Informative neural representations of unseen objects during higher-order processing in human brains and deep artificial networks

Ning Mei

Basque Center on Cognition, Brain and Language, San Sebastian, Spain
ORCID iD:0000-0002-7864-5795

Roberto Santana

Computer Science and Artificial Intelligence Department, University of Basque Country, Spain
ORCID iD: 0000-0002-1005-8535

David Soto

Basque Center on Cognition, Brain and Language, San Sebastian, Spain
Ikerbasque, Basque Foundation for Science, Bilbao, Spain
ORCID iD: 0000-0003-0205-7513

February 1, 2021

Correspondence to: Ning Mei: n.mei@bcbl.eu; David Soto: d.soto@bcbl.eu, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor 20009 San Sebastian

Abstract

Despite advances in the neuroscience of visual consciousness over the last decades, we still lack a framework for understanding the scope of unconscious processing and how it relates to conscious experience. Previous research observed brain signatures of unconscious contents in visual cortex, but these have not been identified in a reliable manner, with low trial numbers and signal detection theoretic constraints not allowing to decisively discard conscious perception. Critically, the extent to which unconscious content is represented in high-level processing stages along the ventral visual stream and linked prefrontal areas remains unknown. Using a within-subject, high-precision, highly-sampled fMRI approach, we show that unconscious contents, even those associated with null sensitivity, can be reliably decoded from multivoxel patterns that are highly distributed along the ventral visual pathway and also involving prefrontal substrates. Notably, the neural representation in these areas generalised across conscious and unconscious visual processing states, placing constraints on prior findings that fronto-parietal substrates support the representation of conscious contents and suggesting revisions to models of consciousness such as the neuronal global workspace. We then provide a computational model simulation of visual information processing/representation in the absence of perceptual sensitivity by using feedforward convolutional neural networks trained to perform a similar visual task to the human observers. The work provides a novel framework for pinpointing the neural representation of unconscious knowledge across different task domains.

Introduction

The neuroscience of consciousness aims to explain the neurobiological basis of subjective experience - the personal stream of perceptions, thoughts and beliefs that make our inner world. Understanding the distinction between conscious and unconscious information processing remains a key unresolved issue. This is paramount for developing a comprehensive neuroscientific account of consciousness and its role in cognition and behaviour. Influential neurocognitive models of visual consciousness such as the global neuronal workspace model propose that conscious awareness is associated with sustained activity in large-scale association networks involving fronto-parietal cortex, making information globally accessible to systems involved in working memory, report and behavioural control (Dehaene, 2014). Unconscious visual processing, on the other hand, is thought to be transient and operate locally in domain-specific systems -i.e. supporting low-level perceptual analysis (V. A. F. Lamme, 2020). Recent studies have however confronted this view with intriguing data suggesting that unconscious information processing is implicated in higher-order operations associated with cognitive control (Van Gaal & Lamme, 2012), memory-guided behaviour across both short- and long-term delays (Soto, Mäntylä, & Silvanto, 2011; Trübutschek et al., 2017; Rosenthal, Andrews, Antoniades, Kennard, & Soto, 2016; Chong, Husain, & Rosenthal, 2014; Wuethrich, Hannula, Mast, & Henke, 2018), and also language computations (Hassin, 2013); however, follow-up work did not support this view (Rabagliati, Robertson, & Carmel, 2018), and even the evidence for unconscious semantic priming has been recently called into question (Kouider & Dehaene, 2007; Stein, Utz, & van Opstal, 2020). The limits and scope of unconscious information processing remain to be determined.

This controversy is likely to originate from the lack of a sound framework to isolate unconscious information processing (Soto, Sheikh, & Rosenthal, 2019). Studies often rely only on subjective measures of (un)awareness (Overgaard, Timmermans, Sandberg, & Cleeremans, 2010) to pinpoint the neural markers of unconscious processing, but these measures are sensitive to criterion biases for deciding to report the presence or absence of awareness (Peters & Lau, 2015), hence making it impossible to determine whether "subjectively invisibility" is truly associated with unconscious processing. Previous studies reported brain signatures of unconscious contents in visual cortex (Sterzer, Haynes, & Rees, 2008; Haynes & Rees, 2005; Jiang, Zhou, & He, 2007; Dehaene, Naccache, Cohen, Bihan, et al., 2001), but these signatures have not been identified in a reliable manner (Fang & He, 2005; Hesselmann, Hebart, & Malach, 2011; Ludwig & Hesselmann, 2015; Ludwig, Kathmann, Sterzer, & Hesselmann, 2015). In these studies using objective measures of (un)awareness, perceptual sensitivity tests are collected off-line, outside the original task context, and typically employ a low number of trials per participant to conclusively exclude conscious awareness and meet the null sensitivity requirement (Macmillan, 1986; Newell & Shanks, 2014). The standard, current approach to study unconscious information processing is therefore limited.

Here we present the first high-precision, highly-sampled, within-subject approach to pinpoint the neural representation of unconscious contents, even those associated with null perceptual sensitivity, by leveraging the power of machine learning and biologically plausible computational models of visual processing. Critically, the extent to which unconscious content is represented in high-level processing stages along the ventral visual stream and linked prefrontal areas (Kravitz, Saleem, Baker, Ungerleider, & Mishkin, 2013) remains unknown. Previous functional MRI studies indicate the role of conscious awareness in this regard; object categories of visible stimuli are represented in ventral-temporal cortex (Haxby et al., 2001; Naselaris, Kay, Nishimoto, & Gallant, 2011; Kriegeskorte, 2011) and parieto-frontal cortex is involved in the representation of conscious perceptual content (Ester, Sprague, & Serences, 2015; Christophel, Hebart, & Haynes, 2012; Kapoor et al., 2020). Here we used a high-precision fMRI paradigm to contrast these views. We further asked the extent to which the representation of unconscious content maps onto the representations of the conscious counterparts. This issue remains unsolved (Sterzer et al., 2008; Schurger, Pereira, Treisman, & Cohen, 2010), yet it has ramifications for models of consciousness such as the neuronal global workspace (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006).

Subsequently, we used deep feedforward convolutional neural network models (FCNNs) (Hinton, Vinyals, & Dean, 2015; LeCun & Bengio, 1995; McFee, Salamon, & Bello, 2018) to provide a representational level (Marr, 1982) simulation of visual representations/processing in the absence of perceptual sensitivity. FCNNs were used given their excellent performance in image classification (Hinton et al., 2015; Kriegeskorte & Douglas, 2018; Kietzmann, McClure, & Kriegeskorte, 2019) and given the known similarities between the representational spaces during object recognition in FCNNs and high-level brain regions in ventral visual cortex (Yamins & DiCarlo, 2016; Kriegeskorte, 2015). FCNNs performed the same task given to the human participants using the same images corrupted by different levels of noise.

We asked whether, similar to the brain, the semantic category of the stimulus could be decoded by analysing the activity state of the hidden layer of the FCNN network, despite the network itself had no perceptual sensitivity at identifying the image class.

Results

Observers (N = 7) performed six fMRI sessions across 6 days leading to a total of 1728 trials per subject, allowing us to pinpoint meaningful and reliable neural patterns of conscious and unconscious content within each observer. Observers were presented with gray-scaled images of animate and inanimate objects with a random-phase noise background (Moreno-Martínez & Montoro, 2012). The target images were presented briefly, preceded and followed by a dynamic mask composed of several frames of gaussian noise. On each trial of the fMRI experiment, participants were required to discriminate the image category and to indicate their subjective awareness (i.e. (i) no experience/just guessing (ii) brief glimpse (iii) clear experience with a confident response). The duration of the images was based on an adaptive staircase that was running throughout the experiment (see Methods), and which, based on pilot testing, was devised to obtain a high proportion of unconscious trials. On average, target duration was (i) 25 ms on trials rated as unaware (ii) 38 ms on glimpse trials and (iii) 47 ms on the aware trials.

Figure 1 illustrates an example of a trial.

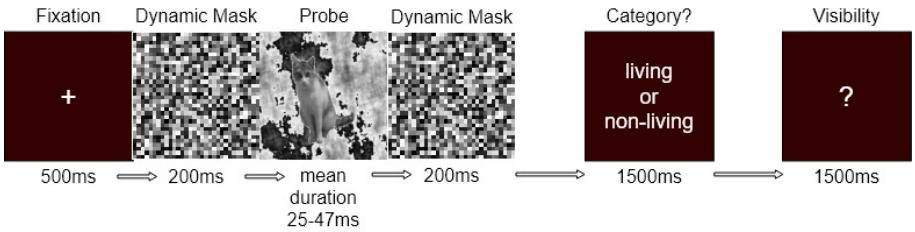


Figure 1: Example of the sequence of events within an experimental trial. Observers were asked to discriminate the category of the masked image (living vs. non-living) and then rate their visual awareness on a trial by trial basis. Example of a trial with a masked image of a cat.

Behavioral performance

We assessed whether observers' performance at discriminating the image category from chance in each of the awareness conditions by using signal detection theoretic measure to index perceptual accuracy, namely, A' (Zhang & Mueller, 2005). Permutation tests were performed to estimate the empirical chance level within each observer (see Methods). All observers displayed above chance perceptual sensitivity in both glimpse and visible trials (highest $p < 0.00001$, permuted p-values). Four of the seven subjects showed null perceptual sensitivity ($0.16 < p < 0.64$) in those trials in which participants reported a lack of awareness of the images. Discrimination performance in two additional participants deviated from chance (p values < 0.02, 0.03) but only one observer clearly showed above chance performance in the unaware trials ($p < 0.00036$). Figure 2 illustrates the distribution of A' values alongside the chance distribution for each participant.

