Superstitious perception in humans and convolutional neural networks

by

Patrick Laflamme

BSc. Psychology, University of Waterloo, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Arts

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Psychology)

The University of British Columbia
(Vancouver)

August 2017

Abstract

The advent of complex Hierarchical Convolutional Neural Networks (HCNNs) has led to great progress in the field of computer vision, with modern implementations of HCNNs rivalling human performance in object recognition tasks. The design of HCNNs was inspired by current understanding of how the neurons of the human visual system are organized to support object recognition. There are researchers who claim that the computations undertaken by HCNNs are approximating those of the human visual system, because of their high accuracy in predicting the neural activity of regions of the brain involved in object classification (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). However, there has been little investigation of how HCNNs and humans compare on other tasks that HCNNs have not been trained on. Our study compared the similarity of one HCNN, AlexNet, and humans on a superstitious perception task that involves falsely recognizing a learned object in the absence of strong evidence for its presence. We began by validating a new technique that quantifies human performance on the superstitious perception task. The first phase of the research revealed that human behaviour in the task is dependent on whether participants employed an active or passive task strategy. Next, the responses of our HCNN to the same images were analyzed in a similar manner. The results showed that HCNNs behaved similarly to humans in some ways and differently in others. Specifically, the classification images generated for the HCNN were similar to those derived from human participants, but the HCNN was also more consistent in its responses than humans. A second finding was that the differences in human participants classification images (created by adopting active versus passive strategies) could not be accounted for by simply altering the proportion of false alarm

responses in the HCNN. This suggests that HCNNs may be using criteria similar to humans’ perception when evaluating the likelihood of an object being present. The higher similarity between humans and HCNN in the passive condition suggests that the criteria similarities are largest when humans recruit minimal central executive resources in the decision making process.

Lay Summary

Some claim that new computational models of vision mimic the human visual system. However, to date, the comparisons have been quite superficial. A time-honored way to study the brain is to identify unexpected ways in which it behaves. For example, visual illusions have often been used to understand the approaches the brain employs in order to identify objects in a visual scene. Superstitious perception, the phenomenon of seeing objects in meaningless noise, is the illusion used here to compare the computational models to the visual system. After verifying that superstitious perception works as expected in humans, we tested the performance of the models on the same task. The models showed similar behaviour to that of humans, although the similarity was reduced when humans actively used executive functions to perform the task.

Preface

All work presented in this thesis is original intellectual property of the author, P. Laflamme, under the supervision of Dr. James Enns. All data reported in chapters 2, 3 and 4 were collected under the UBC Ethics certificate H16-00071 entitled ”Seeing through the noise”.

Table of Contents

List of Tables

List of Figures

Glossary

FMRI	functional magnetic resonance imaging
MEG	magnetoencephalography
CI	Classification Image
SGD	stochastic gradient descent
RSA	representational similarity analysis
HCNN	Hierarchical Convolutional Neural Network
IWC	the Image-Wise Correlation

Acknowledgments

A big thank you to Dr. Jim Enns, whose patience and careful guidance made this work possible. Through the two years that spanned this work, his direction kept my efforts on course. He led by example, showing me how to develop the skills that make a great scientist.

I would also like to thank Ana Pesquita, whose mentorship helped me gain confidence in myself, and Stefan Bourrier, who supported me in too many ways to count.

Last but not least, I would like to thank my incredible parents, without whom I would not have become who I am today. Their support and guidance have been crucial to my professional and personal growth.

Dedication

I dedicate this work to Jenny Wan, who tolerated my busy lifestyle during this work and helped to keep me sane through the process.

Chapter 1: Introduction

1.1 Superstitious perception

When studying a system as complex as the human brain, one method that can be used to understand its inner mechanisms is to investigate the instances in which it performs unexpectedly. One example of unexpected behaviour in the human visual system is the phenomenon of apophenia. Apophenia is defined as the experience of seeing patterns or connections in random or meaningless data, and was first coined by Klaus Conrad, a German neurologist in the early 20th century (Fyfe, Williams, Mason, & Pickup, 2008; Shermer, 2008). In the visual domain, this phenomenon describes instances in which one experiences a percept of an object that is not there. Common examples of visual apophenia include seeing animals in the clouds, or seeing Jesus in a piece of toast (Liu et al., 2014). Importantly, it is highly unlikely that the percept is accurate in these cases. That is to say that Jesus is likely not in the toast, and animals are unlikely to have grown in size and begun floating amongst the clouds. Nonetheless, the experience of apophenia is near ubiquitous. Indeed, it is possible that the visual system is detecting evidence that supports these percepts that were given rise by aphophenia, despite the ground truth being an absence of such perceived objects. The difficulty comes in trying to quantify the extent of such evidence, if any is present.

One human behavioural technique that has been developed to investigate this phenomenon, and allows for quantification of available evidence is the superstitious perception task (Gosselin & Schyns, 2003). This technique involves asking people to identify targets in noisy images (like static on an old TV). Participants were asked to identify a target (e.g., the letter $'s'$) within the noisy images that were presented to them one at a time in the center of a computer screen. Participants were asked to report whether they saw the target of interest within each noise image as it was presented. Unbeknownst to them, the images contained no target at all and consisted of pure random noise. Surprisingly, when Gosselin and Schyns (2003) added together all of the images identified as having a target in them, and then subtracted all of the images that had not been identified, a representation of the target became visible (Gosselin & Schyns, 2003). The image that resulted from such a process was named a classification image (CI). This CI was claimed to be an image of the internal representation of the target that participants were trying to detect within the image (Gosselin & Schyns, 2003).

