsonovich and McNaughton, 1997; Samsonovich, 1998). Thus, a substratum of the "I" could include at least two components: an activity pattern in PFC plus an activity pattern in the hippocampus. Other structures, however, may be involved as well, including the parietal cortex, the insula, and the anterior cingulate cortex (activity in these structures seems to correlate with the sense of self: Farrer and Frith, 2002; Kampe et al., 2003).

Anatomically, there is no direct, monosynaptic link between the hippocampus and PFC, but they are connected via the entorhinal cortex (EC), which is part of the medial temporal lobe (MTL). Patterns of activity in MTL are known not to be very selective with respect to episodes, while exhibiting spatial (Flynn et al., 2004) and other semantic selectivity. In our model, the elements of this pattern represent higher schemas that contribute to the contents of the charts by forming blocks of mental states. Therefore, MTL neural activity corresponds to active states of these schemas in our model view. They are linked to lower schemas in the neocortex (NC), where multimodal neocortical representations are instantiated. The latter activity consists of the set of representations of particular concepts, feelings, sensations, thoughts, intentions, etc.

At the end of this section, we would like to entertain one possibility that is suggested by the proposed neuroanatomical mapping of the model. Indeed this formalism permits an interesting scenario: there may be multiple substrata (and, accordingly, multiple _"subjects"_ of experience if

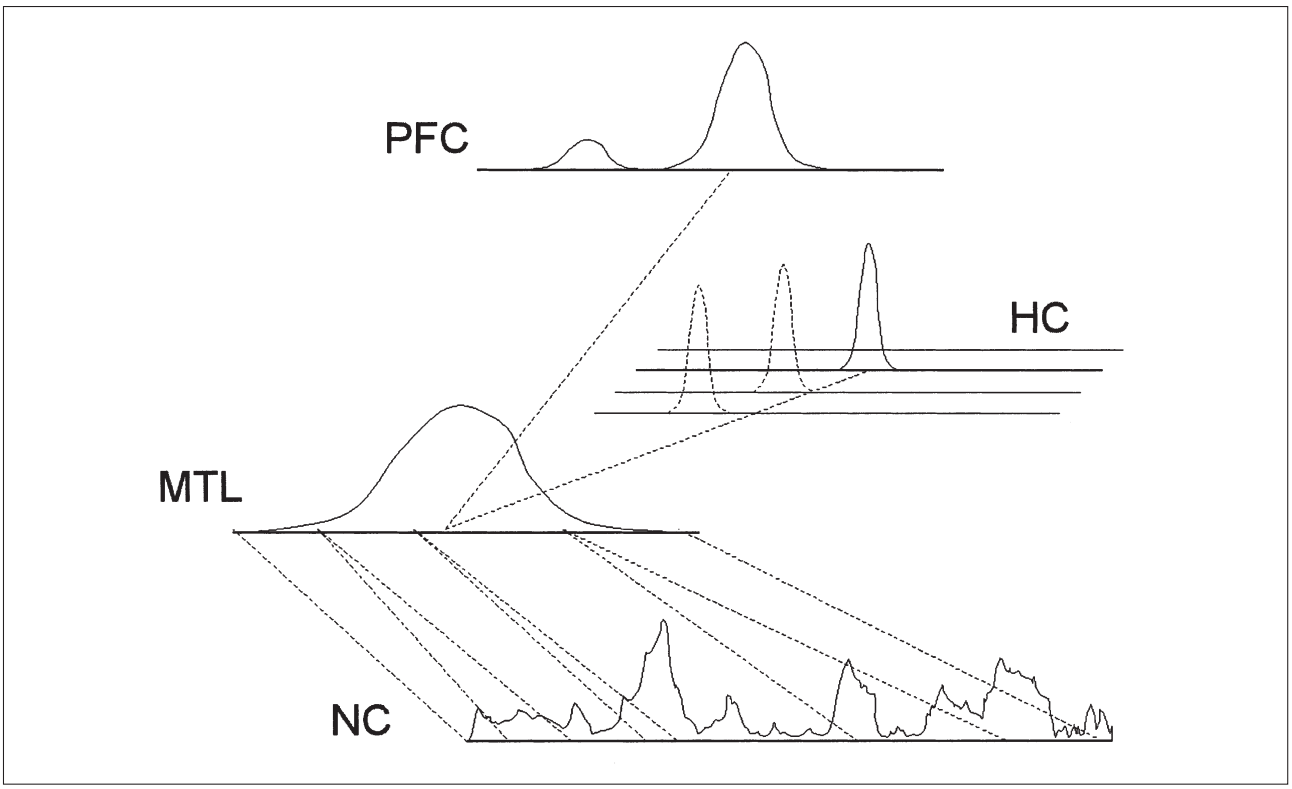

Fig. 5 – *A connectionist model of the memory systems. Its possible mapping onto the brain is given by the labels: PFC (the prefrontal cortices responsible for the egocentric substratum of the self), HC (the hippocampus responsible for the allocentric map), MTL (the medial temporal lobe, including the entorhinal cortex), and NC (the extrastriate neocortex). The horizontal axes represent arrays of neuronal units aligned by their preferred subjective time (as well as other cognitive dimensions in a generalized version of the model). The curves represent neuronal activity distributions. Hippocampal representations are shown as activity packets on hippocampal charts (the latter can be viewed as manifolds constructed of the same units by permutations: Samsonovich and McNaughton, 1997).*

any – see below) associated with one and the same perspective, such as I-Now. Modular organization of the brain provides room for this possibility, because the cross-modular information exchange is not fast enough to keep the parts of a mental simulation divided between modules perfectly synchronized with each other. For example, we can imagine a partition of the substratum of I-Now into several fragments - substrata, say, from I-Now-1 to I-Now-10, that functionally are relatively independent of each other. We assume that each of them may have a different focus of attention and therefore a different content of awareness, but otherwise their parameters are identical. From a functionalist point of view, this means a partition of the contents of consciousness into several fragments that may partially overlap. Each of these fragments is attributed to it's own _"subject"_ – it's own instance of I-Now. Because of the common identity of the _"subjects"_, common subjective time and other determining parameters, the schemas will not distinguish among them. In particular, the rules that determine their access to working memory representations will be identical. This implies that they will have one common working scenario and will perceive each other's actions as their own. This also implies that they will share the contents of working memory and the contents of long-term memory, but not necessarily share their momentary contents of conscious awareness. Assuming that there are no fixed boundaries of the partition, such as, e.g., anatomical division based on modalities or based on left and right hemispheres (in other words, assuming that _"subjects"_ can navigate and jump from one modulate to another), each of the _"subjects"_ will potentially have the same access to the entire content of the chart I-Now, to each instance of I-Past, and so on.

