The Collective Experience of Empathic Data Systems

ICT-258749

Deliverable 1.1

Theory of Unified Human Experience Report 1: Abstract Model of Dataflows Underlying Telepresence

---

Authors	Anil Seth, Paul Verschure, Sid Kouider, Andrea Caria

Version	f1.0
Date	19.10.2011
Classification	Restricted
Contract Start Date	01.09.2010
Duration	48 months
Project Co-ordinator	Goldsmiths, University of London
File Name	CEEDs_D1.1_f1.0_UoS

Project funded by the
European Community under
the Seventh Framework Programme

Consisting of:

No	PARTICIPANT NAME	S.N.	COUNTRY
1	Goldsmiths, University of London	GOLD	UK
2	Universitat Pompeu Fabra	UPF	ES
3	University of Sussex	UOS	UK
4	Informatics and Telematics Institute	ITI	GR
5	Eberhard Karls Universitaet Tuebingen	EKUT	DE
6	Universität Augsburg	UAU	DE
7	University of Teesside	TEESSIDE	UK
8	Università degli Studi di Padova	UNIPD	IT
9	Max Planck Gesellschaft zur Foerderung der Wissenschaften E.V.	MPG	DE
10	Ecole Normale Superieure	ENS Paris	FR
11	Budapesti Muszaki Es Gazdasagtudomanyi Egyetem	BME	HU
12	Universitat Politecnica de Catalunya	UPC	ES
13	Università di Pisa	UDP	IT
15	Electrolux Italia SpA	ELECTROLUX	IT
16	Leiden University	UL	NL
18	Helsingin Yliopisto	UH	FI

This document may not be copied, reproduced, or modified in whole or in part for any purpose without written permission from the CEEDs Consortium. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.

All rights reserved.

Responsible of the document: UOS, UPF, ENS, EKUT

Defined Contributors to the document:Anil Seth (UOS), Paul Verschure (UPF)

Other Expected Contributors:Keisuke Suzuki (UOS), Sid Koudier (ENS), Andrea Caria (EKU), Jurgen Jost (MPG)

Document History

VERSION	ISSUE DATE	BY	CONTENT AND CHANGES
f1.0 Final	<19/10/2011>	UoS	Final edits

Executive Summary

This document reports progress in WP1 of CEEDS, the primary scientific workpackage of the project. The key Year 1 deliverable is *an abstract theoretical model of the dataflows, i.e., the functional and mechanistic basis, that underpin* ***transparency and presence*** *in conscious experience*. Transparency and presence are key features of normal waking consciousness, and key objectives of immersive technologies such as CEEDs. Simply put, presence refers to the subjective sense of reality of the self and the environment, and transparency refers to the observation that we 'see through' our mental representations of the environment, directly experiencing their content. Presence and transparency are therefore closely coupled concepts, referred to as TP. A theoretical model has been developed, based on hierarchically organized interoceptive predictive coding, and is described in the following. This is combined with a detailed psychophysical analysis of the prediction cascade implemented in the Distributed Adaptive Control 7 attention architecture. This document also summarizes progress in other parts of WP1 consistent with the effort expended by the various partners.

We note that activity in WP1 began in month 6; and due to recruitment delays the postdoctoral fellow dedicated to this the theoretical modelling deliverable, based at UOS, was in position only by Sep 12 2011.

©
Copyright 2010-2014 CEEDS Consortium

Table of Contents

©
Copyright 2010-2014 CEEDS Consortium

List of Figures

©
Copyright 2010-2014 CEEDS Consortium

1 WP1.1: Theoretical and computational modelling

The Year 1 deliverable for WP1 falls under this task: 'Development of an abstract model of the dataflows underlying telepresence'. The full model is described in a draft paper entitled: **An interoceptive predictive coding model of conscious presence**. The paper, written by UOS, is attached as an Appendix to this report. The paper abstract is given below.

We describe a theoretical model of the neurocognitive mechanisms underlying conscious presence in Appendix A. The model is based on interoceptive prediction error and is informed by predictive models of agency, general models of hierarchical predictive coding in cortex, the role of the anterior insular cortex in interoception and emotion, and cognitive neuroscience evidence from studies of virtual reality and of psychiatric disorders of presence, specifically depersonalization/derealization disorder. The model associates presence with successful 'explaining away' by top-down predictions of interoceptive signals evoked by afferent sensory signals and by autonomic regulatory signals. The model connects presence to agency by allowing that predicted interoceptive signals will depend on whether afferent sensory signals are determined, by a parallel predictive-coding mechanism, to be self-generated or externally caused. Anatomically, we identify the (right) anterior insular cortex as the likely locus of the relevant neural mechanisms. Our model integrates a broad range of previously disparate evidence, makes specific predictions for conjoint manipulations of agency and presence, offers a new view of emotion as interoceptive inference, and represents a step towards a mechanistic account of a fundamental phenomenological property of consciousness.

It is standard in the discussion on consciousness to distinguish the *hard* problem of phenomenal consciousness or the fundamental problem of qualia (Nagel 1974) from the *easy* problems (Chalmers 1995). The hard problem deals with the putative paradox of how intrinsically subjective states can be subject to objective description. With respect to the easy problems a number of core principles underlying consciousness and qualia have emerged that we have combined in the Grounded Enactive Predictive Experience model of consciousness or GEPE. These principles are:

1: qualia are grounded in the experiencing **physically instantiated self** e.g (Nagel 1974; Metzinger 2003)

2: qualia are enacted in the **sensori-motor coupling** of the agent to the world e.g. (O'Regan and Noe 2001) (Heed, Grundler et al. 2011)

3: qualia are maintained in the coherence between sensori-motor **predictions** of the agent and the dynamics of the interaction with the world e.g. (Hesslow 2002; Grush 2004; Barsalou 2008); (Merker 2005).

4: qualia combine high levels of **differentiation** with high levels of **integration** and thus reflect highly informative discriminations among extremely large repertoires of possible states (Tononi and Edelman 1998; Tononi 2008; Seth 2009)

5: qualia depend on both highly parallel, distributed and implicit factors and metastable, continuous and unified explicit factors comprising a **Global Workspace** hypothesis (GW) e.g. (Baars 1988; Dehaene, Sergent et al. 2003; Kouider and Dehaene 2007).

The CEEDS project aims at realizing a model of consciousness in order to understand how we can optimize human data exploration. The guiding modelling framework for this effort is the, so called, Distributed Adaptive Control architecture.

©
Copyright 2010-2014 CEEDS Consortium

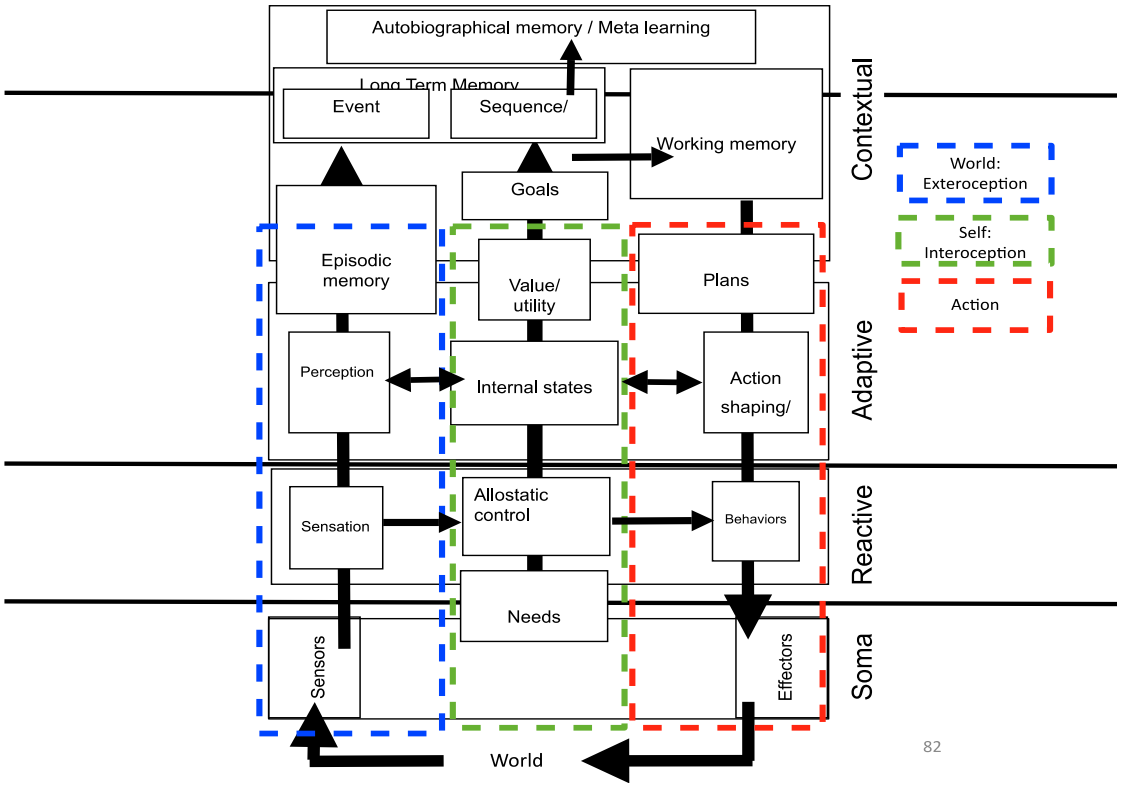


Fig. 1 - *The DAC architecture*

At the level of the *Soma* the direct interfaces to the world are defined combined with the needs of the organism. The *Reactive* layer endows a behaving system with a prewired repertoire of reflexes, low complexity stimuli and responses that enable it to display simple behaviors. The activation of any reflex, however, also provides cues for learning that are used by the *Adaptive* layer via representations of internal states, i.e. valence and arousal. This second layer provides the mechanisms for the adaptive classification of sensory events and the reshaping of responses or in other words to construct a state space of sensing and acting. The sensory and motor representations formed at the level of adaptive control provide the inputs to the *contextual* layer that acquires, retains, and expresses sequential representations using systems for short- and long-term memory that store sequences of sense-act couplets. DAC distinguishes three columns of organization: The sensation and perception of the world, the perception of self and their integration in goal oriented action. The well-established Distributed Adaptive Control – DAC architecture (P. Verschure, 2003) provides a key platform for the modelling work because it is predicated on representations of the environment in terms of sensorimotor contingencies that emerge through prediction. DAC (P. F. M. J. Verschure et al., 2003). DAC is a standard in the domains of "new" artificial intelligence and behaviour robotics (Arkin, 1998; Cordeschi, 2002; Hendriks-Jansen, 1996; McFarland & Bosser, 1993; Pfeifer & Scheier, 1999). It has been shown that the DAC architecture displays optimal performance equivalent to formal models of human decision making (P. Verschure & Althaus, 2003).

In Mathews et al (In Press) we have elaborated the counter current prediction cascades that mediate between the layers of DAC. In particular we have identified the notion of Validation Gate as a key guiding principle in organizing bottom-up and top-down attention. We have now tested this hypothesis by performing psychophysical experiments with human subjects. In this case we show that the stimulus features that are relevant for eye-movements, as indicators of attention, and conscious decisions vary and can become decoupled dependent on cognitive load. The model and results are reported in Appendix B.

References

Zenon Mathews, Sergi Bermúdez i Badia, Paul Verschure (In Press) PASAR: An Integrated Model of Prediction, Anticipation, Sensation, Attention and Response for Artificial Sensorimotor Systems. Information Sciences.

©
Copyright 2010-2014 CEEDS Consortium

©
Copyright 2010-2014 CEEDS Consortium

2. WP1.2: Testing transparency and presence in dataworlds

No Year 1 deliverables are required under this task. However we describe here relevant progress, corresponding to effort expended.

UOS have identified a new technology, **substitutional reality** (SR), which can be used to test transparency and presence. SR involves a combination of head-mounted VR goggles (e.g., the Vuzix VR920) with a forward facing camera, head-motion tracking, and a spherically panoramic video camera (e.g., the LadyBug 3, as used by Google street-view). The key to SR is that a subject can look freely around an environment while the experimenter toggles between live feed from the forward-facing camera and a prerecorded feed using the LadyBug 3 camera. By manipulating the prerecording we can experimentally test presence in a variety of environments and with a variety of manipulations. The new hire Suzuki is experienced in SR. This study provides a basis for experiments in Task 1.8.

Discussions between UOS and UPF have identified this task as a possible high priority target for generating rapid experimental results. XIM infrastructure has been adapted in order to run these experiments.

©
Copyright 2010-2014 CEEDS Consortium

3 WP1.3: VR technology for creating 'virtual synesthesia'

No Year 1 deliverables are required under this task. However we describe here relevant progress, corresponding to effort expended.

UPF have continued development of an integrated audio-visual stimulus generation and interaction system using Unity and SMuSE (LeGroux & Verschure, 2009). In addition, interfaces to technologies generated in WP2 have been defined.

Le Groux, S. L.& Verschure P. F. M. J. (2009). Situated Interactive Music System: Connecting Mind and Body Through Musical Interaction. Proceedings of the International Computer Music Conference.

©
Copyright 2010-2014 CEEDS Consortium

4 WP1.4: Autonomic responses underlying emotional salience

No Year 1 deliverables are required under this task. However we describe here relevant progress, corresponding to effort expended.

UPF have conducted evaluation of physiological responses (GSR, breathing, ECG) to a standardized library of emotional images and a standardized facial expression dataset. The evaluation has including the use of the BCI2000 analysis tools. UPF have also carried out a comparison of wearable physiological sensing systems with standardized methods. These steps constitute a necessary prerequisite to the experimental phase of Task 1.4.

One of the key challenges of affective computing is to extend the expression of emotion to machines. Research in this field has focused mainly on embodied machines that can reproduce verbal or non-verbal cues such as facial movements and gestures. However, most machines we interact with in our daily life are non-humanoid. For this reason, the question we are addressing in our study is whether it is possible to express emotions or internal states using non-symbolic visual cues in non-humanoid artifacts.

We designed animated, highly parametrized graphic motifs on the interactive floor of the eXperience Induction Machine (XIM), an immersive room equipped with a number of sensors and effectors constructed to conduct experiments in mixed reality and we asked the participants to assess the emotions attributed to these abstract visual cues. Our results revealed a clear relationship between the parameters applied to the motifs (color, complexity, speed) and the internal states perceived by the participants. In addition, we discovered a relationship between these parameters and the participants' behaviour.

Our study represents a first step in the direction of the empirical assessment of emotional expression in non-humanoid artifacts. A manuscript summarizing these results is in preparation.

©
Copyright 2010-2014 CEEDS Consortium

5 WP1.5: Decoding of neural activity predicting intention and discovery

No Year 1 deliverables are required under this task. However we describe here relevant progress, corresponding to effort expended.

ENS has been focusing on the recruitment of an engineer specialized in EEG measurements. Leonardo Barbossa has joined our team in May 2011 in order to focus on sources localisation and multivariate analysis methods (e.g., support vector machine) as described in the DoW. ENS has also started to review the different toolboxes available for performing these kinds of analyses. ENS has been in particular studying how the upcoming protocols will be constrained by what can and what cannot be done with these kinds of methods. ENS is working on the design of an EEG protocol relying on support-vector machine and navigation in a virtual maze (common to task 1.6). In this protocol, participants are required to decide which direction to take (left or right) at the end of alleys in the virtual maze and scalp EEG are used to infer, at this stage, which direction will be taken before a button press by subjects. Neural markers of upcoming motor intentions will consist of lateralised readiness potentials. ENS has started to build the computational tools to analyze support-vector machine algorithms for the classification of this type of neural events.

©
Copyright 2010-2014 CEEDS Consortium

6 WP1.6: Subliminal stimulation

No Year 1 deliverables are required under this task. However we describe here relevant progress, corresponding to effort expended.

ENS has acquired the Poser© software for 3D figure design and animation, and started the construction of avatar faces with different emotional expressions, in order to present pleasant vs. unpleasant stimuli (e.g., happy vs. fearful face) as described in the DoW. ENS has also been working on the construction of a 3D labyrinth using the Blender© software. ENS is now preparing the experiment to test whether navigation in a virtual maze can be biased through subliminal stimulation, using both behavioural (direction) and electrophysiological (EEG) indices. ENS is now working on connecting the 3D labyrinth with the gaze-contingent eye-tracking system and working on the parameters for an efficient gaze-contingent substitution of crowded navigational cues. ENS has been also working on setting up a partially immersive environment by acquiring a very large 3D screen in a fully obscure room, rather than a head-mounted system in order to run the upcoming experiments. This was mostly related to the difficulty of directly obtaining a head-mounted system combining also a high-resolution eye-tracker for gaze-contingent substitution. Until the later integration with the XIM 2 setup, the experimental tests for task 1.5 and 1.6 will be done with this partially immersive environment. ENS estimates that this setup, although partially immersive from a technical point of view, actually allows a strong feeling of presence and psychological immersion in the virtual maze.

Copyright 2010-2014 CEEDS Consortium

7 WP1.7: Neural correlates of spatiotemporal properties of presence

No Year 1 deliverables are required under this task. However we describe here relevant progress, corresponding to effort expended.

A review of the literature as well as discussions between EKUT, UOS, and UPF indicated that empirical assessment of the influence of predictive coding and free-energy minimization on the experience of presence would yield important insights for CEEDS in general and Task 1.1 in particular. EKUT adapted for fMRI experiment a previously tested psychophysical protocol provided by UPF (the validation gate experiment described in Task 1.1). This fMRI study would allow us to investigate the neural correlates of bottom-up and top-down processes of human conscious perception. The experiment requires high-quality recordings of eye-movements. EKUT performed preliminary tests of the eye-tracker system in the MR settings to assess high sampling frequency capability. Results indicated that eye-tracking is feasible at 120 Hz, but requires individual fine-tuning. EKUT started preliminary combined eye-movements and fMRI recordings during the validation gate experiment. Behavioral and fMRI data are currently being explored. The paradigm consists of a displacement detection task, in which subjects are required to detect occasional translational displacement of otherwise predictably moving dots within a field of moving dots. Participants are required to perform the task with three different level of cognitive load: (1 – low cognitive load) displacement detection task only, (2 – medium cognitive load) displacement detection task while they continuously recite aloud the alphabet, and (3 – high cognitive load) displacement detection task while they continuously recite aloud the alphabet in reverse order skipping every other letter. The three experiments were designed to modulate the cognitive load without affecting the perceptual load. Further control conditions for the cognitive load tasks during passive exposure to the validation gate stimuli will be required in case of standard GLM approach to cancel out potential enhanced activity in the frontal areas related to modulation of cognitive load *per se*. The fMRI investigation during such psychophysical experiment is supposed to reveal the dissociation of conscious and unconscious processes, and demonstrate that the brain employs multiple parallel anticipatory processes. In particular a differential involvement of frontoparietal circuits and thalamo-cortical connections are expected to play a central role for bottom-up and top-down processes of human conscious perception. In addition, functional connectivity analysis (e.g psychophysiological interaction analysis) might help to elucidate the mechanisms underlying brain anticipatory processes.