The stated theoretical goal of the superstitious perception task (Gosselin & Schyns, 2003) is to measure the internal representation participants are using when they falsely detect target signals in noise. The rationale for this method is built upon the same logic that underpins the use of reverse correlation to estimate the receptive field properties of individual neurons (see Ringach & Shapley, 2004, for a review). In the initial introduction of reverse correlation to that field, researchers stimulated single neurons from the primary visual cortex of cats with random white noise visual input to the retina. Then, they recorded the timing of each of the neuron's spikes in response to the stimulation (Jones & Palmer, 1987). The strength of the white noise input in the moments leading up to a spike allowed researchers to model a given neuron's receptive field, by computing the average stimulus pattern within a specified time window before the neuron spiked. This spike-triggered average was interpreted as an ideal template for the neurons receptive field because it represented the stimulus that was most likely to cause the neuron to fire.

Critics of this interpretation cautioned that before accepting spike-triggered averages computed in this manner, it was important to describe how well the receptive field predicted the neurons activity (DeAngelis, Ohzawa, & Freeman, 1993). This was done by first estimating a neuron's receptive field using reverse correlation. Once the receptive field was estimated, one could predict the neurons firing rate in response to the stimulus used to estimate the receptive field. This is a test of internal validity, and it is illustrated in Figure 1C. Internal prediction was thus a measure of how well the estimated receptive field predicted the neuron's response to the individual stimuli that had contributed to the estimation. If the neural response of individual stimuli was not sufficiently predicted by the receptive field, one could conclude that the calculated receptive field of the neuron was not adequately descriptive of the neurons function.

DeAngelis et al. (1993) performed this internal prediction technique on the receptive fields of cat early visual neurons and was able to show a high degree of predictive accuracy, thus validating the procedure for further use. Reverse correlation, used to estimate the receptive fields of single neurons in this manner, became instrumental for interpreting neuron spike trains and reconstructing representations of the receptive fields of visual and auditory neurons alike (Jones & Palmer, 1987; Ringach, G., & Shapley, 1997; Theunissen et al., 2001, for example).

Ringach and Shapley (2004) suggested another important step to the validation procedure when they argued that one could also use CIs to make predictions of a neuron's response to novel stimuli if its receptive field was already known. This is illustrated in Figure 1D. This could be done by performing an external prediction of a neuron's firing rate. That is, one could evaluate a hypothesized receptive field in response to novel stimuli. The critical difference between the two approaches is whether the stimuli used to generate the receptive field are the same as the stimuli for which you are trying to predict responses (internal prediction) or whether the receptive field was generated with different stimuli (external prediction). When the same stimuli are used for creation of the CI and for its evaluation of predictive power, the measurement procedure is vulnerable to capitalization on chance (or measurement error). When novel stimuli are tested instead, a more stringent test of a receptive field's predictive power is given.

In the domain of superstitious perception, participants do not respond with spikes, but rather with false alarms or correct rejections in a signal detection task. In addition, it is not possible to assess a temporal component of these responses analogous to neural firing times. Instead, discrete responses are made that depend only on the current image. If the response is a false alarm, we gain information implying that the image is target-like in some respects. This is comparable to a neuron's spike. Moreover, if the response is a correct rejection, we gain information implying that the image is not target-like. Note that this is even more information than we get from a neuron, which typically only responds in a positive way by increasing its activity above some baseline level. Gosselin and Schyns (2003) put this extra information to good use in applying reverse correlation to psychophysical data, by creating an average-target image, and an average-non-target image, as shown in Figure 1B.

Furthermore, participants in a behavioural experiment do not necessarily behave in the same way as neuron when it comes to being presented with ambiguous stimuli (Rieth, Lee, Lui, Tian, & Huber, 2011), and instead they require a more elaborate ruse to respond to superstitious perceptions. Rieth et al. (2011) achieved this by having two practice phases: (1) participants were shown targets that were simple to identify in noise — they called this the easy practice session; and (2) participants were asked to identify targets that were very heavily overlaid with noise, making them hard to identify — they called this the hard practice session. The goal here was to encourage participants to see superstitiously. Notably, for our purposes, this also aids in stressing to participants that the targets are homogeneous. The target shown to them at the beginning of the experiment is identical to all targets hidden in the noise, a consideration that helped to increase the strength of our experimental design.

There is one final difference worth noting between the application of reverse correlations in neural and perceptual contexts. In the neural realm, a prediction is calculated based on an aggregate of spikes over time, in the units of firing rate (Hz). This is because while firing rate is predictable, neural firing *time* seems to be stochastic in nature (Dyan & Abbott, 2001). In the superstitious perception realm, we can instead interpret the aggregate prediction as a likelihood of false alarm for a given trial. For this reason, it is best to aggregate trials into bins of comparable target similarity when estimating the extent of explained variance in responses. Since there are no guidelines for bin size, a range of bin sizes will be used throughout the following experiments to ensure that bin size has no effect on our conclusions. The statistical logic behind the reverse correlation technique, however, remains the same as that in the neural spiking domain (Gosselin & Schyns, 2001).

Given the conclusions reached in the reverse correlation literature with respect to the need to test a generated CI to evaluate its predictive capacity (DeAngelis et al., 1993; Ringach & Shapley, 2004), there is a critical step missing in the procedure described by Gosselin and Schyns (2003). Once a CI is generated, one should attempt to predict participants' false alarm rates in response to new white-noise images, called external predictions, in order to estimate the degree to which the CI explains participants' perceptions. Without this crucial step, one does not know the degree to which the CI over-fit the data from which it was generated. With the importance of this step in mind, we intended to implement the external prediction step for the superstitious perception task. In doing so, we aimed to verify that any conclusions made as a result of the CIs generated from this task apply to stimuli beyond the specific noise images from which the CI was generated. Furthermore, we intended to compare the conclusions drawn from internal predictions with those drawn from the novel external prediction procedure.

