Abductive inference and creative thinking under the predictive processing framework

Inferência abdutiva e pensamento criativo sob a ótica da teoria do processamento preditivo

Research proposal for a postdoctoral position at the FFLCH, USP

Supervisor: Prof. Dr. Osvaldo Frota Pessoa Jr

Candidate: Dr. Fabiana Mesquita de Carvalho

Abstract: The problem of the mind-brain correspondence has been a central issue in philosophy of mind and cognitive science in recent decades. Several comprehensive cognitive neuroscience theories have attempted to understand the link between brain states on one hand, and perception and cognition on the other. The situated cognition approach tries to shed light on how cognitive activity spans the agent's perception and action as well as his physical, social, and cultural environment. One of the most influential theories within the situated cognitive science is the predictive processing (PP) framework. It offers a unifying model of perception, cognition and action under the 'Bayesian prediction machine' approach. It states that the agent structure their world and actions in order to fulfil their sensory predictions, and cognition and action emerges from the attempt of reducing prediction errors resulting from the interaction with the environment. The role of philosophy, in this context, is a crucial one. It provides the foundation to reconcile the advances in cognitive science with philosophical theories in the field of phenomenology, logic and language, in order to better understand the relationship between mind, brain and social human behaviour. This project aims to investigate, under the PP framework, how the basic organizing principles of action-oriented predictive processing could be extended to high-level cognition as an attempt to understand the construction of human-specific modes of reasoning to reduce prediction errors. This project will be based on (a) the investigation of the key aspects of human specific high-level cognition, which should include language and social/collective intentionality, (b) the proposal of a formal distinction between different types of inferential processes according to C. S. Peirce's three types of inference, (c) the combination of linguistic and inferential aspects as an attempt to elucidate human-specific ways of reasoning and dealing with agent-level uncertainty, (d) the use of human-specific ways of reasoning to understand the nature of human creativity, and (e) the proposal of neurobiological mechanisms that could underlie human-specific, high-level inferential processes.

Resumo: O problema da correspondência mente-cérebro tem sido um tema central na filosofia da mente e nas ciências cognitivas nas últimas décadas. Várias teorias abrangentes da neurociência cognitiva têm tentado compreender a conexão entre, por um lado, estados cerebrais, e por outro, percepção e cognição. As abordagens da cognição situada tentam elucidar como a atividade cognitiva se estende desde a percepção e cognição do agente até seu meio físico, social e cultural. Uma das teorias mais influentes dentro da ciência da cognição situada é a teoria do processamento preditivo (PP). Ela oferece um modelo unificadbr da percepção, cognição e ação sob a perspectiva da 'máquina de predição Bayesiana'. Ela afirma que o agente estrutura o seu mundo e suas ações de modo a preencher suas predições sensoriais, e a cognição e ação emergem da tentativa de redução dos erros de predição resultantes da interação com o meio. O papel da filosofia, nesse contexto, é crucial. Ela fornece o alicerce para a reconciliação do avanço das ciências cognitivas com as teorias filosóficas nos campos da fenomenologia, lógica e linguagem, de modo a melhor compreender a relação entre mente, cérebro e o comportamento social humano. O presente projeto tem como objetivo investigar, sob a ótica da teoria PP, como os princípios organizacionais básicos do processamento preditivo orientado à ação poderiam ser estendidos a habilidades cognitivas avançadas, no intuito de se compreender a construção de modos especificamente humanos de raciocínio para reduzir erros de predição. O projeto terá como base (a) a investigação de aspectos chave das habilidades cognitivas avançadas especificamente humanas, o que deve incluir a linguagem e intencionalidade social/coletiva, (b) a proposta de uma distinção formal entre diferentes tipos de processos inferenciais de acordo com os três tipos de inferência de C. S. Peirce, (c) a combinação dos aspectos linguísticos e inferenciais na tentativa de elucidar formas de raciocínio e formas de lidar com incerteza especificamente humanas, (d) a utilização das formas de raciocínio especificamente humanas na tentativa de se compreender a natureza da criatividade humana, e (e) a proposta de mecanismos neurobiológicos subjacentes aos processos inferenciais avançados e especificamente humanos.

Content

1 Preface

René Descartes, the best-known classic mind-body dualist, postulated an ontological distinction between the mental and the physical which stated they were categorically different kinds of substances (the consciousness being immaterial) that interact in the living human. The problem that followed this postulation, the so-called "mind-body" problem within the doctrine of dualism in the philosophy of mind, is to understand how mind and body might successfully interact in the agent. The present proposal assumes the philosophical position that the mind-body problem is only an actual problem if we accept Descartes' proposed discrepancy. Otherwise, if one disregards Descartes' premise and acknowledges only one type of substance – the physical – the problems concerning mind-brain interaction are effectively eliminated. However, we are still far from the end of the story. Actually, it is at this very point that we start facing the real philosophical problem of how things we understand as "the mind" – thoughts, ideas, memories, emotions - emerges from the brain and the body.