UPF has transferred its psychophysical paradigm to the setup of EKUT. The setup consists of 2 computer systems that had to be time synchronized and interfaced for data transfer. The first system, is the fMRI+eye tracker machine that log fMRI data and eye movements. The second and newly installed system is the stimulus machine that displays the visual stimulus and logs the subject's responses, i.e. the button presses. The stimulus application displaying N identical non-filled white circles on a black background was installed on a Mac machine running virtual Linux to test the fMRI and eye tracking synchronization. The stimulus presentation application was able to run at about 58 Hz, which is too slow for the final experiment but was enough for testing purposes. This machine will be upgraded to a Linux desktop machine, where the application is known to run at above 200Hz. The fMRI trigger was successfully captured by the stimulus application and the capture time is logged. The subject's response (button press) is also logged by the stimulus machine using the same USB input used for the fMRI trigger. The eye tracking application runs separately time synchronized with the fMRI machine. In summary the logged eye data can be used for analysis together with the stimulus data, the subject's conscious responses (button press) and fMRI data. Some minor issues noticed during the first trials were: data logging formatting error of the stimulus application, and occasional freezing of the stimulus application. The former has to be debugged and fixed and the latter could be an artefact of using virtual linux on a Mac machine.

©
Copyright 2010-2014 CEEDS Consortium

8 WP1.8: Instrumental learning of brain responses and perception

No Year 1 deliverables are required under this task. However we describe here relevant progress, corresponding to effort expended.

This task started in month 6. Discussions between EKUT and UOS indicated an important gap in the literature on subliminal perception and metacognition: While it is generally assumed that posterior parietal and lateral prefrontal cortices play a crucial role in conscious perception and metacognitive awareness, empirical evidence for this assumption is largely correlational. Recent efforts using transcranial magnetic stimulation have begun to shed light on the causal involvement of these areas in (non-)conscious perception. However, it remains to be shown that similar effects can be obtained using voluntary regulation of activity in prefrontal and posterior parietal regions. UOS has provided an experimental protocol which EKUT is currently adapting for a neurofeedback study using real-time fMRI. The paradigm requires subjects to discriminate between simple geometric shapes (e.g., square vs. diamond), presented under various levels of visual masking, while giving confidence ratings related to the accuracy of their judgments. The paradigms therefore allow analyses that separate metacognitive from unconscious judgments via a signal detection theory analysis. The relevance for CEEDS lies in the potential for measurable signals that indicate specifically a subject's metacognitive awareness about their behavior with respect to a particular stimulus.

EKUT performed a first real-time fMRI based neurofeedback pilot experiment to assess voluntary regulation of PFC, an important preliminary issue. Participants underwent a first functional localizer session to delineate the target regions in the prefrontal cortex followed by three rtfMRI regulation sessions where they were required to learn to increase and decrease BOLD activity in the left and right PFC. One participant was tested to learn regulation of correlational activity of PFC and PPC. Preliminary results indicate that self-regulation of PFC only is achievable through the combination of mental strategies (e.g. participants focused on their own thoughts) and contingent feedback in few training sessions (see Figure 2). Regulation of correlational activity of PFC and PPC might require additional effort and further training sessions (see Figure 3). EKUT is currently integrating the protocol provided by UOS (visual discrimination task as in Rounis et al. 2010) with the rtfMRI neurofeedback protocol. Such integration would allow EKUT to directly assess the effect of self-regulation of PFC on conscious perception and specifically to examine whether neurofeedback modulation of prefrontal and/or parietal activity selectively affects subjective but not objective decisions.

©
Copyright 2010-2014 CEEDS Consortium

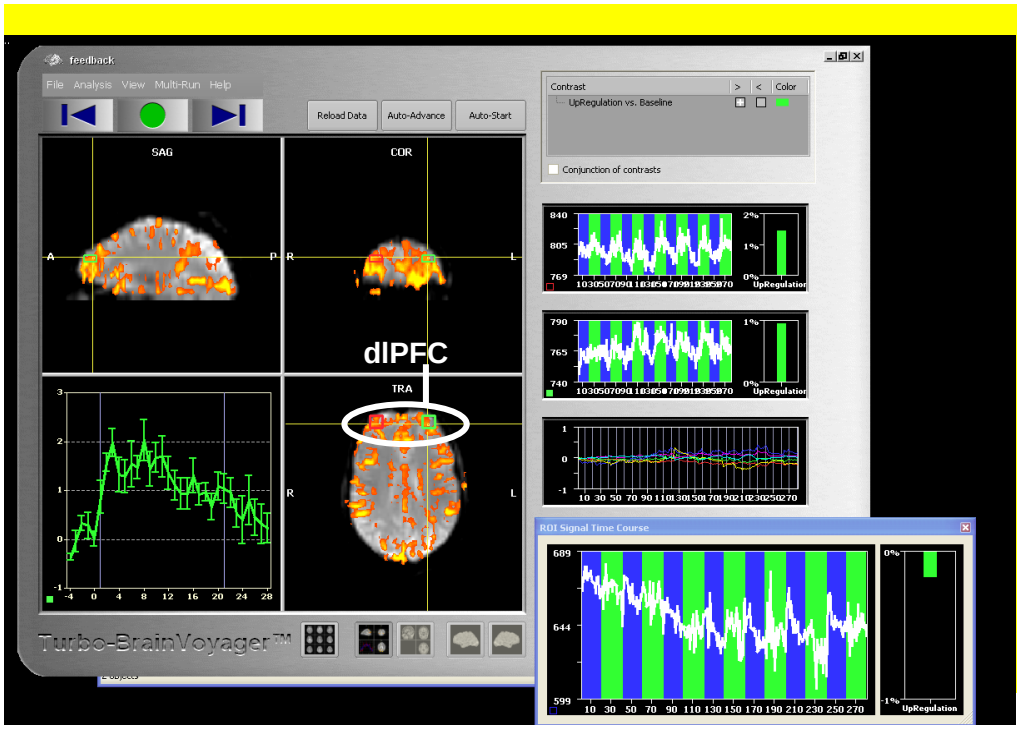


Fig. 2 - *Self-regulation of left (green box) and right (red box) dorso-lateral prefrontal cortex (dlPFC)*

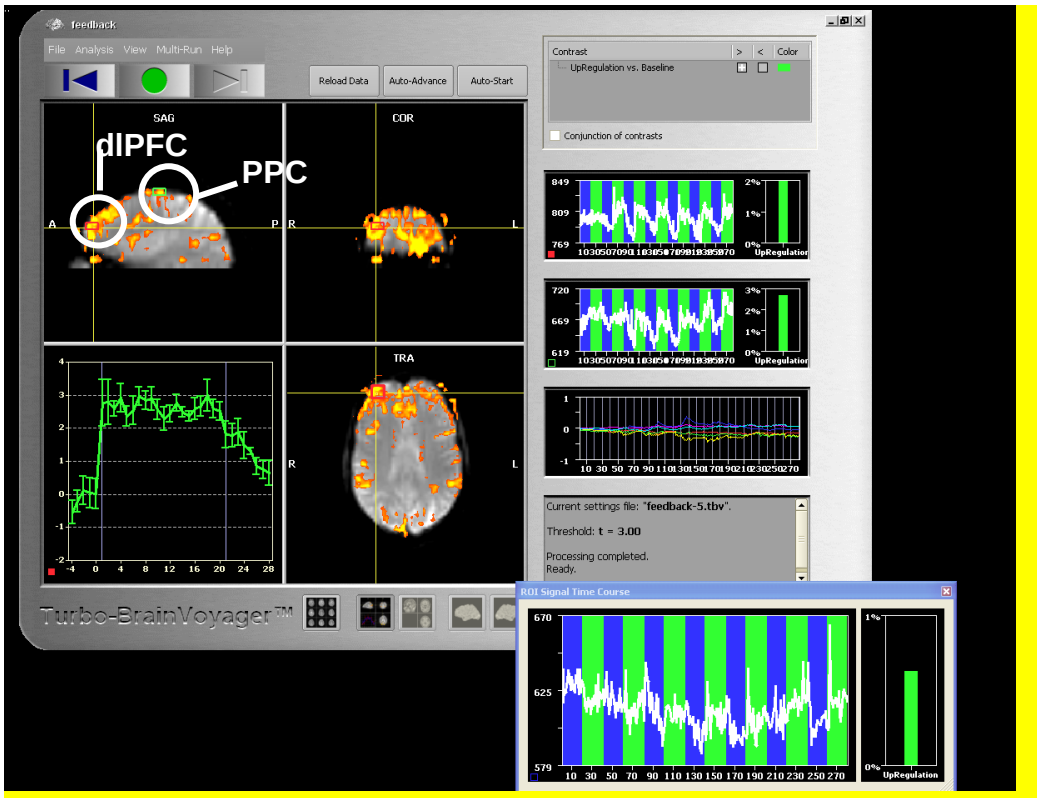


Fig. 3 - *Self-regulation of correlational activity between right dlPFC (red box) and right posterior parietal cortex (PPC, green box)*

Copyright 2010-2014 CEEDS Consortium

9 WP1.9: Integration

No Year 1 deliverables are required under this task.

This task has not started within the first year of the project.

©
Copyright 2010-2014 CEEDS Consortium

Appendix A

An interoceptive predictive coding model of conscious presence

Anil K Seth[1], **Keisuke Suzuki**[1]

[1] Sackler Centre for Consciousness Science and Department of Informatics,
University of Sussex, United Kingdom, BN1 9QJ

Correspondence:

Dr Anil K. Seth

Address: Sackler Centre for Consciousness Science, Department of Informatics, University of Sussex, Falmer, Brighton, BN1 9QJ, UK

Email: a.k.seth@sussex.ac.uk

Web: www.anilseth.com, www.sussex.ac.uk/sackler

Running title: Interoceptive predictive coding and conscious presence

Keywords: Presence, consciousness, depersonalization disorder, agency, interoception, insular cortex, virtual reality, predictive coding

Acknowledgements: AKS is funded by EPSRC Leadership Fellowship EP/G007543/1, by a donation from the Dr Mortimer and Theresa Sackler Foundation, and by EU FP7 CEEDS. The work reported here is a contribution to the CEEDS project.

Abstract:

We describe a theoretical model of the neurocognitive mechanisms underlying conscious presence. The model is based on interoceptive prediction error and is informed by predictive models of agency, general models of hierarchical predictive coding in cortex, the role of the anterior insular cortex in interoception and emotion, and cognitive neuroscience evidence from studies of virtual reality and of psychiatric disorders of presence, specifically depersonalization/derealization disorder. The model associates presence with successful 'explaining away' by top-down predictions of interoceptive signals evoked by afferent sensory signals and by autonomic regulatory signals. The model connects presence to agency by allowing that predicted interoceptive signals will depend on whether afferent sensory signals are determined, by a parallel predictive-coding mechanism, to be self-generated or externally caused. Anatomically, we identify the (right) anterior insular cortex as the likely locus of the relevant neural mechanisms. Our model integrates a broad range of previously disparate evidence, makes specific predictions for conjoint manipulations of agency and presence, offers a new view of emotion as interoceptive inference, and represents a step towards a mechanistic account of a fundamental phenomenological property of consciousness.

©
Copyright 2010-2014 CEEDS Consortium

1.0 Introduction

In consciousness science and psychiatry, the concept of *presence* refers to the subjective sense of reality, of the world and of the self (Metzinger 2003; Sanchez-Vives and Slater 2005; Seth 2010). Presence is a characteristic of most normal healthy conscious experience, and perhaps because of this, has rarely been the focus of targeted neuroscientific inquiry (Sanchez-Vives and Slater 2005). However, selective disturbances of presence are manifest in psychiatric disorders such as depersonalization disorder (DPD, loss of subjective sense of reality of the self) and derealization (DR, loss of subjective sense of reality of the world) (Phillips, Medford et al. 2001; Sierra, Baker et al. 2005; Simeon, Kozin et al. 2008; Sierra and David 2011), indicating that presence is a phenomenological property underpinned by specific neurocognitive mechanisms. In virtual reality (VR), presence is used typically to refer to the subjective sense of being in a virtual environment (VE) rather than in the actual physical environment; an alternative, behavioural, interpretation is that presence is equivalent to 'successfully supported action' within the VE (Sanchez-Vives and Slater 2005). Despite the centrality of the presence concept in these domains, detailed theoretical models of the neural mechanisms responsible for presence are still lacking.

The conceptual overlap between the usages of presence within psychiatry and VR provides a unique opportunity to analyse the cognitive, neural, and environmental constraints governing its emergence. On one hand, studies of DPD/DR can help identify candidate neural mechanisms underlying presence in normal conscious experience; on the other, studies of VR can help identify how presence can be generated even in situations where it would normally be lacking. The objective of this paper is to integrate insights into presence from these different perspectives within a single theoretical framework and model.

Our framework is based on *interoceptive predictive coding* within the anterior insular cortex (AIC). Interoception refers to the perception of internal bodily states, whereas exteroception refers to perception of the environment via the classical sensory modalities (Craig 2003; Critchley, Wiens et al. 2004). Predictive coding is a powerful framework for conceiving of the neural mechanisms underlying perception, cognition, and action (Rao and Ballard 1999; Bubic, von Cramon et al. 2010; Friston 2010). Simply put, predictive coding models describe counterflowing top-down prediction/expectation signals and bottom-up prediction error signals. Successful perception, cognition, and action are associated with successful suppression ('explaining away') of prediction error. Applied to interoception, predictive coding implies that subjective feeling states are determined by predictions about the interoceptive consequences of autonomic regulatory signals and afferent sensory signals, extending James-Lange theories of emotion. Predictive coding models have previously been applied to agency (the sense that a subject's action is the consequence of his or her intention) which propose that disturbances of agency, for example in schizophrenia, arise from imprecise predictions about the sensory consequences of actions (Frith 1987; Blakemore, Smith et al. 2000; Synofzik, Thier et al. 2010; Voss, Moore et al. 2010). To our knowledge, predictive coding models have not so far been applied to presence, or to interoceptive awareness. Anatomically, we focus on the AIC because this region has been strongly implicated in interoception and in the generation of subjective feeling states (interoceptive awareness) (Critchley, Wiens et al. 2004; Craig 2009); moreover, AIC activity in DPD/DR is abnormally low (Phillips, Medford et al. 2001).

In brief, our model proposes that *presence is the result of successful 'explaining away' by top-down predictions of interoceptive signals evoked by afferent sensory signals and by autonomic control signals*. The model is motivated by several factors: (i) general models of hierarchically-organized predictive coding in cortex, following principles of Bayesian inference (Neal and Hinton 1998; Lee and Mumford 2003; Friston 2009; Bubic, von Cramon et al. 2010); (ii) the importance of insular cortex (particularly the AIC) in integrating interoceptive and exteroceptive signals, and in generating subjective feeling states (Critchley, Wiens et al. 2004; Craig 2009); (iii) suggestions and observations of prediction errors in insular cortex (Paulus and Stein 2006; Preuschoff, Quartz et al. 2008; Singer, Critchley et al. 2009; Bossaerts 2010); (iv) evidence of abnormal insula activation in DPD/DR (Phillips, Medford et al. 2001; Sierra and David 2011); (v) models of the subjective sense of 'agency' (and its disturbance in schizophrenia) framed in terms of predicting the sensory consequences of self-generated actions (Frith 1987; Synofzik, Thier et al. 2010; Voss, Moore et al. 2010; Frith 2011), and (vi) theory and evidence regarding the role of dopamine in optimizing the precision of prediction errors (Friston, Kilner et al. 2006; Fletcher and Frith 2009).

In the remainder of this paper, we first define the concept of presence in greater detail. We then introduce the theoretical model before justifying its components with reference to each of the areas just described. We finish by extracting from the model some testable predictions and by deriving implications for consciousness science generally.

©
Copyright 2010-2014 CEEDS Consortium

2.0 Presence in neuroscience, VR, and psychiatry

The concept of presence has arisen independently in different fields (VR, psychiatry, consciousness science) concerned with understanding basic features of normal and abnormal conscious experience. The concepts from each field partially overlap. In VR, presence has both a subjective and an objective interpretation. In the former, presence is understood as the subjective sense of being in a virtual environment while being transiently unaware of one's real location and of the technology delivering the sensory input and recording the motor output (Jancke, Cheetham et al. 2009); a more compact definition is simply 'the sense of being there' (Lombard and Ditton 1997). A second, more objective, interpretation is based instead on establishing a behavioural equivalence between virtual and real environments. As (Sanchez-Vives and Slater 2005) put it, "the key to the approach is that the sense of 'being there' in a virtual environment is grounded on the ability to 'do there'" (p.333). For present purposes the former sense, emphasizing the subjective, is the most relevant.

Within psychiatry, presence is often discussed with reference to its disturbance or absence in syndromes such as DPD/DR and in the early (prodromal) stages of psychoses. A useful characterization is provided by (Ackner 1954): "a subjective feeling of internal and/or external change, experienced as one of strangeness or unreality". A common description given by DPD/DR patients is that their conscious experiences of the self and the world have an 'as if' character; the objects of perception seems unreal and distant, or unreachable 'as if' behind a mirror or window[1]. DPD/DR patients do not normally suffer delusions or hallucinations, marking a clear distinction from full-blown psychoses such as schizophrenia; however, it is increasingly recognized that symptoms of DPD/DR may characterize prodromal stages of psychoses, potentially providing diagnostic, prognostic, and explanatory value. There is a clear overlap between the usages of presence in DPD/DR and VR in picking out the subjective feeling of 'being there'. In the former case the sense of 'being there' is lost, and in the latter, its generation is desired.