FMRI Decoding results

We then used a linear support vector machine with out-of-sample generalization to decode the categories of out-of-sample target images in the unconscious and the conscious trials. Trials in which observers reported a glimpse were a minority and accordingly, we elected to focus on the critical unconscious and conscious trials. The classifier was fed with multi-voxel patterns of BOLD responses in a set of 12 a priori regions of interest comprising the ventral visual pathway and higher-order association cortex (see below and Figure S2). Permutation tests were run within each subject to estimate the reliability of the decoding at the single subject level (see Methods)

In the unconscious trials, the image class was significantly decoded from activity patterns in visual cortex, including high-level areas in the ventral visual cortex and even in prefrontal regions. Specifically, activity patterns in the fusiform cortex allowed for decoding of unconscious contents reliably within each of the four observers showing null perceptual sensitivity, and moreover the unconscious content

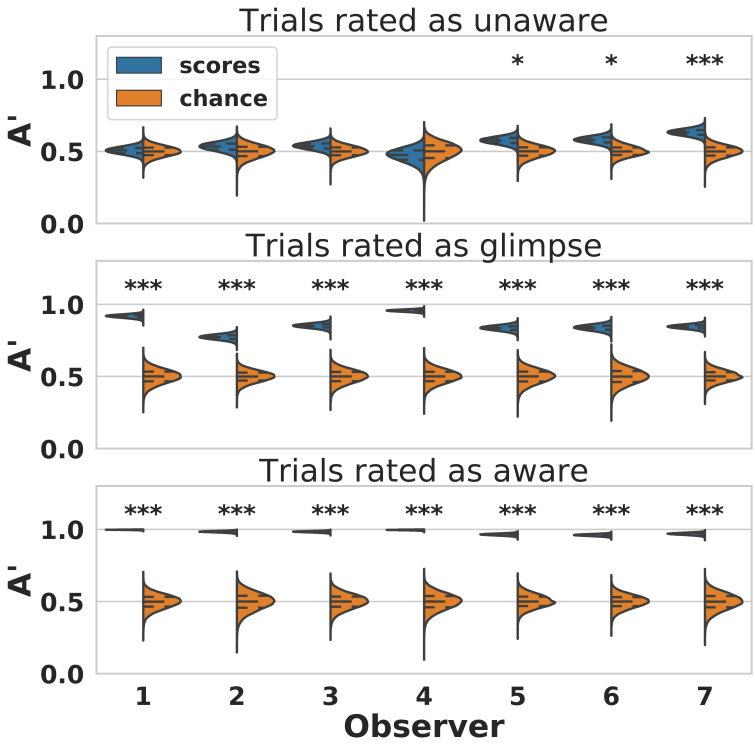


Figure 2: Behavioral performance. Distribution of within-observer A' scores with mean, first and third quartile, and the corresponding empirical chance distributions for each observer and awareness state. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

could be decoded from prefrontal areas in middle and inferior gyrus in these observers (observers 1 - 4). Figure 3 illustrates the decoding results. A similar pattern was observed in the participants whose perceptual sensitivity deviated from chance (observers 5 - 7). Yet, there were no apparent differences between the pattern of decoding performance among the observers that were at chance vs. those showing above chance discrimination performance on the trials rated as unaware, even for the observer whose perceptual sensitivity was clearly above chance (observer 7). Further inspection of the distribution of decoding accuracies of the observers that were at chance and that deviated from chance did not reveal any advantage for the observers whose sensitivity was above chance in the trials rated as unaware (see Supplementary Information).

In the conscious trials, the level of decoding accuracy was highly reliable in all subjects tested, across all visual ROIs in the ventral visual pathway, its linked higher-order regions in the inferior frontal cortex (Kravitz et al., 2013), and also inferior and superior parietal cortex. Due to the online adaptive staircase that was running throughout the experiment to achieve null sensitivity on the unconscious trials, the signal to noise ratio of the image was higher in the conscious trials. Note that if stimulus properties were kept constant throughout the experiment it would have been impossible to obtain trials associated with null perceptual sensitivity alongside the conscious trials. Then, using transfer learning we investigated whether the multivoxel brain representation of perceptual contents in the visible, conscious trials was similar to those of the unconscious trials. Accordingly, we trained the classifier in the conscious trials and then performed out-of-sample cross-validation in the unconscious trials. The fact that conscious and unconscious stimulus differed in signal strength actually makes this generalization test stronger.

The results showed that a decoder trained in the conscious trials using multivoxel patterns in fusiform gyrus, lateral occipital cortex, and precuneus generalised well to predict the target image in the unconscious trials, remarkably, in all subjects. Also, a decoder trained in the conscious trials with BOLD activity patterns in inferior parietal lobe, inferior temporal lobe, lingual gyrus, middle frontal gyrus, and superior parietal gyrus generalised to the unconscious trials in 6 out of 7 subjects. There was some variability across observers in the generalization from conscious to unconscious representations in prefrontal areas, but this was successful in all observers in either the middle/inferior prefrontal cortex, except for one the participants showing null perceptual sensitivity in the unconscious trials. Across all observers, we observed that multivoxel patterns in the inferior frontal cortex, and also in pericalcarine cortex, gen-

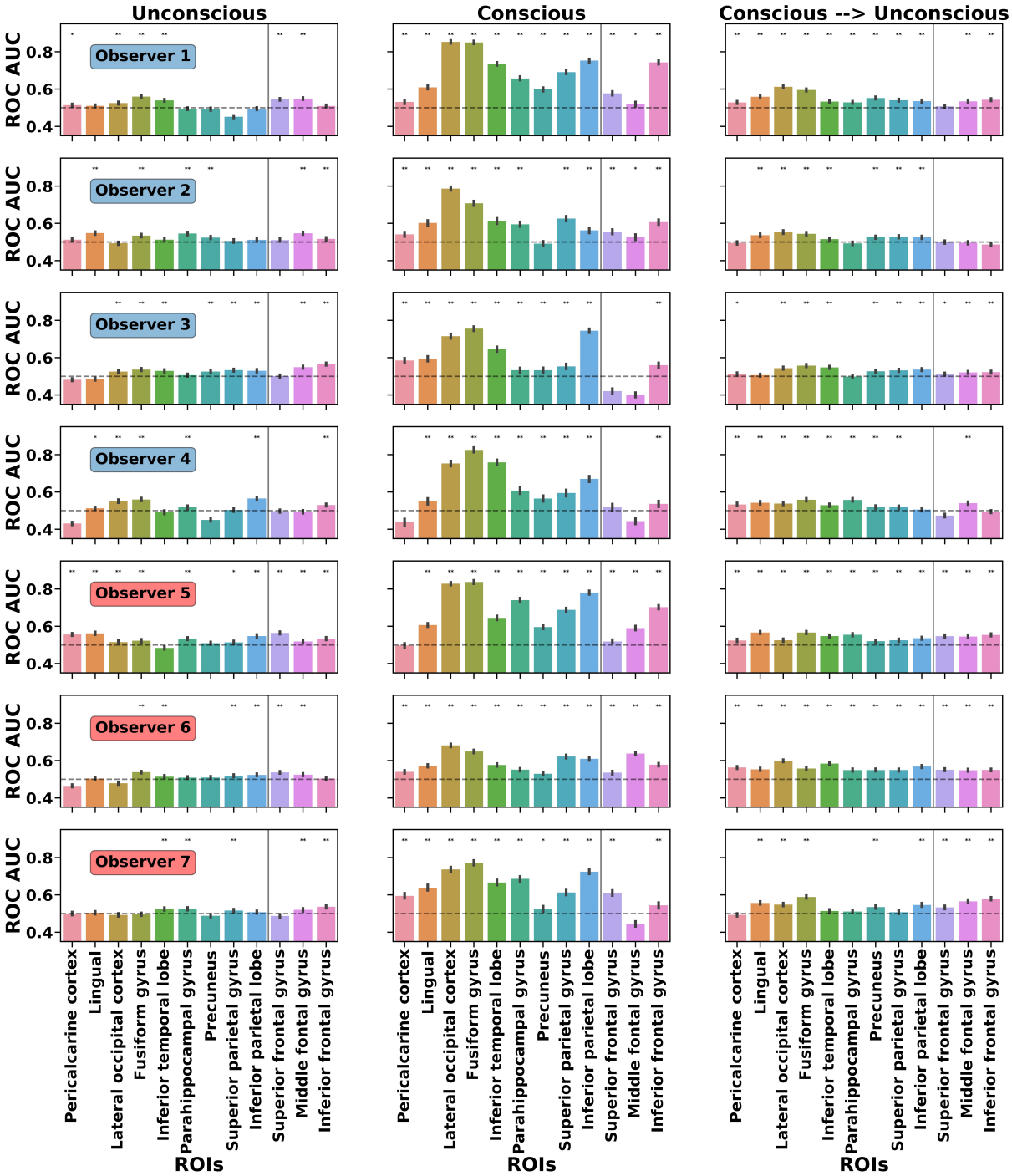


Figure 3: Decoding accuracy for out-of-sample images for each observer across the unconscious and conscious trials. *:p < 0.05, **:p < 0.01, ***:p < 0.001, after multiple comparison correction for the number of ROIs tested for each observer. Error bars represent the standard error of the mean.

eralised from conscious to unconscious in 5 of them. Taken together this pattern of results indicates the presence of invariant multivariate patterns in both the visual areas and the frontal regions for the same item categories in both conscious and unconscious conditions.