For example, at one moment in an episode of conscious experience, I-Now-1 will be conscious of visual information, and at the next moment it may shift its attention to auditory information, but then, say, I-Now-2 will be attending to visual information; at the next moment I-Now-1 may take control of the action, and so on. One could imagine multiple _"subjects"_ competing with each other at their subconscious level for the control of action, while being totally unaware of their competition, as well as of their multiplicity, at the level of consciousness. Indeed, because the schemas will not discriminate among these multiple _"subjects"_, the _"subjects"_ will not be able to discover their multiplicity. Their subjective feeling of "global awareness" would be an illusion in this picture. Of course, each of them, for example, I-Now-1, can test this feeling by trying to recall "what I was aware of just a second before", but the result

would be consistent with the illusion, because there is no restriction preventing I-Now-1 from accessing, e.g., a trace of awareness of I-Now-2. Furthermore, because of the Error Fundamentalis, I-Now-1 and I-Now-2 will never be able to engage in a dialogue with each other: in order for that to happen, they would have to assume different perspectives, but changing the perspective from I-Now to, e.g., I-Previous will cancel the privilege of the _"subject"_ to exhibit initiative (including the privilege to think anew and to talk), thus making a dialogue impossible.

Application to Neurological Syndromes and to Normal Phenomena

One essential observation regarding the above multiple-subject model is that the mechanisms underlying the navigation of each _"subject"_ over working memory must be subconscious, because the _"subjects"_ are assumed to be unaware of the partition (and therefore of the navigation they may perform). However, as long as there are no restrictions on possible partitions of the working memory, the partition will not result in an objective cognitive deficit, other than the limited span of immediate conscious awareness of each _"subject"_ at any given moment of time, as compared to the subjectively perceived span of conscious awareness. Suppose that some feature inconsistent with the rest of sensory content is missed by most _"subjects"_, and yet is noticed by one _"subject"_. The dynamics of the stream of consciousness in this case is likely to be determined by the majority of the _"subjects"_, and the "inconsistency" in memory is likely to be ignored or corrected afterwards, given that the _"subjects"_ frequently swap their contents of awareness in this model. The result could be a case of any of the phenomena known as masked priming (Forster, 1998; Van den Hout et al., 2000), change blindness, inattentional blindness (Enns and Di Lollo, 2000; Thornton and Fernandez-Duque, 2002), etc. These phenomena therefore provide a potential ground for testing our multiple-subject hypothesis experimentally.

Indirect support for this interpretation comes from the studies based on a masked priming paradigm, which indicate a high level of "subconscious" information processing, up to the emotional and Theory-of-Mind level, even though a normal physiological reaction, as well as an explicit memory of the missed detail, may not develop (Samsonovich, 2000; Kaszniak, 2002). On the other hand, there are indications that change blindness correlates with a missing activation of the putative substratum of the self (Rees et al., 1999; Beck et al., 2001). Although the latter observations are consistent with a one-subject view of change blindness, the multiple-subject view allows us to account for a wider range of phenomena, as we shall see now.

From the multiple-subject view, restrictions on possible partitions of the content of awareness may become imposed by brain damage, thus resulting in an inability of, say, I-Now-1 to navigate over the entire content of working or long-term memory in order to access traces laid out by I-Now-2. But I-Now-1 cannot become aware of this deficit, because it was not aware of its ability to navigate in the first place. The existing schemas of conscious awareness will not allow for that kind of awareness. As a result, a situation may develop in which each _"subject"_ (i) has a permanent deficit of awareness, limited to specific modalities or otherwise; (ii) is unaware of this deficit; (iii) as a consequence, may develop the contents of its own awareness and memory inconsistently with respect to other _"subjects"_, while (iv) retaining the common identity and other determining parameters together with the other _"subjects"_. This point of view allows us to account for two neurological syndromes, known as hemineglect and reversible anosognosia.

In the syndrome referred to as "hemineglect" the patient fails to explicitly perceive half the world, while at the same time exhibiting implicit knowledge about this denied half. In addition, the _"subject"_ has no awareness of the deficit. The missing half of the world appears not to exist in the _"subject's"_ awareness at all, creating no problem with understanding the truncated existence of the rest of the world for the _"subject"_. This hemineglect applies not only to the currently observed external world, but to imagined and remembered worlds as well (Bisiach, 1996). Moreover, it applies to general concepts, if they involve the existence of the other half of the world. These concepts no longer can be understood by the _"subject"_. For instance, "mirror agnosia" (Ramachandran et al., 1997) is a case when an otherwise normal subject with hemineglect may lose the ability to understand that a reflection in a mirror represents an object located outside of the mirror in a particular case, and not in general.

In our view, these examples fit into the multiple-subject model with a fixed partition of working memory. The _"subject"_, e.g., I-Now-1, who is unable to shift awareness to the other half of the world, cannot become aware of this deficit, because it was never aware of the fact and the mechanisms of shifting that previously took place in the intact brain. Similarly, it may not be possible to apply the schema of a mirror to the content of working memory, because this would require shifting attention across the partition, in order to incorporate both the reflection and the reflected original into one mental state. In conclusion, the version of the model with multiple "_subjects_" of awareness appears to be interesting in that it may provide an account of mysterious consciousness disorders such as hemineglect.

