A Novel Framework for Unconscious Processing

David Soto[a,b,*], Usman Ayub Sheikh[a], Clive R. Rosenthal[c]

[a] *Basque Center on Cognition, Brain and Language, San Sebastián, Spain*
[b] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*
[c] *Nuffield Department of Clinical Neurosciences, University of Oxford, United Kingdom*

Abstract

Understanding the distinction between conscious and unconscious cognition remains a priority in psychology and neuroscience. A comprehensive neurocognitive account of conscious awareness will not be possible without a sound framework to isolate and understand unconscious information processing. Here we provide a brain-based framework that allows the identification of unconscious processes even with null effects on behaviour.

Keywords: conscious awareness, visual cognition, memory, unconscious processes, neuroimaging, machine learning, pattern analyses, computational models

*Correspondence to: d.soto@bcbl.eu, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor 20009 San Sebastián

March 3, 2019

Prior research has implicitly assumed that demonstrations of unconscious information processing require a behavioural effect triggered by unconscious stimulation. Fundamental challenges however remain to isolate unconscious information processing from behavioural tests. Here we first review these challenges. Then we propose a novel framework that leverages recent advances in neuroimaging technology and computational models to isolate unconscious information processing from brain activity patterns. In this framework, behavioural effects from unconscious stimuli are dispensable to demonstrate the existence of unconscious representations. Our framework can be used to predict whether or not unconscious stimuli may influence behaviour.

Behavioural tests of unconscious processing

Over the last decade or so, there has been a palpable shift towards the use of subjective reports to identify states of visual unawareness [1]. Trials on which observers report no awareness of a stimulus are used to infer the properties of unconscious processing mechanisms across multiple cognitive domains, including visual perception [2], learning and memory [3].

Relying on subjective measures of awareness to understand unconscious information processing remains the subject of ongoing strong criticism [4]. One major concern is that weak but above chance sensitivity may involve conscious knowledge in the service of behaviour even when observers report no awareness of the information. Another is that subjective reports of (un)awareness may not exhaustively reveal all relevant knowledge (i.e. observers may not report knowledge held with very low confidence).

Current behavioural paradigms (e.g. subliminal priming, force-choice discrimination of masked items followed by (in)visibility ratings), do not decisively meet all the different criteria to measure awareness appropriately, namely, reliability, sensitivity, relevance, inmediacy and exhaustiveness [4]. While subjective ratings are undeniably useful for understanding conscious visual awareness, it seems clear that signal detection measures offer the best principled way to demonstrate that an observer lacks awareness of the critical information (i.e. null sensitivity). Figure 1 shows an example of the behavioural tests of visual awareness using subjective ratings and objective measures based on signal detection theory, and further illustrates the possibility that brain measures can reveal unconscious information processing even with null sensitivity at the behavioural level.

A brain-based framework

Unconscious events associated with null perceptual or mnemonic sensitivity may only produce weak effects on the guidance of behaviour. Importantly, demonstrating the absence of behavioural effects from unconscious stimuli may be taken as evidence against the existence of unconscious processing. Crucially, in the framework we propose here this need not be the case. We contend that the representation of unconsciously processed items can be isolated through fine-grained analyses of brain activity patterns even in the context of behavioural protocols that are associated with null sensitivity. Moreover, this approach can reveal types of unconscious processes that would be missed if only studying conditions

that yield significant effects of unconscious items on behaviour. Leveraging the power of advanced neuroimaging technology, developments in machine learning and computational models, has the potential to uncover the properties of unconscious representations and unconscious information processing from analyses of brain activity patterns. In line with our framework, a number of reports have revealed brain markers of unconscious information processing in the absence of effects on behaviour, in syntactic and semantic processing [5, 6] and short-term memory tasks [7], but these have thus far been largely confined to standard uni/multi-variate contrasts.

The framework we propose also has the potential to circumvent problems associated with the use of subjective reports in the study of unconscious processing. In particular, when unconscious processes are assessed under protocols with above chance objective and/or diagnostic subjective measures, the neural correlates of unconscious information processing may be confounded with the neural correlates of attendant cognitive functions that are associated with (conscious) access – for example, search, decision making, verbal report, and response selection (see [8] for a counterpart contention in the evaluation of the neural correlates of consciousness).

Exploiting machine learning, pattern analyses and computational models

Figure 2 illustrates some of the approaches currently used in our lab to uncover the brain representation of unconsciously processed information. These include the use of transfer learning based on models pre-trained using conscious items that are re-used to characterise unconscious knowledge. The potential of transfer learning lies in the re-use of pre-trained models from different experiments/domains/task/stimuli to address the common and distinct brain representational spaces across different states of (un)awareness. Another key area is the analysis of the dis-similaries between the brain representations of conscious and unconscious items through the use of computational models. These can also be used to characterize the dynamics of the brain representations across states of (un)awareness (see Figure 2 for details).