Feedforward convolutional neural network (FCNN) model simulations

Then, we sought to provide a representational model simulation using artificial FCNNs (Fukushima, 1980; LeCun & Bengio, 1995; McFee et al., 2018; Hinton et al., 2015)). The goal was to simulate the observation of informative neural representations despite null perceptual sensitivity. FCNNs were

trained to perform a similar visual task to the human observers using the same images across different levels of gaussian noise. FCNNs are known to be excellent in image classification (Hinton et al., 2015; Kriegeskorte & Douglas, 2018; Kietzmann, McClure, & Kriegeskorte, 2019). We expected the level of classification accuracy of the FCNN network to drop with increasing levels of noise in the image.

84,000 FCNN model simulations were performed, resulting from combining 5 pre-trained FCNN configurations, 7 hidden layer units, 4 dropout rates, 6 hidden layer activation functions, 2 output layer activation functions, and 50 noise levels (see Methods). Because we were interested in those poorly performing FCNNs, we only attempted to decode the stimulus category from the hidden layer in those cases in which the FCNN ROC-AUC classification performance was lower than 0.55. We observed that 61,086 of the FCNN models showed a ROC-AUC score below 0.55 under conditions of increasing noise in the image.

Informative hidden layer representations in the FCNN models

Then, we asked whether, despite the FCNN failing to classify the image, the semantic category of the stimulus could still be decoded by analysing the activities of the hidden layer of the network. To test this, a linear SVM was applied to the hidden layer representation for decoding the image class across different levels of noise, even when the FCNN model classification performance was at chance (see Methods). Previous studies modeled visual recognition using FCNN (Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & van Gerven, 2015), demonstrating that the last hidden layer of FCNNs has representational spaces that are similar to those in high-level regions in ventral visual cortex (Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & van Gerven, 2015; Kriegeskorte et al., 2008). Therefore, we focused our analyses on the very last hidden layer of the FCNN in the current study, also considering limitations in computational resources due to the large number of simulations (see Methods).

Figure 4[a] shows the classification performance of the FCNN models (black) and also the decoding accuracy SVM applied to the hidden layer representation of the FCNN (blue) as a function of the level of noise and the different factors. When the level of noise was low, FCNN models could classify the category of the images very well reaching ROC-AUC scores higher than 0.9 but performance dropped with the level of gaussian noise. The observed logarithmic downward trend could be due to the exponential sampling of noise levels (see Methods).

Remarkably, when the FCNN models failed to classify the noisy images ($p > 0.05$; $N = 32{,}435$), we observed that the hidden layer representation of these FCNN models contained information that allowed a linear SVM classifier to decode the image category above chance levels reliably in 12,777 of the simulations ($p < 0.05$, one sample permutation test). Figure 5 illustrates the decoding results based on the hidden layer representation when the FCNN was at chance.

It is noted that even when the noise level was relatively low, some FCNN models such as AlexNet and ResNet did not perform well in the image classification task. Inspection of these models indicated poor performance in the validation phase of the training (prior to testing), which suggests that particular combinations of hidden layer units, activation function, and dropout rate in AlexNet and ResNet impeded learning the classes properly.

When the noise level was relatively low and the FCNN models failed to discriminate the noisy images ($N = 7{,}841$), 75.39% of the linear SVMs could decode the FCNN hidden layers and the difference in decoding performance between SVM and FCNN was significantly greater than zero ($p < 0.001$, one sample permutation test).

Remarkably, even when the noise level was higher and the FCNN classified the images at chance level, 27.91% of the linear SVMs could decode the image category from the FCNN hidden layers. Crucially, the comparison of SVM decoding from the hidden layer and the FCNN classification performance including those 24,584 cases in which the FCNN was at chance again showed a significant difference (permutation $p < 0.05$), demonstrating that the hidden layer of the FCNN contained informative representations despite the FCNN classification performance was at chance.

MobileNet produced more informative hidden representations that could be decoded by the SVMs compared to other candidate model configurations. It was observed that the classification performance of ResNet models trained with different hidden units, dropout rates, etc did not fall to chance level until the noise level was relatively high (closer to the dashed line) and the proportion of SVMs being able to decode FCNN hidden layers was higher compared to other model configurations (50.62% v.s. 46.92% for MobileNet, 35.51% for AlexNet, 31.35% for DenseNet, and 30.22% for VGGNet). Additionally, we observed that even when the noise level was high, the MobileNet models provided a higher proportion of

[a] High definition figure: https://tinyurl.com/y6dhls7c

Figure 4: The black dots illustrate the classification performance of the FCNN models, as a function of noise level, type of pre-trained model configuration (column) and activation functions (row). The blue dots illustrate the classification performance of the linear SVMs applied to the hidden layer of the FCNN model when the FCNN classification performance was lower than 0.55.

hidden representations that were decodable by the linear SVMs (34.77% v.s. 29.53% for ResNet, 27.84% for DenseNet, 26.35% for AlexNet, and 21.62% for VGGNet, see Figure 5).

Then, we sought to further understand the influence of the components of the FCNN architecture (i.e. dropout rate, number of hidden units) on decoding performance. We used a random forest classifier to compute the feature importance of the different FCNN components for predicting whether or not the SVM decoded the image class based on the hidden layer representation. The classification performance was estimated by random shuffle stratified cross-validation with 100 folds (80/20 splitting). In each fold, a random forest classifier was fit to predict whether or not the hidden representation was decodable on the training set, and then the feature importance was estimated by a permutation procedure on the test set (Fisher, Rudin, & Dominici, 2018; Altmann, Toloşi, Sander, & Lengauer, 2010). Briefly, for a given component (i.e. hidden layer activation function), the order of instances was shuffled while the order of the instances of other components was not changed, in order to create a corrupted version of the data. The dropped classification performance indicated how important a particular feature was. Figure 6 shows that the noise level in the image was the best indicator of whether a hidden representation was decodable, followed by model architecture, followed by the number of hidden units, and by the type of hidden activation and output activation functions. The least important feature was the dropout rate. A one-way ANOVA assessed the contribution of the noise level, model architecture, number of hidden units, type of hidden activation function, type of output activation function, and dropout rate, on the feature importance. There were significant differences between the components of the network models tested ($F(5, 594) = 2215.57$, $p < 0.001$). Post-hoc t tests showed that all the pairwise comparisons were reliable (lowest $p < 0.015$, Bonferroni corrected for multiple comparisons).

Discussion

We tested a high-precision, within-subject framework to provide a representational account of the scope of information processing for unseen items, even those associated with null perceptual sensitivity (Soto