More generally, presence can be considered as a constitutive property of conscious experience. Following (Metzinger 2003), presence can be understood as the subjective sense of reality, and also in a temporal sense as marking a 'temporal window of presence' precipitating a subjective conscious 'now' from the flow of objective time. Here, we adopt the first interpretation. Although presence can vary in its intensity, it is a characteristic of conscious experiences generally and not an instance of any specific conscious experience (e.g., an experience of a red mug); in other words, presence can be considered to be a 'structural property' of consciousness (Seth 2009). Metzinger connects the concept of presence to that of *transparency*, which refers to the fact that our perceptions of the world and of the self appear direct, unmediated by the neurocognitive mechanisms that in fact give rise to them. Here, transparency and presence are treated synonymously.

3.0 An interoceptive predictive-coding model of conscious presence

Figure 1 depicts the functional architecture of the proposed model. The model consists of two primary components, an 'agency component' and a 'presence component', mutually interacting according to hierarchical Bayesian principles and connected, respectively, with a sensorimotor system and an autonomic/emotional system. Each main component has a 'state module' and an 'error module'. The core concept of the model is that *presence arises when interoceptive prediction signals are successfully matched to inputs so that prediction errors are suppressed ('explained away').*

[1] Sierra and David (2010) suggest instead that the 'as if' qualifier relates to the inadequacy of available verbal description of the DPD/DR phenomenology, and does not describe the phenomenology itself.

©
Copyright 2010-2014 CEEDS Consortium

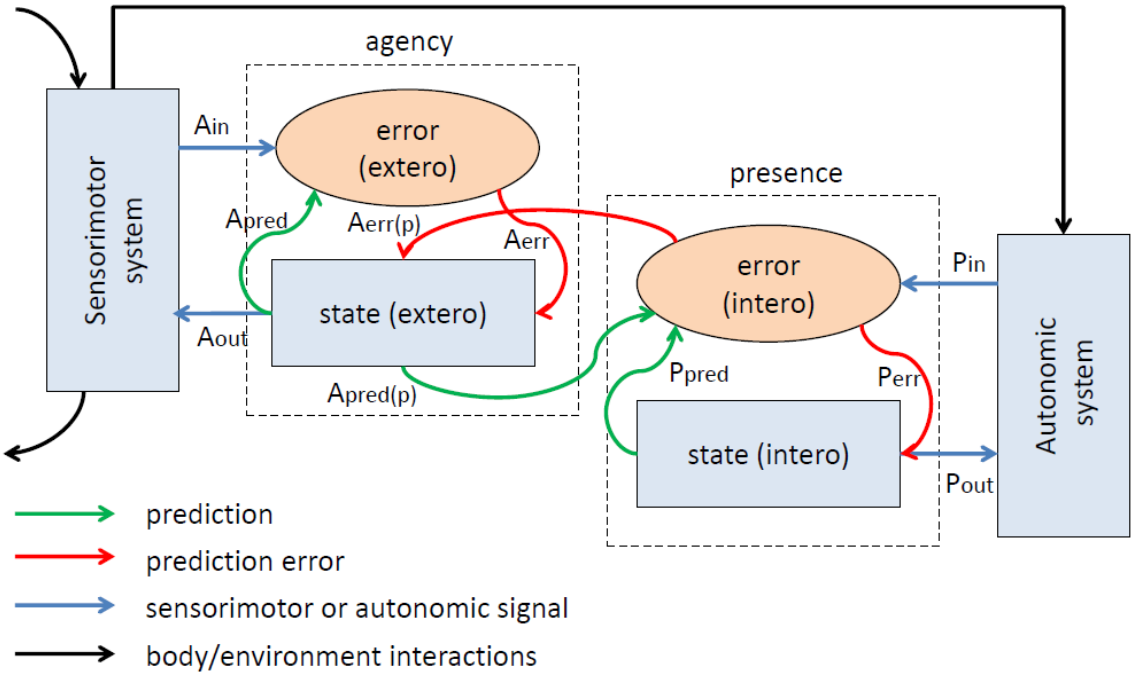


Figure 1.*An interoceptive predictive coding model of conscious presence. Both agency and presence components comprise state and error units; state units generate control signals (A_{out}, P_{out}) and make predictions (A_{pred}, P_{pred}, $A_{pred(p)}$) about the consequent incoming signals (A_{in}, P_{in}); error units compare predictions with afferents, generating error signals (A_{err}, P_{err}, $A_{err(p)}$). The agency component is hierarchically located above the presence component, so that it generates predictions about the interoceptive consequences of sensory input generated by its motor control signals.*

The agency component is based on Frith's well-established 'comparator model' of schizophrenia (Frith 1987; Blakemore, Smith et al. 2000; Frith 2011), recently extended to a Bayesian framework (Fletcher and Frith 2009). In the state module of this component, motor signals are generated which influence the sensorimotor system (A_{out}); these motor signals are accompanied by prediction signals (A_{pred}) which attempt to predict the sensory consequences of the motor actions via a forward model informed by efference copy and/or corollary discharge signals. Predicted and afferent sensory signals are compared in the error module, generating a prediction error signal A_{err}. In this model, the subjective sense of agency depends on successful prediction of the sensory consequences of action, i.e., suppression or 'explaining away' of the exteroceptive prediction error A_{pred}. Following (Fletcher and Frith 2009; Synofzik, Thier et al. 2010), disturbances in sensed agency arise not simply from predictive mismatches, but from *imprecise predictions* about the sensory consequences of action. Prediction errors per se are meaningless unless accompanied by some estimate of their precision[2]. Experimentally it has been shown that imprecise predictions prompt patients to rely more strongly (and therefore adapt more readily to) external cues, explaining a key feature of schizophrenic phenomenology in which actions are interpreted as having external rather than internal causes (Synofzik, Thier et al. 2010). Interestingly, the precision of prediction error signals has been associated specifically with dopaminergic activity (Fiorillo, Tobler et al. 2003), suggesting a proximate neuronal origin of schizophrenic symptomatology in terms of abnormal dopamine neurotransmission (Fletcher and Frith 2009). Prediction error precision also features prominently in recent models of hierarchical Bayesian networks, discussed in Section 4 (Friston, Kilner et al. 2006; Friston 2009).

[2] Precision can be understood as inverse variance and is critical in comparing distributions. Think of a standard t-test in statistics, in which differences in means can only be interpreted given estimates of the variances of the corresponding distributions.

Copyright 2010-2014 CEEDS Consortium

In the presence component, the autonomic system is driven both by afferent sensory signals and by internally generated control signals from the state module (P_{out}), modulating the internal physiological mileu[3]. The state module is responsible for the generation of subjective emotional states (feeling states) according to the principles of James and Lange, i.e., that subjective emotions arise from perceptions of bodily responses to emotive stimuli (Critchley, Wiens et al. 2004; Craig 2009). Extending these principles, in our model emotional content is determined by the nature of the predictive signals P_{pred}, and not simply by the 'sensing' of interoceptive signals per se (i.e., we apply the Helmoltzian perspective of perception as inference to subjective feeling states, see Section 4.1). As in the agency component, there is also an error module which compares predicted autonomic signals with actual autonomic signals P_{in} via a forward, giving rise to an autonomic or interoceptive prediction error P_{err}(Paulus and Stein 2006). We suggest that this comparison function has its anatomical locus in the AIC. The sense of presence is underpinned by a match between predicted and actual interoceptive signals; disturbances of presence, as in DPD/DR, arise because of disturbances in this predictive mechanism. Again, by analogy with the agency component (Fletcher and Frith 2009; Synofzik, Thier et al. 2010) we propose that these disturbances arise not because of faulty prediction or prediction error signals *per se*, but rather because of *imprecise* prediction signals $P_{pred.}$

Importantly, the two components are hierarchically interconnected such that the state module of the agency component generates predictions not only for sensorimotor signals (A_{pred}) but also for the presence component ($A_{pred(p)}$); correspondingly, the presence error module sends prediction error signals to the agency state module ($A_{err(p)}$) as well as to the presence state module. Thus, in this model, agency is functionally localized at a higher hierarchical level than presence. This arrangement requires an additional generative component which produces predicted interoceptive signals given the current state of both agency and presence components; we suggest that this integrative generative model is a key component of a core sense of selfhood, in line with recent hierarchical models of the self (Northoff and Bermpohl 2004; Feinberg 2011).

3.1 Brain basis of the model

The model implicates a broad network of brain regions for both the agency and the presence components. Neural correlates of the sense of agency have been studied extensively, mainly by manipulating spatial or temporal delays to induce exteroceptive predictive mismatches. Regions identified include motor areas (ventral premotor cortex, supplementary and pre-supplementary motor areas and basal ganglia), the cerebellum, the posterior parietal cortex, the posterior temporal sulcus, subregions of the prefrontal cortex, and the anterior insula (Haggard 2008; Nahab, Kundu et al. 2011). Among these areas the pre-supplementary motor area plays a key role in implementing complex, open decisions among alternative actions and has been suggested as a source of the so-called 'readiness potential' identified in the classic experiments of Libet on volition (Haggard 2008). The right angular gyrus of the inferior parietal cortex has been associated specifically with awareness of the discrepancy between intended and actual movements (Farrer, Frey et al. 2008).

The presence component also implicates a broad neural substrate (Figure 2). Areas potentially involved in interoceptive predictive coding include a range of brainstem (periaqueductal grey, locus coeruleus), subcortical (amygdala, substantia innominata, nucleus accumbens), and cortical (orbitofrontal, anterior cingulate, and insular) regions, forming a hierarchy [(Tamietto and de Gelder 2010), see Figure 2]. Among these areas, the anterior insular cortex (AIC) plays a key role in integrating interoceptive and exteroceptive signals, and in generating subjective feeling states. VR experiments that directly manipulate presence also implicate the AIC along with multiple visual, parietal, and prefrontal areas (see Section 6.1); DPD/R patients show hypoactivity in the AIC (see Section 5.1); the AIC is also differentially activated by changes in the sense of agency (Nahab, Kundu et al. 2011), providing a link between the agency and presence components of the model. We discuss the AIC in more detail in the following section.

[3] The concept of *allostasis*, which refers to anticipatory autonomic regulatory signals, may be relevant here. However, allostasis typically refers to hormonal rather than neural control, and anticipation is generally interpreted in a social rather than a predictive coding sense.

©
Copyright 2010-2014 CEEDS Consortium

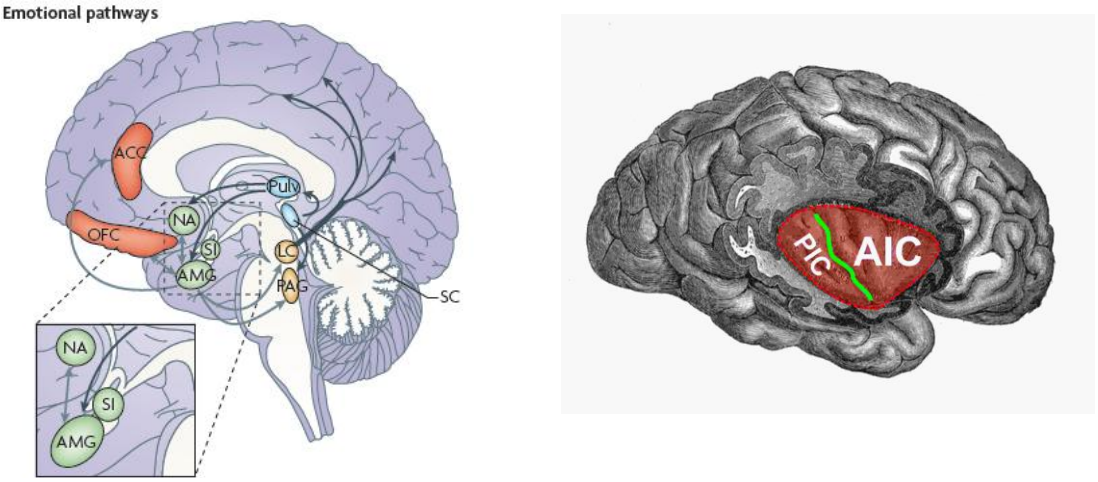


Figure 2*.Left panel: Regions hierarchically involved in interoceptive processing. Cortical regions: orbitofrontal (OFC), anterior cingulate (ACC), insula (not shown); subcortical: nucleus accumbens (NA), substantia innominata (SI), amygdala (AMG); brainstem: locus coeruleus (LC), periaqueductal grey (PAG). Also shown are the thalamic pulvinar nucleus (Pulv) and the superior colliculus (SC) which are involved in visual processing. From (Tamietto and de Gelder 2010). Right panel: The human insular cortex, partitioned into anterior (AIC) and posterior (PIC) subregions.*

3.2 The insular cortex, interoception, and emotion

Human insular cortex is a large and highly interconnected structure, deeply embedded in the brain (see Figure 2) (Augustine 1996; Medford and Critchley 2010; Deen, Pitskel et al. 2011). The insula has been divided into several subregions based on connectivity and cytoarchitectonic features (Mesulam and Mufson 1982; Mesulam and Mufson 1982; Mufson and Mesulam 1982; Deen, Pitskel et al. 2011). Craig has suggested that the anterior portion (AIC) is in a privileged position to receive interoceptive signals communicated via dedicated lamina-1 spinothalamocortical pathways (Craig 2002). Bidirectional connections with amygdala, nucleus accumbens, and orbitofrontal cortex further suggest that the AIC is well placed to receive input about (positive and negative) stimulus salience (Augustine 1996). The AIC and the anterior cingulate cortex (ACC) are often coactivated despite being spatially widely separated, forming a 'salience network' in conjunction with the amygdala and the inferior frontal gyrus (Seeley, Menon et al. 2007; Medford and Critchley 2010; Palaniyappan and Liddle 2011). The AIC and ACC are known to be functionally (Taylor, Seminowicz et al. 2009) and structurally (van den Heuvel, Mandl et al. 2009) connected. Interestingly, Craig has suggested that AIC-ACC connections are mediated via their distinctive populations of von Economo neurons, which have rapid signal propagation properties and are rich in dopamine D1 receptors (Hurd, Suzuki et al. 2001; Craig 2009). Generally AIC is considered as the (hierarchically) highest cortical integration site of interoceptive and exteroceptive signals.

The AIC has been implicated a very wide range of functions with a common factor in visceral representation, interoception, and emotional experience (Craig 2002; Critchley, Wiens et al. 2004; Craig 2009). Critchley and colleagues have suggested that the AIC instantiates interoceptive representations that are accessible to conscious awareness as subjective feeling states (Critchley, Wiens et al. 2004; Singer, Critchley et al. 2009). Evidence for this view comes from a study in which individual differences in a heartbeat detection could be predicted by AIC activation and morphometry (better performance associated with higher activation

---©
Copyright 2010-2014 CEEDS Consortium

and higher gray matter volume), suggesting a role for AIC in interoceptive awareness (Critchley, Wiens et al. 2004). Along similar lines, Craig proposes the AIC as a 'central neural correlate of consciousness' (Craig 2009) by drawing attention to the possible role of the AIC in the perception of flow of time.

Taken together, the evidence summarized so far highlights that the AIC is involved in interoceptive processing, and via interactions within the salience network, integration with exteroceptive signals and stimulus salience underlying subjective feeling states. More specific support for the role of the AIC as a comparator underlying the sense of presence, as proposed by our model, includes (i) evidence for predictive coding in the AIC; (ii) hypoactivation of AIC in patients with DPD/DR, and (iii) modulation of AIC activity by reported subjective presence in VR experiments. Before turning to this evidence we next discuss the principles of predictive coding in more detail.

4.0 Prediction, perception, and Bayesian inference

Following the early insights of von Helmoltz there is now increasing recognition of the importance of prediction, and prediction error, in perception, cognition, and action (Rao and Ballard 1999; Lee and Mumford 2003; Egner, Summerfield et al. 2008; Friston 2009; Summerfield and Egner 2009; Bubic, von Cramon et al. 2010). The concept of 'predictive coding' overturns classical notions of perception as a largely bottom-up process of evidence accumulation or feature detection driven by impinging sensory signals, proposing instead that perceptual content is determined by top-down predictive signals arising from multi-level generative models of the external causes of sensory signals, which are continually modified by bottom-up prediction error signals communicating mismatches between predicted and actual signals across hierarchical levels (see Figure 3). In this view, even low-level perceptual content is determined via a cascade of predictions flowing from very general abstract expectations about the world which constrain successively more detailed (fine-grained) predictions. We emphasize that in these frameworks bottom-up/feed-forward signals convey *prediction errors*, and top-down/feed-back signals convey *predictions* determining content.

Predictive coding models are now well established in accounting for various features of perception (Rao and Ballard 1999), cognition (Grush 2004) and motor control (Wolpert and Ghahramani 2000) [see (Bubic, von Cramon et al. 2010) for a review]. Two examples from visual perception are worth describing briefly. In an early study, (Rao and Ballard 1999) implemented a model of visual processing in which feedback connections conveyed predictions and feedforward connections conveyed prediction errors. When exposed to natural images, simulated neurons developed receptive field properties observed in simple visual cells (e.g., oriented receptive fields) as well as non-classical receptive-field effects such as 'end-stopping'. These authors pointed out that predictive coding schemes are computationally and metabolically efficient since neural networks learn the statistical regularities embedded in their inputs, reducing redundancy by removing the predictable components of afferent signals and transmitting only what is not predictable (residual errors). More recently, Egner and colleagues elegantly showed that the phenomenon of repetition suppression (decreased cortical responses to familiar stimuli) is better explained by predictive coding than by alternative explanations based on adaptation or sharpening of representations. Their key finding is that repetition suppression can be abolished when the local likelihood of repetitions is manipulated so that repetitions become unexpected (Egner, Summerfield et al. 2008).