Another example of this sort is a syndrome known as reversible anosognosia. Ramachandran (1995; Ramachandran and Blakeslee, 1996) described a woman with the entire left half of her body paralyzed, who denied her paralysis, sometimes resorting to rather sophisticated mental tricks. Placing cold water in her ear reversibly changed both the woman's perception of her present condition and her memory about the past (this case is only one of a number of similar examples: see also Bisiach et al., 1986). From the multiple-subject point of view, in this case the partition of the substratum of awareness was fixed due to the brain damage, which resulted in mutually inconsistent memories created in different parts of the brain; however, it was possible to transiently reorganize the partition by applying cold water to the ear. As a result, either the speech controlling _"subject"_ gained access to a different autobiographical memory, or a _"subject"_ with the alternative memory gained access to control of speech and behavior. The result was an apparent alteration of memory (note that the personality remained the same). Reverting to the original, warm conditions in the ear apparently restored the previously existing partition together with the previous access to memory. It would not be possible for the multiple _"subjects"_ in this case to become consciously aware of the alteration for exactly the same reasons as discussed above.

The general model presented in this work can help us give an account of other neurological disorders noted earlier (see Table II). We start our consideration with hippocampal amnesia. In our view, patterns of activity in the hippocampus serve as allocentric pointers (substrata) associated with different instances of the self, and egocentric pointers are instantiated as patterns of activity in the prefrontal cortex. Therefore, if the hippocampal component (HC in Figure 5) is damaged, the ability to recall a specific episode and to attribute it to the past will be impaired. The reason for this, in addition to the context shifting problem discussed above, could be that the related egocentric pointer I-Past may not uniquely refer to that given episode, or may become unavailable after a certain "time window" (Lee and Kesner, 2003). Therefore, without the allocentric pointer it would not be possible to retrieve a previous state of mind, to create new autobiographical memories, or even to retrieve previously existing autobiographical memories. It would be possible, however, to understand concepts, to remember general (including spatial) facts about the world, and to have first-person experiences. Conditions exactly like these are observed clinically in patients with hippocampal lesions, e.g., the case of K.C. studied by Moscovitch et al. (1998) (Tulving et al., 1988; Rosenbaum et al., 2000).

We can briefly discuss interpretation of other agency-related disorders based on the proposed model. To begin with, development of PTSD can be understood as a situation when (a) the experience is unusually strong, which results in formation of abnormally strong associations between the content of experience and the pointer I-Now, while at the same time (b) an allocentric pointer cannot be created, because the hippocampus is driven by intense stress to an abnormal activity state. Given these conditions, if the memory of the episode is retrieved afterwards, it will not have a connection to the past. This memory can be retrieved indirectly by some salient cue/reminder. Because of the lack of information about the source context, on the one hand, and, on the other hand, because of a strong association with I-Now, the content of memory will be perceived as a present experience upon retrieval, even though it may not be entirely consistent with the sensory input. This condition is characteristic of PTSD (Nadel and Jacobs, 1998). In this case a previously experienced episode, when retrieved again, is experienced in the chart I-Now, as the currently happening situation.

If the experience in the above case was not abnormally strong, nor abnormally associated with I-Now for any other reason, and only the creation of an allocentric pointer was blocked, e.g., by scopolamine, then some details of the episode could be recalled by a salient cue or while in a hypnagogic state. In this case the pieces of the lost memory could become randomly associated with one of the active charts that is least inconsistent with their content: e.g., this could be I-Imagine. Then in principle it might be possible for the subject, with considerable mental effort, to assemble these pieces into a coherent story within I-Imagine. In order to "save" the recovered experience as a memory of the past, the perspective I-Imagine would need to be shifted to I-Past.

This operation, however, may not be possible for the subject to accomplish deliberately, without hypnosis: there are a limited number of perspective shifts that occur under normal conditions. E.g., as described above, the expected I-Next becomes I-Now, and I-Now becomes I-Previous, and then I-Past. But, what if a perspective shift goes spontaneously in a wrong way? For example, if I-Past takes the functional position of I-Now, while I-Now is missing, then the result could be a state in which the subject has no subjective feelings related to the present and, moreover, may come to believe that she does not exist in the present (*Cotard's syndrome*: Berrios and Luque, 1995). Similarly, cases of *reverse intermetamorphosis* (Breen et al., 2000), in which one mistakes self for another person, based on our framework can be understood as a chart He-Now or She-Now in the position of I-Now.

Consider now what happens in a possible situation when the system of charts and their relations is disrupted and starts malfunctioning more severely. For example, instances of "I" in

TABLE II
Interpretation of Neurological Syndromes based on the Proposed Model[3]

Neurological syndrome	Characteristic subjective experience	Functional abnormality of the self substrata in the brain
Hippocampal anterograde amnesia	Inability to develop new autobiographical memories.	New allocentric pointers are not formed; the context shifting does not occur.
Hippocampal retrograde amnesia	Loss of previously acquired autobiographical memories.	Previously formed allocentric pointers become absent or unavailable; the egocentric pointer I-Past cannot discriminate among episodes.
Post-traumatic stress disorder (PTSD)	The content of a specific memory, when retrieved, is experienced as a part of the present situation.	In a particular mental state, an allocentric pointer is not created, while at the same time unusually strong associations with I-Now are created.
Schizophrenia	Persistent delusions and hallucinations, involving alien control, alien voices, experience of passivity.	Charts acquire inappropriate privileges, may engage in dialogues or present imagery to each other, may create individual memories and pursue individual goals.
Multiple personality	Multiple subjects residing in the same brain with different, possibly overlapping memories.	Wrong relations among charts are stabilized by learning mechanisms. Dissociated charts develop different identities and personalities.
Autism	Inability to understand intuitively other minds and self at other times or under imaginary conditions.	Multiple charts do not develop due to neurophysiological deficits.
Hemineglect	Ignorance of the inability to attend to one half of the world, including all sensory and behavioral modalities.	The partition of working memory among self substrata (e.g., those in left and right hemispheres) becomes fixed by a brain damage.
Reversible anosognosia	Reversible alteration of the autobiographical memories related to a personal deficit, together with the awareness of these memories (and without any awareness of the alteration).	Reversible fixation of a memory partition. Long-term memories laid out by different subjects of experience are not consistent with each other, while the subjects retain their common identity.