In recent decades, the debate over situated cognition has become a central issue in cognitive science and philosophy of mind. In addition to the narrow mind-brain correspondence, situated cognition broad approach tries to shed light on how cognitive activity spans the agent's perception and action as well as his physical, social, and cultural environment. The central question of extended cognition accounts can be summed up in 'where does the mind stop and the rest of the world begin?' as cognition is often taken to be continuous with processes in the environment (Clark, 1997; Clark & Chalmers, 1998; Varela, 1999).

The situated cognition approaches cannot be accommodated within traditional cognitive science and associated philosophical views of the mind. Within the cognitive science, one of the most prominent situated cognition theories is the predictive processing (PP) framework (Geisler & Kersten, 2002; Friston, 2003, 2005; Clark, 2013), which promises to bring perception, cognition and action together under the same basic neurobiological principle. This principle is based on empirical Bayes, which can be seen as the subjective view of probability. Models in Bayesian statistics start with the idea that the agent implicitly infers or represents the causes of its sensory samples by combining the actual sensory input and prior beliefs about these causes. This is a way of updating previous belief about the world (priors, or *a priori* probability distribution of causes of data) by the current sensory evidence (likelihood, or the probability of sensory data, *given* their causes). Bayesian models are becoming increasingly prominent in science and philosophy as it provides an alternative approach to purely logical or frequentist views in understanding how problems of induction can be solved in the human mind (Griffiths *et al.*, 2008; Oaksford & Chater, 2009). The central question of how the agent goes beyond the data of experience has been the source of many deep problems and debates in modern epistemology and philosophy of science.

The PP framework and its Bayesian formalism (Friston, 2005) accommodates nicely the fundamental problem the cognitive system confronts: that of coping with uncertainty by reducing prediction error resulting from our

exchanges with the environment. Taking into account the recent explosion of empirical work supporting the PP framework, situated cognition is to mark a significant departure from the traditional views in cognitive neuroscience as well as philosophy of science. The present proposal addresses how the basic principles of the PP framework could be extended to high-level cognition as an attempt to understand the construction of human-specific modes of reasoning to reduce prediction errors.

2 Introduction to the problem

According to a recent and increasingly influential approach in computational and cognitive neuroscience, the human brain is essentially a hierarchical prediction machine (Rao & Ballard, 1999; Geisler & Kersten, 2002; Lee & Mumford, 2003; Friston, 2003, 2005; Hohwy, 2007; Clark, 2013). This approach – initially called predictive coding and more recently predictive processing (PP) framework - states that the primary function of the cerebral cortex is to infer the causes of sensory input (sensation and perception), to learn the relationship between input and its causes (cognition) and to act accordingly to what has been inferred. The PP framework ambitious attempt to unify behaviour and the structure of the world retraces back to Hermann von Helmholtz's original writings on the "unconscious inference" theory of perception (Helmholtz, 1866/1962), and to R. L. Gregory's paper on the importance of perceptual errors in hypothesis-generation procedures (Gregory, 1980). At the heart of the PP framework is the key notion that perception is a probabilistic inference process based on premises and preferences.

The cerebral cortex has a hierarchical organization and an activity-dependent plasticity, the two key anatomical and physiological features for the PP framework statistical scheme operation, which is based on empirical Bayes (Friston, 2005). The simplicity of the empirical Bayes strategy is that it allows the brain to use the best model at one cortical level as priors for the level below. Each level tries to predict the neural activity pattern at the level below via top-down (backward) connections. Where the activity is correctly predicted, no further action is required – the sensory input has been explained away by high-level predictions. But where there is a mismatch between the causes inferred by the generative model and the actual sensory stimulus, a prediction error occurs. The prediction error is used both to estimate the causes at the level above in a feed-forward cycle and to reconfigure the generative models for future predictions. The feed-forward, error-correction cycle goes on until the system stabilizes as it reaches a progressive minimization of the overall surprisal.

This is to say that the statistical structure of the world is represented at different levels of abstraction organized on different levels of the cortical hierarchy. Our beliefs, or memories, are used to generate expectations about the world, which are carried down the hierarchy by top-down connections. The expectations are simply our 'initial guess'. The recognition of what is 'out there' in the world can be seen as a matching issue – the system must find an analogy between the actual sensory signal and its most similar representation in memory. And it is the divergence between what is expected and what is actually been

observed, the surprisal, that might result in learning, i.e., updating of the prior distribution to more closely match the posterior distribution. In short, the brain function could be summarized as minimization of prediction error in a bidirectional cascade of cortical processing.

In its present format, the PP framework can be seen as the description of a core cortical processing mechanism. Action follows the same logic principle as perception and learning (Friston, 2010, 2011; Clark, 2013). The agent is depicted as a statistical model of its environment since the agent embodies an optimal model of its environment. As Clark (2013) puts it, "this is both good news and bad news" (p.13). It is good because the PP framework can indeed be used to explain cognition and action of all mammals, for instance – since they all have a hierarchical and plastic neocortex – and can accommodate a myriad of different behavioural procedures. The downside is, if we intend to focus on understanding the specific functional nature and format of human neural representation, the gap between the PP current general formulations and the very specific features of human cognition is still gigantic (Clark, 2013; Roepstorff, 2013). The major hurdle here is to find out whether a single corpus of predictive information could seamlessly support both low-level perception and high-level human cognition.