©
Copyright 2010-2014 CEEDS Consortium

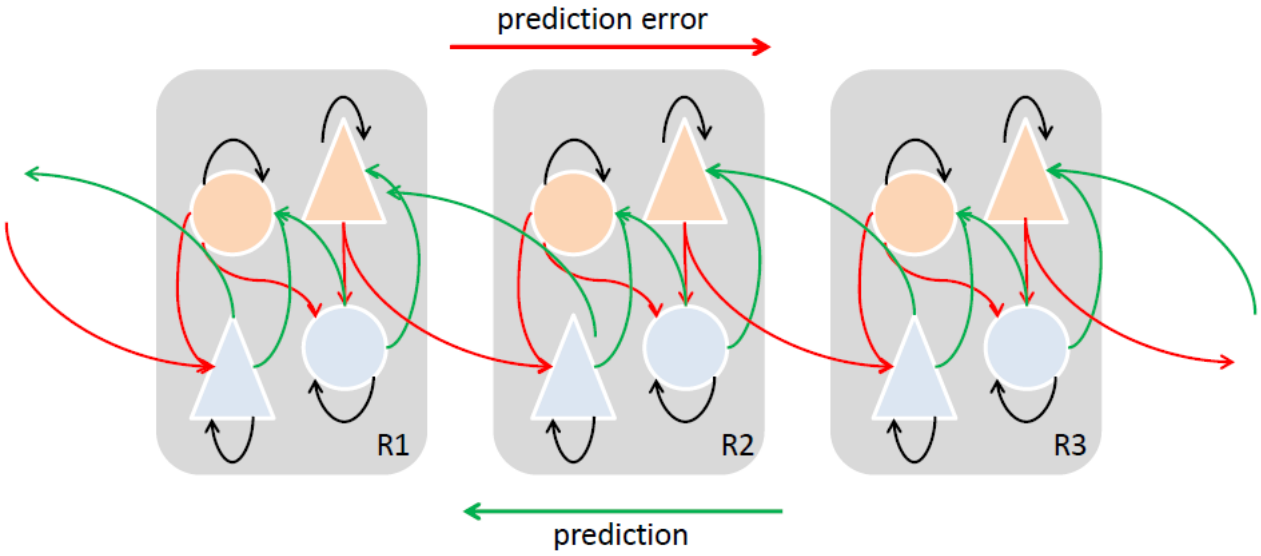


Figure 3.*A schematic of hierarchical predictive coding across three cortical regions; the 'lowest' (R1) on the left and the 'highest' (R3) on the right. Light blue cells represent state units, orange cells represent error units. Note that predictions and prediction errors are sent and received from each level in the hierarchy. Feedforward signals conveying prediction errors originate in superficial layers and terminate in deep (infragranular) layers of their targets, are associated with gamma-band oscillations, and are mediated by fast AMPA receptor kinetics. Conversely, feedback signals conveying predictions originate in deep layers and project to superficial layers, are associated with beta-band oscillations, and are mediated by slow NMDA receptor kinetics. Adapted from (Friston 2009). See also (Wang 2010).*

Theoretically, computational accounts of predictive coding have now reached high levels of sophistication (Dayan, Hinton et al. 1995; Rao and Ballard 1999; Lee and Mumford 2003; Friston, Kilner et al. 2006; Friston 2009). A key feature of these accounts is that they leverage the hierarchical organization of cortex to show how generative models underlying top-down predictions can be induced empirically via hierarchical Bayesian inference. Bayesian methods provide a computational mechanism for estimating the probable causes of data (posterior distribution) given the observed conditional probabilities of the data and associated priors; in other words, Bayes' theorem relates a conditional probability (which can be observed) to its inverse (which cannot be observed but knowledge of which is desired).

As illustrated in Figure 3, in these models each layer attempts to 'explain away' activity in the layer immediately below, as well as within the same layer, and passes prediction errors related to its own activity both internally and to the layer immediately above. From a Bayesian perspective, top-down influences constitute empirically-induced priors on the causes of their input. Advances in machine learning theory based on hierarchical Bayesian inference (Dayan, Hinton et al. 1995; Neal and Hinton 1998; Lee and Mumford 2003; Friston, Kilner et al. 2006; Friston 2009) show how these schemes may operate in practice. Attention has recently focused on Friston's 'free energy' principle (Friston, Kilner et al. 2006; Friston 2009) which shows how generative models can be hierarchically induced from data by assuming that the brain minimizes a bound on the evidence for a model of the data.[4] The machine learning algorithms able to perform this minimization are based on so-called 'variational Bayes' worked out by (Neal and Hinton 1998) among others; these algorithms have plausible neurobiological implementations (Friston, Kilner et al. 2006; Friston 2009).

[4] A key element of Friston's framework, distinguishing it from other predictive coding models, is that predictions act as a sort of self-fulfilling prophecy; that is, the brain does not only passively match expected to actual signals by changing generative models, but it actively attempts to fulfil existing predictions by selectively sampling the environment. Thus, for Friston, perception and action become unified within the overarching framework of free energy minimization.

Copyright 2010-2014 CEEDS Consortium

Interestingly, the precision of prediction error signals plays a key role in Friston's theory on the grounds that hierarchical models of perception require optimization of the relative precision of top-down predictions and bottom-up evidence. This process corresponds to modulating the gain of error units at each level, implemented by neuromodulatory systems. While for exteroception Friston emphasizes the role of cholinergic neurotransmission (Yu and Dayan 2005), linking this process to attention; for interoception, proprioception, and value-learning prediction error precision is suggested to be encoded by dopamine (Friston 2009). The role of dopamine in the present model is discussed further in Section 5.3.

It is important to emphasize that in predictive coding frameworks, predictions and prediction errors interact over rapid (synchronic) timescales providing a constitutive basis for the corresponding percepts, cognitions, and actions. This timescale is distinct from the longer (diachronic) timescales across which the brain might learn temporal relations among stimuli (Schultz and Dickinson 2000), or form expectations about the timing and nature of future events (Suddendorf and Corballis 2007).

In summary, the predictive coding framework may capture a general principle of cortical functional organization. It fluently explains a broad range of evidence (though a key prediction, that of distinct 'state' and 'error' neurons in different cortical laminae, remains to be established) and has attractive computational properties, at least in visual perception. It has been applied to agency, where by extending Frith's comparator model it suggests that disorders of agency arise from pathologically imprecise predictions about the sensory consequences of self-generated actions. However the framework has not yet been applied to interoception or to presence. We now turn to these issues.

4.1 Interoception as inference: A new view of emotion?

Predictive coding models of interoceptive processing have not yet been elaborated. Such a model forms a key component of our model of presence, offering a starting point for predictive models of interoception and emotion generally. Here, we sketch the outlines of interoception as inference.

Interoceptive theories of emotion originated with James and Lange who argued that emotions arise from interoception of physiological changes in the body. This basic idea has gained substantial currency over the last century, underpinning more recent frameworks for understanding emotion such as the 'somatic marker' hypothesis of Damasio (Damasio 2000), the 'sentient self' model (Craig 2002; Craig 2009), and 'interoceptive awareness' (Critchley, Wiens et al. 2004). Despite the substantial advances embedded in these frameworks, interoception remains generally understood along 'feed-forward' lines, similar to classical feature-detection or evidence-accumulation theories of visual perception. However, it has been known for nearly half a century that cognitively explicit beliefs about the causes of physiological changes can influence subjective feeling states. Schachter and Singer (1962) famously demonstrated that injections of adrenaline, proximally causing a variety of significant physiological changes, could give rise to either anger or elation depending on the concurrent context (an irritated or elated confederate), an observation formalized in their 'two factor' theory in which subjective emotions are determined by a combination of cognitive factors and physiological conditions (Schachter and Singer 1962).

Though it involves expectations, Schachter and Singer's theory falls considerably short of a full predictive coding model of emotion. Drawing a parallel with models of perception, predictive interoception would involve hierarchically cascading top-down interoceptive predictions counterflowing with bottom-up interoceptive prediction errors, with subjective feeling states being determined by the joint content of the top-down predictions across multiple hierarchical levels. In other words, according to the model emotional content is determined by a suite of hierarchically organized generative models predicting interoceptive responses to external stimuli and/or internal physiological control signals (Figure 4).

It is important to distinguish interoceptive predictive coding from more generic interactions between prediction and emotion. As already mentioned, predictive coding involves prediction at synchronic, fast time-scales, such that predictions (and prediction errors) are constitutive of mental content. By contrast, several previous studies have examined how predictions can influence emotion over longer, diachronic, timesecales. For example, (Gilbert and Wilson 2009) suggest that the brain instantiates simulations which are used to forecast the emotional consequences of future events. Similarly, (Ploghaus, Tracey et al. 1999; Porro, Cettolo et al. 2003; Ueda, Okamoto et al. 2003) identify brain networks involved in emotional predictions across time; the areas identified include prefrontal and anterior cingulate cortices.

©
Copyright 2010-2014 CEEDS Consortium

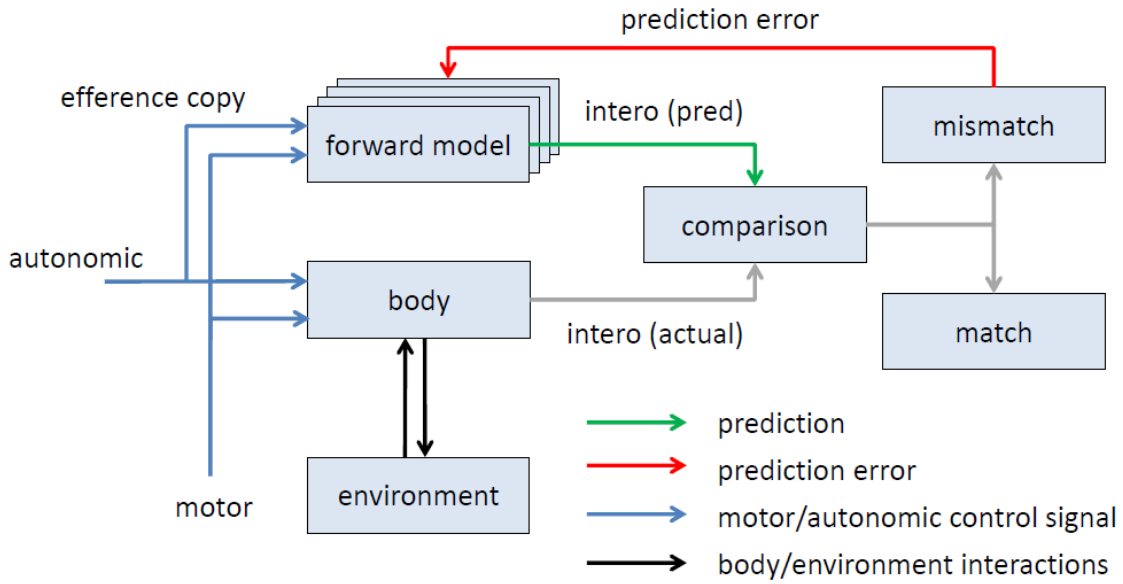


Figure 4.*Predictive coding applied to interoception. Motor control and autonomic control signals evoke interoceptive responses [intero(actual)] either directly (autonomic control) or indirectly via the musculoskeletal system and the environment (motor control). These responses are compared to predicted responses [intero(pred)], which are generated by hierarchically organized forward/generative models informed by motor and autonomic efference copy signals. The comparison, which may take place in AIC, generates a prediction error which refines the generative models. Subjective feeling states are associated with predicted interoceptive signals intero(pred). The figure is adapted from a general schematic of predictive coding in (Bubic, von Cramon et al. 2010).*

4.2 Predictive coding in the AIC

A key requirement of our model is that the AIC participates in interoceptive predictive coding. In a model of anxiety, (Paulus and Stein 2006) suggested that insular cortex compares predicted to actual interoceptive signals, with subjective anxiety associated with heightened interoceptive prediction error signals. In line with their model they found that highly anxious individuals showed increased AIC activity during emotion processing. Direct experimental evidence of insular predictive coding, though not specifically regarding interoceptive signals, comes from a study by (Preuschoff, Quartz et al. 2008) who recorded fMRI responses during a gambling task. They found activity encoding both predicted risk, and risk prediction error, in spatially separate subregions of the AIC (the former localizing to a region slightly more superior and anterior). The risk prediction error signal exhibited a fast onset, whereas the risk prediction signal exhibited a slow onset; these dynamics are consistent with their respective bottom-up and top-down origins in predictive coding models. Consistent with these findings, (d'Acremont, Lu et al. 2009) found in a neuroimaging study of the Iowa gambling task that AIC responses reflected risk prediction error; however, *reward* prediction errors were localized to the striatum. In an earlier study (Pessiglione, Seymour et al. 2006) found reward prediction error signals in both striatum and AIC; striatal responses were positively correlated with reward prediction error, but AIC responses were negatively correlated with reward prediction error and were evident during 'loss' trials only, possibly reflecting an aversive prediction error. Risk, reward, and interoception are closely linked, as underlined by theories of decision-making that emphasize the importance of internal physiological responses in supporting apparently rational behaviour (Bechara, Damasio et al. 1997; Damasio 2000). These links are also reflected in the structural and functional interconnectivity of AIC with orbitofrontal cortex and other reward-related and decision-making structures (see Section 3.2).

A different source of evidence for interoceptive predictive coding comes from exogenous manipulations of interoceptive feedback. In a heartbeat detection task, (Gray, Harrison et al. 2007) used false physiological feedback to experimentally induce a mismatch between predicted and actual interoceptive signals. They found that right AIC was more active for asynchronous (i.e., false) feedback, suggesting sensitivity to predictive mismatches and consistent with a role as a comparator. They also found, in the false feedback

Copyright 2010-2014 CEEDS Consortium

condition, an increased emotional salience attributed to previously unthreatening stimuli, consistent with a revision of top-down interoceptive predictions in the face of unexplained error.

In summary, there is accumulating evidence for predictive signalling in AIC relevant to risk and reward, as well as limited evidence for interoceptive predictive coding arising from false feedback evidence. Direct evidence for interoceptive predictive coding in the AIC has not yet been obtained and stands as a key test of the present model.

5.0 Disorders of agency and presence

A useful model should be able to account for features of relevant disorders. The 'comparator model' forming the present 'agency component' was in fact motivated by disorders of agency, namely schizophrenia (Frith 2011). As discussed, schizophrenic delusions of control are well explained by this model in terms of problems with kinematic and sensory aspects of the forward modelling component. Specifically, lower precision of exteroceptive predictions coincides with greater delusions of control, consistent with abnormal dopaminergic neurotransmission (Synofzik, Thier et al. 2010) [see also Section 5.3]. Other first-rank symptoms, for example thought insertion, are less well accounted for by current comparator models (Frith 2011). Here, we focus on the less extensively discussed issue of disorders of presence.

5.1 Depersonalization and derealisation

DPD/DR manifests as a disruption of conscious experience at a very basic, preverbal level, most colloquially as a 'feeling of unreality' which can be equally interpreted as the absence of normal feelings of presence. According to DSM-IV, DPD is characterized by "alteration in the perception or experience of the self so that one feels detached from and as if one is an outside observer of one's own mental processes". DPD/DR is common as brief transient phenomenon in normals, but may occur as a chronic disabling condition, either as a primary disorder, or secondary to other neuropsychiatric illness such as panic disorder, post-traumatic stress disorder, and depression. Recent surveys of clinical populations suggest that DPD/DR may be the third most common psychiatric symptom after anxiety and low mood, potentially affecting about 1-2% of the general population without gender bias. Although DPD/DR has a complex phenomenology, encompassing abnormalities of bodily sensation and emotional experience, it can be summarized as a psychiatric condition marked by the selective diminution of the sense of subjective reality of the self and world; a *presence deficit*. Notably, DPD/DR is often accompanied by alexithymia, which refers to a deficiency in understanding, processing, or describing emotions – more generally a deficiency of conscious access to subjective emotional states (Simeon, Giesbrecht et al. 2009)

Neuroimaging studies of DPD/DR, though rare, reveal significantly lower activation in AIC (and bilateral cingulate cortex) as compared to normal controls when viewing aversive images (Phillips, Medford et al. 2001).[5] It has been proposed that DPD is associated with a suppressive mechanism grounded in fronto-limbic brain regions, and in particular the AIC, which "manifests subjectively as emotional numbing, and disables the process by which perception and cognition become emotionally coloured, giving rise to a subjective feeling of unreality" (Sierra and David 2011); this mechanism could account for comorbid alexithymia as well.

In our model, DPD/DR could correspond to abnormal interoceptive predictive coding dynamics. Whereas anxiety has been associated with heightened prediction error signals (Paulus and Stein 2006), we suggest that DPD/DR is instead associated with imprecise interoceptive prediction signals P_{pred} in analogy with predictive models of disorders of agency (Fletcher and Frith 2009; Synofzik, Thier et al. 2010). Our model is consistent with that of (Paulus and Stein 2006). Chronically high anxiety could result from chronically elevated interoceptive prediction error signals, leading to overactivation in AIC resulting from failure to suppress these signals. In contrast, the imprecise interoceptive prediction signals associated with DPD/DR could result in hypoactivation of AIC since there is an excessive but undifferentiated 'explaining away' of error signals.

5.2 From hallucination and dissociation to delusion

[5] Interestingly DPD/DR patients showed *increased* insula activation (as compared to controls) when viewing neutral scenes.

Copyright 2010-2014 CEEDS Consortium

Both psychotic and dissociative disorders encompass disorders of perception and disorders of belief (delusions). In psychoses such as schizophrenia, disordered perceptions arise as hallucinations and delusions are characterized by bizarre or irrational beliefs such as thought insertion by aliens or government agencies (Maher 1974; Fletcher and Frith 2009). In dissociative disorders, disordered perceptions are characterized by negative symptoms as in DPD/DR; dissociative delusions may include conditions such as Cotard syndrome in which patients believe that they are dead. Fletcher and Frith have argued that, at least for positive symptoms in psychoses, a Bayesian perspective can accommodate hallucinations and delusions within a common framework (Fletcher and Frith 2009). In this view, a shift from hallucination to delusion reflects readjustment of top-down predictions within successively higher levels of cortical hierarchies, in successive attempts to explain away residual prediction errors.

Apparently, a similar explanation could apply to a transition from non-delusional interoceptive dissociative symptoms in DPD/DR to full-blown delusions in Cotard and the like. To the extent that imprecise predictions at low levels of (interoceptive) cortical hierarchies are unable to suppress interoceptive prediction error signals, imprecise predictions will percolate upwards through the hierarchy eventually leading not only to generalized imprecision across hierarchical levels but to re-sculpting of abstract predictive models underlying delusional beliefs.