1.2 Task strategy and its influence on task outcome

The influence of task strategy has been well documented with respect to its effects on task outcomes for a variety of cognitive tasks (Jacoby & Brooks, 1984; Marcel, 1983; Smilek, Enns, Eastwood, & Merikle, 2006; Van Selst & Merikle, 1993; Whittlesea & Brooks, 1994). In the realm of categorization, participants were observed to be more accurate during an item classification task when adopting a feature-focussed, analytical strategy when compared to a more wholistic, non-analytical strategy (Jacoby & Brooks, 1984; Whittlesea & Brooks, 1994). Even when percepts never reached conscious awareness, the effects of unconscious stimuli on task outcome were stronger when participants let the stimulus 'pop' into mind, as opposed to trying to determine whether a stimulus was present in a preceding trial (Van Selst & Merikle, 1993). In the field of visual search, participants' search efficiency can be modulated by the instructions they are given before the task. Participants were faster and more accurate in their search when told to let the target of their search ‘pop’ into their minds as opposed to when they were told to actively search for the target of interest (Smilek et al., 2006). While the literature makes it clear that task strategy influences task outcome, the optimal strategy appears to vary by task, with analytical, active strategies being the better strategy in some tasks, but not others.

Some researchers suggest that the differences in task outcome are due to differences in the recruitment of the various cognitive systems across task strategies (Smilek et al., 2006). For instance, a passive search strategy in a visual search task may recruit executive functions to a smaller degree than an active strategy (Smilek et al., 2006). Such an interpretation would be in line with the previous findings (Jacoby & Brooks, 1984; Marcel, 1983; Van Selst & Merikle, 1993; Whittlesea & Brooks, 1994), where the optimal strategy is consistent for a given task, but varies between different tasks. This is could be because tasks differ with respect to the requirement of various cognitive functions to effectively complete the task in question. For example, a memory task might benefit more from an increase in recruitment of memory functions, while an executive functions task may see less of an improvement from the same change in recruitment. In this way, identifying the optimal strategy for a given task can give important insights into the cognitive functions upon which the task depends.

To our knowledge, there has yet to be an investigation into the effects of task strategy on the outcomes of the superstitious perception task. Given that the superstitious perception task is intended to measure internal representations (Gosselin & Schyns, 2003), identifying the optimal task strategy for the task may give insight into the functions from which those internal representations are derived. To this end, we investigated the effect of two task strategies on the outcomes of the task. In a passive task strategy, participants were asked to let the target ‘pop’ into their mind when viewing the trial images. In an active task strategy, participants were asked to actively search for evidence of the targets’ presence in each trial image. If a passive task strategy is more effective than an active task strategy, the internal representations are likely intrinsic to the visual system. However, if the active task strategy is most effective, then executive functions are likely involved in the manifestation of these internal representations.

1.3 Hierarchical convolutional neural networks and the human visual system

Recent advances in computing technology have expanded the capabilities of computational models used for simulating neural connections in the human brain. The advent of highly parallelized processing means that we can increase the number of simultaneously simulated neurons. From these new technologies, a particular set of models have emerged that model the way that the human visual system translates points of light detected in the retina into a representation of the world around it. These models, known as Hierarchical Convolutional Neural Networks (HCNNS), have even been used to predict neural activity relating to object recognition with remarkable success (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). However, we are not aware of any attempts to test these models by using their outputs to predict human behaviour beyond testing categorization accuracy. Comparing the behaviours of humans and those of these models at a deeper, more fundamental level will allow us to probe the accuracy of the models. Do HCNNs behave similarly to humans on a task that probes the boundaries of the abilities of the human visual system to detect patterns? In this paper, we hope to provide an answer to such a question by creating a neural network that is able to perceive superstitiously. We then hope to compare the outcomes of the superstitious perception task between humans and HCNNs to identify their similarities and differences.

The hierarchical nature of HCNNs mimic our current understanding of the architecture of the human visual system, with neurons at higher levels of the processing hierarchy having receptive fields that are larger relative to the visual input, and that select for more complex visual features (Dyan & Abbott, 2001; Yamins & DiCarlo, 2016). These HCNNs are a class of neural network models that search for patterns in incoming data using filters, labelling sections of data that match a specific pattern through high levels of activation in the individual nodes representing that area of the input. In order to distinguish the biological neurons of the human brain from the simulated neurons of the HCNNs, the simulated neurons will be referred to as nodes. In the first layer of the network, a set of filters are applied to the input image, to detect patterns of interest (e.g., Gabor filters of various orientations). For each section of the image, there is a node whose activity represents the degree to which the input image matches one of the filters. When taken together, this means that the first layer of an HCNN is a set of spatial maps, one for each filter, indicating the locations in the image where patterns in the input image match the pattern in a filter. The activations of the nodes in the first layer are then used as the input data for the second layer, where the same procedure is followed. In essence, the first layer of the network searches for specific patterns in the image, the second layer searches for specific second-order patterns in the collection of patterns identified by the first layer, the third layer searches for specific patterns in the collection of patterns identified by the second layer, etc.

One example of an implemented HCNN is that of Krizhevsky, Sutskever, and Hinton (2012), commonly known as AlexNet, which consists of 8 layers. In the first 5 layers (layers 1-5), the filters are smaller than the width of the previous layer, meaning that the spatial locations are preserved from one layer to the next. The next 2 layers (layers 6 & 7) have filters that cover the entirety of the previous layer, meaning that each node has a unique filter - each node is sensitive to patterns in the entirety of the image. The final layer is another layer where the filters are the size of the previous layer except that it is used as the classification layer. In the classification layer, each node corresponds to a class that the neural network is trying to identify in the images. The activation of nodes in this layer can be interpreted as the networks' confidence in the presence of that object in the input image. This procedure is similar to the human visual system in that the receptive fields of neurons in the visual system increase in size relative to the visual input as one increases from low-level areas to higher level areas (Dyan & Abbott, 2001).