In addition to the distance between the PP framework general mechanisms and high-level human cognition, it is still quite cumbersome to try to pull apart the agent-level surprise from the more perceptual sense of surprisal (implausibility of some sensory state given a model of the world). As active agents, we are continuously trying to make sense of the world by linking sights, sounds, and situations that we encounter to representations stored in our memories (recognition by analogy – Bar & Neta, 2008; Bar, 2009; Yardley *et al.*, 2012). Prediction errors are yielded by a mismatch between the sensory signals encountered and those anticipated. It comes as a perceptual error at first. For instance, we expect carrots to be orange, to see a shirt upon opening the closet door, and a chocolate candy to be sweet. If these expectations are broken – if we happen to see a blue carrot in a clear day – it can be understood as a very improbable percept. But it is still the one that best respects the current combination of driving inputs, and reflects the agent's brain high degree of confidence in the sensory signal (surely it is a blue carrot). Nonetheless, for the agent, the percept still emerges as surprising and demands an explanation. The inferential process must be able to minimize surprisal "for" the brain as well as surprise for the agent in any highly surprising and unexpected context. However, the PP framework still lacks a formal distinction between types of inferential processes used to reduce the surprise/surprisal pair (Hirsh *et al.*, 2012; Clark, 2013; Anderson & Chemero, 2013).

The main aim of the present proposal is to try to extend the basic principles of action-oriented prediction to agent-level cognition as an attempt to understand the construction of human thought and reason. How the predictive processing framework relates to high-level human cognition is still an open question.

3 Specific aims

The following topics will be explored in this research project:

(1) The understanding of the key aspects of human specific high-level cognition (characteristic of humans as a life form).

(2) The proposal of a formal distinction between different types of inferential processes.

(3) The combination of (1) and (2) as an attempt to elucidate human-specific ways of reasoning and dealing with agent-level uncertainty (surprise).

(4) The use of (3) to understand the nature of human intellect and creativity.

(5) The proposal of neurobiological mechanisms that could underlie human-specific, high-level inferential processes.

4 Significance: Philosophical reflections as part of progress in Neuroscience

The PP framework is a deeply important development for neuroscience, artificial intelligence and philosophy of mind. It offers a unifying model of brain function which proposes that perception, learning, cognition and action might all be constructed out of the same base materials (prediction and prediction error minimization).

However, how this may be achieved is still far from well established. Clark (2013) refers to this problem by stating that "a full account of human cognition cannot hope to 'jump' directly from the basic organizing principles of action-oriented predictive processing to an account of the full (and in some ways idiosyncratic) shape of human thought and reason" (p. 21). Indeed, it is necessary an interdisciplinary approach to try to illuminate how an intelligent agent might emerges from the first principles proposed by the PP framework. The approach we propose here combines computational insights with empirical neuroscience models with a full appreciation that philosophy is central to the interdisciplinary investigation of the human mind.

Philosophical ideas are important stimulators of scientific investigations through the well-known heuristic role that it plays in theory construction and in creative thought. For instance, Daniel Dennett's views of intentional action (intentional stance – Dennett, 1978) inspired a research line in developmental psychology concerned with children's judgments about false beliefs. The Kantlan 'a priori' principles of reason provided the basis for Konrad Lorenz's influential theory on innate behaviour patterns in the field of ethology.

The PP framework operates at the edge of what is known, walking on a delicate and important line as it access general issues about the nature of knowledge and reality. Thagard (2009) argues that philosophy crucially contributes to cutting-edge cognitive science theories by bringing generality and normativity on board. Philosophical generality allows a broader reflection on complex questions that cross multiple areas of

investigation. It helps to bring down the wall between disciplines and to solve terminology and method problems, thereby unifying what otherwise appears to be diverse approaches to understanding human cognition. Normativity pushes science toward the investigation of the nature of concepts and models before assuming that such concept or model ought to be the correct norm for a given process. Working on establishing the normative view (rather than descriptive, as it is common in science) helps cognitive science to not take for granted common norms in the field. Specifically within the scope of PP framework and the present project, philosophy of mind, mainly the fields of logic and language, could be a very powerful instrument in understanding the uniqueness of human language and proposing different types of inferential processes

On the other hand, cognitive sciences theories can be hugely relevant to philosophical issues. Naturalist approaches see both philosophy and science as pursuing essentially the same ends, and progress in philosophy would require close attention to scientific developments. Regarding the PP framework, as it has been currently presented, has led to arguments on whether it constitutes an explanatory partner for embodied mind approaches (e.g., active externalism – Clark & Chalmers, 1998; situated agency – Noë 2004; enactivism – Varela, 1999), which is the intended job, or fails to do so by working as a premise of a too indirect (e.g., epistemic internalism, representationism) or too direct (as fantasy-like perception) mind-world relation (Froese & Ikegami, 2013; Anderson & Chemero, 2013; Paton *et al.*, 2003). This misleading interpretation of the PP framework has confounded the task of placing it in its proper philosophical niche. The present proposal questions whether this misleading interpretation stems from the inappropriate conflation of key conceptual terms in the theory, such as *inference*.