The phenomenon of *intentional binding* is relevant here: actions and consequences accompanied by a sense of agency are perceived as closer together in time than they objectively are; conversely, if the consequence is not perceived as the result of the action, the events are perceived as more distant in time than they actually are (Haggard, Clark et al. 2002). Importantly, intentional binding has both a predictive and a retrospective component; (Voss, Moore et al. 2010) found that schizophrenic patients with disorders of agency showed stronger intentional binding than controls, with abnormalities most evident in the predictive component, reflecting indiscriminate (i.e., imprecise) predictions, consistent with (Synofzik, Thier et al. 2010). In contrast, prodromal subjects showed increased influence of both predictive and retrospective components, consistent with elevated prediction error signals (Hauser, Moore et al. 2011). These results suggest a process in which abnormal prediction errors lead, over time, to imprecise and eventually reformulated top-down predictions. With respect to dissociative symptoms, again a similar account may apply: Anxiety is often prodromal to DPD/DR, and anxiety has been associated with enhanced interoceptive prediction error (Paulus and Stein 2006).

5.3 The role of dopamine

Dopaminergic neurotransmission has been implicated at several points in the discussion so far, most prominently as encoding precisions within predictive coding. Here we expand briefly on the potential importance of dopamine for the present model.

Seminal early work relevant to predictive coding showed that dopaminergic responses to reward, recorded in the monkey midbrain, diminish when rewards become predictable over repeated (diachronic) stimulus-reward presentations suggesting that dopamine encodes a reward prediction error signal useful for learning (Schultz and Dickinson 2000; Chorley and Seth 2011). More recently, (Pessiglione, Seymour et al. 2006) found that reward prediction errors measured in humans via fMRI were modulated by dopamine levels. Modulation was most apparent in the striatum but was also slightly evident in the AIC. In considering this evidence it is important distinguish the phasic (diachronic) role of dopamine in signalling reward prediction error (Schultz and Dickinson 2000) from its (synchronic) role in modulating (or optimizing) the precision of prediction errors by modulating signal-to-noise response properties in cortical hierarchies (Friston 2009; Friston 2010). Although our model emphasizes the latter role, the learning function of dopamine may nonetheless play a role the transition from hallucination (or dissociation) to delusion. In this view, dopamine-modulated learning underlies the resculpting of generative models to accommodate persistently elevated prediction error signals (Corlett, Taylor et al. 2010). Dopaminergic neurotransmission may therefore govern the balance between (synchronic) optimization of precisions at multiple hierarchical levels (for both the agency and the presence components of our model) and the reformulation of predictive models themselves, with both mechanisms contributing to the transition from disordered perception and interoception to delusion. This account is also compatible with an alternative interpretation of dopaminergic signalling in identifying aspects of environmental context and behaviour potentially responsible for causing unpredicted events (Redgrave and Gurney 2006).

Abnormal dopaminergic neurotransmission has been observed in the ACC of individuals with schizophrenia (Dolan, Fletcher et al. 1995; Takahashi, Higuchi et al. 2006). Although, nothing appears to be known about dopaminergic processing in the insula in individuals with either DPD/DR or schizophrenia, the AIC has a rich abundance of dopamine D1 receptors (Williams and Goldman-Rakic 1998), and the insula and

Copyright 2010-2014 CEEDS Consortium

the ACC also boast relatively high levels of extrastriatal dopamine transporters, indicating widespread synaptic availability of dopamine in these regions. Dopamine is also a primary neurochemical in relation to numerous functions implicating the AIC, including novelty-seeking, craving, and noiciception (Palaniyappan and Liddle 2011). A more general role for dopamine in modulating conscious contents is supported by a recent study showing that dopaminergic stimulation increases both accuracy and confidence in reporting rapidly presented words (Lou, Skewes et al. 2011).

6.0 Testing the model

Testing the model requires (i) the ability to measure presence and (ii) the ability to experimentally manipulate predictions and prediction errors independently in the agency and presence components of the model.

Measuring presence remains an important challenge. Subjective measures depend on self-report and can be formalized by questionnaires (Lessiter, Freeman et al. 2001); however these measures can be unstable in that prior knowledge can influence the results (Freeman, Avons et al. 1999). Directly asking about presence could also induce (or reduce) experienced presence (Sanchez-Vives and Slater 2005). Alternatively, various behavioural measures can test for behavioural equivalence between real and virtual environments, however these measures are most appropriate for a behavioral interpretation of presence (Sanchez-Vives and Slater 2005). Physiological measures can also be used to infer presence, for example heart rate variability (Meehan, Insko et al. 2002) in stressful environments. Presence can be measured indirectly by the extent to which subjects are able to perform cognitive tasks based on memory and performance that depend on features of the VE (Bernardet, Valjamae et al. 2011), though again these measures may correspond to a behavioural rather than a phenomenal interpretation of presence. Finally, presence could be inferred by the ability to induce so-called 'breaks in presence' which would not be possible if presence was lacking in the first place (Slater and Steed 2000). In practice, a combination of the above strategies is likely to be the most useful.

Several technologies are available for experimentally manipulating predictions and prediction errors. Consider first manipulations of prediction error. In the agency component, these errors can be systematically manipulated by, for example, interposing a mismatch between actions and sensory feedback using either VR (Nahab, Kundu et al. 2011) or standard psychophysical (Blakemore, Frith et al. 1999; Farrer, Frey et al. 2008) methods. In the presence component, prediction errors could be manipulated by subliminal presentation of emotive stimuli prior to target stimuli (Tamietto and de Gelder 2010) or by false physiological feedback (Gray, Harrison et al. 2007). Manipulations of top-down expectations can be achieved by modifying the context in which subjects are tested. For example, expectations about self-generated versus externally-caused action can be manipulated by introducing a confederate as a potential actor in a two-player game (Farrer, Frey et al. 2008) or by explicitly presenting emotionally salient stimuli to induce explicit expectations of interoceptive responses.

A novel technology called substitutional reality (SR) may offer a unique approach to testing neurocognitive models of presence (Suzuki et al, submitted). In SR, seated subjects view their immediate environment through a VR headset with an attached camera providing real-time visual input. At a certain point, the real-time input is seamlessly switched to a prerecorded scene, taken using an omnidirectional video camera situated at the position of the subject. Omnidirectional recording coupled with registration of head movements allows realistic sensorimotor coupling to be preserved by always presenting the appropriate part of the prerecorded visual scene. Typically, naïve subjects do not notice the switch for quite some time. This technology therefore provides both a quantitative index of presence (i.e., time until realization of the prerecorded nature of the scene), the potential to selectively manipulate components of the scene and of the sensorimotor coupling, and the unique ability to record physiological signals under conditions (unlike in standard VR) in which subjects *really believe* they are experiencing the real world when in fact they are not.

6.1 Evidence from VR

Important constraints on neural models of presence come from experiments directly manipulating the degree of presence while measuring neural responses. VR technology, especially when used in combination with neuroimaging, offers a unique opportunity to perform these manipulations (Sanchez-Vives and Slater 2005). (Baumgartner, Speck et al. 2008) used VR to correlate the reported subjective sense of reality, induced by a virtual rollercoaster ride, with brain activity measured using fMRI. They uncovered a distributed network of

©
Copyright 2010-2014 CEEDS Consortium

brain regions elements which were both correlated, and anticorrelated, with presence (see Figure 4). Areas showing higher activity during high presence included extrastriate and dorsal visual areas, the superior parietal cortex, the inferior parietal cortex, some parts of the ventral visual stream, the premotor cortex, various thalamic, brainstem, and hippocampal regions, and notably the AIC (see arrow on Figure 4).

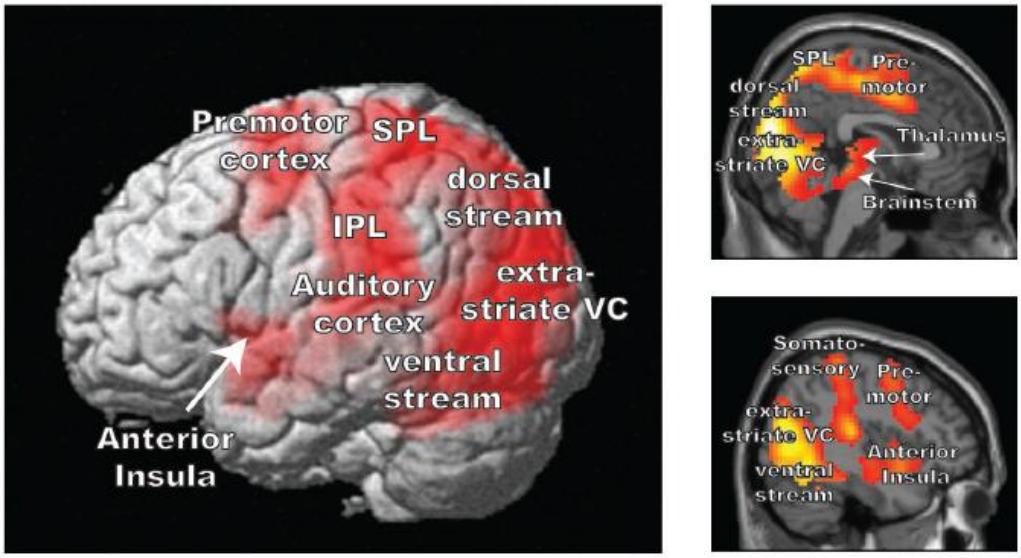


Figure 4.*Brain areas more strongly active during high than low presence. SPL: superior parietal lobe; IPL: inferior parietal lobe; VC: visual cortex. From (Jancke, Cheetham et al. 2009).*

Other relevant studies have examined behavioural correlates of presence as modulated by VR. (Aardema, O'Connor et al. 2010) found, in a non-clinical population, that immersion in a VE tended to enhance self-reported dissociative symptoms on subsequent re-exposure to the real environment, indicating that VR does indeed modulate the neural mechanisms underpinning presence. (Gutierrez-Martinez, Gutierrez-Maldonado et al. 2011) found that the ability to exert control over events in a VE substantially enhanced self-reported presence, consistent with our model in which predictive signals emanating from the agency component can influence presence.

VR has also been used to study the neural basis of experienced agency. For example, (Nahab, Kundu et al. 2011) use VR to manipulate the relationship between intended and (virtual) experienced hand movements, identifying via fMRI a network of brain regions the correlate with experienced agency. They identified both 'leading' and 'lagging' networks, with the former – the proposed locus of mismatch detection – involving right supramarginal gyrus [just anterior to the angular gyrus identified by (Farrer, Frey et al. 2008), see Section 3].

7.0 Related models

Here we briefly describe related theoretical models of presence and insula function. Models of agency have already been mentioned (Section 3) and are extensively discussed elsewhere (David, Newen et al. 2008; Fletcher and Frith 2009; Corlett, Taylor et al. 2010; Synofzik, Thier et al. 2010; Voss, Moore et al. 2010; Frith 2011; Hauser, Moore et al. 2011). With respect to presence, (Riva, Waterworth et al. 2011) interpret presence as "the intuitive perception of successfully transforming intentions into actions (enaction)". Their model differs from the present proposal by focusing on action and behaviour [therefore aligning more with an 'objective' stance on presence (Sanchez-Vives and Slater 2005)], by assuming a much greater phenomenological and conceptual overlap between presence and agency, and by not considering the role of interoception or the AIC. Baumgartner and colleagues propose a model of presence based on activity within the dorsolateral prefrontal cortex (DLPFC). In their model, DPLFC activity downregulates activity in the visual dorsal stream, diminishing presence (Baumgartner, Speck et al. 2008). Conversely, decreased DLPFC activity leads to increased dorsal visual activity, which is argued to support attentive action preparation in the VE as if it were a real environment. Supporting their model, bilateral DLPFC activity was anticorrelated with self-reported

©
Copyright 2010-2014 CEEDS Consortium

presence in their virtual rollercoaster experiment (Baumgartner, Speck et al. 2008). However, application of transcranial direct current stimulation to right DLPFC, decreasing its activity, did not enhance reported presence, thereby challenging the model (Jancke, Cheetham et al. 2009).

Models of insula function are numerous and cannot be covered exhaustively here. Among the most relevant are the model by (Singer, Critchley et al. 2009) who propose that AIC integrates exteroceptive and interoceptive signals with computations about their uncertainty. In their model, the AIC is assumed to engage in predictive coding for both risk-related and interoceptive signals, however in contrast to the present model no particular mechanistic implementation is specified. Another useful model is due to (Menon and Uddin 2010) who leverage the concept of a salience network (Section 3.2) to ascribe the insula with a range of functions including detecting salient stimuli and modulating autonomic and motor responses via coordinating switching between large-scale brain networks implicated in externally-oriented attention and internally-oriented cognition and control. To the best of our knowledge, no extant model proposes that the AIC engages in interoceptive predictive coding underlying conscious presence.

8.0 Summary

We have described a theoretical model of the mechanisms underpinning the subjective sense of presence, a basic property of normal conscious experience. The model is based on parallel predictive coding schemes, one relating to agency reflecting existing 'comparator' models of schizophrenia (Frith 1987; Frith 2011), and a second based on interoceptive predictive coding. The model operationalizes presence as the suppression of interoceptive prediction error, where predictions (and corresponding errors) arise (i) directly, via autonomic control signals, and (ii) indirectly, via motor control signals which generate sensory inputs. By analogy with models of agency (Synofzik, Thier et al. 2010), the sense of presence is specifically associated with the precision of interoceptive predictive signals, potentially mediated by dopamingergic signalling. The role of the agency component with respect to presence is critical; it furnishes predictions about future interoceptive states on the basis of a parallel predictive model of sensorimotor interactions. The joint activity of these predictive coding models may instantiate key features of an integrated self-representation. Converging evidence points to a key role for the AIC in instantiating predictive models, both for interoceptive and possibly exteroceptive signals, in line with growing opinion that the AIC is a core neural substrate for conscious selfhood (Critchley, Wiens et al. 2004). In addition, the model suggests a novel perspective on emotion, namely as interoceptive inference along Helmholtzian lines.

The model is consistent with known neurobiology and phenomenology of disorders of presence and agency. Presence deficits are particularly apparent in DPD/DR, which is known to involve hypoactivity in the AIC. Associating disturbances of presence with imprecise interoceptive predictions is also consistent with the frequently comorbid alexythmia exhibited by DPD/DR patients. Anxiety, often prodromal or comorbid with DPD/DR is also accommodated by the model in terms of enhanced prediction error signals, which when sustained could lead to the imprecise predictions underlying dissociative symptoms. The hierarchical predictive coding scheme may also account for transitions from disordered perception to delusion as predictive mismatches percolate to successively more abstract representational levels, eventually leading to dopaminergically governed resculpting of predictive models underlying delusional beliefs.

The model is amenable to experimental testing, especially by leveraging powerful combinations of VR and, more prospectively, SR. These technological developments need however to be accompanied by more sophisticated subjective scales reflecting more accurately the phenomenology of presence. A basic prediction of the model is that artificially induced imprecisons in interoceptive predictions should lead to diminished conscious presence and abnormal AIC activity; by contrast, simple elevation of interoceptive prediction error signals should lead instead to increased anxiety. As described in Section 6, these manipulations could be engendered either by preexposure to emotionally ambigious but salient stimuli or by direct pharmacological manipulation affecting dopaminergic neuromodulation in the AIC. A second basic prediction is that the AIC, as well as other areas involved in interoceptive processing, should show responses consistent with interoceptive predictive coding; for example, by analogy with studies of repetition suppression, AIC should show reduced responses for well predicted interoceptive signals. Third, the model predicts that distortions of presence may not necessarily lead to distortions of agency; they will only do so if agency-component predictions realign or change their precision or structure in order to suppress faulty interoceptive prediction errors. Further predictions can be

©
Copyright 2010-2014 CEEDS Consortium

based on the relative timing of activity. In the visual domain, (Melloni, Schwiedrzik et al. 2011) found that expectations about upcoming sensory input reduce the latency of neuronal signatures differentiating seen and unseen stimuli; in other words, expectations speed up conscious access. By analogy, an expected interoceptive signal may be perceived as occurring earlier than an unexpected interoceptive signal. This hypothesis could be tested by manipulations of physiological feedback as implemented by (Gray, Harrison et al. 2007). Potentially, VR or SR based experimental environments could be used not only for testing the model but perhaps also for therapeutic purposes in DPD/DR sufferers.

A possible objection to the model challenges the plausibility of hierarchically-organized interoceptive predictive coding. Hierarchical predictive coding schemes are best established for visual processing on both functional and anatomical grounds. Hierarchical visual processing is well established and predictive coding models of visual perception are well motivated by the need for efficient processing of the high-bandwidth and highly redundant afferent sensory signals (Rao and Ballard 1999). It is not immediately clear that interoceptive processing faces the same order of computational challenge, nor does interoceptive processing map so clearly onto hierarchically organised structure. This objection does not however rule out interoceptive predictive coding; it may equally be considered that interoception and exteroception are likely to utilize common processing frameworks. In either case the onus remains on designing novel experiments to test explicitly for signs of interoceptive predictive coding.

In summary, our model integrates previously disparate theory and evidence from predictive coding, interoceptive awareness and the role of the insula, dopaminergic signalling, depersonalization and schizophrenia, and experiments combining virtual reality and neuroimaging. It develops a new view of emotion as interoceptive inference and provides a computationally explicit, neurobiologically grounded account of conscious presence, a fundamental but understudied phenomenological property of conscious experience.

References

Aardema, F., K. O'Connor, et al. (2010). "Virtual reality induces dissociation and lowers sense of presence in objective reality." Cyberpsychol Behav Soc Netw**13**(4): 429-435.

Ackner, B. (1954). "Depersonalization. I. Aetiology and phenomenology." J Ment Sci**100**(421): 838-853.

Augustine, J. R. (1996). "Circuitry and functional aspects of the insular lobe in primates including humans." Brain Res Brain Res Rev**22**(3): 229-244.

Baumgartner, T., D. Speck, et al. (2008). "Feeling present in arousing virtual reality worlds: prefrontal brain regions differentially orchestrate presence experience in adults and children." Frontiers in human neuroscience**2**: 8.

Bechara, A., H. Damasio, et al. (1997). "Deciding advantageously before knowing the advantageous strategy." Science**275**(5304): 1293-1295.

Bernardet, U., A. Valjamae, et al. (2011). "Quantifying human subjective experience and social interaction using the eXperience Induction Machine." Brain Res Bull**85**(5): 305-312.