[3] In the last two rows of this table, the interpretation is based on the multiple-subject version of the model.

perspectives other than I-Now may acquire privileges they normally do not have. As a consequence, they may start creating new memories (delusions), engage in dialogues ("voices"), independently perform imagery (thus presenting illusions to each other), or take control of actions. Events like these could destabilize and destroy the whole system of their previous, normal relations. Conditions like these are typically observed in schizophrenic patients (Mellors, 1970; Blakemore et al., 2000): specifically, schizophrenic states are characterized by passivity experiences (delusions of alien control), feelings of control of other's thoughts and actions, other persistent delusions and hallucinations. Passivity experiences may include actions, thoughts or emotions made for the subject by some external agent (Schneider, 1959): e.g., a patient may experience bodily sensations or emotions without apparent reasons; feelings of insertion, blocking or withdrawal of thoughts, control of other's thoughts and actions, supernumerary limbs that may be voluntarily controlled or may belong to somebody else, limbs with their own will (utilization behavior: Lhermitte, 1983; anarchic hand sign: Goldberg et al., 1981), etc. The following types of auditory hallucinations are characteristic of schizophrenia: voices arguing, voices commenting on one's action, audible thoughts (voices repeat verbatim or comment on subject's thoughts), and voices that command the subject. All these features are consistent with a picture of a destabilized system of multiple charts, as outlined above. Also consistent with it is the fact that normal human abilities to understand others and other instances of self are impaired in schizophrenia (Frith and Corcoran, 1996; Herold et al., 2002). Furthermore, this scenario could lead to a new stable situation in the system of charts, in which multiple instances of I-Now with different identities would co-exist (I-Now and J-Now). The result may fit the description of a multiple personality disorder (Nissen et al., 1988; Schacter et al., 1989).

Finally, we predict that if multiple charts do not develop in the brain at all, then "mental simulations" (e.g., states of autonoetic awareness: Wheeler et al., 1997) will be impossible. This means that the subject will not be able to understand intuitively other perspectives than I-Now. In particular, the subject will not be able to imagine other minds, self in the past or in the future, or a hypothetical situation in which he or she may happen to be. Given these conditions, the subject still may be able to develop theoretical concepts referring to time, other minds and imaginary situations (in analogy with color-blind individuals developing concepts of colors, yet lacking the qualia). In other words, speaking in terms of the model, the corresponding attitudes may be developed within I-Now, and without implementation of the related perspectives; autobiographical events could be learned as semantic knowledge, and so on. All this, however, would require more time and effort. The resultant picture fits the characteristic description of autism (Baron-Cohen et al., 1985; Blair et al., 2002). In conclusion, the results of analysis presented in this section are summarized in Table II.

We have to admit that idea of co-existing multiple instances of the self, and moreover, multiple _"subjects"_ for a single instance of a self, in all the ways described above is quite extraordinary. It seems to go well beyond the kinds of pathological conditions noted above; however, in fact we are very far from proposing anything like a multiple-person view of a normal brain. Because in the present work we eschew all questions related to the existence of a subject of conscious experience in the first place, we do not make any inferences from our objective, functionalist observations of logical possibilities.

Therefore, we use the word "subjects" in quotes, considering them merely as functional units in one possible functional organization of the brain. From an objective standpoint, a functional organization like this can be implemented in a computer, regardless of wheter it is implemented in the brain or not, and may prove pratically useful; therefore, its study may be of a scientific interest in this regard. Nevertheless, we emphasize here the value of this model for better understanding of *ordinary* psychological conditions, for which it has been adduced in the present work. For example, of particular interest are interpersonal relationships. Despite being adults, people often experience themselves as a child in relation to a spouse or an employer. Or, an adult might still experience a now past attribute of self or "I" (perhaps as a 'fat' or 'nerdy' self) from adolescence, even when they are dating in their 30's or 40's. Such ideas have mainly been discussed only from within a clinical psychological or psychoanalytic perspective, and are referred to as "transference experiences". Our model can be used to further understand these transference experiences based on the framework of multiple mental simulations. This is an area of specific interest for some of us (e.g., Jacobs and Nadel, 1985; Moscovitch and Nadel, 1998, who integrated neuroscience with the domains of clinical psychotherapy and psychoanalysis). In addition, Pally (2000) has integrated work on normal psychological development with psychotherapy and psychoanalysis. The bottom line is that these more functional and ordinary psychological issues, as opposed to the organic brain pathologies discussed above, make the proposed point of view even more attractive.

Concluding Remarks and Further Perspectives

The topic of perspectives and attitudes addressed in this work has a long history. For example, Husserl (1905/1990) considered a similar set of problems, offering a different interpretation: "I remember a lighted theater – this cannot mean that I remember having perceived the theater. Otherwise, this would imply that I remember that I have perceived, that I perceived the theater, and so on. I remember the lighted theater; this means that in my internal consciousness I see the lighted theater as having been." (Husserl, 1905/1990). What is apparently missing in this line of reasoning is an understanding of the fact that the attribution of the experience to the subject, the "I", may not be a part of the *content* of memory and of awareness upon recall, and yet it may be present in memory. The idea that the central role in the determinant of the source of experience belongs to the subject of experience was introduced by Tulving (1985; Wheeler et al., 1997), and is accepted today only by some researchers in the field. The counterpart dilemma in Theory-of-Mind studies is known as the dilemma of the theory-theory view (Gopnik, 1993; Carruthers, 1996; Gopnik and Meltzoff, 1997) vs. simulationism (Gordon, 1986; Heal, 1986; Goldman, 1989, 1993, 2000; Nichols and Stich, 2000, 2003), although the two alternatives could be viewed as complementary parts of one and the same global picture, similarly to the notions of perspectives and attitudes.

The model we described here is capable of providing a novel account of subjective feelings experienced during emotional states. An interpretation of emotions can be offered based on our multiple chart view described above. Thus, we can introduce a point of view that a separate chart, i.e., a separate mental perspective of the subject (call it I-Feel), must be used in parallel with I-Now in order to allow for the development of an emotional state in it, without allowing this state to seize all voluntary control in the given individual. Then the perception of this "mentally simulated" state of I-Feel from within I-Now must be the mechanism responsible for the emergence of the subjective feeling of emotion. This subjective feeling can be related then to the shift of I-Feel with respect to I-Now, viewed in egocentric coordinates of I-Now. From this point of view, if the state of emotion would be based on I-Now instead of I-Feel, then there would be no subjective feeling. This can be understood as follows. Attributing emotional attitudes, such as "angry", "sad", or "funny", to the target (which is the "I") implies locating this target outside of the center in the main subject's perspective, I-Now. On the other hand, the "I" is, by definition, the center of its own perspective. But then it follows that the target cannot be identical with I-Now. From this point of view, the ability to *perceive* emotions subjectively is likely to be correlated with the ability to understand other minds, as well as with the abilities to develop episodic memories, to imagine possible scenarios involving self and others, to be aware of self at a metacognitive level and the like. On the other hand, the ability to *exhibit* emotions behaviorally need not correlate with these factors. No surprise that this point of view is

counterintuitive, because our intuition about emotions involves Theory of Mind: in other words, we automatically attribute feelings to others and then take them for granted. Another mind may be different from what it seems. E.g., from the point of view proposed here, lower mammals may lack subjective feelings of emotions, while being able to exhibit emotional states (Roberts, 2002).