As stated before, at the core of the PP framework is the premise that what the brain does is to *infer* the causes of changes in its sensory inputs. Essentially, we make sense of the world (including ourselves – the agents) by making *inferences* based on what we already know. However, little consideration has hitherto been given to present a conceptual distinction between different types of inference (or prediction). We believe that such formal distinction might offer important insights into central questions in the theory, mainly on understanding the difference between knowledge-free and knowledge-rich types of inference (Anderson & Chemero, 2013; Sloman, 2013), and the separation and relatedness of the surprise/surprisal pair.

5 Research method

The first specific objective concerns the exploration of the key aspects of human high-level cognition. What makes the human mind unique? The understanding of the human unique ability to co-construct (material and symbolic) shared worlds involves the elucidation of the following questions:

(a) How central is the role played by human language in the complex interplay between internal biological resources and external non-biological resources? What differs human lexical and syntactic components from other animals language structure? We humans have the unique ability to produce and make use of language as material symbols which participate in cognitive processes.

There are two compelling evidences regarding the uniqueness of human language presented by Chomsky & Berwick we intend to explore (Berwick & Chomsky, 2010; Berwick *et al.*, 2013). First, human lexical terms do not refer solely to something external and mind-independent as it happens in symbolic systems of other animals. Humans do not only classify or categorize aspects of the environment which are present - we create internal concepts which are updated also during "offline" imagination which adds context beyond the moment. As Jane Goodall describes, for the chimpanzees "the production of a sound in the *absence* of the appropriate emotional state seems to be an almost impossible task". Chimpanzees can classify and categorize, but since their language is keyed to emotional states which are automatically linked to objects or events in the external world they would never be able to mentally resume last week's unsolved problem or construct counterfactuals. Second, the hierarchical syntactic structure of human language has no equivalent in any nonhuman species. It allows us to produce an infinite range of meaningful structured expressions by simply merging syntactic elements. Even chimpanzees that have been taught sign language lack this combinatorial ability.

It is a very strong viewpoint that language should be understood as a cognitive computational system which primary role is internal as 'instrument for thought'. Communication should be seen as an element of language externalization, and secondary to its key function as 'Inner tool'. This viewpoint states that language did not evolve as a communication system - the overwhelming use of language is internal, for thought. It is supported by the majority of biolinguists (e.g., Fodor, Chomsky, Tomasello, and Hurford), and its evolutionary aspects are well developed by Chomsky & Berwick in several articles (under Chomsky's 'Strong Minimalist Thesis'). Thus, these two unique characteristics of human language (the ability to construct internal concepts and to combine them by means of a recursive generative procedure) made possible the emergence of new forms of reasoning which have promoted the development of abstract and productive thinking.

(b) How central is the role played by collective intentionality in human social learning and cognition? Here we aim to show that the attribution of collective intentionality to nonhuman primates is unjustified. We will investigate the developmental roots of cooperation in social cognition, by means of a comparative, cross-species methodology driven by an enactivist approach - which includes aspects of the brain, the body and the environments of an organism, in a temporal and spatial extension (Tomasello *et al.*2005; Penn *et al.*, 2008; Premack, 2010; Rosati *et al.*, 2014).

In cognitive anthropology, one attempt to relate collective intentionality and co-constructed shared worlds can been found in studies which depicts human cultural practices as particular unfoldings of temporality.

(Roepstorff *et al.* 2010; Malafouris, 2013). We will assess whether the concept of chronesthesia, or mental time travel (Tulving, 2002), could be incorporated into the high-level human cognition within the PP framework. It refers to the ability that allows humans to be constantly aware of the past and the future. Chronestesia could be a key element of human-specific shared material and social cognition, the reason why its exploitation in the present project can help us to understand the establishment of 'collective priors', or sets of collective expectations that shape social perception and guide action.

(c) Are the differences in language and social cognition between nonhuman primates and humans qualitative or quantitative? We will review the recent literature on the neurobiological roots of language and comparative cognitive psychology to investigate whether the mechanisms by which human and nonhuman animals access higher-order, relational capabilities are qualitatively distinct (Premack, 2007; Penn *et al.*, 2008; Berwick *et al.*, 2013; Bolhuis *et al.*, 2014; Bornkessel-Schlesewsky *et al.*, 2015).

II

The second specific objective focuses on a proposal of a formal distinction between different types of inferential processes. Here, we will use Charles S. Peirce deep investigation on the properties and mutual relations amongst his three types of inference (deduction, induction and abduction) as a starting point (Peirce, 1931–1958). We want to demonstrate that the PP framework can encompass all three types of inference as unconscious inferential process and conscious inferential reasoning.