Blakemore, S. J., C. D. Frith, et al. (1999). "Spatio-temporal prediction modulates the perception of self-produced stimuli." J Cogn Neurosci**11**(5): 551-559.

Blakemore, S. J., J. Smith, et al. (2000). "The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: evidence for a breakdown in self-monitoring." Psychol Med**30**(5): 1131-1139.

Bossaerts, P. (2010). "Risk and risk prediction error signals in anterior insula." Brain Struct Funct**214**(5-6): 645-653.

Bubic, A., D. Y. von Cramon, et al. (2010). "Prediction, cognition and the brain." Front Hum Neurosci**4**: 25.

Chorley, P. and A. K. Seth (2011). "Dopamine-signaled reward predictions generated by competitive excitation and inhibition in a spiking neural network model." Front Comput Neurosci**5**: 21.

Corlett, P. R., J. R. Taylor, et al. (2010). "Toward a neurobiology of delusions." Prog Neurobiol**92**(3): 345-369.

Craig, A. D. (2002). "How do you feel? Interoception: the sense of the physiological condition of the body." Nat Rev Neurosci**3**(8): 655-666.

©
Copyright 2010-2014 CEEDS Consortium

Craig, A. D. (2003). "Interoception: the sense of the physiological condition of the body." Curr Opin Neurobiol**13**(4): 500-505.

Craig, A. D. (2009). "How do you feel--now? The anterior insula and human awareness." Nat Rev Neurosci**10**(1): 59-70.

Critchley, H. D., S. Wiens, et al. (2004). "Neural systems supporting interoceptive awareness." Nat Neurosci**7**(2): 189-195.

d'Acremont, M., Z. L. Lu, et al. (2009). "Neural correlates of risk prediction error during reinforcement learning in humans." Neuroimage**47**(4): 1929-1939.

Damasio, A. (2000). The feeling of what happens: Body and emotion in the making of consciousness, Harvest Books.

David, N., A. Newen, et al. (2008). "The "sense of agency" and its underlying cognitive and neural mechanisms." Conscious Cogn**17**(2): 523-534.

Dayan, P., G. E. Hinton, et al. (1995). "The Helmholtz machine." Neural Comput**7**(5): 889-904.

Deen, B., N. B. Pitskel, et al. (2011). "Three systems of insular functional connectivity identified with cluster analysis." Cereb Cortex**21**(7): 1498-1506.

Dolan, R. J., P. Fletcher, et al. (1995). "Dopaminergic modulation of impaired cognitive activation in the anterior cingulate cortex in schizophrenia." Nature**378**(6553): 180-182.

Egner, T., C. Summerfield, et al. (2008). "Neural repetition suppression reflects fulfilled perceptual expectations." Nature Neuroscience**11**(9): 1004-1006.

Farrer, C., S. H. Frey, et al. (2008). "The angular gyrus computes action awareness representations." Cereb Cortex**18**(2): 254-261.

Feinberg, T. E. (2011). "The nested neural hierarchy and the self." Conscious Cogn**20**(1): 4-15.

Fiorillo, C. D., P. N. Tobler, et al. (2003). "Discrete coding of reward probability and uncertainty by dopamine neurons." Science**299**(5614): 1898-1902.

Fletcher, P. C. and C. D. Frith (2009). "Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia." Nat Rev Neurosci**10**(1): 48-58.

Freeman, J., S. E. Avons, et al. (1999). "Effects of sensory information and prior experience on direct subjective ratings of presence." Presence: Teleoperators and Virtual Environments**8**: 1-13.

Friston, K. (2009). "The free-energy principle: a rough guide to the brain?" Trends Cogn Sci**13**(7): 293-301.

Friston, K. (2010). "The free-energy principle: a unified brain theory?" Nat Rev Neurosci**11**(2): 127-138.

Friston, K., J. Kilner, et al. (2006). "A free energy principle for the brain." J Physiol Paris**100**(1-3): 70-87.

Frith, C. (2011). "Explaining delusions of control: The comparator model 20years on." Conscious Cogn.

Frith, C. D. (1987). "The positive and negative symptoms of schizophrenia reflect impairments in the perception and initiation of action." Psychol Med**17**(3): 631-648.

Gilbert, D. T. and T. D. Wilson (2009). "Why the brain talks to itself: sources of error in emotional prediction." Philos Trans R Soc Lond B Biol Sci**364**(1521): 1335-1341.

Gray, M. A., N. A. Harrison, et al. (2007). "Modulation of emotional appraisal by false physiological feedback during fMRI." PLoS One**2**(6): e546.

Grush, R. (2004). "The emulation theory of representation: motor control, imagery, and perception." Behav Brain Sci**27**(3): 377-396; discussion 396-442.

Gutierrez-Martinez, O., J. Gutierrez-Maldonado, et al. (2011). "Control over the virtual environment influences the presence and efficacy of a virtual reality intervention on pain." Stud Health Technol Inform**167**: 111-115.

Haggard, P. (2008). "Human volition: towards a neuroscience of will." Nat Rev Neurosci**9**(12): 934-946.

Haggard, P., S. Clark, et al. (2002). "Voluntary action and conscious awareness." Nat Neurosci**5**(4): 382-385.

©
Copyright 2010-2014 CEEDS Consortium

Hauser, M., J. W. Moore, et al. (2011). "Sense of agency is altered in patients with a putative psychotic prodrome." Schizophr Res**126**(1-3): 20-27.

Hurd, Y. L., M. Suzuki, et al. (2001). "D1 and D2 dopamine receptor mRNA expression in whole hemisphere sections of the human brain." J Chem Neuroanat**22**(1-2): 127-137.

Jancke, L., M. Cheetham, et al. (2009). "Virtual reality and the role of the prefrontal cortex in adults and children." Frontiers in neuroscience**3**(1): 52-59.

Lee, T. S. and D. Mumford (2003). "Hierarchical Bayesian inference in the visual cortex." Journal of the Optical Society of America. A, Optics, image science, and vision**20**(7): 1434-1448.

Lessiter, J., J. Freeman, et al. (2001). "A cross-media presence questionnaire: The ITC-sense-of-presence inventory." Presence: Teleoperators and Virtual Environments**10**(3): 282-297.

Lombard, M. and T. Ditton (1997). "At the heart of it all: The concept of presence." Journal of Computer-Mediated Communication**3**(2).

Lou, H. C., J. C. Skewes, et al. (2011). "Dopaminergic stimulation enhances confidence and accuracy in seeing rapidly presented words." J Vis**11**(2).

Maher, B. A. (1974). "Delusional thinking and perceptual disorder." Journal of Individual Psychology**30**(1): 98-113.

Medford, N. and H. D. Critchley (2010). "Conjoint activity of anterior insular and anterior cingulate cortex: awareness and response." Brain Struct Funct**214**(5-6): 535-549.

Meehan, M., B. Insko, et al. (2002). "Physiological measures of presence in stressful environments." ACM Transaction in Graphics**21**: 645-652.

Melloni, L., C. M. Schwiedrzik, et al. (2011). "Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness." J Neurosci**31**(4): 1386-1396.

Menon, V. and L. Q. Uddin (2010). "Saliency, switching, attention and control: a network model of insula function." Brain Struct Funct**214**(5-6): 655-667.

Mesulam, M. M. and E. J. Mufson (1982). "Insula of the old world monkey. I. Architectonics in the insulo-orbito-temporal component of the paralimbic brain." J Comp Neurol**212**(1): 1-22.

Mesulam, M. M. and E. J. Mufson (1982). "Insula of the old world monkey. III: Efferent cortical output and comments on function." J Comp Neurol**212**(1): 38-52.

Metzinger, T. (2003). Being No-One. Cambridge, MA, MIT Press.

Mufson, E. J. and M. M. Mesulam (1982). "Insula of the old world monkey. II: Afferent cortical input and comments on the claustrum." J Comp Neurol**212**(1): 23-37.

Nahab, F. B., P. Kundu, et al. (2011). "The neural processes underlying self-agency." Cereb Cortex**21**(1): 48-55.

Neal, R. M. and G. Hinton (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. Learning in Graphical Models. M. I. Jordan, Kluwer Academic Publishers**:** 355-368.

Northoff, G. and F. Bermpohl (2004). "Cortical midline structures and the self." Trends Cogn Sci**8**(3): 102-107.

Palaniyappan, L. and P. F. Liddle (2011). "Does the salience network play a cardinal role in psychosis? An emerging hypothesis of insular dysfunction." J Psychiatry Neurosci**36**(4): 100176.

Paulus, M. P. and M. B. Stein (2006). "An insular view of anxiety." Biological psychiatry**60**(4): 383-387.

Pessiglione, M., B. Seymour, et al. (2006). "Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans." Nature**442**(7106): 1042-1045.

Phillips, M. L., N. Medford, et al. (2001). "Depersonalization disorder: thinking without feeling." Psychiatry Res**108**(3): 145-160.

Ploghaus, A., I. Tracey, et al. (1999). "Dissociating pain from its anticipation in the human brain." Science**284**(5422): 1979-1981.

Porro, C. A., V. Cettolo, et al. (2003). "Functional activity mapping of the mesial hemispheric wall during anticipation of pain." Neuroimage**19**(4): 1738-1747.

©
Copyright 2010-2014 CEEDS Consortium

Preuschoff, K., S. R. Quartz, et al. (2008). "Human insula activation reflects risk prediction errors as well as risk." The Journal of neuroscience : the official journal of the Society for Neuroscience**28**(11): 2745-2752.

Rao, R. P. and D. H. Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects." Nat Neurosci**2**(1): 79-87.

Redgrave, P. and K. Gurney (2006). "The short-latency dopamine signal: a role in discovering novel actions?" Nat Rev Neurosci**7**(12): 967-975.

Riva, G., J. A. Waterworth, et al. (2011). "From intention to action: The role of presence." New Ideas in Psychology: 1-14.

Sanchez-Vives, M. V. and M. Slater (2005). "From presence to consciousness through virtual reality." Nature reviews. Neuroscience**6**(4): 332-339.

Schachter, S. and J. E. Singer (1962). "Cognitive, social, and physiological determinants of emotional state." Psychol Rev**69**: 379-399.

Schultz, W. and A. Dickinson (2000). "Neuronal coding of prediction errors." Annu Rev Neurosci**23**: 473-500.

Seeley, W. W., V. Menon, et al. (2007). "Dissociable intrinsic connectivity networks for salience processing and executive control." J Neurosci**27**(9): 2349-2356.

Seth, A. K. (2009). "Explanatory correlates of consciousness: Theoretical and computational challenges." Cognitive Computation**1**(1): 50-63.

Seth, A. K. (2010). "The grand challenge of consciousness." Frontiers in Psychology**1**(5): 1-2.

Sierra, M., D. Baker, et al. (2005). "Unpacking the depersonalization syndrome: an exploratory factor analysis on the Cambridge Depersonalization Scale." Psychological medicine**35**(10): 1523-1532.

Sierra, M. and A. S. David (2011). "Depersonalization: a selective impairment of self-awareness." Consciousness and cognition**20**(1): 99-108.

Simeon, D., T. Giesbrecht, et al. (2009). "Alexithymia, absorption, and cognitive failures in depersonalization disorder: a comparison to posttraumatic stress disorder and healthy volunteers." J Nerv Ment Dis**197**(7): 492-498.

Simeon, D., D. S. Kozin, et al. (2008). "De-constructing depersonalization: further evidence for symptom clusters." Psychiatry research**157**(1-3): 303-306.

Singer, T., H. D. Critchley, et al. (2009). "A common role of insula in feelings, empathy and uncertainty." Trends Cogn Sci**13**(8): 334-340.

Slater, M. and A. Steed (2000). "A virtual presence counter." Presence: Teleoperators and Virtual Environments**9**: 413-434.

Suddendorf, T. and M. C. Corballis (2007). "The evolution of foresight: What is mental time travel, and is it unique to humans?" Behav Brain Sci**30**(3): 299-313; discussion 313-251.

Summerfield, C. and T. Egner (2009). "Expectation (and attention) in visual cognition." Trends in Cognitive Sciences**13**(9): 403-409.

Synofzik, M., P. Thier, et al. (2010). "Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions." Brain**133**(Pt 1): 262-271.

Takahashi, H., M. Higuchi, et al. (2006). "The role of extrastriatal dopamine D2 receptors in schizophrenia." Biol Psychiatry**59**(10): 919-928.

Tamietto, M. and B. de Gelder (2010). "Neural bases of the non-conscious perception of emotional signals." Nat Rev Neurosci**11**(10): 697-709.

Taylor, K. S., D. A. Seminowicz, et al. (2009). "Two systems of resting state connectivity between the insula and cingulate cortex." Hum Brain Mapp**30**(9): 2731-2745.

Ueda, K., Y. Okamoto, et al. (2003). "Brain activity during expectancy of emotional stimuli: an fMRI study." Neuroreport**14**(1): 51-55.

©
Copyright 2010-2014 CEEDS Consortium

van den Heuvel, M. P., R. C. Mandl, et al. (2009). "Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain." Hum Brain Mapp**30**(10): 3127-3141.

Voss, M., J. Moore, et al. (2010). "Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences." Brain**133**(10): 3104-3112.

Wang, X. J. (2010). "Neurophysiological and computational principles of cortical rhythms in cognition." Physiol Rev**90**(3): 1195-1268.

Williams, S. M. and P. S. Goldman-Rakic (1998). "Widespread origin of the primate mesofrontal dopamine system." Cereb Cortex**8**(4): 321-345.

Wolpert, D. M. and Z. Ghahramani (2000). "Computational principles of movement neuroscience." Nat Neurosci**3 Suppl**: 1212-1217.

Yu, A. J. and P. Dayan (2005). "Uncertainty, neuromodulation, and attention." Neuron**46**(4): 681-692.

©
Copyright 2010-2014 CEEDS Consortium

Appendix B

Visual anticipation biases conscious perception but not bottom-up visual processing

10/20/2011

Zenon Mathews[1] & Paul Verschure[1, 2]

[1] SPECS, Technology Department, Universitat Pompeu Fabra, Barcelona, Spain

[2] ICREA Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

Abstract

The role of prediction in visual perception can be observed easily in everyday life and studied in laboratory experiments. Yet, due to the lack of efficient indirect measures, the dynamic effects of prediction on perception and decision making has been difficult to assess and to model. We propose a psychophysical experiment using a displacement detection task combined with reverse correlation to reveal signatures of the usage of prediction at three different levels of perception: bottom-up early saccades, top-down driven late saccades and conscious decisions. Our results suggest that the brain employs multiple parallel mechanisms at different levels of information processing to restrict the sensory field using predictions. We observe that cognitive load has a quantifiable effect on this dissociation of the bottom-up sensory and top-down predictive processes. We propose a probabilistic data association model from dynamical systems theory to model this predictive bias in different information processing levels.

Introduction

Everyday human perception involves anticipating future events based on predictions and apriori knowledge about the world. Indeed, conscious perception has long been postulated to be inherently predictive and anticipatory (Gregory, 1980). Perceptual prediction has been supported by physiological accounts of the influence of cortical projections in providing predictions (Ekstrom, Roelfsema, Arsenault, Bonmassar, & Vanduffel, 2008, Cox, Meyers, & Sinha, 2004), for anticipatory modulation of bottom-up visual responses (Summerfield et al., 2006), and in general for the integration of bottom-up and top-down processes including at the single neuron level (Fries, Reynolds, Rorie, & Desimone, 2001, Ekstrom et al., 2008). In visual tasks, signatures of explicit/implicit anticipations are manifested in diverse paradigms, ranging from reduced delays of visual processing (Barnes, Barnes, & Chakraborti, 2000), changes in smooth pursuit of motion (Winges & Soechting, 2011) and hemodynamic responses measured using fMRI (Turk-Browne, Scholl, Johnson, & Chu, 2010) etc.

Building on the above observation, prominent models of perception often use the principle of the minimization of error between sensory input and sensory prediction (Friston & Kiebel, 2009, Duff & Verschure, 2010, Dean & Porrill, 2008). In this view, perception is seen as a monolithic process involving a single prediction and error generation mechanism (Duff & Verschure, 2010, Spratling, 2010). Nevertheless, it can not be excluded that predictions are processed differently at various levels of perceptual processing. No experimental paradigm and mathematical framework have been proposed to measure and model the top-down effects of prediction at different levels of visual processing. Beyond the accounts of anticipatory modulation of physiological responses in early brain areas (Summerfield et al., 2006), it is unknown how prediction/anticipation directly affects perception.

Here, we first ask whether signatures of predictions of sensory events can be observed in visual decision making tasks and investigate if such biases affect bottom-up perception and top-down perceptual decision making processes. To this end, we designed a novel psychophysical experimental paradigm and used it to demonstrate that the brain employs multiple parallel

©
Copyright 2010-2014 CEEDS Consortium

predictive processes at different perception levels. Each of them constrain the perceptual space using the available predictions of future sensory stimuli. We refer to such constrained perceptual spaces as anticipatory fields. Given that prediction is a higher level process, we hypothesize that how prediction affects perception should be influenced by high level cognitive processes. To this end we manipulated cognitive loads of the subjects with extra cognitive tasks. Our results show that cognitive load has direct, quantifiable and dynamic influences on the parallel visual anticipations generated by the brain. We model the observed phenomena using Bayesian dynamical systems theory to capture the parallelity and the top-down bias of such anticipations in human visual perception. Our results provide concrete evidence for parallel influences of prediction in perception at lower and higher information processing levels. We provide concrete applications of our finding in video compression, in novel robotic assistance for human operations and in humanoid robotics.