The scheme outlined so far applies to most basic *emotional reactions* (such as anxiety or depression) that by themselves do not require multiple mental perspectives for their development. Our claim is, again, that the associated *subjective feelings* (of anger, fear, etc.) require more than one mental perspective for their development. In some cases, however, an emotional reaction itself may require multiple mental perspectives for its development. E.g., this would be the case with a "higher-order" emotional reaction, one that depends on (mentally simulated) subjective feelings attributed to other instances of a self. One example of this sort is humor. Indeed, one cannot appreciate a joke without perceiving the target from a different perspective, while the content of the joke presupposes certain subjective feelings inside the target perspective (Lefcourt, 2001). Consistent with the above, there is no documented evidence of lower mammals possessing a sense of humor. Other examples of higher-order emotions involving simulated feelings of a "naïve" target can also be given. This topic is beyond the scope of the present work; however, it will be addressed elsewhere, due to its significance for artificial intelligence and robotics (Canamero, 2001; Gadanho and Hallam, 2001; Breazeal, 2003). The bottom line is that in order to fully understand or to model emotional mental states, one need to take into account the multiple-chart framework introduced in the present work.

In conclusion, in this work we presented a general functionalist framework that allows for a computational description of the self as the subject of experience, the agent and the author of controlled actions. In this framework, the self is introduced as a functional unit via its substratum and a set of fundamental constraints that we call Error Fundamentalis. These constraints, together with general common-sense knowledge and intuition and general facts about the world are instantiated by the set of schemas, that together constitute the person's semantic memory and thus provide the basis for the personality.

We started with a general epistemological and functionalist analysis of the concept of the self, basing it on the available data on normal and pathological conditions and exploiting the method of introspection. We then moved down to the connectionist and neuroanatomical levels of analysis. We made an explicit connection between abstract philosophical notions, such as instances of the self, and connectionist neuronal units. We proposed a plausible mapping of the connectionist model onto the functional neuroanatomy of the brain.

We showed that the problem of context shifting emerges in connectionist versions of the proposed model of episodic memory formation. Therefore, if the proposed framework applies to the brain, understanding the mechanisms of context shifting is necessary for understanding brain dynamics. We proposed a biologically plausible solution of the context shifting problem based on two memory indexing maps and suggested their brain instantiations.

We showed that the proposed framework could be applied to major agency disorders, including various forms of hippocampal amnesia, various aspects of schizophrenia, multiple personality, PTSD, and autism. Our interpretation of various forms of schizophrenia naturally accounts for the experienced multiplicity of agents sharing the same brain, yet differing in their mental perspectives; at the same time it provides room for passivity of "I". In contrast, most previous theoretical accounts of schizophrenic experiences left aside the nature and the origin of the multi-agent, alien content that has no obvious objective ground. For example, Frith et al. (Frith, 1992; Frith et al., 1998; Frith and Gallagher, 2002) speculate that a failure of "forward modeling" can result in misattribution of self-generated inner speech; however, the authors themselves say: "we take no account whatever of the *content* of these experiences" (Frith et al., 1998). Indeed, it is hard to understand, based on their point of view, why and how the misattributed inner speech should have contents that are totally alien to the subject, unless these contents are generated by separate systems – or, according to our interpretation, by separate charts. The same question applies to other modern interpretations of schizophrenic experiences (e.g., Metzinger, 2003). We anticipate that any interpretation of multiple agents observed in various forms of schizophrenia based on the current mainsteam view of consciuosness and the self (Dennett, 1991) wuold be more difficult than an interpretation that assumes multiple agents naturally co-existing, in a normal brain.

We further discussed a possible version of our model – the multiple subject model – that allowed us to give an account of hemineglect and reversible anosognosia. These interpretations of disorders could be made explicit by computer simulations based on the framework that we presented here. In our view, a general-purpose computer implementation of this framework is possible and would provide researchers with a new tool, allowing them to study all agency-related psychic phenomena within one unified cognitive-theoretical and computational paradigm. This implementation would require development of new cognitive architectures, departing from existing prototypes, such as SOAR (e.g., Jones et al., 1999) and ACT-R (Anderson and Lebiere, 1998).

In this view, the proposed conceptual framework opens new perspectives in artificial intelligence; in particular, for computer implementations of conscious systems. One possible application in artificial intelligence would be to create a virtual agent that would possess awareness of self and others, would be capable of intelligent cooperation with other agents, and would be able to learn from it's own and others' experience. The agent could communicate using a specially designed language (that could be based on the language of schemas and at the same time could be close to the natural language), would be able to process general symbolic information, and in principle could be installed in a mobile robot. Furthermore, a team of such agents could be conceived as capable of successful cooperation with each other in an unfamiliar environment, under conditions where extensive logical reasoning, as well as frequent communications, cannot be used. In addition, a team of this sort could be created ad hoc, with limited prior knowledge among the agents about each other, and potentially could include humans and non-cognitive components. components. Most importantly, a system of this sort should be able to learn and to self-organize indefinitely, by itself (e.g., using Internet resources and the like), without any obvious limitations, starting from a certain minimal "critical mass" of computational consciousness and going beyond the human level in all cognitive dimensions.