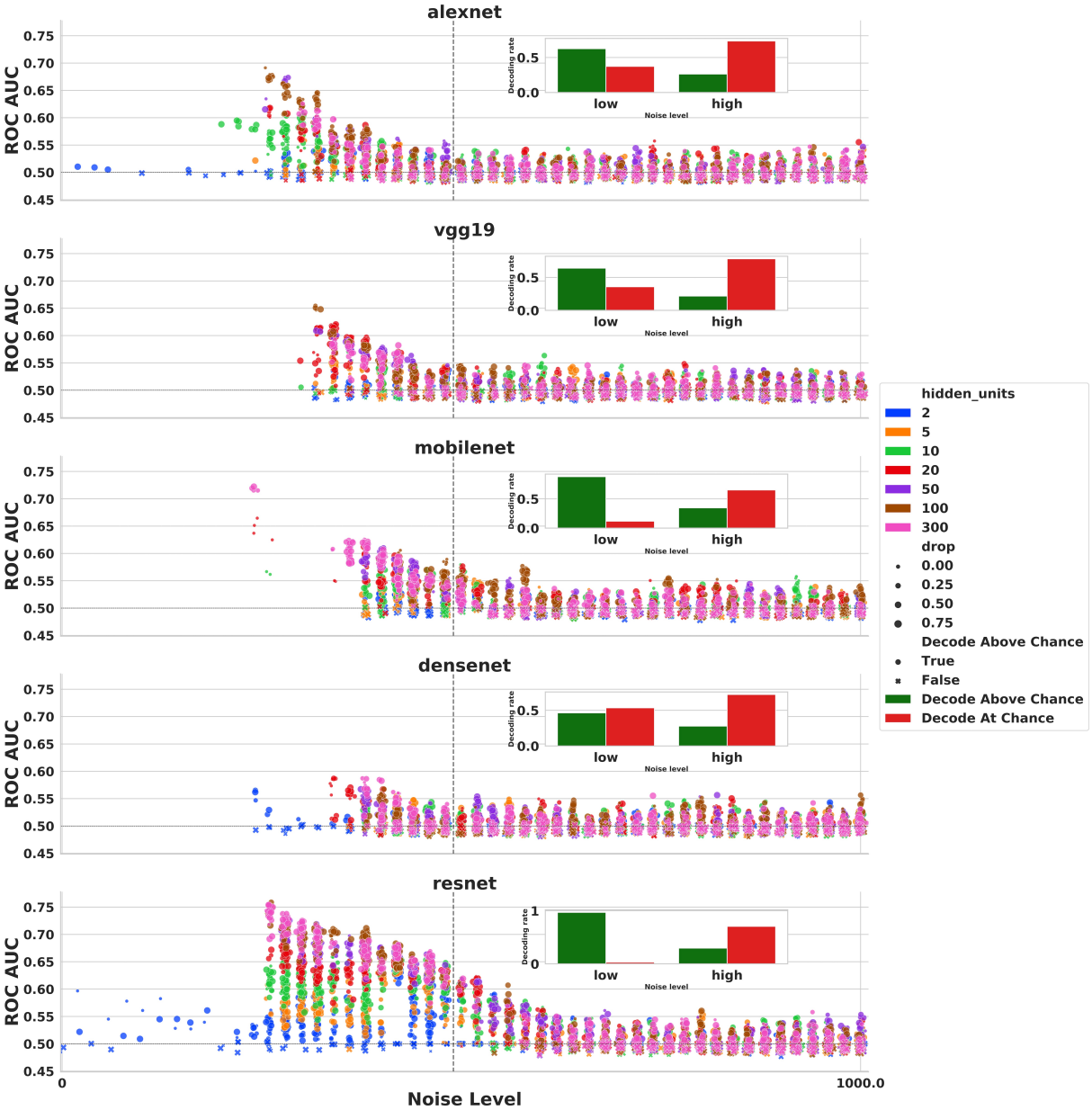


Figure 5: Image classification performance of the linear SVMs applied to the FCNN hidden layers when a given FCNN failed to discriminate the living v.s. nonliving categories, as a function of the noise level. The superimposed subplot depicts the proportion of times in which the linear SVM was able to decode the FCNN hidden layers as a function of low and high noise levels. The blue bar represents the proportion of linear SVMs being able to decode the FCNN hidden layers, while the orange bar represents the proportion of linear SVMs decoding the FCNN hidden layers at chance level.

et al., 2019) in both brains and deep artificial neural networks. Isolating the brain representation of unconscious contents has been difficult to achieve in systematic and reliable fashion in previous work, with low numbers of trials and signal detection theoretic constraints (Macmillan, 1986) not allowing to decisively discard conscious perception (Fang & He, 2005; Hesselmann et al., 2011; Gayet et al., 2020; Ludwig et al., 2015), and, critically, when unconscious content could be decoded, this was restricted to visual cortex - see also (Ludwig & Hesselmann, 2015). The current results demonstrate that when human participants and FCNNs models fail to recognise the image content, there remain informative representations of the unseen items in a hidden state of the network during high-level stages of information processing. These hidden representations allow for classification of the semantic category of unseen perceptual contents. Notably, the fMRI results from our high-precision, highly-sampled, within-subject approach showed that unconscious contents can be reliably decoded from multivoxel patterns that are

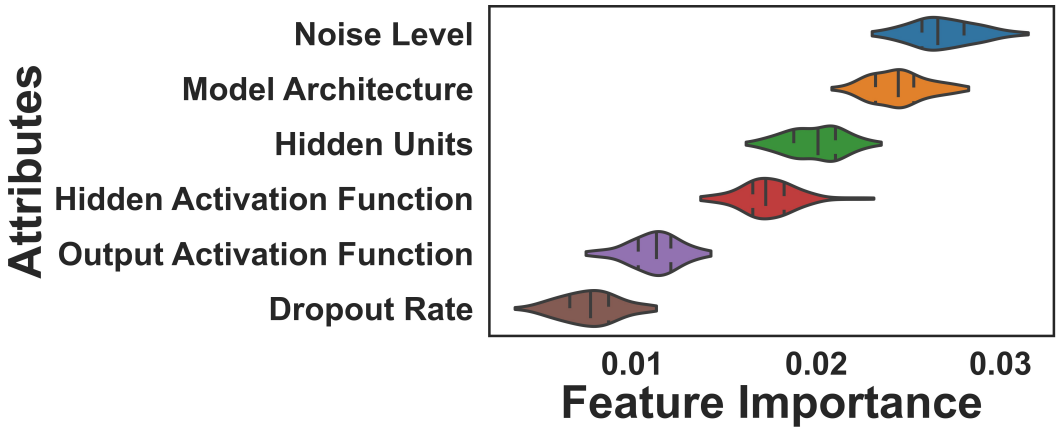


Figure 6: Feature importance of FCNN components that were manipulated in computational modeling. Feature importance was measured in arbitrary units. The number of hidden layer units, noise levels, and pre-trained configurations influenced the decoding performance of the image class based on the hidden layer of the FCNN when its classification performance was at chance.

highly distributed along the ventral visual pathway and also involving prefrontal substrates in middle and inferior gyrus. High-precision fMRI decoding paradigms can thus provide a richer information-based approach (Kriegeskorte, Goebel, & Bandettini, 2006) to reveal meaningful feature representations of unconscious content, and that otherwise would be missed.

The current findings have implications for models proposing that unconscious information processing is local and restricted to sensory cortex (V. A. F. Lamme, 2020). For instance, according to the neural global workspace model (Dehaene & Naccache, 2001), distributed activity patterns in fronto-parietal cortex are a marker of conscious access (Dehaene & Changeux, 2004). Both the middle frontal gyrus and inferior frontal areas have been implicated in the coding of visible items during working memory tasks (Ester et al., 2015) and also in binocular rivalry paradigms used to track moment-to-moment changes in the contents of consciousness (Kapoor et al., 2020). The inferior frontal cortex forms part of the ventral visual pathway that links extrastriate, and inferior temporal areas that is crucial for object recognition (Kravitz et al., 2013). Remarkably, the fMRI decoding results demonstrated the overlap between the neural representation of conscious and unconscious contents in these areas, with a decoder trained in the conscious trials generalising to predict unconscious contents. Visual consciousness may be associated with neural representations that are more stable across different presentations of the events (Schurger et al., 2010), but our data indicates that the underlying representational patterns in terms of perceptual content are to a significant extent invariant and generalised across awareness states, despite the non-linear dynamic changes in the intensity of the neural response that occur in fronto-parietal cortex during conscious processing (Dehaene & Changeux, 2004; Dehaene, Naccache, Cohen, Le Bihan, et al., 2001; Haynes, Driver, & Rees, 2005; Beck, Rees, Frith, & Lavie, 2001; Pessoa & Ungerleider, 2004; Kranczioch, Debener, Schwarzbach, Goebel, & Engel, 2005). Previous studies using lowly sampled fMRI designs could not reveal evidence consistent with this view (Sterzer et al., 2008; Schurger et al., 2010). The current observation that unconscious content is represented in high-order visual processing stages including the prefrontal cortex and that it does so in a similar way to the conscious counterparts points to revisions to the neuronal global workspace model (Dehaene, 2014).

The generalization of the feature representations across visibility states is also supported by the deep neural network model simulations. FCNNs initially trained with clear visible images, subsequently produced informative feature representations in the hidden layer when they were exposed to noisy images. Prior work showed that FCNNs are a good computational model of the ventral visual pathway (Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & van Gerven, 2015; Yamins, Hong, Cadieu, & DiCarlo, 2013; Yamins & DiCarlo, 2016; Kriegeskorte, 2015). FCNNs performed well the perceptual identification task with clear images, also in keeping with prior studies (Geirhos et al., 2017; Wichmann et al., 2017; Geirhos et al., 2018; Ghodrati, Farzmahdi, Rajaei, Ebrahimpour, & Khaligh-Razavi, 2014). FCNNs are sensitive to image perturbations (Kubilius et al., 2019) and accordingly, FCNNs classification performance dropped as the noise level increased and eventually fell to chance levels. Crucially, in these conditions, the hidden representation of the FCNN contained informative representations of the target class despite the classification accuracy was at chance.