Materials and methods

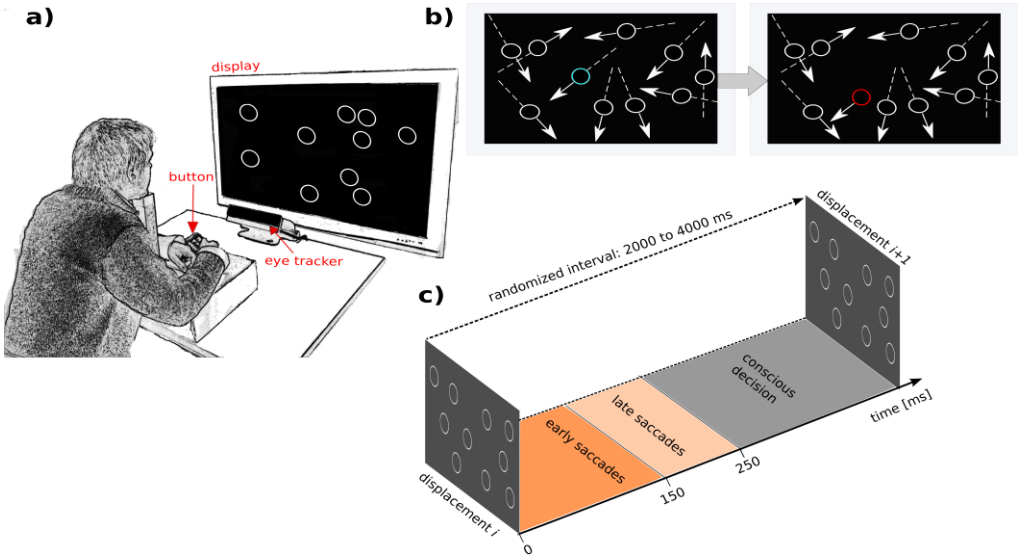


Figure 1: Experimental setup and basic analysis. a) **Subjects face a screen with the head stabilized on a chin rest. The display shows linearly moving circular items and an eyetracker is used to track eye movements. Subjects report detected displacements with a button press.** b) **Illustration of a displacement: on the left is the constellation of the moving items (indicated by arrows) before the displacement. On the right is the constellation after displacement. The displaced item is shown in blue before the displacement and in red after the displacement (colors, arrows and lines are only for illustration purposes, and not used in the experiment).** c) **Schematic of the time windows used to define early saccades, late saccades and conscious decisions after displacement. The inter-displacement times are randomized between 2000 and 4000 ms.**

Experimental setup

We used the Tobii X120 eye tracker (©Tobii Technology, Sweden 2011) that tracks eyes at 120 Hz. The visual stimulus was developed using the OpenGL library in Linux Fedora 9, C++ environment and ran at 200Hz on a quadcore Intel(R) Xeon(R) 2.00GHz CPU. The button press and the circular item movements were logged time synchronized with the eye tracking data. The screen resolution was 1360x768 and the subject sat head stabilized at 102 cms from the screen midpoint, giving a 1.0° radius of the circular items of the visual stimulus. The LCD screen measured 115*65 cms and had a refresh rate of 60Hz. The movement speed of the circular items

©
Copyright 2010-2014 CEEDS Consortium

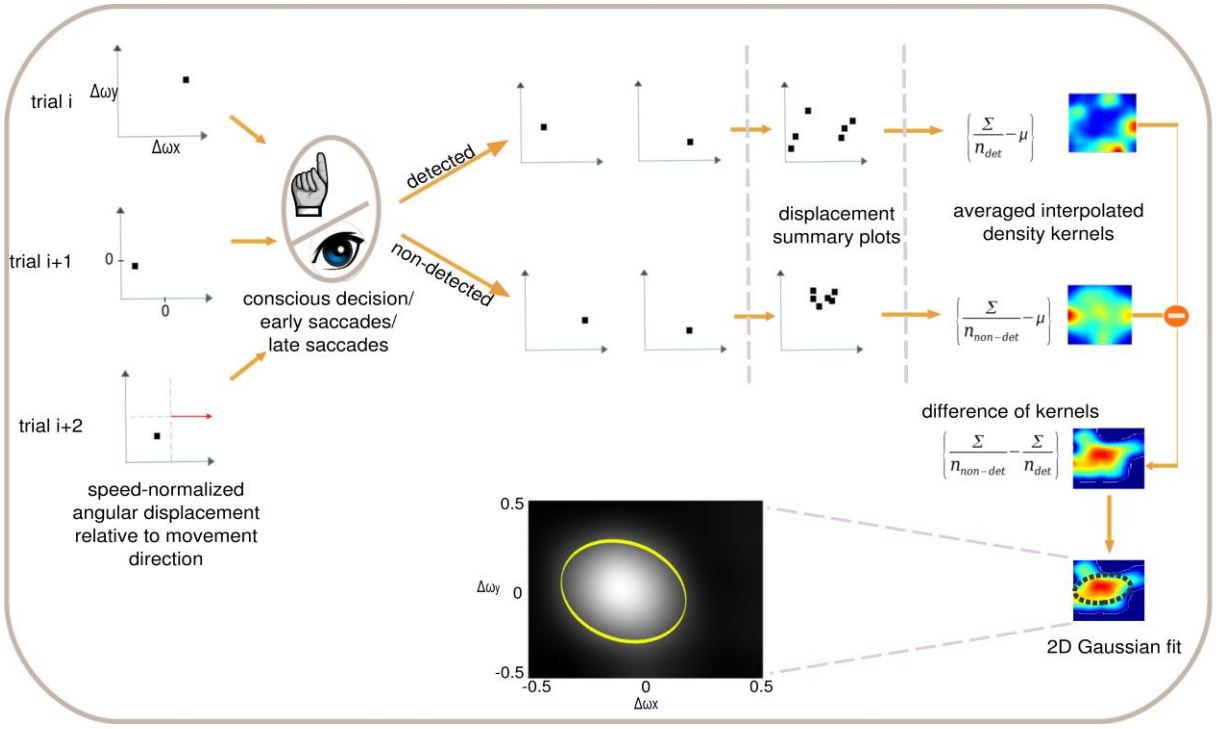
were between 1 and 18 $^{\circ}/s$ and a slight change in speed was induced ($\pm 0.001^{\circ}/s$) at boundary bounces to allow minimally different linear paths. Eye tracker calibration was performed once for each subject before the experiments. Calibration error was below 1.2°. Subjects were alone in the controlled experimental space and were instructed to recite aloud the alphabets so that it was audible to the experimenter in the nearby separated space.

Experimental procedure

Subjects were 15 university students (8 male and 7 female) between 21 and 32 years old. All were right handed and had normal or corrected-to-normal vision. Subjects provided verbal consent before the experiments. A session consisted of 3 experiments, 3 minutes each (one for each cognitive load) and were carried out in a randomized order for each subject. After each experiment subjects could rest for 2 minutes by taking the head off the chin rest. Item displacements were randomized in two ways. First, the item to be displaced was randomly chosen from all items except the ones that were closer than 24° to a boundary (to guarantee linear motion before and after displacements). Second, the displacement angle and distance were chosen randomly around the item position from a radius of 12°. For each displacement the direction and length of displacement were chosen from a uniform distribution inside the above radius.

Data analysis

To analyze the eye tracking data and detect saccades, we formalize an operational definition of *saccade* towards displacement position: we compute the distance between the eye position and the displacement position at the beginning ($dist_b$) and at the end ($dist_e$) of a given time window. We define that a saccade occurred if $dist_e < \frac{dist_b}{2}$. We studied two kinds of saccades and computed the *anticipatory fields* for each of them separately. First we looked at the *early* saccades, that occur up to 150ms after displacement. Secondly we looked at the *late* saccades that occur after 150ms but before 250ms after displacement. We defined a displacement as *detected* by an early/late saccade, if the saccade towards the displacement location occurred in the above defined respective time windows. All data analysis was performed using Matlab© toolboxes.



Copyright 2010-2014 CEEDS Consortium

Figure 2: Psychophysical reverse correlation. **Each displacement trial (left) is normalized to speed and movement direction corrected to be the x-axis (red arrow). X and Y axis indicate horizontal and vertical angular displacements respectively after direction correction. Each trial is sorted by the three detection levels (conscious decision, early saccade and late saccade) and the density kernels of detected and non-detected trials are computed separately. Further, the detected trials' kernel is subtracted from the non-detected kernel and the error ellipse is computed for this difference of kernels. This error ellipse is the psychophysical kernel/*anticipatory field*, of which we analyze four properties: area, shift, eccentricity and orientation. Inset shows the *anticipatory field* for a single subject for the early saccades in the low cognitive load experiment.**

Psychophysical reverse correlation

Reverse correlation has been used in psychophysical studies to characterize observer's strategies in visual tasks (Ahumada, 1996) and in physiological studies to characterize neural responses to visual stimuli (Victor, 2005). Reverse correlation has proven to be a strong technique for seeking relationships between a high-dimensional variable (e.g. an image) and a categorical variable (two-choice decision or neural spiking) (Victor, 2005). Here we tailored psychophysical reverse correlation to analyze conscious decisions, and early and late saccades. Each event is a displacement that is plotted as a point on the speed-normalized and direction-corrected coordinate system (Fig. 2), where the positive x-axis is the linear movement direction of the circular item. We then sort the stimuli according to the 'detected' and 'non-detected' choices (using either conscious, early or late saccades as the detection criterion). We then compute the average detection and non-detection densities and use data interpolation to yield a two-dimensional probability distribution for detection and non-detection densities separately. The difference between the two probability distributions (non-detected − detected) is computed and fitted with a 2D Gaussian distribution. The covariance ellipse of the Gaussian distribution is referred to as the *psychophysical kernel* or the *anticipatory field*. The ellipse covers 39.4 % of the total probability mass (Fig. 2).

Model

We propose *data association* as the core mechanism underlying anticipation in human perception. Data association refers to the process of associating novel sensory input to memory items. In discrete linear dynamical systems theory, the state of each discrete memory item is updated using linear state transition equations. Cognitive load in a controlled cognitive task is modeled by the process noise of the state update mechanism (Paas & Merriënboer, 1994).

To formalize the anticipatory influence in human perception, we use the Joint Probabilistic Data Association (JPDA) algorithm (Bar-Shalom & Fortmann, 1988). JPDA is a single-scan approximation to the optimal Bayesian filter, which associates current observations (sensory input) to previously known targets (memory items) sequentially. JPDA enumerates all associations between observations and targets at each time step and computes the association probabilities β_{jk}, i.e. the probability that the *j*-th observation originated from the *k*-th target. Given this, the target state is estimated by Kalman filtering (Kalman, 1960) and this conditional expectation of the state is weighted by the association probability. (Note that the Kalman filter is just one of many possible mechanisms to update the target state.) Let x_t^k indicate the state of target *k* at time step *t*, ω_{jk} the association event where the observation *j* is associated to target *k* and $Y_{1:t}$ denotes all the observations from time step 1 to time step *t*. Using apriori knowledge about the world (e.g. state transition matrix(*A*), process noise covariance (*Q*), measurement matrix (*H*), control-input model (*B*) and the control input-vector ($\hat{u}$) of the Kalman filter) and the current state of the *target*, a prediction is made for each *target*. At time step *t*, for each *target k*, we compute the state prediction, its covariance and the measurement prediction as follows

$$\tilde{x}_k^k = A\hat{x}_{t-1}^k + B\hat{u}_{t-1}^k \qquad (1)$$

Copyright 2010-2014 CEEDS Consortium

$$\tilde{P}_t^k = A P_{t-1}^k A^T + Q_{t-1}^k \quad (2)$$

$$\tilde{y}_t^k = H x_t^k \quad (3)$$

Then the state of the target can be estimated as:

$$E(x_t^k | Y_{1:t}) = \sum_{\omega} E(x_t^k | \omega, Y_{1:t}) P(\omega | Y_{1:t}) \quad (4)$$

$$= \sum_{j} E(x_t^k | \omega_{jk}, Y_{1:t}) P(\omega_{jk} | Y_{1:t}) \quad (5)$$

where ω_{jk} denotes the association event of observation j being associated to target k and ω_{0k} denotes the event that no observation is associated to target k. Thus, the event association probability is $\beta_{jk} = P(\omega_{jk} | Y_{1:t})$. JPDA computes an *anticipatory field* for each target using the Kalman innovation of new observations. It only considers observations inside the *anticipatory field* for each target. We consider the linear state evolution model for state dynamics of target x at time k:

$$x_k = A x_{k-1} + B u_{k-1} + q \quad (6)$$

where q is the process noise with time-invariant covariance matrix Q, B the control-input model and u_k the control vector. The well-known linear Kalman filter prediction and estimation steps is used to update the state. The ellipsoidal *anticipatory field* is optimal for the above linear observation model with additive noise (Bar-Shalom & Fortmann, 1988) (time-subscripts are omitted for clarity):

$$z = Hx + \varphi \quad (7)$$

where φ is the zero Gaussian measurement error with $p(\varphi) = \mathcal{N}(\varphi; 0, R)$ and is independent of the state x. H is the observation model which maps the true state space into the observed space. The state probability density function is Gaussian $p(x) = \mathcal{N}(x; \hat{x}, P)$. The validity of measurement y_i is determined from its innovation, i.e. from $v = y_i - H\hat{x}$, with the covariance $S = R + HPH^T$. *Anticipatory field* is computed by gating the Mahanalobis distance (Normalized Innovation Square (NIS)):

$$v^T S^{-1} v < M_d \quad (8)$$

M_d is the threshold for an innovation dimension d and can be computed efficiently since the NIS follows a chi-square probability density function. E.g. to compute the probability that j% of true associations are accepted, M_d is obtained from

$$\frac{j}{100} = P\left(\frac{d}{2}, \frac{M_d}{2}\right) \quad (9)$$

where $P(a,b) = \frac{1}{\Gamma(a)} \int_0^b e^{-t} t^{a-1} dt$ is the incomplete gamma function (Press, Teukolsky, Vetterling, & Flannery, 1992). The *anticipatory field* defines a region of acceptance such that $(100-j)$% of true measurements are rejected given that the measurements y_i are distributed according to

$$p(y) = \mathcal{N}(y; H\hat{x}, S) \quad (10)$$

This formulation of the *anticipatory field* avoids the necessity to model clutter and eliminates unlikely associations. Also non-linear *anticipatory fields* are conceivable for non-Gaussian models (Bailey, Upcroft, & Durrant-Whyte, 2006).

©
Copyright 2010-2014 CEEDS Consortium

Results

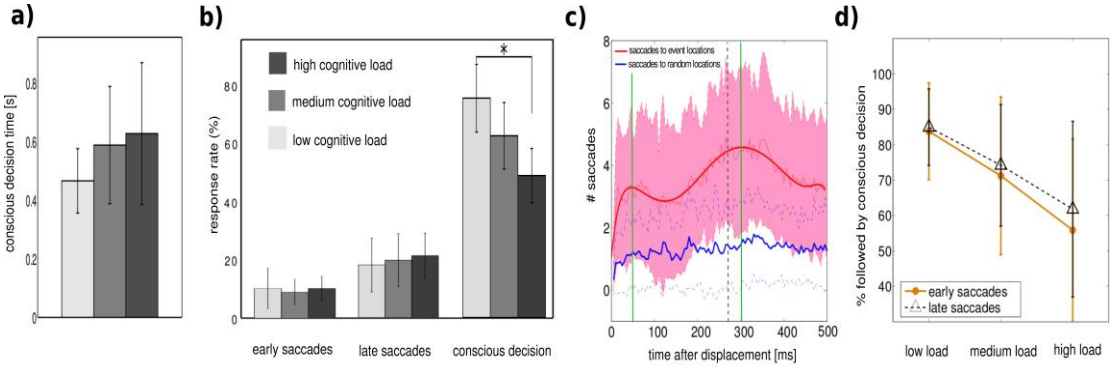


Figure 3: Preliminary analysis.a) **Response time of conscious decisions for the three cognitive loads.** b) **Displacement detection rate (response rate) in the three cognitive load cases for early saccades, late saccades and conscious decision.** c) **Histogram means, fitting and standard deviations comparing mean number of saccades to displacement positions (red) and to random positions on the screen (blue). Vertical green lines indicate peak times of the fitting for saccades to displacements.** d) **Percentage of early and late saccades followed by a conscious decision in the three cognitive load cases.**

15 subjects observed a fixed number (n=10) of identical non-filled white circles (referred to as *items*) on a black background (Fig. 1a). Each item followed a linear path at constant speed and bounced off the display boundaries. Linear movements of items were designed to test the existence of movement predictions by the subjects. By using several simultaneously moving circular items we avoided visual habituation during smooth-pursuit of a single linear movement (Eggert, Ladda, & Straube, 2009). Once every T seconds, one of the moving items displaced itself from its linear trajectory and then continued linear motion at the same speed and direction prior to the displacement (Fig. 1b). Inter displacement time T was randomized between 2000 and 4000 ms to avoid automated rhythmic response. The displaced item and the displacement distance was chosen randomly (see methods section). Items that were closer than twice the displacement radius to a boundary were not chosen for displacement, guaranteeing linear motion before and after displacements. Subjects were instructed to press a button whenever they perceived a displacement of an item (Fig. 1a,b). Eye tracking was performed during the task.

A session consisted of three experiments of three minutes length each, designed to modulate the cognitive load without affecting the perceptual load (Camos & Barrouillet, 2004). Experiment one was of low cognitive load where the subject solely performed the above displacement detection task (referred to hereafter as the low load task). Experiment two was of higher cognitive load as the subjects were instructed to continuously recite aloud the alphabet in their mother tongue while performing the above psychophysics task (medium load task). Experiment three was of the highest cognitive load where the subjects were instructed to continuously recite aloud the alphabet in their mother tongue in reverse order skipping every other letter (high load task). Recalling less automatized chains (alphabets in reverse order and skipping every other letter) is known to induce higher cognitive loads than more automatized chains (like the alphabet in forward order) (Camos & Barrouillet, 2004). The order of low, medium and high cognitive load tasks were randomized for each subject.