The similarities between HCNNs and the human visual system go further.

AlexNet, trained to identify a wide array of object classes, shows early receptive fields that show a surprising similarity to receptive fields of neurons in V1 (Krizhevsky et al., 2012). These computational receptive fields of the HCNN were not determined intentionally, but were settled upon as the most effective means of classifying the images it was presented through a learning process known as stochastic gradient descent (SGD; Bottou, 1991, 2010). In short, SGD involves showing an example image to the network and estimating how wrong the neural network was in classifying that image. This margin of error is then reduced by modifying the strength of the connections between nodes within the network, either strengthening or weakening each connection, so that the network becomes closer to being correct in its classification of the objects in the images it was shown. This process is repeated with new images until a certain error threshold is met and the researchers conclude that the model has reached the approximate global minimum in classification error (i.e. it is as accurate as its going to get without over-fitting to the specific images that it has been shown).

Early attempts to model the primate visual system have now been recognized as falling under the umbrella of HCNN models (e.g., HMAX; Yamins & DiCarlo, 2016). These early models share similar functional architecture as the human visual system and AlexNet, but were constructed by attempting to mimic the filters, or receptive fields, of visual processing neurons measured by neural recordings in the early visual areas of various mammals (Poggio & Riesenhuber, 1999). This differs from AlexNet in that it does not adapt to best distinguish between the stimuli it is presented. Instead, all features for which the system is sensitive are hand-designed to match those observed in the human visual system.

It is possible that the higher order representations are similar in the human visual system and in HCNNs, but as the neural signal is abstracted from the raw image input, it becomes harder to interpret its receptive field, thus making it harder to compare with human neurons. This is mostly because of the difficulty in visualizing and interpreting the receptive fields of these higher-order artificial neurons. In the first layer of the HCNN, the filters are directly referencing the image itself. As a result, the receptive fields of those neurons can be directly interpreted, as the strength of connection between a given pixel and a given neuron is directly related to what the neuron is detecting in the image. However, once you abstract beyond the first layer of the network, the interpretability of the strength of connections between pixel and image weakens. This is because the upper layers receive the activation of the lower level layers as input.

The similarity in higher order representations of visual stimuli, in both HCNNs and the human visual system, can be compared using an analysis known as representational similarity analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008). This technique involves comparing activity within a system in response to a wide variety of stimuli. From there, the similarity in activation between each combination of pairs of stimuli gives an indication of the degree to which the system sees those two stimuli as being similar. For example, consider a system that displays the same pattern of activation in response to a house and a car. Due to the similarity in activation, one can conclude that the system represents houses and cars as similar objects. On the other hand, if the system shows very different patterns of activity in response to those stimuli, one could say that the system represents cars and houses very differently. If two systems show similar patterns of representational similarity between a wide variety of stimuli, one can consider them to be similar with respect to how they represent objects.

In a HCNN, the activity of each layer of the network can be saved directly to memory during the simulation, since its activity is being calculated during the simulation. As such, HCNN activity is easy to acquire, and is easily reproducible. Since there is no noise introduced in most HCNNs, the resulting activation for a given image will be exactly the same each time that it is presented. In addition, it is important to note that the unit of such activity is arbitrary, typically ranging from 0 to 1, or -1 to 1. However, it can be interpreted as a relative firing rate for the purposes of these types of comparisons, with the minimum value indicating the least activation, and the maximum value indicating the most activation.

In contrast to the HCNNs, neural activity in the primate visual system is much harder to measure as there is no direct access to neural activity, but it can be done by a variety of approaches, including functional magnetic resonance imaging (FMRI) (e.g., Kriegeskorte et al., 2008), magnetoencephalography (MEG) (e.g., Cichy et al., 2016), or acfEEG (e.g., Cichy et al., 2016), each of which has advantages and disadvantages. Critically for RSA analysis, though, is that the analysis technique would be more or less the same regardless of the method of neural measurement - the only thing that would change is the interpretation of the results. In fMRI, which has good spatial resolution but poor temporal resolution when measuring neural activity, the representational similarity matrix generated from the data will be predominantly interpreted as a measurement of similarity in *spatial* encoding (i.e. which neurons fired). For MEG or EEG, which have better temporal resolution but worse spatial resolution than fMRI, the data will be predominantly interpreted as a measurement of similarity in *temporal* encoding (i.e. when neurons fired).

Recently, researchers have attempted to compare activity in the macaque visual cortex with that of the high-order layers of a HCNN using RSA (Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016). Yamins and DiCarlo (2016) showed that a performance optimized HCNN was able to predict spatial patterns in neural activity in the macaque inferotemporal cortex more accurately than other leading computational models of human vision (Yamins & DiCarlo, 2016; Yamins et al., 2014). In addition, recent comparisons of the primate visual system and HCNNs showed that low-level activity in a similar HCNN model predicted low level visual activity in the macaque visual system, both temporally and with respect to spatial activity (Cichy et al., 2016). Taken together, these observations have led some to believe that they are the best models of human vision currently available (Yamins & DiCarlo, 2016). At a behavioural level, we see similar trends in the comparison between human vision and our computational models. Recent HCNN models have been able to achieve accuracy in view-variant classification tasks that are indistinguishable from human performance on the same task (Kheradpisheh, Ghodrati, Ganjtabesh, & Masquelier, 2016). All of this evidence suggests that the HCNN model may be well on its way to accurately modelling human vision by looking at the internal activations of the two systems while performing the task for which the HCNN was trained.