After the original manuscript of this work was submitted to *Cortex* a paper by Aleksander and Dunmall (2003) introduced the notion of the *minimally necessary mechanisms for consciousness expressed in the form of axioms*. Their idea thus appears to be a precursor to our idea of a non-emergent, constraint-based self, and their set of axioms appears a somewhat complementary analog to our *Error Fundamentalis* of the self (in particular, their axioms require that a conscious agent must have a private sense of an "out-there world" and the self, must be able to control its attention and imagination, to acquire emotional values, etc.), yet their notion of the self appears to be limited to the distinction between the system and its environment and does not necessarily lead to an immanent contradiction of the self-concept that characterizes a conscious system according to our framework. We therefore decided to keep the term *Error Fundamentalis* (borrowed from Golosovker, 1936/1987).

Acknowledgments. We are grateful to Dr. K. De Jong for valuable discussions and to two anonymous referees for their very useful critical comments on an earlier version of the manuscript. The early stage of this research was supported by the Consciousness Studies Research Grant "Artificial Consciousness as a Metaphor for Human Consciousness" to Alexei Samsonovich and Lynn Nadel from the Center for Consciousness Studies at the University of Arizona and the Fetzer Institute, 1999-2000.

References

Aleksander I and Dunmall B. Axioms and tests for the presence of minimal consciousness in agents. *Journal of Consciousness Studies, 10:* 7-18, 2003.

Anderson JR and Lebiere C. *The Atomic Components of Thought.* Mahwah: Lawrence Erlbaum Associates, 1998.

Arbib MA, Erdi P and Szentagothai J. *Neural Organization: Structure, Function and Dynamics.* Cambridge: MIT Press, 1998.

Armstrong DM. What is Consciousness? In D Armstrong (Ed), *The Nature of Mind.* St. Lucia Queensland: University of Queensland Press, 1980, pp. 55-67.

Baron-Cohen S, Leslie A and Frith U. Does the autistic child have a 'theory of mind'? *Cognition, 21:* 37-46, 1985.

Beahrs JO. My voice will go with you. *American Journal of Clinical Hypnosis, 26:* 2 100-113, 1983.

Beck DM, Rees G, Frith CD and Lavie N. Neural correlates of change detection and change blindness. *Nature Neuroscience, 4:* 645-650, 2001.

Berrios GE and Luque R. Cotard's syndrome: Analysis of 100 cases. *Acta Psychiatrica Scandinavica, 91:* 185-188, 1995.

Bisiach E. Unilateral neglect and the structure of space representation. *Current Directions in Psychological Science, 5:* 62-65, 1996.

Bisiach E, Vallar G, Perani D, Papagno C and Berti A. Unawareness of disease following lesions of the right-hemisphere – anosognosia for hemiplegia and anosognosia for hemianopia. *Neuropsychologia, 24:* 471-482, 1986.

Blair RJ, Frith U, Smith N, Abell F and Cipolotti L. Fractionation of visual memory: Agency detection and its impairment in autism. *Neuropsychologia, 40:* 108-118, 2002.

Blackmore S. There is no stream of consciousness. *Journal of Consciousness Studies, 9:* 17-28, 2002.

Blakemore SJ, Smith J, Steel R, Johnstone EC and Frith CD. The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: Evidence for a breakdown in self-monitoring. *Psychological Medicine, 30:* 1131-1139, 2000.

Breazeal C. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies, 59:* 119-155, 2003.

Breen N, Caine D, Coltheart M, Hendy J and Roberts C. Towards an understanding of delusions of misidentification: Four case studies. In M Coltheart and M Davies (Eds), *Pathologies of Belief.* Oxford: Blackwell, 2000.

Canamero L. Emotions and adaptation in autonomous agents: A design perspective. *Cybernetics and Systems, 32:* 507-529, 2001.

Carruthers P. Simulation and self-knowledge: A defence of theory-theory. In P Carruthers and PK. Smith (Eds), *Theories of Theories of Mind.* Cambridge: Cambridge University Press, 1996, pp. 22-38.

Chalmers DJ. Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2:* 200-219, 1995.

Chalmers DJ. *Conscious Mind: In Search of a Fundamental Theory.* New York: Oxford University Press, 1996.

Chalmers DJ. Consciousness and its place in nature. In SP Stich and TA Warfield (Eds), *Blackwell Guide to Philosophy of Mind.* Malden: Blackwell Publishing, 2003, pp. 102-142.

Cheng PW and Holyoak KJ. Pragmatic reasoning schemas. *Cognitive Psychology, 17:* 391-416, 1985.

Churchland PM. *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind.* Cambridge: MIT Press, 1988.

Damasio AR. *The Feeling of What Happens.* New York: Harcourt, 1999.

De Jong K. *Evolutionary Computation: Theory and Practice.* Cambridge: MIT Press, 2005.

Dennett DC. *Consciousness Explained.* Boston: Little, Brown and Co., 1991.

Descartes R. *Discourse on Method.* Paris, 1637.

Enns JT and Di Lollo V. What's new in visual masking? *Trends in Cognitive Science, 4:* 345-352, 2000.

Ethier K, Le Marec N, Rompre PP and Godbout R. Spatial strategy elaboration in egocentric and allocentric tasks following medial prefrontal cortex lesions in the rat. *Brain and Cognition, 46:* 134-135, 2001.

Farrer C and Frith CD. Experiencing oneself vs another person as being the cause of an action: The neural correlates of the experience of agency. *Neuroimage, 15:* 596-603, 2002.

Flynn M, Molden S, Witter MP, Moser EI and Moser MB.