It has been proposed that modes of reasoning are mediated by model-sparse and/or model-rich forms of cortical processing (Anderson & Chemero, 2013; Sloman, 2013). However, despite providing a description of both, we intend to focus only on model-rich, high-level (or agent-level) perception and cognition. Also, only inductive and abductive inferences should be triggered by a mismatch between actual and predicted sensory signals (agent-level surprise), and both should engage mechanisms of model-rich cognition. According to Peirce, both induction and abduction are modes of ampliative (or content-increasing) reasoning, but they do it in distinct ways. Induction is constrained by considerations of similarity, being its conclusion simply the generalization of the content of the premises. The extra content generated here does not carry new ideas. On the other hand, abduction, as Peirce puts it "infers very frequently a fact not capable of direct observation". It is not constrained by similarity and its inferred conclusion is totally dissimilar to the facts that suggested it in the first place. As Psillos (2011) sums it up, "it seems reasonable to claim that the chief difference between H [hypothesis, which latter would be called abduction] and I [induction] is that Induction involves what we have called 'horizontal extrapolation' [extra content by generalization], whilst Hypothesis involves (or allows for) 'vertical extrapolation', viz., hypotheses whose content is about unobservable causes of the phenomena. Indeed, as has been stressed already, the very rationale for Hypothesis is that it makes possible the generation of new content or new ideas."

III

The third specific objective is an attempt to elucidate human-specific ways of reasoning and dealing with agent-level uncertainty (surprise). This will be done by taking into account the specific-human features in language and social cognition and related topics considered beforehand [(1) and (2)].

Here, we intend to investigate the problem of whether abductive inference may be seen as a human-specific mode of inference. Nonhuman animals (at least those ones that meet the anatomical and physiological requirements necessary to be put under the PP framework umbrella) also exhibit sophisticated, model-rich Bayesian prediction. Evidence exists for the formation of at least three sorts of categories (or classes) by nonhuman animals: perceptual or basic level category (classification of stimuli according to the shared physical properties); associative category or functional equivalence (association between an object and its several symbolic representations); and relational category (relationship among stimuli through comparison, such as larger than or better than) (Zentall *et al.*, 2014). However, the inferential process by which they access implicitly present information sometimes necessary for categorization and for dealing with agent-level uncertainty should be inductive. This is to say that animals also are capable of doing inference to the best explanation, but are limited to induction, i.e., by projecting observed regularities to unobserved instances (all snakes are poisonous, all red fruits are sweet). We aim to explore the argument that the way nonhuman animals expand their knowledge should be solely by generalization/analogy and not by developing new ideas (unifying unrelated domains of knowledge).

IV

The fourth specific objective concerns the use of human-specific ways of reasoning and dealing with agent-level uncertainty (surprise) to understand the nature of human intellect and creativity.

As said before, there is strong evidence that only humans can generate internal concepts by combining online and offline contextual properties. This allows ever increasing levels of conceptual abstraction. Abstract concepts are fluid – they have flexible boundaries and can adapt to unexpected circumstances (it has been described as 'cluster concept' by Gasking and as 'concept with blurred edges' by Wittgenstein). Even when we are not able to understand a puzzling situation (surprise) at the moment it occurs, it is amazing how humans can revisit the same thought over and over and generate several abductive hypothetic explanations. By doing this, humans combine concepts from different domains, play with existing knowledge and take fresh look at situations. Fluidity of concepts can blur human cognition into human creativity. [NOTE: It is important to notice that (1) we encounter puzzlement not only as a perceptual implausibility but also during mind-wandering, and (2) the plausibility of the abductively generated beliefs should be subjected to further testing by ways of induction and deduction, admitting subjective prior probabilities in an iterative and corrective logical process].

Abductive reasoning is always initiated by the awareness of an anomalous situation and could be seen as a movement towards transformation of the agent's conceptual space in order to overcome agent-level uncertainty and keep internal entropy at a manageable level. I think this movement could be well visualized in the terms Sloman (2007) describes as logical geography and logical topography, from Gilbert Ryle's notion of logical geography. Logical geography means "the network of relationships between a collection of connected concepts -- as they are *currently* used". On the other hand, logical topography regards not only connections in existing usage, but also a variety of possible types of states. The space defined by the logical topography could be explored in several ways revealing the different conceptual relationships that can occur within it (or different logical geographies). Thus, one logical topography supports several possible logical geographies. Abduction modifies the agent's conceptual space by moving from actual to possible concept usage defined in connection with the subject matter that is being investigated (triggered by uncertainty). This process should be facilitated by concept abstractness and should oscillate between stable, robust, beliefs and the new beliefs that will eventually substitute the old ones. [NOTE: In his paper, Sloman mentions the role of abduction as a non-empirical task in what he calls deep science. However, abduction is pervasive and it is invariably employed in everyday life as well as in science.]

V

The fifth specific objective is a proposal of neurobiological mechanisms that could underlie human-specific, high-level inferential processes, which includes creative invention, one of the stepping stones of human cumulative cultural evolution.

First of all, it is very important to mention that there have been quite robust empirical findings at the system-level (mainly in the visual system) that support the existence both of surprise (error) signals and of the hierarchical interplay between expectation and surprise, the PP framework core propositions. The evidence supporting knowledge-sparse mechanisms comes from human functional magnetic resonance imaging (fMRI) (den Ouden *et al.* 2009; Alink *et al.*, 2010; Smith & Muckli, 2010), electroencephalographic (EEG) (Summerfield *et al.* 2011; Wacongne *et al.*, 2012), and magnetoencephalographic (MEG) (Todorovic *et al.* 2011; Wacongne *et al.*, 2011) recordings. In contrast, little is known about knowledge-rich ways of minimizing agent-level surprisal. This is where the PP framework becomes very open-ended (Roepstorff, 2013).