Neurocognitive theories of consciousness propose that unconscious processing reflects feed-forward processing only, while local recurrent connections in sensory cortex are critical for bringing unconscious content into conscious awareness (V. A. Lamme & Roelfsema, 2000). Visual signals that are embedded in noise and visually masked, are more likely to trigger feedforward processing only (Fahrenfort, Scholte, & Lamme, 2007). The neural network models used in our computer simulations of the visual task were all feedforward (DiCarlo, Zoccolan, & Rust, 2012; Yamins & DiCarlo, 2016), which may lack the capacity to preserve visual features across higher-order layers, so that any useful information might be left to local processes operating within each layer (Nayebi et al., 2018). Therefore, in the presence of image perturbations (i.e. added gaussian noise) the last readout layer of the FCNN may not fully exploit the information from previous layers to guide the perceptual decision. Likewise, in the human brain, unconscious feedforward processing may be able to produce information-rich representations in higher-order regions of the ventral visual pathway and even prefrontal cortex, but without feed-back connections those representations are unable to guide behaviour and lead to conscious sensation. Recurrent feedback is thought to be critical for conscious experience (Bullier, 2001; V. A. Lamme & Roelfsema, 2000; Pascual-Leone & Walsh, 2001), though importantly, recent evidence indicates that long-range feedback connections from prefrontal cortex, rather than local feedback loops in visual cortex are more critical for visual consciousness (L. Huang et al., 2020). The role of recurrent processing in unconscious information processing, however, remains unclear (M. A. Cohen & Dennett, 2011; Soto & Silvanto, 2014), and there is suggestive evidence of a link between recurrency and unconscious processing too (Melloni et al., 2007; Koivisto, Mäntylä, & Silvanto, 2010; M. X. Cohen, Van Gaal, Ridderinkhof, & Lamme, 2009). Recent modeling work indicates that recurrent neural networks (RNNs) that incorporate feedback connections provide better representations than FCNN models in object recognition tasks at different levels of image noise (Zwickel, Wachtler, & Eckhorn, 2007; Spoerer, McClure, & Kriegeskorte, 2017), which are better at explaining brain activity compared to FCNN models (Shi, Wen, Zhang, Han, & Liu, 2018; Nayebi et al., 2018; Kietzmann, Spoerer, et al., 2019; Spoerer et al., 2017). It will be relevant for future modeling work to investigate whether the addition of recurrent connections to the FCNN model can improve the read-out of the hidden representations by the decision layer and hence improve classification performance of noisy images, or whether recurrent connections improve the informativeness of the hidden layer representation despite the FCNN classification performance remains at chance level with noisy images. We conclude that unconscious information processing in visual domain, including processing without sensitivity, can lead to meaningful but hidden representational states that are ubiquitous in brains and biologically plausible models based on deep artificial neural networks. The work thus provides a framework for testing novel hypotheses regarding the scope of unconscious processes across different task domains.

Methods

Participants

Following informed consent, seven participants (mean = 29 years; SD = 2; 6 males) took part in return of monetary compensation. All of them had normal or corrected-to-normal vision and no history of psychiatric or neurological conditions. The study conformed to the Declaration of Helsinki and was approved by the BCBL Research Ethics Board.

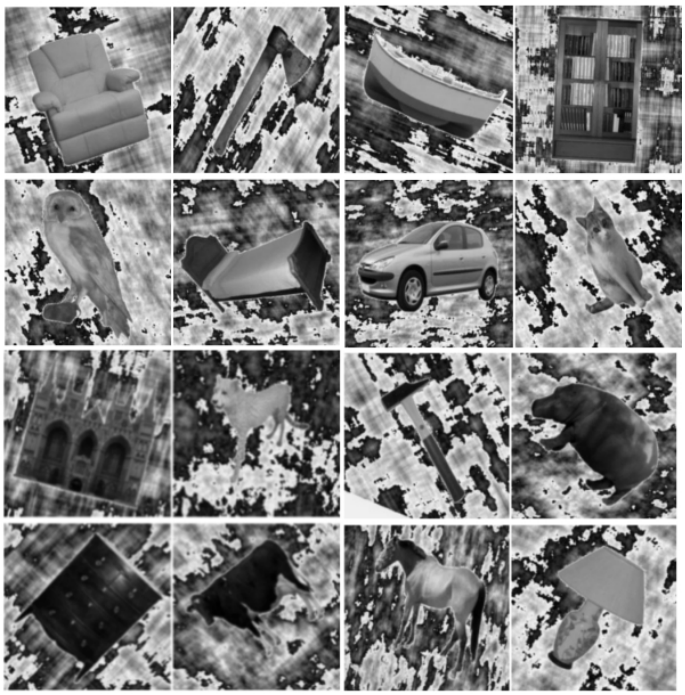
Experimental procedure and stimuli

Subjects (N = 7) were presented with images of animate and inanimate objects (Moreno-Martínez & Montoro, 2012). We selected 96 unique items (48 animate and 48 inanimate, i.e. cat, boat) for the experiment. These images could also be grouped by 10 subcategories (i.e. animal, vehicle) and 2 categories (i.e. living v.s. nonliving). The experiment was an event-related design. On each day, subjects carried out nine blocks of 32 trials each. Each block was composed of 16 animate and 16 inanimate items. Subjects performed six fMRI sessions of around 1 hour in separate days. There were hence 288 trials per day and 1728 trials in total per observer. The probe images were gray-scaled and presented in different orientations. The images were previously augmented using Tensorflow-Keras (Chollet et al., 2018). A random-phase noise background generated from the images was added to the target image before the experiment to facilitate masking.

The experiment was programmed using Psychopy v1.83.04 (Peirce, 2007). The experiment was carried out on a monitor with a refresh rate of 100 Hz. A fixation point appeared for 500 ms, followed by a blank screen for 500 ms. Twenty frames of gaussian noise masks were then presented and followed by

the probe image, which was followed by another twenty frames of gaussian noise masks. Then, there was a jittered blank period (1500 - 3500 ms) with a pseudo-exponential distribution in 500 ms steps (sixteen 1500 ms, eight 2000 ms, four 2500 ms, two 3000 ms, and two 3500 ms), selected randomly and without replacement on each block of 32 trials. Following the jittered blank period, participants were required (i) to identify the category of the image (ii) to rate the state of visual awareness associated with the image. There was a 1500 ms deadline for each response. For the categorization decision task, two choices were presented on the screen - living (V) and nonliving (nV) - i.e. "V nV" or "nV V" with the left-right order of the choices randomly selected for each trial. Subjects pressed "1" (left) or "2" (right) to indicate the probe condition. For the awareness decision task, there were 3 choices: (i) "I did not see anything that allowed me to categorize the item, I was completely guessing"; henceforth, the unconscious trials, (ii) partially unconscious ("I saw a brief glimpse but I am not confident of the response"), and (iii) conscious ("I saw the object clearly or almost clearly and I am confident of the categorization decision"). The inter-trial interval then followed with a jittered blank period of 6000 - 8000 ms with a pseudo-exponential distribution in 500 ms steps. The asynchrony between probe images across successive trials therefore ranged between 11.5 and 15.5 seconds.

The duration of the probe image was based on an adaptive staircase that was running throughout the trials. Specifically, based on pilot tests, we elected to use an staircase to get a high proportion of unconscious trials while ensuring that perceptual sensitivity was not different from chance level. If the observer reported "glimpse", the number of 10 ms frames of stimulus presentation was reduced by one frame for the next trial, unless it was already only one frame of presentation; if the observer reported "conscious", the number of frames of presentation would be reduced by two or three frames for the next trial, unless it was less than two to three frames, in which case it would be reduced by one frame; if the observer reported "unconscious", the number of frames increased by one or two frames, randomly, for the next trial. Examples of probe images were shown in Figure S1.



Supplementary Figure 1: Examples of the images used in the fMRI experiment.

Analysis of behavioral performance

We assessed whether the level of discrimination accuracy of the image departed from chance level in each of the awareness conditions. The metric to measure accuracy was A', based on the area under the receiver operating curve (ROC-AUC) (Zhang & Mueller, 2005). A response was defined as a "true positive" (TP) when "living" was both responded and presented. A response was defined as a "false positive" (FP) when "living" was responded while "nonliving" was presented. A response was defined as a "false negative" (FN) when "nonliving" was responded while "living" was presented. A response

was defined as a "true negative" (TN) when "nonliving" was both responded and presented. Thus, a hit rate (H) was the ratio between TP and the sum of TP and FN, and a false alarm rate (F) was the ratio between FP and the sum of FP and TN. A' was computed with different regularization based on 3 different conditions: 1) F <= 0.5 and H >= 0.5, 2) H >= F and H <= 0.5, 3) anything that were not the first two conditions. We first calculated A' associated with the individual behavioral performance within each of the different states of awareness (henceforth called the experimental A'). Then, we applied permutation tests to estimate the empirical chance level. We bootstrapped trials for a given awareness state with replacement (Horowitz, 2001); the order of the responses was shuffled while the order of the correct answers remained the same to estimate the empirical chance level. We calculated the A' based on the shuffled responses and correct answers to estimate the chance level of the behavioral performance, and we called this the chance level A' .

This procedure was repeated 10,000 times to estimate the distribution of the empirical chance level A' for each awareness state and each observer. The probability of empirical chance level A' being greater or equal to the experimental A' was the statistical significance level (one-tailed p-value, (Ojala & Garriga, 2010)). Hence, we determined whether the A' of each individual was above chance across the different awareness states (Bonferroni corrected for multiple tests).

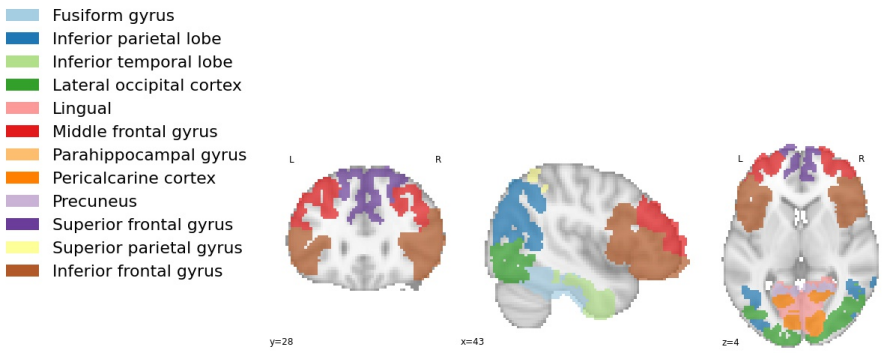
fMRI acquisition and preprocessing

A 3-Tesla SIEMENS's Magnetom Prisma-fit scanner and a 64-channel head coil was used. In each fMRI session, a multiband gradient-echo echo-planar imaging sequence with an acceleration factor of 6, resolution of 2.4 x 2.4 x2.4 mm^3, TR of 850 ms, TE of 35 ms, and bandwidth of 2582 Hz/Px was used to obtain 585 3D volumes of the whole brain (66 slices; FOV = 210mm). For each observer, one high-resolution T1-weighted structural image was also collected. The visual stimuli were projected on an MRI-compatible out-of-bore screen using a projector placed in the room adjacent to the MRI-room. A small mirror, mounted on the head coil, reflected the screen for presentation to the subjects. The head coil was also equipped with a microphone that enabled the subjects to communicate with the experimenters in between the scanning blocks.