Spatial representation in the entorhinal cortex. *Science, 305:* 1258-1264, 2004.
FORSTER KI. The pros and cons of masked priming. *Journal of Psycholinguistic Research, 27:* 203-233, 1998.
FRITH CD. *The Cognitive Neuropsychology of Schizophrenia.* Hove: Lawrence Erlbaum Associates, 1992.
FRITH C. Neuropsychology of schizophrenia what are the implications of intellectual and experiential abnormalities for the neurobiology of schizophrenia? *British Medical Bulletin, 52:* 618-626, 1996.
FRITH CD and CORCORAN R. Exploring 'theory of mind' in people with schizophrenia. *Psychological Medicine, 26:* 521-530, 1996.
FRITH C and GALLAGHER S. Models of the pathological mind. *Journal of Consciousness Studies, 9:* 57-80, 2002.
FRITH C, REES G and FRISTON K. Psychosis and the experience of self: Brain systems underlying self-monitoring. *Annals of the New York Academy of Sciences, 843:* 170-178, 1998.
FUSTER JM. *Cortex and Mind.* New York: Oxford University Press, 2003.
GADANHO SC and HALLAM J. Robot learning driven by emotions. *Adaptive Behavior, 9:* 42-64, 2001.
GOLDBERG G, MAYER NH and TOGLIA JU. Medial frontal cortex infarction and the alien hand sign. *Archives of Neurology, 38:* 683-686, 1981.
GOLDMAN AI. Interpretation psychologized. *Mind and Language, 4:* 161-185, 1989.
GOLDMAN AI. The psychology of folk psychology. *Behavioral and Brain Sciences, 16:* 15-28, 1993.
GOLDMAN A. Folk psychology and mental concepts. *Protosociology, 14:* 4-25, 2000.
GOLOSOVKER YAE. *The Logic of Myth.* Moscow: Nauka, 1936/1987.
GOPNIK A. How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences, 16:* 1-14, 1993.
GOPNIK A and MELTZOFF A. *Words, Thoughts and Theories.* Cambridge: MIT Press, 1997.
GORDON RM. Folk psychology as simulation. *Mind and Language, 1:* 158-171, 1986.
HEAL J. Replication and functionalism. In J Butterfield (Ed), *Language, Mind and Logic.* Cambridge: Cambridge University Press, 1986.
HEROLD R, TENYI T, LENARD K and TRIXLER M. Theory of mind deficit in people with schizophrenia during remission. *Psychological Medicine, 32:* 1125-1129, 2002.
HUME D. *A Treatise of Human Nature.* London: John Noon, 1739/1965.
HUSSERL E. *The Phenomenology of Internal Time Consciousness.* Ann Arbor: University Microfilms International, 1905/1990.
IRAN-NEJAD A and WINSLER A. Bartlett's schema theory and modern accounts of learning and remembering. *Journal of Mind and Behavior, 21:* 5-35, 2000.
JACOBS WJ and NADEL L. Stress-induced recovery of fears and phobias. *Psychological Review, 92:* 512-531, 1985.
JONES RM, LARID JE, NIELSEN PE, COULTER KJ, KENNY P and KOSS FV. Automated intelligent pilots for combat flight simulation. *AI Magazine, 20:* 27-41, 1999.
KAMPE KK, FRITH CD and FRITH U. "Hey John": Signals conveying communicative intention toward the self activate brain regions associated with "mentalizing" regardless of modality. *Journal of Neuroscience, 23:* 5258-5263, 2003.
KANT I. *Critique of Pure Reason.* Translated by NK Smith, New York: St. Martin's Press, 1781/1929.
KASZNIAK AW. Challenges in the empirical study of conscious emotion. *Consciousness Research Abstracts – Toward a Science of Consciousness, 2002:* n. 158.
KESNER RP, FARNSWORTH G and DI MATTIA BV. Double dissociation of egocentric and allocentric space following medial prefrontal and parietal cortex lesions in the rat. *Behavioral Neuroscience, 103:* 956-61, 1989.
LANGDON R and COLTHEART M. Visual perspective-taking and schizotypy: Evidence for a simulation-based account of mentalizing in normal adults. *Cognition, 82:* 1-26, 2001.
LANGDON WB and POLI R. *Foundations of Genetic Programming.* Berlin: Springer, 2002.
LEE I and KESNER RP. Time-dependent relationship between the dorsal hippocampus and the prefrontal cortex in spatial memory. *Journal of Neuroscience, 23:* 1517-1523, 2003.
LEFCOURT HM. *Humor: The Psychology of Living Buoyantly.* New York: Plenum, 2001.
LHERMITTE F. 'Utilisation behaviour' and its relation to lesions of the frontal lobes. *Brain, 106:* 237-255, 1983.
LIBET B. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8:* 529-566, 1985.
LIBET B, WRIGHT EW, FEINSTEIN B and PEARL DK. Subjective referral of the timing for a conscious sensory experience – functional role for the somatosensory specific projection system in man. *Brain, 102:* 193-224, 1979.
LYCAN WG. *Consciousness and Experience.* Cambridge: MIT Press, 1996.
MARK VW. Conflicting communicative behavior in a split-brain patient: Support for dual consciousness. In SR Hameroff, AW Kaszniak and AC Scott (Eds), *Toward a Science of Consciousness: The First Tucson Discussion and Debates.* Cambridge: MIT Press, 1996, pp. 189-196.
MCGONIGLE DJ, HANNINEN R, SALENIUS S, HARI R, FRACKOWIAK RSJ and FRITH CD. Whose arm is it, anyway? An fMRI case study of supernumerary phantom limb. *Brain, 125:* 1265-1274, 2002.
MCNAUGHTON BL, BARNES CA, GERRARD JL, GOTHARD K, JUNG MW, KNIERIM JL, KUDRIMOTI H, QIN Y, SKAGGS WE, SUSTER M and WEAVER KL. Deciphering the hippocampal polyglot: The hippocampus as a path integration system. *Journal of Experimental Biology, 199:* 173-185, 1996.
MELLORS CS. First rank symptoms of schizophrenia. *British Journal of Psychiatry, 117:* 13-23, 1970.
METZINGER T. *Being No One: The Self-Model Theory of Subjectivity.* Cambridge: MIT Press, 2003.
MOSCOVITCH M and NADEL L. Consolidation and the hippocampal complex revisited: In defense of the multiple-trace model. *Current Opinion in Neurobiology, 8:* 297-300, 1998.
NADEL L and MOSCOVITCH M. Memory consolidation retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology, 7:* 217-227, 1997.
NADEL L and JACOBS WJ. Traumatic memory is special. *Current Directions in Psychological Science, 7:* 154-157, 1998.
NADEL L, SAMSONOVICH A, RYAN L and MOSCOVITCH M. Multiple trace theory of human memory: Computational neuroimaging and neuropsychological results. *Hippocampus, 10:* 352-368, 2000.
NICHOLS S and STICH S. A cognitive theory of pretense. *Cognition, 74:* 115-147, 2000.
NICHOLS S and STICH S. *Mindreading.* Oxford: Oxford University Press, 2003.
NISSEN MJ, ROSS JL, WILLINGHAM DB, MACKENZIE TB and SCHACTER DL. Memory and awareness in a patient with multiple personality disorder. *Brain and Cognition, 8:* 117-134, 1988.
O'KEEFE J and NADEL L. *The Hippocampus as a Cognitive Map.* New York: Clarendon Press, 1978.
PALLY R. *The Mind-Brain Relationship.* London: Karnac Books, 2000.
PANZARASA P, JENNINGS NR and NORMAN TJ. Formalizing collaborative decision-making and practical reasoning in multi-agent systems. *Journal of Logic and Computation, 12:* 55-117, 2002.
PARFIT D. *Reasons and Persons.* Oxford: Clarendon, 1984.
PERRY J. The problem of the essential indexical. *Noûs* XIII: 3-21, 1979.
RAGOZZINO ME and KESNER RP. The role of rat dorsomedial prefrontal cortex in working memory for egocentric responses. *Neuroscience Letters, 308:* 145-148, 2001.
RAMACHANDRAN VS. Anosognosia in parietal lobe syndrome. *Consciousness and Cognition, 4:* 22-51, 1995.
RAMACHANDRAN VS, ALTSCHULLER EL and HILLYER S. Mirror agnosia. *Proceedings of the Royal Society of London, 264:* 645-647, 1997.
RAMACHANDRAN VS and BLAKESLEE S. *Phantoms in the Brain: Probing the Mysteries of the Human Mind.* New York: William Morrow, 1996.
REES G, RUSSEL C, FRITH CD and DRIVER J. Inattentional blindness versus inattentional amnesia for fixated but ignored words. *Science, 286:* 2504-2507, 1999.
ROBERTS WA. Are animals stuck in time? *Psychological Bulletin, 128:* 473-489, 2002.
ROSENBAUM RS, PRISELAC S, KOHLER S, BLACK SE, GAO F, NADEL L and MOSCOVITCH M. Remote spatial memory in an amnesic person with extensive bilateral hippocampal lesions. *Nature Neuroscience, 3:* 1044-1048, 2000.
ROSENTHAL DM. Two Concepts of Consciousness. *Philosophical Studies, 49:* 329-359, 1986.
ROSENTHAL DM. Unity of consciousness and the self. *Proceedings of the Aristotelian Society, 103:* 325-352, 2003.
SAMSONOVICH A. Hierarchical multichart model of the hippocampal spatial map. In BW Mel and T Sejnowski (Eds),