We would like to propose a neurobiological mechanism underlying the abductive inferential process, and it is my opinion that we may gain insight if we look at the level of activity of neuronal ensembles. Construction of brain graphs from structural and functional neuroimaging data is beginning to reveal the brain network topology as a combination of high clustering and high efficiency in a modular and hierarchical organization which is able to deliver both specialized (short-range, local integration within modules) and distributed (long-range, global integration) information processing (Bullmore & Sporns, 2012; van den Heuvel & Sporns, 2013). High cognitive function can be described as global integration of local, densely intra-connected, modules and

depends on coupling parameters (connection strengths and timing) between modules. According to the global workspace framework (Baars, 2012), the mechanism underlying awareness of a particular perceptual element is the emergence of synchronized oscillations in large ensembles of distributed modules which integrates content that is relevant (dominant) at a given moment or context. Such an integrated workspace can rapidly exchange information between many specialized modules that are now part of a globally distributed network. Workspaces are formed and disintegrated on demand (cognitive load) in a dynamic process.

The neuronal ensembles compose probabilistic generative models of incoming sensory input. This is the Information that is being distributed in the workspace – information that is integrated and adjusted (updated) by tuning edge strength (or synaptic efficacy) between and within modules. Thus, information is encoded in a distributed and probabilistic manner across neuronal ensembles that bind together the properties of a lexical concept. The global availability of information distributed in the ensemble is subjectively experienced so as to become aware of something (e.g., a concept).

Sloman's logical geography and logical topography of the conceptual space could be used as a link between abductive reasoning (put as transformation of the conceptual space) and the modulation of neuronal ensembles by changing the strength of their synchronous activity. We aim to demonstrate that abductive inference is a cognitive process that evaluates which are the critical variables of the environmental data (input signal) by comparing and combining it with internal representations (exploration of the agent's conceptual space). These critical variables correspond to the essential features that are to be bound together to became an explicit new concept. During this process, some original perceptual data is discarded (explained away) and others are maintained and integrated to existing internal concepts. I think this process is mediated by an interplay between local and long-range synchronization. As the long-range synchronization between high-level association cortices increases (binding-related activity), information coming from local bottom-up regions starts being more and more suppressed (explained away). The representation of a concept in part derives from other representations depending on how similar are the ensembles of neurons that they evoke (associative memory). At this point, Annette Karmiloff-Smith's representational redescription model (Karmiloff-Smith, 1992) will be used to understand how the conceptual distance between two concepts might not always be conscious. It is possible that, in some cases, two conceptual representations overlap but the correlation between them has never been explicitly noticed - they may have been encoded at different times and under different situations. A global neuronal ensemble representing a concept does not only carry information about the content of lower-level module representations but also about the agent's confidence in those representations. Thus, an unexpected situation could be described as a situation with a great level of neural competition and ambiguity due to a low degree of precision and confidence in the interpretation of sensory input (lack of a clear dominant hypothesis). This means decreasing synaptic gain and blunting neuronal representations, which could reveal associations that are implicitly present within the possibilities of the brain logical topography and would enables us to go beyond the explicit conceptual knowledge and generate new ideas.

In addition to how synchronization within a workspace could operate in the generation of abstract concept, we also intend to investigate whether the neural circuitry of nonhuman primates is insufficient to support the fundamental computational architecture that gives rise to higher-order language processing and reasoning. We will consider the human prefrontal cortex (PFC) circuitry re-organization in relation to the nonhuman primate PFC. Two large-scale networks that occupy the expanded portions of PFC and other association cortices - the default mode network and the frontoparietal network - are good candidates to underpin human-specific forms of prediction and control as their functional properties (including dynamic, context-specific modulation of other networks) enable the emergence of key characteristics of high-order inference:

(a) Combination of online (actual signal - environment) and offline (imagination – internal representations) contextual properties in a cyclic process;
(b) Capacity of high-level conceptual abstraction;
(c) Integration of emotional- and reward-valuation into the cognitive model (anxiety and long-term planning which are good candidates in sustaining the abductive reasoning process (Hirsh *et al.*, 2012).

The default mode network is involved in spontaneous cognition (mind-wandering) and monitoring the environment (low-level focus of attention for unexpected events). The frontoparietal network is involved in deciding and planning actions not only in known context, but its flexibility allow it to generalize its function to many types of novel and uncertain situations – where further analysis of the stimulus in relation to mnemonic/internal representations is required (Spreng *et al.*, 2013).

So far, comparative analysis of network organization has failed to reveal a structural or functional equivalent of the human frontoparietal network in macaques and chimpanzee (it seems that the rostrolateral PFC has no homologue in the macaque brain). A recent study presented strong evidence suggesting that human high-order relational thinking relies on the frontoparietal network (Vendetti & Bunge, 2014). A homologous set of default mode network-like areas has been reported in macaques, however, with considerable lower activation in the mediai frontal cortex. MPFC activation has been related to processing of self-projection on goal states through mental construction of alternate realities (Mantini & Vanduffel, 2013). The default mode network can flexibly couple with the frontoparietal network if the environment demands combination of internally-focused cognition (imagining future accomplishments) and goal-directed cognition (problem-solving to attain personal goals) (Spreng *et al.*, 2010).