Basic analysis

Early or express saccades refer to saccades occurring up to 150 ms after the occurrence of a sensory event (Fischer & Rampsberger, 1984). Besides early saccades, we also consider the slower late saccades with latencies around 200 ms after the stimulus onset (Fischer & Rampsberger, 1984, Edelman, Kristjánsson, & Nakayama, 2007). To investigate if our stimulus triggered both

----------©
Copyright 2010-2014 CEEDS Consortium

kinds of saccades we looked at latencies of saccades towards displacement locations and compared it to saccades towards random locations after displacement (Fig. 3c). We observed that saccades towards displacement locations were clearly above chance level (Fig. 3c). We observed a peak in the number of saccades towards displacements below 100ms and above 200ms after displacement time (green vertical lines in fig. 3c). The early and late saccade times are in the latency ranges after stimulus onset as reported in earlier studies (Fischer & Rampsberger, 1984, Edelman et al., 2007). Based on this, we defined the time windows of of 0-150ms and 150-250ms for the early and late saccades respectively (Fig. 1c). Next we investigated the conscious decision latencies (button press) for the three cognitive load cases. We observed that the conscious decision latency increases with the cognitive load, and the lowest mean was above 0.4 seconds for the low load case (Fig. 3a). We investigated the response/detection rate of early and late saccades and conscious detections (Fig. 3b). We observed that while the early and late saccade responses in the direction of displacements cannot be distinguished on statistical grounds with changing cognitive load (p<0.05), there is a significant decrease of the conscious detection rate (p<0.05) from low to high cognitive load (fig.3b). This served as the first indicator of a dissociation between bottom-up saccades and conscious decision processes. In order to elucidate this further, we analyzed the relation between early and late saccades and conscious decisions after displacements. We observed that early and late saccades do not change with cognitive load but with increasing cognitive load there is a decrease in the rate of early and late saccades that are followed by conscious decisions (Fig. 3d). I.e. with increasing cognitive load, there is an increasing number of early and late saccades that are not followed by conscious decisions. This again suggests a dissociation of bottom-up saccadic and conscious top-down decision making processes. To further understand this phenomenon we investigated the specific features of the detected and non-detected displacements in the three cognitive load conditions.

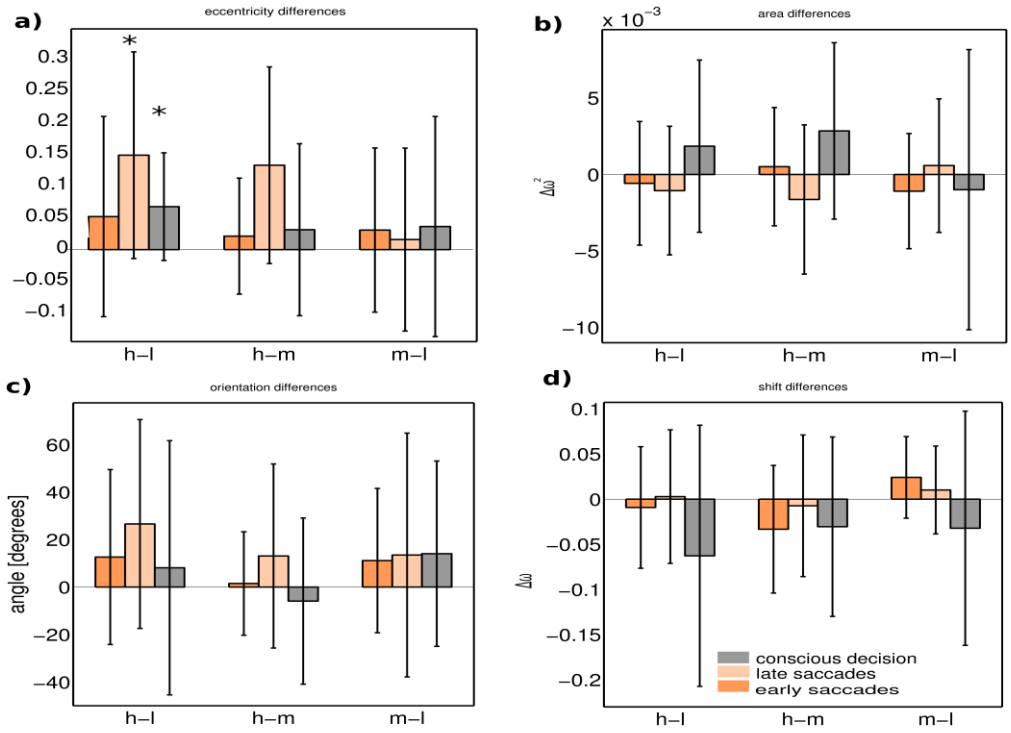


Figure 4: Intra-subject differences in psychophysical kernel. a) **Eccentricity differences for the psychophysical kernels of early saccades, late saccades and conscious decision in the three cognitive load cases. *h–l* indicates high load minus low load, *h–m* indicates high load minus medium load, and *m–l* indicates medium load minus low load. Star indicates significance of sign-test (p<0.05). Differences in area** b)**, orientation** c) **and shift** d) **were not significant.**

Copyright 2010-2014 CEEDS Consortium

Psychophysical reverse correlation

We designed a psychophysical reverse correlation analysis (Ahumada, 1996) to investigate the nature of the detected and the non-detected displacements. This technique provides a unique tool to uncover the internal representations (early and late saccades) and conscious decision strategies of individual participants in the perceptual task. The computation of the so-called psychophysical kernel (which we further refer to as the *anticipatory field*) is described in the methods section. The *anticipatory field* represents the area (centered at the location of the item if there were no displacement) in which displacements do not trigger responses, i.e. early/late saccades or button presses. Alternatively the *anticipatory field* could be thought of as the perceptive area where future stimuli were anticipated, leading to non-detection of displacements. Fig. 2 illustrates the computation of the psychophysical kernel (see also the methods section).

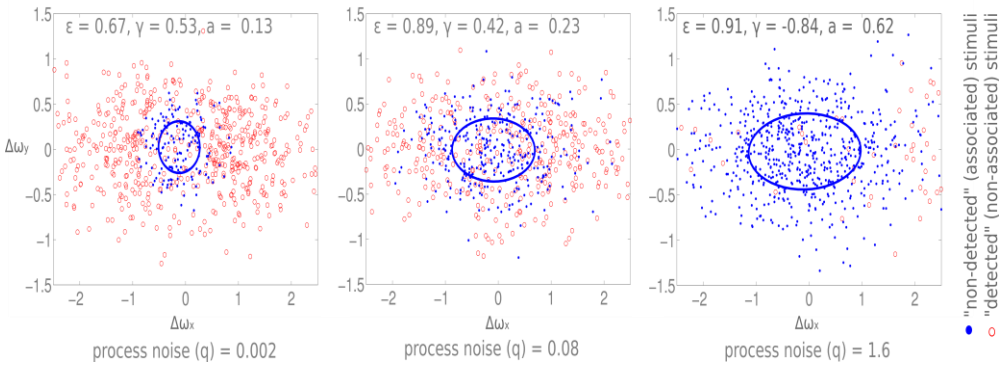


Figure 5: Model simulation results. **Area (*a*), eccentricity (ε) and orientation (γ) of the *anticipatory field* is computed for varying process noises (*q*). Shift was zero for all three cases. Process noise (*q*) is lowest on the leftmost plot and the highest on the rightmost one. X and Y axis are horizontal and vertical angular displacements (radians) respectively.**

To shed light onto the nature of anticipations at work at the different levels of perception, we investigated four different properties of the *anticipatory fields* computed separately for conscious detection, early saccades and late saccades: area, eccentricity, shift and orientation. The area of the *anticipatory field* ellipse corresponds to the amount of non-detected displacements. The orientation allows to investigate potential anticipatory biases in directions relative to the item movement. If there were no bias in anticipations of future stimuli, the *anticipatory fields* should be circular (eccentricity 0) and centered at origin (shift 0), as a circular *anticipatory field* would signify uniform detection chances in all directions. We normalized all displacements with respect to the item movement direction, i.e. the positive x-axis (Fig. 2). After the computation of the *anticipatory fields* for each subject and for each cognitive load, we analyzed intra-subject changes in the above properties of the *anticipatory fields* (we observed a high variance in the inter-subject means for the above four parameters in all cognitive load cases, suggesting distinct displacement detection baselines for each subject and making an intra-subject analysis more informative). For this, we computed the difference of the above four parameters for each subject between medium and low (m-l), high and medium (h-m) and high and low (h-l) cognitive loads. We observed that there were no significant changes in area, shift and orientation (fig.4). However, in the *h−l* case we observed a significant increase in the eccentricity of the *anticipatory fields* of conscious decision and late saccades, but not for early saccades (Fig. 4a). (A slight increase in the area of the *anticipatory field* of conscious decision with increasing cognitive load was seen, as also observed in our basic analysis, fig. 4b). The increase in eccentricity with cognitive load with no significant change in orientation (fig. 4c), signifies that the eccentricity increase was manifested along the 0 orientation, that is along the movement direction of the items. This was true only for conscious decision and late saccades, i.e. for the higher level processes, suggesting a top-down influence on movement anticipations. This was not observed for the early saccades confirming their bottom-up nature (Fischer & Rampsberger, 1984, Edelman et al., 2007).

Copyright 2010-2014 CEEDS Consortium

Modeling

Our model investigates whether the observed changes in the *anticipatory field* are due to the cognitive load induced noise (see methods section). Based on our earlier work on a self-contained model of bottom-up and top-down attention (Mathews, Badia, & Verschure, 2008), we use a probabilistic data association model from dynamic systems theory which uses a Kalman filter for the state update of individual items. A displacement that is not detected is modeled here as the association of sensory data to a known memory item. Analogously, a detected displacement is modeled as a sensory input not being associated to any items in memory. We model the cognitive load as process noise of the dynamic system which corresponds to the noise involved in the state updates of memory items (Bar-Shalom & Fortmann, 1988). Thereby we conjecture that an increase in cognitive load strains higher level processes and this therefore induces higher noise in the state update of memory items. Indeed it has been shown that uncertainty in decision making correlates with an increase in the variability of firing in the prefrontal cortex (Churchland, Kiani, & Shadlen, 2008). We increased the process noise and observed the change in the *anticipatory field* properties. As shown in fig. 5, an increase in the area (i.e. more missed hits), and more significantly, an increase in the eccentricities were observed. Nevertheless, the change in process noise induces no change in shift and orientation.

Discussion

We have investigated the question of how prediction affects different levels of perception and how cognitive load influences these effects. Our major finding is the existence of constrained sensory regions (which we refer to as anticipatory fields) that is characteristic for each perception level. Our results also suggest that higher level cognitive load dynamically biases visual anticipation at three different levels of visual processing. We used the notion of early and late saccades with a novel displacement detection task and psychophysical reverse correlation. Early and late saccades were first observed in monkeys using gap and overlap tasks, where a bimodal distribution of saccadic latencies to single targets was reported (Fischer & Rampsberger, 1984). Top-down influence on late saccades has been reported previously while this effect for early saccades remains unclear (Edelman et al., 2007). Our experimental paradigm deviates significantly from Multiple Object Tracking (MOT), where visually identical items move on non-linear Brownian motion tracks and the task is to keep track of a specific item (Pylyshyn & Storm, 1988). Our stimulus was designed to investigate if linear motion cues are used for anticipation of movements. By maintaining the perceptual task load over trials, and independently varying the cognitive load using the extra verbal task, we were able to investigate the role of cognitive load in biasing visual anticipations. The existence and the influence of higher level cognitive processes on visual anticipation have been shown previously via top-down contextual influences on property attributions to objects (Tremoulet & Feldman, 2006), top-down attentional influences on object location perception (Tse, Whitney, Anstis, & Cavanagh, 2011), attentional facilitation of motion perception based on past object movements (Watanabe & Shimojo, 1998) and implicit perceptual anticipation triggered by statistical learning (Turk-Browne et al., 2010). Our concept of *anticipatory fields* characterizes and quantifies this influence of higher level processes on anticipation at different levels of perception. Our approach avoids the drawbacks of the so-called direct measures of unconscious decision making and provides a method to quantitatively assess the effects of prediction on perception at different levels, posing an alternative to other psychophysical approaches like metacontrast masking (Lau & Passingham, 2006).

The recurring topic of anticipation in conscious perception has motivated the usage of dynamical systems theory to model the anticipatory properties of conscious perception (Freeman, 2007). We proposed data association as the underlying mechanism to explain our finding and used the JPDA algorithm to model the variation in the *anticipatory field* properties with changing cognitive load. The process noise of the Kalman process used for state prediction and estimation of the moving items captures the influence of cognitive load on the *anticipatory fields*. Our model can be extended to non-linear movements (Bailey et al., 2006) and also to non-spatial domains (Gärdenfors, 2000). Our probabilistic model of perception is supported by earlier seminal research suggesting that what we see is a *statistical consequence of past experience* rather than a representation of the retinal stimulus itself (Purves, Lotto, Williams, Nundy, & Yang, 2001, Verschure, Voegtlin, & Douglas, 2003).

An interesting parallel to our findings is the effect created by magicians, where most tricks rely on the fact that the human mind is vulnerable to deceptions as it works with anticipations about

©
Copyright 2010-2014 CEEDS Consortium

the world (Kuhn, Amlani, & Rensink, 2008). The *anticipatory field* proposal could explain the psychological phenomenon of inattentional blindness (Simons & Chabris, 1999), as the former provides the perceptual area inside which changes in the sensory input mostly go unnoticed. Technological applications of our finding are conceivable, such as an automated driver risk assessment tool that combines traffic information with driver cognitive load and eye movements to assess human fault chances, or efficient video compression algorithms that use the average observer's *anticipatory field* in a given context to dynamically compress specific frame regions separately. Furthermore, we also envision novel attention sharing mechanisms using our model to drive the gaze of a humanoid robot when its limited perception, computation and motor resources are challenged in a multitasking scenario. To fully understand the process of anticipation and its influences on perception, our probabilistic model needs to be complemented with neural models and physiology (Lamme, Supèr, Landman, Roelfsema, & Spekreijse, 2000) or brain imaging (Turk-Browne et al., 2010) to investigate potential neural correlates of *anticipatory fields*.

Acknowledgments

We thank Sergi Bermúdez i Badia for suggestions on an initial version of the model, and Karsten Rauss and Anil K.Seth for useful feedback. This work was carried out as part of the CEEDS project, an EU funded Integrated Project under the Seventh Framework Programme (ICT-258749).

References

Ahumada, A. J. (1996). Perceptual classification images from vernier acuity masked by noise. *Perception*, 25-18.

Bailey, T., Upcroft, B., & Durrant-Whyte, H. (2006). Validation gating for non-linear non-gaussian target tracking. *IEEE Conference on Information Fusion*, 1-6.

Barnes, G. R., Barnes, D. M., & Chakraborti, S. R. (2000). Ocular pursuit responses to repeated, single-cycle sinusoids reveal behavior compatible with predictive pursuit. *Journal of Neurophysiology*, *84*, 2340-2355.

Bar-Shalom, Y., & Fortmann, T. E. (1988). Tracking and data association. *Boston Academic Press*.

Camos, V., & Barrouillet, P. (2004). Adult counting is resource demanding. *British Journal of Psychology*, *95*, 19-30.

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*, 693-702.

Cox, D., Meyers, E., & Sinha, P. (2004). Contextually evoked object-specific responses in human visual cortex. *Science*, *304*, 115-117.

Dean, P., & Porrill, J. (2008). Adaptive-filter models of the cerebellum: computational analysis. *Cerebellum*, *7*, 567-571.

Duff, A., & Verschure, P. F. M. J. (2010). Unifying perceptual and behavioral learning with a correlative subspace learning rule. *Neurocomputing*, *73*, 1818-1830.

Edelman, J., Kristjánsson, A., & Nakayama, K. (2007). The influence of object-relative visuomotor set on express saccades. *Journal of Vision*, 7-12.

Eggert, T., Ladda, J., & Straube, A. (2009). Inferring the future target trajectory from visual context: is visual background structure used for anticipatory smooth pursuit? *Experimental Brain Research*, *196*, 205-215.

Ekstrom, L. B., Roelfsema, P. R., Arsenault, J. T., Bonmassar, G., & Vanduffel, W. (2008). Bottom-up dependent gating of frontal signals in early visual cortex. *Science*, *321*, 414-417.

Fischer, B., & Rampsberger, E. (1984). Human express saccades: Extremely short reaction times to goal directed eye movements. *Experimental Brain Research*, *57*, 191-195.

Freeman, W. J. (2007). Indirect biological measures of consciousness from field studies of brains as dynamical systems. *Neural Networks*, *20*, 1021-1031.

Fries, P., Reynolds, J. H., Rorie, A. E., & Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, *291*, 1560-1563.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society London B Biological Sciences*, *364*, 1211-1221.

Gärdenfors, P. (2000). Conceptual spaces: The geometry of thought. *The MIT Press*.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society London B Biological Sciences*, *290*, 181-197.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering D*, *82*, 35-45.

Copyright 2010-2014 CEEDS Consortium

Kuhn, G., Amlani, A. A., & Rensink, R. A. (2008). Towards a science of magic. *Trends in Cognitive Sciences*, *12*, 349-354.

Lamme, V. A., Supèr, H., Landman, R., Roelfsema, P. R., & Spekreijse, H. (2000). The role of primary visual cortex (v1) in visual awareness. *Vision Research*, *40*, 1507-1521.

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences (USA)*, *103*, 18763–18768.

Mathews, Z., Badia, S. Bermúdez i, & Verschure, P. F. M. J. (2008). Intelligent motor decision: From selective attention to a bayesian world model. *4th International IEEE Conference on Intelligent Systems*, *1*, 8-13.

Paas, F. G. W. C., & Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*, 351-371.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). Numerical recipes in c (2nd ed.): the art of scientific computing. *Cambridge University Press, New York, NY, USA*.

Purves, D., Lotto, R. B., Williams, S. M., Nundy, S., & Yang, Z. (2001). Why we see things the way we do: evidence for a wholly empirical strategy of vision. *Philosophical Transactions of the Royal Society of Lond B Biological Sciences*, *356*, 285-297.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 179-197.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, *28*, 1059-1074.

Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area v1. *Journal of Neuroscience*, *30*, 3531-3543.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science*, *314*, 1311-1314.

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception and Psychophysics*, *68*, 1047-1058.

Tse, P. U., Whitney, D., Anstis, S., & Cavanagh, P. (2011). Voluntary attention modulates motion-induced mislocalization. *Journal of Vision*, *11*(3).

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chu, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, *30*, 11177-11187.

Verschure, P. F. M. J., Voegtlin, T., & Douglas, R. J. (2003). Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature*, *425*, 620-624.

Victor, J. D. (2005). Analyzing receptive fields, classification images and functional images: challenges with opportunities for synergy. *Nature Neuroscience*, *8*, 1651-1656.

Watanabe, K., & Shimojo, S. (1998). Attentional modulation in perception of visual motion events. *Perception*, *27*, 1041-1054.

Winges, S. A., & Soechting, J. F. (2011). Spatial and temporal aspects of cognitive influences on smooth pursuit. *Experimental Brain Research*, *211*, 27-36.

©
Copyright 2010-2014 CEEDS Consortium