On the other hand, there is also evidence suggesting that HCNNs do not fully model the functional processes that underlie human vision. For instance, recent attempts to compare human image content similarity judgments to those made by an HCNN suggest that parallels in HCNN performance may not extend beyond image classification (Peterson, Abbott, & Griffiths, 2016). Moreover, in experiments trying to compare the view-variant classification accuracy between the model and the human visual system, visual masks were used to reduce re-entrant processing effects (Kheradpisheh et al., 2016). This procedure, suggests that the HCNN visual model is an incomplete model of the human visual system, through the implication that the re-entrant processes in human vision are not included in the HCNN model. When examining the architecture of HCNNs, processing proceeds linearly and unidirectionally, with information being passed in only one direction up the hierarchy, from early layers to late layers. Re-entrance involves information also being passed down from high-order processing areas to lower-order areas (Fahrenfort, Scholte, & Lamme, 2007). However, it is important to note that these findings do not preclude the possibility that the HCNN is simply an incomplete, but accurate, model on human vision. Further comparisons such as those made by Peterson et al. (2016) and Kheradpisheh et al. (2016) should be considered in order to identify the limits of the HCNN-human visual system comparison. In the following experiments, such comparisons were used to identify whether the human visual system and the HCNNs use similar criteria to perform object recognition.

There are some notable similarities between the human visual system and HCNN models. Often, they are compared at the level of representational similarity in neural activity (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016). To our knowledge, there has been relatively little comparison between the behavior of human and computational models of vision. The present work aimed to further the process of making such behavioural comparisons between systems. Specifically, we aimed to compare the behavior of HCNN and human participants in a way that did not mean directly testing the neural network on a task for which it was trained, using a superstitious perception paradigm (Gosselin & Schyns, 2003).

In summary, the primary goal of this manuscript was to evaluate the validity of the superstitious perception technique (Gosselin & Schyns, 2003) with respect to two assumptions. The first assumption is that the Classification Images (CIS) generated using this technique reflect a measure of the internal representations of a participants' visual system. In order to test this assumption, we attempted to predict participants' false alarm rates when they attempt to detect a target within novel white noise stimuli. If the predictions were accurate, we can conclude that the technique met this assumption. However, if prediction of participants' false alarm rates was unsuccessful, then it suggests that the CI does not represent the internal representation of a target. The second assumption is with respect to the task procedure. Specifically, in the original experiments by Gosselin and Schyns (2003), participants were not given a particular strategy to use during the task. We believe that by identifying the optimal task strategy, we can locate the cognitive functions from which the internal representations are derived. Here, we investigate the effect of task strategy on both internal and external prediction ability.

In addition to the primary goal, we also compared the behaviour of an HCNN to that of the human visual system during the superstitious perception task. The goal was to probe the ability of a HCNN to model human vision by testing its ability to perceive superstitiously as humans do. This test has a number of advantages over previous comparisons between HCNNs and humans. Most notably, the HCNN will be trained on data that are not generated by human responses, as is the case for object recognition in natural scenes. If the HCNN bears similarities in response to humans despite its training on objective, non-human data, then we can conclude that the model itself, and not the training on human data, is what is leading to the HCNN - human visual system similarities.

Chapter 2: Experiment 1

The overall rationale and design of Experiment 1 are illustrated in Figure 1. The main goal of this experiment was to replicate the methodology of Gosselin and Schyns (2003) in a context in which the participant's search images were known. As such, participants were assigned to search for one of the four specific targets (shown in Figure 1A) in noisy images. Each participant was shown an image of their specific target prior to beginning the task, as is illustrated for the X-target (shown in Figure 1A). This differed from Gosselin and Schyns (2003), who did not control the target shapes participants searched for beyond telling them to "look for the letter ′S′, present in half of the following images." The four search targets in the experiment were selected so that participants could be assigned to targets that differed from one another only a little (e.g., the X and the italic X share a high degree of similarity) or by a large margin (e.g., the X and the two plus signs have a low degree of similarity).

Given that there was a defined target for participants to use as an internal representation, we reasoned that fewer trials were necessary to achieve the same strength of test. The goal here was to bring the relative power of the two experiments into a comparable range. Since the targets are more clearly defined in our paradigm, fewer trials would be required to reduce the statistical noise to levels comparable to that of Gosselin and Schyns (2003). As such, just under half as many trials were collected per target in this experiment, 8,000, relative to the Gosselin & Schyns experiment, 20,000.