Default mode network could be seen as the one at the top of the hierarchy as it continuously monitor the environment and suppress exogenous and endogenous stimuli: bottom-up prediction-errors from its subordinates systems (Friston & Carhart-Harris, 2010) - i.e., frontoparietal network for exogenous and also the limbic/paralimbic system (emotion & reward valuation) for endogenous signals.

In the present proposal, we would like to investigate whether these two large-scale networks might correspond to major hubs, working as the resilient core of the "global neuronal workspace" cited above (Baars, 2012). If so, they could serve to integrate several pieces of information coming from other large-scale networks through dynamic connectivity between networks forming discrete spatiotemporal patterns of activity. Exploring this integrative mechanism could be extremely useful in understanding how creative thinking can arise from combining neural patterns into ones that are potentially novel and useful.

6 References

Alink A., Schwiedrzik C. M., Kohler A., Singer W., Muckli L. (2010) Stimulus predictability reduces responses in primary visual cortex. Journal of Neuroscience 30:2960–2966.

Anderson M.L., Chemero T. (2013) The problem with brain GUTs: conflation of different senses of "prediction" threatens metaphysical disaster. Behavioral and Brain Sciences 36(3):204-205.

Baars B.J. (2002) The conscious access hypothesis: origins and recent evidence. Trends in cognitive sciences 6:47-51.

Bar M. (2009) The proactive brain: memory for predictions. Philosophical Transactions of the Royal Society of London B: Biological Sciences 364(1521):1235-1243.

Bar M., Neta M. (2008) The proactive brain: Using rudimentary information to make predictive judgments'. Journal of Consumer Behavior 7:319-330.

Berwick R., Chomsky N. (2011) The biolinguistic program: the current state of its development. In *The Biolinguistic Enterprise*, ed. A. M. Di Sciullo, C. Boeckx, pp. 19–41. Oxford University Press.

Berwick R.C., Friederici A.D., Chomsky N, Bolhuis J.J. (2013). Evolution, brain, and the nature of language. Trends in Cognitive Sciences 17(2):89-98.

Bolhuis J. J., Tettersall I., Chomsky N., Berwick R. C. (2014) How could language have evolved? PLOS Biology 12, e1001934.

Bornkessel-Schlesewsky I., Schlesewsky M., Small S.L., Rauschecker J.P. (2015) Neurobiological roots of language in primate audition: common computational properties. Trends in Cognitive Sciences 19(3):142-150.

Bullmore E., Sporns O. (2012) The economy of brain network organization. Nature Reviews Neuroscience 13:336–349.

Carhart-Harris R. L., Friston K. (2010) The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. Brain 133(4):1265-1283.

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

Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36(3):181-204.

Clark A., Chalmers D. (1998) The extended mind. Analysis 58:10–23.

Dennett D. C. (1978) Beliefs about beliefs. Behavioural and Brain Sciences 1:568–570.

den Ouden H. E. M, Friston K. J., Daw N. D., McIntosh A. R., Stephan K. E. (2009) A dual role for prediction error in associative learning. Cerebral Cortex 19:1175–1185.

Friston K. (2003) Learning and inference in the brain. Neural Networks 16(9):1325– 1352.

Friston K. (2005) A theory of cortical responses. Philosophical Transactions of the Royal Society of London B: Biological Sciences 360(1456):815–836.

Friston K. (2011) Embodied inference: Or I think therefore I am, if I am what I think. In *The implications of embodiment (Cognition and Communication)*, ed. W. Tschacher, C. Bergomi, pp. 89–125. Imprint Academic.

Friston K. (2010) The free-energy principle: A unified brain theory? Nature Reviews Neuroscience 11(2):127–138.

Froese T., Ikegami T. (2013) The brain is not an isolated "black box," nor is its goal to become one. Behavioral and Brain Sciences 36(3):213-214.

Geisler W. S., Kersten D. (2002) Illusions, perception and Bayes. Nature Neuroscience 5(6):508–510.

Gregory R. L. (1980) Perceptions as hypotheses. Philosophical Transactions of the Royal Society of London B 290(1038):181–197.

Griffiths T.L., Kemp C., Tenenbaum J. B. (2008) Bayesian models of cognition. In *The Cambridge Handbook of Computational Psychology*, ed. R. Sun, pp. 59–100, Cambridge University Press.

Helmholtz, H. 1860/1962 Handbuch der physiologischen optic, ed. Southall J. P. C., vol. 3. New York: Dover (English trans.).

Hirsh J. B., Mar R. A., Peterson, J. B. (2012) Psychological entropy: A framework for understanding uncertainty-related anxiety. Psychological Review 119 (2):304–20.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Lee T. S., Mumford D. (2003) Hierarchical Bayesian inference in the visual cortex. Journal of Optical Society of America, A 20(7):1434–1448.

Malafouris K. (2013). *How Things Shape the Mind: A Theory of Material Engagement*. Cambridge, MA: MIT Press.