The first 10 volumes of each block were discarded to ensure steady state magnetization; to remove non-brain tissue, brain extraction tool (BET, (Smith, 2002)) was used; volume realignment was performed using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002); minimal spatial smoothing was performed using a gaussian kernel with FWHM of 3 mm. Next, Independent component analysis based automatic removal of motion artifacts (ICA-AROMA) was used to remove motion-induced signal variations (Pruim et al., 2015) and this was followed by a high-pass filter with a cutoff of 60 sec. The scans were aligned to a reference volume of the first session. All the processing of the fMRI scans were performed within the FSL (FMRIB Software Library; v6.0, (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012)) framework and were executed using NiPype Python library (Gorgolewski et al., 2011). Details of the NiPype preprocessing pipeline can be found in `https://tinyurl.com/up2txma`.

For each observer, the relevant time points or scans of the preprocessed fMRI data of each run were labeled with attributes such as (i.e. cat, boat), category (i.e. animal, vehicle), and condition (i.e. living vs. nonliving) using the behavioral data files generated by Psychopy (v1.84, (Peirce, 2007)). Next, data from all sessions were stacked and each voxel's time series was block-wise z-scored (normalized) and linear detrended. Finally, to account for the hemodynamic lag, examples were created for each trial by averaging the 3 or 4 volumes between the interval of 4 s and 7 sec after image onset.

For a given awareness state, examples of BOLD activity patterns were collected for each of the 12 regions of interest (ROIs). There were 12 ROIs for each hemisphere (see Figure S2). The ROIs included the lingual gyrus, pericalcarine cortex, lateral occipital cortex, fusiform gyrus, parahippocampal gyrus, inferior temporal lobe, inferior parietal lobe, precuneus, superior parietal gyrus, superior frontal gyrus, middle frontal gyrus, and inferior frontal gyrus (comprising pars opercularis gyrus, pars triangularis gyrus, and pars orbitalis gyrus). Automatic segmentation of the high-resolution structural scan was done with FreeSurfer's automated algorithm `recon-all` (v6.0.0). The resulting masks were transformed to functional space using 7 degrees of freedom linear registrations implemented in FSL FLIRT (Jenkinson et al., 2002) and binarized. All further analyses were performed in native BOLD space within each observer.



Supplementary Figure 2: The figure shows the selected regions of interest. Twelve bilateral ROIs were extracted, comprising the lingual gyrus, pericalcarine cortex, lateral occipital, fusiform, parahippocampal cortex, precuneus, inferior temporal lobe, inferior and superior parietal lobe, superior frontal, middle frontal gyrus, and inferior frontal gyrus.

Multivariate pattern analysis: decoding within each awareness state

Multivariate pattern analysis (MVPA) was conducted using Scikit-learn (Pedregosa et al., 2011) and Nilearn (Abraham et al., 2014). A linear support vector machine (SVM) classifier. SVM has limited complexity, hence reducing the probability of over-fitting (model performs well in training data but bad in testing data) and it has been shown to perform well with fMRI data (Pereira & Botvinick, 2011; Lewis-Peacock & Norman, 2014). We used an SVM with L1 regularization, nested with invariant voxels removal and feature scaling between 0 and 1 as preprocessing steps. The nested preprocessing steps were fit in the training set and applied to the testing set. Note that these preprocessing steps are different from the detrending and z-scoring of the BOLD signals and represent conventional machine learning practices (Bruha, 2000).

During cross-validation, trials corresponding to one living (i.e. cat) and one non-living (i.e. boat) item for a given awareness state (i.e. unconscious) were left-out as the test set and the rest was used to fit the machine learning pipeline. With 96 unique items, 2256 cross-validation folds could be performed in principle. However, because the awareness states were randomly sampled for each unique item (i.e. cat), the proportion of examples for training and testing were not equal among different folds. Some subjects had less than 96 unique items for one or more than one of the awareness states. Thus, less than 2256 folds of cross-validations were performed in these cases.

To get an empirical chance level of the decoding, the same cross-validation procedures were repeated by replacing the linear SVM classifier with a "dummy classifier" as implemented in Scikit-learn, which makes predictions based on the distribution of the classes of the training set randomly without learning the relevant multivariate patterns. The same preprocessing steps were kept in the pipeline.

The mean difference between the true decoding scores and the chance level decoding scores was computed as the experimental score. To estimate the null distribution of the performance differences, we performed permutation tests. First, we concatenate the true decoding scores and the chance level decoding scores and then shuffle the concatenated vector. Second, we split the concatenated vector into a new 'decoding scores' vector and a new 'chance level decoding scores' vector. The mean differences between these two vectors were computed. This procedure was repeated 10,000 times to estimate the null distribution of the performance differences. The probability that the experimental score was greater or equal to the null distribution was the statistical significant level (one-tailed p-value, corrected for the number of ROIs using Bonferroni).

Multivariate pattern analysis: generalization across awareness states

Here the classifier was trained from data in a particular awareness state (the 'source'; e.g. on conscious trials) and then tested on a different awareness state (the 'target'; e.g. on unconscious trials) on top of the cross-validation procedure described above. Similar to the decoding analysis within each awareness state, instances corresponding to one living and one non-living item in both 'source' and 'target' were left out, but only the left-out instances in the 'target' were used as the test set. The rest of the instances in 'source' were used as the training set to fit the machine learning pipeline (preprocessing + SVM) as described above. The performance of the fitted pipeline was estimated by comparing the predicted labels

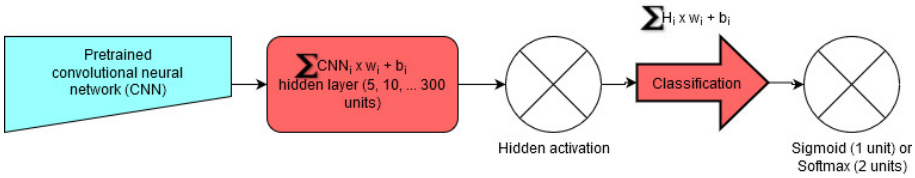
and the true labels using ROC-AUC for the test set.

To get an empirical chance level of the decoding, a similar procedure to that described above with a 'dummy classifier' was used here. Similar permutation test procedures were used to estimate the empirical null distribution of the difference between the experimental and chance level ROC-AUC and the estimation was repeated 10,000 times. The probability that the experimental score was greater or equal to the null distribution was the statistical significant level (one-tailed p-value).

Computational model simulation

Different FCNN models (i.e. AlexNet, VGGNet, ResNet, MobileNet, and DenseNet) implemented in Pytorch V1.0 (?, ?), learned to perform the same visual discrimination task as the human observers. The FCNNs were trained with clear images and then tested under different levels of noise in the image. The goal here was to emulate the pattern observed in the fMRI study (i.e. decoding of the noisy image in the absence of perceptual sensitivity) using a FCNN. To control for the initialization state of the FCNN models, we fine-tuned some of the popular FCNN pre-trained models, such as AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGGNet (Simonyan & Zisserman, 2015), ResNet (He, Zhang, Ren, & Sun, 2016), MobileNet (Howard et al., 2017), and DenseNet (G. Huang, Liu, Van Der Maaten, & Weinberger, 2017), which were pre-trained using the ImageNet dataset (Deng et al., 2009) and then were adapted to our experiment using a transfer learning (fine-tuning) procedure (Yosinski, Clune, Bengio, & Lipson, 2014). After fine-tuning the FCNN models on the clear images used in the experiment, the models were tested on images with different noise levels.

As shown in Figure S3, pretrained FCNN models using ImageNet (Deng et al., 2009) were stripped of the original fully-connected layer while weights and biases of the convolutional layers were frozen and not updated further (Yosinski et al., 2014). An adaptive pooling (McFee et al., 2018) operation was applied to the last convolutional layer so that the output of this layer became a one-dimensional vector, and a new fully-connected layer took the weighted sum of these outputs (i.e. the 'hidden layer'). The number of artificial units used in the hidden layer could be any positive integer, but for simplicity, we took 300 as an example and we explored how the number of units (i.e. 2, 5, 10, 20, 50, 100, and 300) influenced the pattern of results. The number of hidden layer units determined the number of new weights, w_i, for training. The outputs of the hidden layer were passed to an activation function (Specht, 1990), which could be linear (i.e. identical function) or nonlinear (i.e. rectified function). A dropout was applied to the hidden layer during training but not during testing. Different dropout rates were explored (i.e. 0, 0.25, 0.5, and 0.75), where zero dropout rate meant no dropout was applied. The dropout operation was varied to investigate how feature representations could be affected by a simple regularization.