Figure 1B illustrates the process of creating a classification image from the

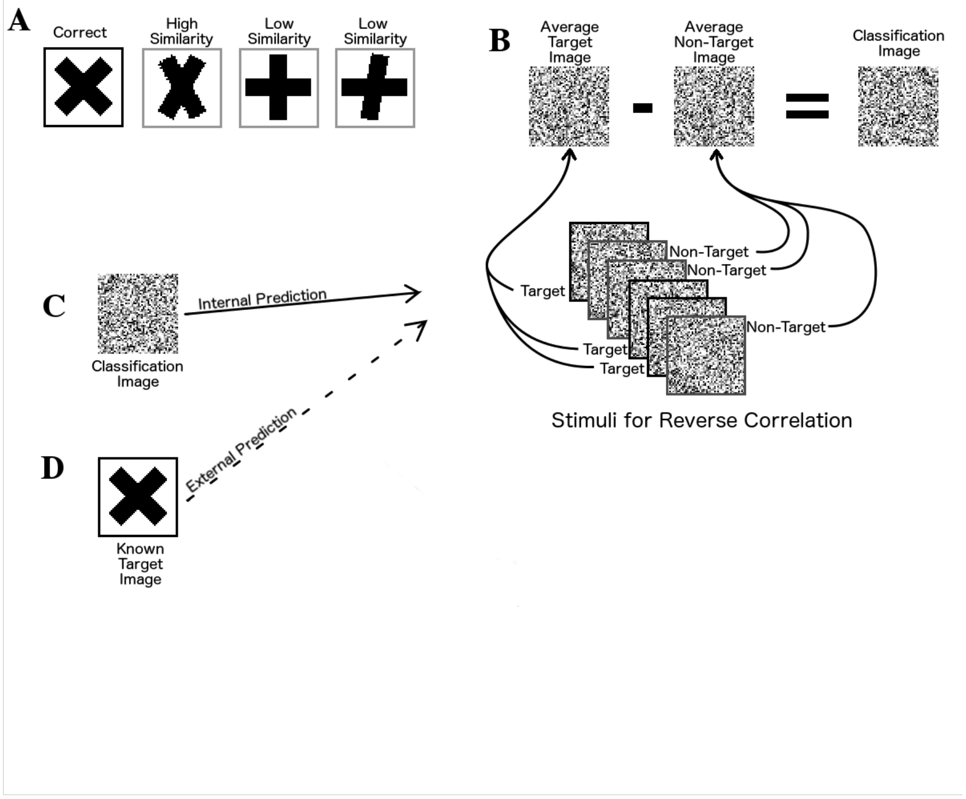

Figure 1: A visual summary depicting the difference between reverse correlation, internal prediction, and external prediction. (A) Participants were assigned to one of four possible targets. (B) Reverse correlation is the process of generating the CI by looking at the sum of stimuli that led to a false alarm, and subtracting the sum of stimuli that did not lead to a false alarm. (C) Internal prediction is the process of using a generated CI to predict participants responses to the stimuli used to generate the CI. This process is referred to as backward prediction in the neural spike train literature. Such a process will lead to an over-estimation of the CI's goodness of fit. (D) External prediction is mathematically identical to internal prediction, but differs in using a generated CI to predict responses to novel white noise stimuli. This process is referred to as forward prediction in the neural spike train literature, and serves to give a better estimation of goodness of fit. In this experiment, however, we will use participants' assigned target images in lieu of a generated CI.

responses of participants who were trying to detect a given target. The four classification images, corresponding to each of the four targets, could then be evaluated to see how much they differed from one another. The average difference in correlation magnitude between target CI and low versus high similarity non-targets was thus a way to quantify the extent to which differences in target images were reflected in differences in the corresponding classification images.

In Experiment 1, we first repeated the Gosselin and Schyns (2003) method of preparing CIs using 50 pixel by 50 pixel images and participants' responses to a signal detection task using the reverse correlation logic. But we did not end our analysis there, at steps A and B in Figure 1, as Gosselin and Schyns (2003) did. Stopping here implies that the success of a superstitious perception study is evaluated primarily by an intuitive visual comparison made by the experimenter (and/or the reader). Rather, we went on to use the CIs to predict participants' responses in two more ways, in an effort to put the evaluation of superstitious perception results onto a more objective footing.

First, we used the CIs to make internal predictions, following the logic of DeAngelis et al. (1993) research on single cell receptive fields (Figure 1C). This allowed us to estimate how much variance in the responses made to the individual stimuli giving rise to the CI were explained by this model. Second, we followed the suggestion of Ringach and Shapley (2004) to compute external predictions for comparison with the internal predictions (Figure 1D). Any discrepancies between these two measures would provide an estimate of the extent to which reverse correlations on their own are capitalizing on chance variation in the data, as suggested by DeAngelis et al. (1993).

In summary, Experiment 1 had three goals. The first goal was to replicate the Gosselin & Schyns procedure in a context where the shape of the participant's target images are known in advance and can be compared for their similarity to one another. The second goal was to examine the internal prediction accuracy of the classification images to estimate the degree of predictive accuracy achievable by the generated image, as argued by DeAngelis et al.. The third goal was to compare these internal predictions with external predictions, in order to estimate the degree to which reverse correlations are fortuitously capitalizing on measurement error. The addition of the internal and external prediction analyses removes the visual-interpretive role of the experimenter in estimating the strength of a superstitious perception effect.

The design of Experiment 1 leads to two sets of hypotheses. The first set concern the CIs and their correlation with the images that gave rise to them (reverse correlations). If the CIs generated in the superstitious perception task differentiate between the low-similarity target images assigned to participants (e.g., X versus plusses), then the superstitious perception technique of Gosselin and Schyns can be considered to have low-resolution discriminability between differing internal representations. If the resulting CIs are further able to differentiate between between high-similarity targets (e.g., X from italic X), then the technique must have high-resolution discriminability. The alternative outcome of no discriminability among the four target shapes would imply that the procedure cannot differentiate between internal representations that differ in the way a + differs from an x, thus seriously undermining the logic of the superstitious perception task.

A second set of critical predictions concern the relative strength of internal and external predictions. If the external predictions are similar in accuracy to the internal predictions, then it suggests that the CI is a reliable estimate of a participant's internal representation. Alternatively, if the external prediction accuracy is much smaller, it suggests that the CIs generated from the superstitious perception task are over-fitting the data that comes from presenting noise images to participants and, as such, are less accurate estimates of participant's internal representations than has previously been assumed (Gosselin & Schyns, 2003).

2.1 Methods

2.1.1 Participants

Nine participants (five females) between the ages of 18 and 25 were recruited from a paid subjects pool at the University of British Columbia. The only restriction on participation was that participants must have had normal- or corrected-to-normal vision, verified by self-report upon participants' arrival to the lab. All participants were tested one at a time on the same computer.

2.1.2 Stimuli and procedures

Participants came into the lab twice and repeated the same set of procedures each time. This was done in order to collect 4000 experimental trials from each participant, without participants becoming overly fatigued during the experiment. Each time the participants came into the lab, they completed a two-phase task. This task included a practice phase to acquaint participants with the procedures, and an experimental phase in which there were no targets, and from which a CI was constructed for each participant. The practice phase was further broken into two conditions: (1) an easy condition, intended to allow for practice identifying the target for which they were searching; and (2) a hard condition, intended to encourage participants to identify images in the experimental phase by giving them the impression that targets were present but very difficult to identify. This practice phase procedure is similar to that used by Reith et. al 2011, who were able to obtain reliable results in a similar task.