Mantini D., Vanduffel W. (2013) Emerging roles of the brain's default network. Neuroscientist 19(1):76-87.

Noë A. (2004) *Action in perception*. MIT Press.

Oaksford M., Chater N. (2009). Précis of Bayesian rationality: the probabilistic approach to human reasoning. Behavioral and Brain Sciences 32:69-120.

Paton B., Skewes J., Frith C., Hohwy J. (2013) Skull-bound perception and precision optimization through culture. Behavioral and Brain Sciences 36(3):222.

Peirce C. S. (1931-1958) Collected Papers of Charles Sanders Peirce, ed. C. Hartshorne & P. Weiss (volumes 1-6) and A. Burks (volumes 7 and 8), Cambridge MA: Belknap Press.

Penn D.C., Holyoak K.J., Povinelli D.J. (2008) Darwin's mistake: explaining the discontinuity between human and nonhuman minds. Behavioral and Brain Sciences 31(2):109-130.

Premack D. (2010) Why humans are unique: three theories. Perspectives on Psychological Science, 5:22–32.

Psillos S. (2011) An explorer upon untrodden ground: Peirce on abduction. In *The handbook of the history of logic,* ed. D. Gabbay, S. Hartmann, J. Woods, Vol. 10: inductive logic, pp. 117–151. Elsevier B.V.: Oxford.

Rao R. P. N., Ballard D. H. (1999) Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2(1):79–87.

Roepstorff A. (2013) Interactively human: Sharing time, constructing materiality. Behavioral and Brain Sciences 36(3):224-225.

Rosati A.G., Wobber V., Hughes K., Santos L.R. (2014) Comparative developmental psychology: how is human cognitive development unique? Evolutionary Psychology, 12(2):448-73.

Sloman A. (2013) What else can brains do? Behavioral and Brain Sciences 36(3):230-231.

Sloman, A. (2007). Two Notions Contrasted: 'Logical Geography' and 'Logical Topography' (Variations on a theme by Gilbert Ryle: The logical topography of 'Logical Geography'.) (Tech. Rep. No. COSY-DP-0703). Birmingham, UK: School of Computer Science, University of Birmingham. Available from (http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0703)

Smith F. W., Muckli L. (2010) Nonstimulated early visual areas carry information about surrounding context. Proceedings of the National Academy of Sciences USA 16:20099–20103.

Spreng R. N., Stevens W. D., Chamberlain J. P., Gilmore A. W., Schacter D. L. (2010) Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. Neuroimage 53(1):303–317.

Spreng R. N., Sepulcre J., Turner G. R., Stevens W. D., Schacter D. L. (2013) Intrinsic architecture underlying the relations among the default, dorsal attention, and fronto-parietal control networks of the human brain. Journal of Cognitive Neuroscience 25:74–86.

Summerfield C., Wyart V., Johnen V. M., De Gardelle, V (2011) Human scalp electroencephalography reveals that repetition suppression varied with expectation. Frontiers in Human Neuroscience 5:67.

Thagard P. (2009) Why cognitive science needs philosophy and vice versa. Topics in Cognitive Science 1(2):237-254.

Todorovic A., van Ede F., Maris E., de Lange F. P. (2011) Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: An MEG study. Journal of Neuroscience 31:9118–9123.

Tomasello M., Carpenter M., Call J., Behne T., Moll H. (2005) Understanding and Sharing Intentions: The Origins of Cultural Cognition. Behavioural and Brain Sciences 28:675-735

Tulving, E. (2002) Chronesthesia: conscious awareness of subjective time. In *Principles of Frontal Lobe Function*, ed. D. T. Stuss & R. T. Knight, pp. 311–325. Oxford University Press.

van den Heuvel M. P., Sporns O. (2013) Network hubs in the human brain. Trends in cognitive sciences 17(12):683-696.

Varela F. J. (1999) The specious present: A neurophenomenology of time consciousness. In *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science*, ed. J. Petitot, F. J. Varela, B. Pachoud, J.M. Roy, pp. 266–317. Stanford University Press.

Vendetti M. S., Bunge S. A. (2014) Evolutionary and Developmental Changes in the Lateral Frontoparietal Network: A Little Goes a Long Way for Higher-Level Cognition. Neuron 84(5):906-917.

Wacongne C., Changeux J.P., Dehaene S. (2012) A neuronal model of predictive coding accounting for the mismatch negativity. J Neuroscience 32(11):3665-3678.

Wacongne C., Labyt E., van Wassenhove V., Bekinschtein T., Naccache L., Dehaene S. (2011) Evidence for a hierarchy of predictions and prediction errors in human cortex. Proceedings of the National Academy of Sciences USA, 108(51):20754-20759.

Yardley H., Perklovsky L., Bar M. (2012). Predictions and incongruency in object recognition: A cognitive neuroscience perspective. In *Detection and identification of rare audiovisual cues. Studies in computational intelligence series*, ed. D. Weinshall, J. Anemüller, L. van Gool, Vol. 384, pp. 129–136. Berlin Heidelberg: Springer Publishing.

Zentall T. R., Wasserman E. A., Urcuioli P. J. (2014) Associative concept learning in animals. Journal of the Experimental Analysis of Behavior 101:130–151.