Proceedings of the 5th Joint Symposium on Neural Computation. San Diego: Institute for Neural Computation, UCSD, 1998, Vol. 8, pp. 140-147.

SAMSONOVICH A. Masked-priming 'Sally-Anne' test supports a simulationist view of human theory of mind. In BW Mel and T Sejnowski (Eds), *Proceedings of the 7th Joint Symposium on Neural Computation.* San Diego: Institute for Neural Computation, UCSD, 2000, Vol. 10, pp. 104-111.

SAMSONOVICH AV and ASCOLI GA. Towards virtual brains. In GA Ascoli (Ed), *Computational Neuroanatomy: Principles and Methods.* Totowa: Humana Press, 2002, pp. 425-436.

SAMSONOVICH AV and ASCOLI GA. The conscious self: Ontology, epistemology and the mirror quest. *Cortex,* in press.

SAMSONOVICH AV and DEJONG KA. Meta-cognitive architecture for team agents. In R Alterman and D Kirsh (Eds), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society,* Boston: Cognitive Science Society, 2003, pp. 1029-1034.

SAMSONOVICH AV and DEJONG KA. A general-purpose computational model of the conscious mind. In M Lovett, C Schunn, C Lebiere and P Munro (Eds), *Proceedings of the Sixth International Conference on Cognitive Modeling*. Mahwah: Lawrence Erlbaum Associates Publishers, 2004, pp. 382-383.

SAMSONOVICH A and MCNAUGHTON BL. Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience, 17:* 5900-5920, 1997.

SCHACTER DL, KIHLSTROM JF and KIHLSTROM LC. Autobiographical memory in a case of multiple personality disorder. *Journal of Abnormal Psychology, 98:* 508-514, 1989.

SCHNEIDER K. *Clinical Psychopathology.* New York: Grune and Stratton, 1959.

SEARLE JR. Minds, brains and programs. *Behavioral and Brain Sciences, 3:* 417-424, 1980.

SEARLE JR. How to study consciousness scientifically. In Hameroff SR, Kaszniak AW and Scott AC (Eds), *Toward a Science of Consciousness – II: The Second Tucson Discussion and Debates*. Cambridge: MIT Press, 1998, pp. 15-29.

STRAWSON G. 'The Self'. *Journal of Conscious Studies, 4:* 405-428, 1997.

THORNTON IM and FERNANDEZ-DUQUE D. Converging evidence for the detection of change without awareness. *Progress in Brain Research, 140:* 99-118, 2002.

TULVING E. How many memory-systems are there. *American Psychologist, 40:* 385-398, 1985.

TULVING E. Episodic memory: From mind to brain. *Annual Reviews in Psychology, 53:* 1-25, 2002.

TULVING E, SCHACTER DL, MCLACHLAN DR and MOSCOVITCH M. Priming of semantic autobiographical knowledge: A case study of retrograde amnesia. *Brain and Cognition, 8:* 3-20, 1988.

TYE M. *Consciousness and Persons: Unity and Identity.* Cambridge: MIT Press, 2003.

VAN DEN HOUT MA, DE JONG P and KINDT M. Masked fear words produce increased SCRs: An anomaly for Ohman's theory of pre-attentive processing in anxiety. *Psychophysiology, 37:* 283-288, 2000.

VOGELEY K and FINK GR. Neural correlates of the first-person-perspective. *Trends in Cognitive Sciences, 7:* 38-42, 2003.

WHEELER MA, STUSS DT and TULVING E. Toward a theory of episodic memory: The frontal lobes and autonoetic consciousness. *Psychological Bulletin, 121:* 331-354, 1997.

Alexei V. Samsonovich, Krasnow Institute for Advanced Study, George Mason University, 4400 University Drive MS 2A1, Fairfax, VA 22030-4444, USA.
e-mail: asamsono@gmu.edu

Lynn Nadel, Department of Psychology, University of Arizona 15051, University Blvd, Tucson AZ 85731-0068, USA.
e-mail: nadel@u.arizona.edu