While two phases were completed per session, the core task throughout the experiment was the same. All participants were asked to view images, one at a time, in the centre of the screen. They were then asked to identify whether a target, shown to them earlier, was present in the noisy image in front of them. They were then asked to indicate its presence or absence with different key presses.

Participants were randomly assigned to one of four groups: X, italic X, +, and italic + such that there were at least two participants per group. The group names denote the type of target the participants were given during the experiment. All participants completed the task on an iMac late 2009 model, and the task was constructed using Matlab 2010a and PsychToolBox v11 (Brainard & Vision, 1997; Kleiner et al., 2007; Pelli, 1997). Images presented to participants occupied about 2° of visual angle, and were 50px X 50px in resolution.

The target images in the practice phase were generated by overlaying white pixel noise over top of the target image in question. Non-targets were generated by simply overlaying pixel noise onto a solid grey square with the same average luminance as the target image. The contrast in the target image was adjusted between the easy and hard practice phases in order to make the target harder to identify. The overlay noise was generated by sampling from a normal distribution with a mean of 0, and a standard deviation of 0.2 for each pixel in the image which was then added to the target image. Notably, the average luminance of each trial image stayed constant throughout the experiment, with minor variations due to random sampling of the noise. During the experimental phase of the experiment, all stimuli were generated by overlaying white pixel noise atop the solid grey square, whose luminance matched that of the average luminance in the target.

2.1.3 Data analysis

Sensitivity (d$'$) and decision bias (lnβ) was used to assess the degree to which participants understood the task and participated as per the instructions. One participant who showed near zero sensitivity to the target in the easy and hard conditions was removed from subsequent analyses.

Classification Images (CIS) were generated for each participant based on the responses they gave for each trial image during the test phase. This was performed by adding together all of the trial images in which a target was identified and then subtracting all of the images in which a target was not identified. This procedure is exactly as performed by Gosselyn & Schyns 2003. When visualising the images, a gaussian filter was used to smooth the image in order to eliminate random noise, yielding a fully processed CI for each target. Since the spatial frequency of the targets presented in this experiment were at most 3 cycles per image, the gaussian filter was configured to eliminate frequencies higher than 3 cycles per image. This spatial smoothing procedure is as performed in previous literature (Gosselin & Schyns, 2003).

These CIs were then assessed with regards to their similarity to the target for which each participant was searching. This was done by taking the IWC between each participant's CI and each of the 4 targets. The predicted target for each participant was the target with which their CI was most correlated. CIs that discriminated the participant's assigned target from the other, irrelevant targets in the experiment were taken to be accurate.

Once the CIs were created, and their accuracy was quantified by using them to predict the participants' targets, another approach was used to assess consistency in responses across participants. We attempted to internally predict the likelihood that each participant would select a given trial image as containing a target. This was done first by computing Image-Wise Correlations (IWCS) between the CI and the given trial image. These IWCs were interpreted as an index of the degree to which a given trial image matched the participant's generated CI. They were generated for every single image that was presented to participants. Then the trials were sorted from least to greatest with respect to their IWCs, and then split into 400 bins in order to estimate the likelihood of an image with a given IWC being selected as containing a target. The number of bins was selected due to its allowance for enough trials in each bin for the resulting proportion to be considered continuous. To ensure that bin size had no effect on the reliability of any resulting relationships, the same procedures were also conducted using 50 and 1000 bins. The proportion of images identified as having a target was then calculated for each bin, along with the mean IWC value. The relationship between these two variables was then observed and interpreted. This procedure, predicting participants' responses using the CI will be referred to as internal prediction.

A similar procedure was used to estimate predictive power of the target image that was shown to participants during the training phase of the experiment. To do this, the IWC between the participants' respective target image and each trial image was calculated. This yielded a single number for each trial image, interpreted as the degree of similarity between the trial image and the target the participants were assigned to identify. As performed with the CIs, the same binning procedure was used to generate the proportion of images selected as containing a target, as the degree of similarity between the image and the target varies. This procedure, predicting participants' responses using the known target they were trying to detect, will be referred to as external prediction.

2.2 Results

2.2.1 Practice phase

Analysis of participants' individual performance on the practice phase was assessed by examining sensitivity (d$'$) and decision bias (lnβ) for both the easy and hard conditions. The purpose of this preliminary assessment is to identify participants who did not understand the task. One participant was removed due to having near zero sensitivity in both the easy and hard conditions.

Figure 2A shows that there were generally very high sensitivity scores (d$' > 4$), and that there was no difference between easy targets (mean d' = 4.15, SE = 0.24) and hard targets (mean d$'$ = 4.34, SE = 0.25), *t*(15.95) = -0.56, *p* = 0.59, 95% CI = [-0.92 , 0.54]. Figure 2B shows participants' average decision bias, as assessed by log transformed β (lnβ) in both the easy and the hard conditions. The results show that participants were generally more liberal (more willing to say 'target present' when uncertain) in the easy condition (mean lnβ = -0.77, SD = 0.3889410) than in the hard condition (M = 0.59, SD = 0.44), *t*(15.73) = -2.31, *p* = 0.035.