One example of the empirical and theoretical questions that this framework can tackle is the extent to which the representation of unconscious items map onto the brain activity patterns associated with conscious counterparts. How unconscious and conscious states differ during information processing remains unclear. For instance, the global neuronal workspace model proposes distinctive functional architectures supporting conscious processes (global, large-scale, dynamic and sustained) and non-conscious processing (local, domain-specific and transient) [1]. Our proposed framework can be used to identify the common and distinct brain activity patterns that support (un)conscious representations, under conditions that mitigate the confounding effects from downstream processes related to subjective reporting on the contents of consciousness. More broadly, this framework can be used to determine whether a particular brain activity pattern is sufficient to be associated with a particular level of (un)conscious processing and/or representational state.

Unconscious processing across domains

A key issue is the extent to which unconscious information processing in this framework can be isolated across multiple cognitive domains of perception, learning and memory, language and emotion. For instance, implicit learning studies involve acquiring knowledge about complex spatiotemporal regularities for visible items. Concluding that the knowledge is unconscious has often been based on subjective awareness tests that do not meet stringent criteria (Figure 1; see [4]). One solution would be to investigate brain markers of implicit learning both in the absence and presence of behavioural sensitivity in the knowledge tests [12].

Addressing the dynamic transitions between unconscious and conscious knowledge states during learning remains a challenge. This research will benefit from the development of algorithms for multidimensional time series analyses to discover the properties of the brain representations as they transition between conscious and unconscious informational states during learning.

More generally we believe the impact of our framework is likely to be most substantial when studying higher-order cognition because here it has been difficult to investigate above chance performance without awareness.

Caveats

Our framework does not fully address the extent to which unconscious information processing influences behaviour. We think our proposal is a special case in which a reductionist neuroscience approach is needed to circumvent the long-held concerns with existing approaches to unconscious information processing. The framework however could be developed to test whether or not and when an unconscious representation is associated with behaviour when a particular brain pattern is present or absent, irrespectively of first-person report.

Another caveat is that by enforcing null behavioural sensitivity we run the risk of reducing the detectability of the brain signal for decoding unconscious information processing. Developments in behavioural protocols will thus be needed to optimise the signal in brain measures that can be obtained under conditions of null behavioural sensitivity.

Acknowledgements

D.S. acknowledges support from the Spanish Ministry of Economy and Competitiveness, through the 'Severo Ochoa' Programme for Centres/Units of Excellence in R & D (SEV-2015-490) and project grants PSI2016-76443-P from MINECO and PI-2017-25 from the Basque Government.

[1] Stanislas Dehaene. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking, 2014.

[2] Moti Salti, Simo Monto, Lucie Charles, Jean-Remi King, Lauri Parkkonen, and Stanislas Dehaene. Distinct cortical codes and temporal dynamics for conscious and unconscious percepts. *Elife*, 4:e05652, 2015.

[3] Clive R. Rosenthal and David Soto. The anatomy of non-conscious recognition memory. *Trends in Neurosciences*, 39(11):707–711, nov 2016.
[4] Ben R Newell and David R Shanks. Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, 37(1):1–19, 2014.
[5] Vadim Axelrod, Moshe Bar, Geraint Rees, and Galit Yovel. Neural correlates of subliminal language processing. *Cerebral Cortex*, 25(8):2160–2169, 2014.
[6] Usman Ayub Sheikh, Manuel Carreiras, and David Soto. Decoding the meaning of unconsciously processed words using fMRI-based MVPA. *NeuroImage*, feb 2019.
[7] Fredrik Bergström and Johan Eriksson. Neural evidence for non-conscious working memory. *Cerebral Cortex*, 28(9):3217–3228, aug 2017.
[8] Naotsugu Tsuchiya, Melanie Wilke, Stefan Frässle, and Victor A.F. Lamme. No-report paradigms: Extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences*, 19(12):757–770, dec 2015.
[9] Christiane Lemke, Marcin Budka, and Bogdan Gabrys. Metalearning: a survey of trends and technologies. *Artificial Intelligence Review*, 44(1):117–130, jul 2013.
[10] Jean-Rémi King, Niccolo Pescetelli, and Stanislas Dehaene. Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron*, 92(5):1122–1134, dec 2016.
[11] Stefano Anzellotti and Marc N. Coutanche. Beyond functional connectivity: Investigating networks of multivariate representations. *Trends in Cognitive Sciences*, 22(3):258–269, mar 2018.
[12] Clive R. Rosenthal, Indira Mallik, Cesar Caballero-Gaudes, Martin I. Sereno, and David Soto. Learning of goal-relevant and -irrelevant complex visual sequences in human v1. *NeuroImage*, 179:215–224, oct 2018.