Supplementary Figure 3: A simplified scheme of fine-tuning a pre-trained feedforward convolutional neural network model. The task is to classify the living vs. non-living category of the images (without noise) used in the fMRI experiment. The blue architecture was frozen and the weights were not updated, while weights of the red architectures were updated during training.

A new fully-connected layer, namely, the classification layer, took the outputs processed by the activation function of the hidden layer to compose the classification layer. The number of artificial units used in the classification layer depended on the activation function applied to the outputs of the layer. If the activation function was sigmoid (formula 1), one unit was used, while if the activation was a softmax function (formula 2), two units were used. Under subscripts 'i' denotes the ith output of a given artificial unit.

$$\psi(x_i) = \frac{1}{1 + e^{-x_i}} \tag{1}$$

$$\psi(x_i) = \frac{e^{x_i}}{\sum e^{x_i}} \tag{2}$$

The re-organized FCNN was trained on the gray-scaled and augmented (flipped or rotated) experimental images and validated on images that were also gray-scaled but different degrees of augmentation.

The loss function was binary cross-entropy. The optimizer was Adam (Kingma & Ba, 2014) with a learning rate of 1e-4 without decay. The validation performance was used to determine when to stop training, and the validation performance was estimated every 10 training epochs. The FCNN model was trained until the loss did not decrease for five validation estimations.

As noted, after training, the weights and biases of the FCNN model were frozen to prevent the model changing during the test phase. During the test phase, gaussian noise was added to the images to reduce the FCNN classification performance. Similar augmentations as in the validation set were fed to the testing image sets. The noise added to the images was sampled from a gaussian distribution centered at zero and different variance (σ). The level of noise was defined by setting up the variance at the beginning of each test phase.

The noise levels ranged from 0 to 1000 in steps of 50 following a logarithmic trend. For a given noise level, 20 sessions of 96 images with a batch size of 8 were fed to the FCNN model, and both the outputs of the hidden layer and the classification layer were recorded. The outputs of the classification layer were used to determine the "perceptual sensitivity" of the FCNN model, and the outputs of the hidden layer were used to perform subsequent decoding analyses with a linear SVM classifier, in keeping with the fMRI analysis.

To determine the significance level of the FCNN model performance, the order of the true labels in each session was shuffled while the order of the predicted labels remained the same. The permuted performance was calculated for the 20 sessions. This procedure was repeated 10,000 times to estimate the empirical chance level of the FCNN model. The significance level was the probability that the performance of the FCNN model was greater or equal to the chance level performances (one-tailed test against 0.05). If the p-value is greater or equal to 0.05, we considered that FCNN performance was not different from the empirical chance level.

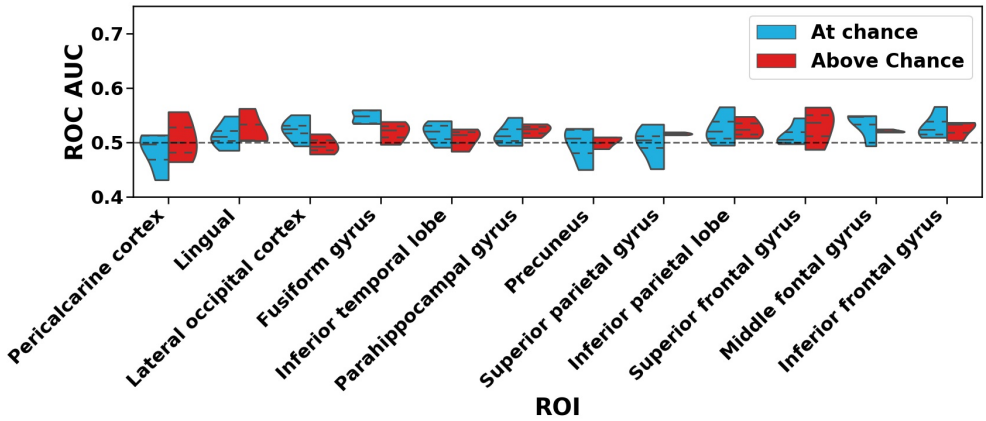
We then assessed, for a given noise in the image, whether the hidden layer of the FCNN (i.e. following the last convolutional layers), contained information that allowed decoding of the category of the image (living vs non-living). A linear SVM used the information contained in the FCNN hidden layer to decode the image class across different levels of noise, even when the FCNN model classification performance was at chance. The outputs and the labels of the hidden layer from the 20 sessions were concatenated. A random shuffle stratified cross-validation procedure was used in the decoding experiments with 50 folds to estimate the decoding performance of the SVM. The statistical significance of the decoding performance was estimated by a different permutation procedure to the FCNN, which here involved fitting the SVM model and testing the fitted SVM with 50-fold cross-validation in each iteration of permutation, and it was computational costly (`https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.permutation_test_score.html`). On each permutation iteration, the order of the labels was shuffled while the order of the outputs of the hidden layer remained unchanged before fitting the SVM model (Ojala & Garriga, 2010). The permutation iteration was repeated 100 times to estimate the empirical chance level. The significance level was the probability of the true decoding score greater or equal to the chance level. Because we were interested in those poorly performing FCNNs, we only attempted to decode the stimulus category from the hidden layer in those cases in which the FCNN classification performance was lower than 0.55 ROC-AUC .

Supplementary Information

Across the different ROIs, we pooled the decoding accuracy of the four participants that displayed null perceptual sensitivity on trials rated as unaware and likewise for those participants that displayed above chance sensitivity. The Figure S4 illustrates these data. There are no apparent and consistent differences across the different ROIs.

Author contributions

N.M. and D.S. designed the study; N.M. analysed the data under the guidance of R.S. and D.S. N.M. prepared a first draft of the paper. All authors discussed the results and contributed towards the writing of the final version of the manuscript; D.S. supervised the project.



Supplementary Figure 4: Distribution of decoding accuracies across the participants whose perceptual sensitsupportiivity was at chance and those who deviated from chance.

Data availability statement

Analysis scripts are available at https://github.com/nmningmei/unconfeats. The neuroimaging data will be made fully open access at Openneuro upom the publication of the paper.

Acknowledgements

D.S. acknowledges support from the Basque Government through the BERC 2018-2021 program, 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 also from project grants PSI2016-76443-P and PID2019-105494GB-I00 from MINECO. R.S. acknowledges support by the Basque Government (IT1244-19 and ELKARTEK programs), and the Spanish Ministry of Economy and Competitiveness MINECO (project TIN2016-78365-R).

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., ... Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*, 14.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

Beck, D. M., Rees, G., Frith, C. D., & Lavie, N. (2001). Neural correlates of change detection and change blindness. *Nature Neuroscience*, *4*(6), 645–650.

Bruha, I. (2000). From machine learning to knowledge discovery: Survey of preprocessing and postprocessing. *Intelligent Data Analysis*, *4*(3-4), 363–374.

Bullier, J. (2001). Feedback connections and conscious vision. *Trends in Cognitive Sciences*, *5*(9), 369—370.

Chollet, F., et al. (2018). Keras: The python deep learning library. *ASCL*, ascl–1806.

Chong, T. T.-J., Husain, M., & Rosenthal, C. R. (2014). Recognizing the unconscious. *Current Biology*, *24*(21), 1033–1035.

Christophel, T. B., Hebart, M. N., & Haynes, J.-D. (2012, September). Decoding the contents of visual short-term memory from human visual and parietal cortex. *Journal of Neuroscience*, *32*(38), 12983–12989. Retrieved from https://doi.org/10.1523/jneurosci.0184-12.2012 doi: 10.1523/jneurosci.0184-12.2012

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358–364.

Cohen, M. X., Van Gaal, S., Ridderinkhof, K. R., & Lamme, V. (2009). Unconscious errors enhance prefrontal-occipital oscillatory synchrony. *Frontiers in Human Neuroscience*, *3*, 54.

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin.

Dehaene, S., & Changeux, J.-P. (2004). Neural mechanisms for access to consciousness. *The Cognitive Neurosciences*, *3*, 1145–58.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1–37.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J.-F., Poline, J.-B., & Rivière, D. (2001, July). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*(7), 752–758. Retrieved from https://doi.org/10.1038/89551 doi: 10.1038/89551

Dehaene, S., Naccache, L., Cohen, L., Le Bihan, D., Mangin, J.-F., Poline, J.-B., & Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. *Nature Neuroscience*, *4*(7), 752–758.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 248–255).

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Ester, E. F., Sprague, T. C., & Serences, J. T. (2015, August). Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron*, *87*(4), 893–905. Retrieved from https://doi.org/10.1016/j.neuron.2015.07.013 doi: 10.1016/j.neuron.2015.07.013

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience*, *19*(9), 1488–1497.

Fang, F., & He, S. (2005). Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature Neuroscience*, *8*(10), 1380–1385.

Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv:1801.01489*.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Gayet, S., Guggenmos, M., Christophel, T. B., Haynes, J.-D., Paffen, C. L., Sterzer, P., & Van der Stigchel, S. (2020). No evidence for mnemonic modulation of interocularly suppressed visual input. *NeuroImage*, 116801.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in neural information processing systems* (pp. 7538–7550).

Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2014). Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in Computational Neuroscience*, *8*, 74.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hassin, R. R. (2013, February). Yes it can. *Perspectives on Psychological Science*, *8*(2), 195–207. Retrieved from https://doi.org/10.1177/1745691612460684 doi: 10.1177/1745691612460684

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Haynes, J.-D., Driver, J., & Rees, G. (2005). Visibility reflects dynamic changes of effective connectivity between v1 and fusiform cortex. *Neuron*, *46*(5), 811–821.

Haynes, J.-D., & Rees, G. (2005, April). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*(5), 686–691. Retrieved from https://doi.org/10.1038/nn1445 doi: 10.1038/nn1445

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hesselmann, G., Hebart, M., & Malach, R. (2011). Differential bold activity associated with subjective and objective reports during "blindsight" in normal observers. *Journal of Neuroscience*, *31*(36), 12936–12944.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Horowitz, J. L. (2001). The bootstrap. In *Handbook of econometrics* (Vol. 5, pp. 3159–3228). Elsevier.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

Huang, L., Wang, L., Shen, W., Li, M., Wang, S., Wang, X., ... Zhang, X. (2020, November). A source for awareness-dependent figure–ground segregation in human prefrontal cortex. *Proceedings of the National Academy of Sciences*, 201922832. Retrieved from https://doi.org/10.1073/pnas.1922832117 doi: 10.1073/pnas.1922832117

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. M. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–841.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782–790.

Jiang, Y., Zhou, K., & He, S. (2007, April). Human visual cortex responds to invisible chromatic flicker. *Nature Neuroscience*, *10*(5), 657–662. Retrieved from https://doi.org/10.1038/nn1879 doi: 10.1038/nn1879

Kapoor, V., Dwarakanath, A., Safavi, S., Werner, J., Besserve, M., Panagiotaropoulos, T. I., & Logothetis, N. K. (2020). Decoding the contents of consciousness from prefrontal ensembles. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2020/01/28/2020.01.28.921841 doi: 10.1101/2020.01.28.921841

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*(11).

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019, 1). Deep neural networks in computational neuroscience. *Oxford Research Encyclopaedia of Neuroscience*. Retrieved from https://oxfordre.com/neuroscience/view/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-46" doi: 10.1093/acrefore/9780190264086.013.46

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koivisto, M., Mäntylä, T., & Silvanto, J. (2010, June). The role of early visual cortex (v1/v2) in conscious and unconscious visual perception. *NeuroImage*, *51*(2), 828–834. Retrieved from https://doi.org/10.1016/j.neuroimage.2010.02.042 doi: 10.1016/j.neuroimage.2010.02.042

Kouider, S., & Dehaene, S. (2007, April). Levels of processing during non-conscious perception: a critical review of visual masking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 857–875. Retrieved from https://doi.org/10.1098/rstb.2007.2093 doi: 10.1098/rstb.2007.2093

Kranczioch, C., Debener, S., Schwarzbach, J., Goebel, R., & Engel, A. K. (2005). Neural correlates of conscious perception in the attentional blink. *Neuroimage*, *24*(3), 704–714.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, *17*(1), 26–49.

Kriegeskorte, N. (2011). Pattern-information analysis: from stimulus decoding to computational-model testing. *Neuroimage*, *56*(2), 411–421.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... others (2019). Brain-like object recognition with high-performing shallow recurrent anns. In *Advances in neural information processing systems* (pp. 12805–12816).

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Lamme, V. A. F. (2020, February). Visual functions generating conscious seeing. *Frontiers in Psychology*, *11*. Retrieved from https://doi.org/10.3389/fpsyg.2020.00083 doi: 10.3389/fpsyg.2020.00083

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, *3361*(10), 1995.

Lewis-Peacock, J. A., & Norman, K. A. (2014). Multi-voxel pattern analysis of fmri data. *The Cognitive Neuroscience*, *512*, 911–920.

Ludwig, K., & Hesselmann, G. (2015). Weighing the evidence for a dorsal processing bias under continuous flash suppression. *Consciousness and Cognition*, *35*, 251–259.

Ludwig, K., Kathmann, N., Sterzer, P., & Hesselmann, G. (2015). Investigating category-and shape-selective neural processing in ventral and dorsal visual stream under interocular suppression. *Human Brain Mapping*, *36*(1), 137–149.

Macmillan, N. A. (1986, March). The psychophysics of subliminal perception. *Behavioral and Brain Sciences*, *9*(1), 38–39. Retrieved from https://doi.org/10.1017/s0140525x00021427 doi: 10.1017/s0140525x00021427

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, *2*(4.2).

McFee, B., Salamon, J., & Bello, J. P. (2018). Adaptive pooling operators for weakly labeled sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(11), 2180–2193.

Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W., & Rodriguez, E. (2007, March). Synchronization of neural activity across cortical areas correlates with conscious perception. *Journal of Neuroscience*, *27*(11), 2858–2865. Retrieved from https://doi.org/10.1523/jneurosci.4623-06.2007 doi: 10.1523/jneurosci.4623-06.2007

Moreno-Martínez, F. J., & Montoro, P. R. (2012). An ecological alternative to snodgrass & vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PloS One*, *7*(5), e37527.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage*, *56*(2), 400–410.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., ... Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. In *Advances in neural information processing systems* (pp. 5290–5301).

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*(1), 1–19.

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, *11*(Jun), 1833–1863.

Overgaard, M., Timmermans, B., Sandberg, K., & Cleeremans, A. (2010). Optimizing subjective measures of consciousness. *Consciousness and Cognition*, *19*(2), 682–684.

Pascual-Leone, A., & Walsh, V. (2001). Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science*, *292*(5516), 510–512.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.

Pereira, F., & Botvinick, M. (2011). Information mapping with pattern classifiers: a comparative study. *Neuroimage*, *56*(2), 476–496.

Pessoa, L., & Ungerleider, L. G. (2004). Neural correlates of change detection and change blindness in a working memory task. *Cerebral Cortex*, *14*(5), 511–520.

Peters, M. A. K., & Lau, H. (2015, October). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *eLife*, *4*. Retrieved from https://doi.org/10.7554/elife.09651 doi: 10.7554/elife.09651

Pruim, R. H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J. K., & Beckmann, C. F. (2015). Ica-aroma: A robust ica-based strategy for removing motion artifacts from fmri data. *Neuroimage*, *112*, 267–277.

Rabagliati, H., Robertson, A., & Carmel, D. (2018). The importance of awareness for understanding language. *Journal of Experimental Psychology: General*, *147*(2), 190.

Rosenthal, C. R., Andrews, S. K., Antoniades, C. A., Kennard, C., & Soto, D. (2016). Learning and recognition of a non-conscious sequence of events in human primary visual cortex. *Current Biology*, *26*(6), 834–841.

Schurger, A., Pereira, F., Treisman, A., & Cohen, J. D. (2010). Reproducibility distinguishes conscious from nonconscious neural representations. *Science*, *327*(5961), 97–99.

Shi, J., Wen, H., Zhang, Y., Han, K., & Liu, Z. (2018). Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision. *Human Brain Mapping*, *39*(5), 2269–2282.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, S. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*(3), 143–155.

Soto, D., Mäntylä, T., & Silvanto, J. (2011). Working memory without consciousness. *Current Biology*, *21*(22), 912–913.

Soto, D., Sheikh, U. A., & Rosenthal, C. R. (2019). A novel framework for unconscious processing. *Trends in Cognitive Sciences*, *23*(5), 372–376.

Soto, D., & Silvanto, J. (2014). Reappraising the relationship between working memory and conscious awareness. *Trends in Cognitive Sciences*, *18*(10), 520–525.

Specht, D. F. (1990). Probabilistic neural networks and the polynomial adaline as complementary techniques for classification. *IEEE Transactions on Neural Networks*, *1*(1), 111–121.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551.

Stein, T., Utz, V., & van Opstal, F. (2020, February). Unconscious semantic priming from pictures under backward masking and continuous flash suppression. *Consciousness and Cognition*, *78*, 102864. Retrieved from https://doi.org/10.1016/j.concog.2019.102864 doi: 10.1016/j.concog.2019.102864

Sterzer, P., Haynes, J. D., & Rees, G. (2008, November). Fine-scale activity patterns in high-level visual areas encode the category of invisible objects. *Journal of Vision*, *8*(15), 10–10. Retrieved from https://doi.org/10.1167/8.15.10 doi: 10.1167/8.15.10

Trübutschek, D., Marti, S., Ojeda, A., King, J.-R., Mi, Y., Tsodyks, M., & Dehaene, S. (2017). A theory of working memory without consciousness or sustained activity. *Elife*, *6*, e23871.

Van Gaal, S., & Lamme, V. A. (2012). Unconscious high-level information processing: implication for neurobiological theories of consciousness. *The Neuroscientist*, *18*(3), 287–301.

Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging*, *2017*(14), 36–45.

Wuethrich, S., Hannula, D. E., Mast, F. W., & Henke, K. (2018). Subliminal encoding and flexible retrieval of objects in scenes. *Hippocampus*, *28*(9), 633–643.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356.

Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. In *Advances in neural information processing systems* (pp. 3093–3101).

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

Zhang, J., & Mueller, S. T. (2005). A note on roc analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*(1), 203–212.

Zwickel, T., Wachtler, T., & Eckhorn, R. (2007). Coding the presence of visual objects in a recurrent neural network of visual cortex. *Biosystems*, *89*(1-3), 216–226.