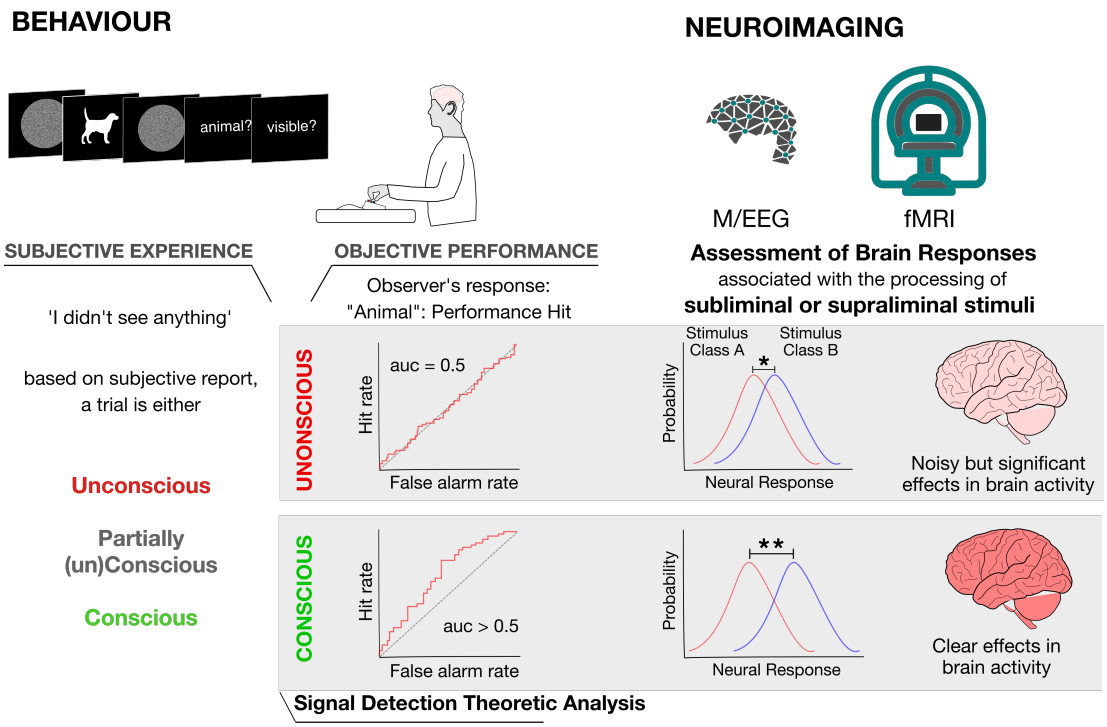

Figure 1: **Testing unconscious information processing with behavioural and brain measures**. In a typical scenario, an observer provides subjective reports about a target stimulus using different measures of awareness (e.g. perceptual visibility or confidence ratings) and also provides specific categorical judgements ("animal"). Different measures of the observer's awareness are collected (e.g. visibility and confidence ratings, post-decision wagers etc). However, meeting all the criteria for a thorough assessment of the observer's awareness -e.g. the reliability, sensitivity, exhaustiveness of tests [4]- can be challenging. Signal detection theory (SDT) alongside carefully designed experiments and analytical methods can meet the different criteria that are needed to assess awareness precisely. Within a SDT model, the area under the curve (auc) that relates the rate of response hits and false alarms is 0.5 when the observer lacks sensitivity and randomly classifies the relevant information (auc = 1 means optimal sensitivity for classification). It should be noted that because signal detection thresholds may vary across different testing sessions, any awareness test using signal detection measures must be contained in the same experimental session that aims to demostrate behavioural or brain effects from unconscious items, rather than post-hoc as typically happens in subliminal priming studies [1]. A combination of moment-to-moment subjective reports of (un)awareness with signal detection measures is also useful to probe the lack of sensitivity for items rated as unaware by observers. We propose that behavioural measures of consciousness are dissociable from brain measures. In particular, it is possible that even when the observer lacks sensitivity to the relevant information (auc = 0.5), measures of brain activity can be used to decode relevant information from the unconscious stimuli.

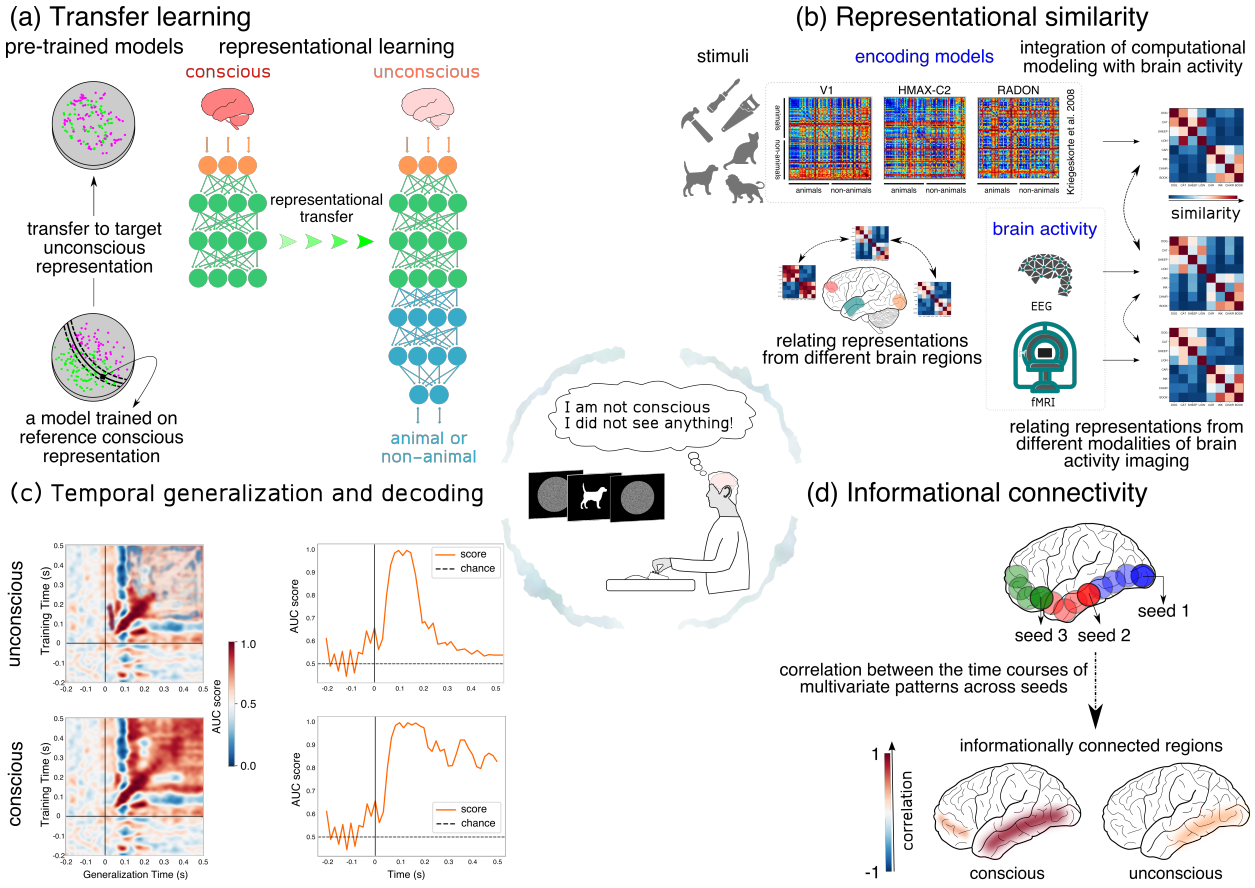

Figure 2: **Uncovering unconscious brain representations with computational models**. A: Using transfer learning, the decoding of unconscious representations may be improved through the transfer of knowledge from related conscious representations. This can be investigated through the re-use of pre-trained classification models or unsupervised representational learning. The latter is usually achieved through a multi-layered neural network [9]. As shown, the final weights of the source network trained on brain data from conscious items are used to initialize the first few layers (in green) of the target network and extract relevant features for the learning of unconscious items. This target network may also comprise a few untrained layers (in blue) that are fine-tuned using brain data from unconscious items. B: Representational dissimilarity analyses can address which computational (i.e. encoding) models (e.g. V1,HMAX-C2, RADON or semantic models such as Word2vec) exhibit the strongest correlation with the brain activity patterns associated with unconscious items, and assess how they differ from the representations of conscious counterparts.C: Temporal generalization methods can reveal dynamic properties of unconscious representations over time (i.e. metastability) by training a classifier to discriminate the relevant information content at a given time point and testing it in the remaining time points [10]. This approach has provided insight into the maintenance in working memory of items masked from visual awareness [10], although this remains to be tested for items associated with null sensitivity. (*continued on the next page*)

Figure 2 (*previous page*): D: Informational connectivity analyses can inform how inter-areal exchange of information differ between conscious and unconscious processes. This is given by the temporal correlations in classification accuracy revealed by multivoxel pattern classifiers across brain areas [11].