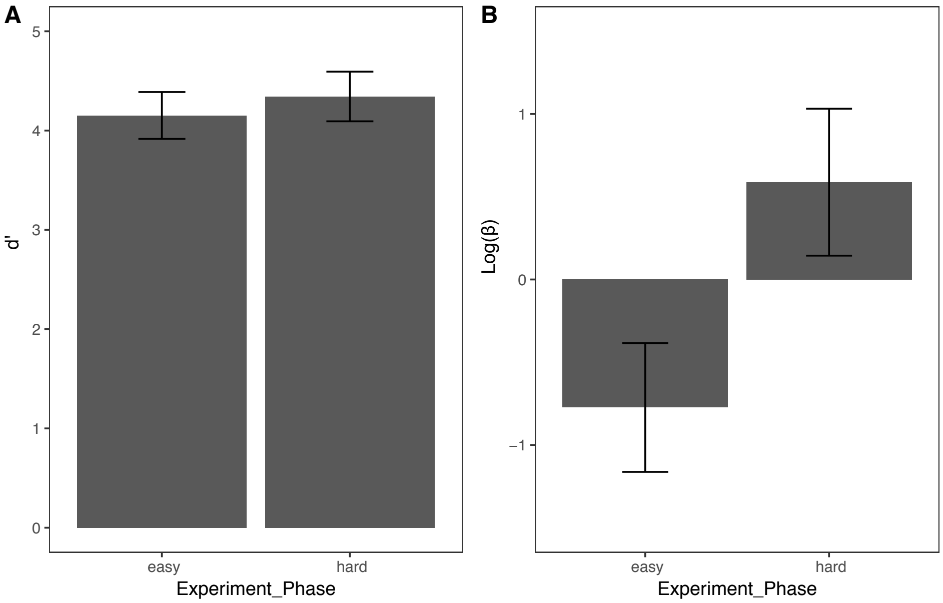

Figure 2: Summary of sensitivity and decision bias in the practice phase of Experiment 1. A) Mean sensitivity by target visibility. B) Mean decision bias by target visibility. Error bars represent the standard error of the mean.

2.2.2 Test phase

Participants detected targets in the test phase of the experiment, where no true targets were presented, at a rate of 32.8% on average (SD = 18.0%). This compares to a false alarm rate of 4.9% (SD = 5.0%) in the easy and 1.3% (SD = 1.5%) in the hard conditions of the practice phase, suggesting that participants were indeed seeing superstitiously throughout the task. They also took considerably longer to respond in the test phase (M = 1661ms, SD = 1022ms), than in the practice phase (M = 793ms, SD = 145ms). They responded to an average of 2729 trials (SD = 854) within their 2 sessions of 1 hour. No sensitivity or criterion measures could be computed in the test phase because no targets were presented, meaning that all responses were correct rejections or false alarms.

Figure 3 shows the CIs generated for each target. CIs were calculated for each target by adding together all of the images in which the target was identified and then subtracting the images in which no target was identified, as performed by Gosselin & Schyns (2003). In order to display them as an image, linear transformations were performed on the data, subtracting the minimum value, and then dividing by the range. This was performed so that the minimum pixel luminance was 0, and that the maximum pixel luminance was 1. Finally, a spatial filter was applied to allow spatial frequencies around 3 cycles per image, as performed by Gosselin and Schyns.

Target CIs were correlated with the 4 possible targets that were assigned to participants to estimate the degree to which they could be used to discriminate between the target it was generated to represent and those that it was not. To this end, an image-wise correlation was calculated between the CI and each target image for each CI. These IWCs were grouped into two groups: those generated with the target associated with the given CI, those generated with the target that had high similarity to the target associated with the given CI, and those that had low similarity to the target associated with the given CI. For example, for the 'x' CI, the IWC associated with the 'x' target image was grouped as a "correct target", the IWC associated with the italic 'x' was grouped as a "High Similarity Target", and the two associated with the "+" targets were grouped as "Low Similarity Targets". Overall, the average IWC in the "Correct Target" group, *r* = .33 (SD = .04), was the same as the average IWC in the "High Similarity Target" group, *r* = .33 (SD = 0.06). The average IWC in the "Low Similarity Target" group , *r* = 0.25 (SD = .05), was significantly smaller than the "Correct Target" group, *t* = 2.30, *p* = 0.045.

These data, therefore, replicates important aspects of previous studies that used superstitious perception as a technique (Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2017; Gosselin & Schyns, 2003, to name a few). The question we turn to now is whether the data can be used to make predictions about which specific images participants will be most likely to report a false alarm (report a "target" when it is not present). The slope of relationship when trying to make predictions with participants' CIs will serve as a measure of the degree of sensitivity for the target participants' mental representations. The steeper the slope, the more consistent

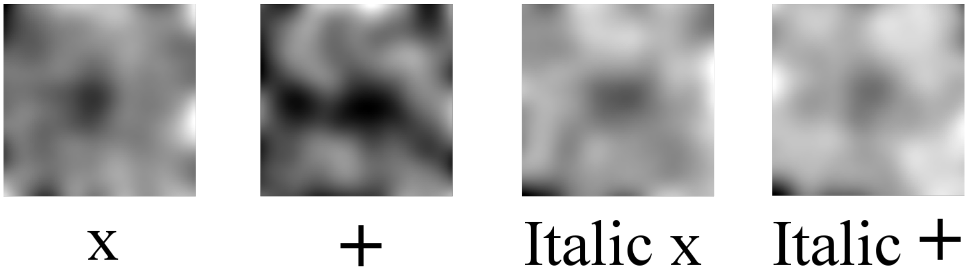

Figure 3: Classification Images generated across participants with each target. Note that the target that participants were trying to detect is difficult to discern, but present.

participants were in their selections.

Evaluation of internal prediction accuracy

Next, each image that was shown to a participant was correlated with the participant's CI in order to obtain an Image-Wise Correlation. The IWCs were calculated by computing the correlation between the two images at the pixel-by-pixel level. This meant that each image shown to a participant in the experiment was assigned a correlation value indicating the degree to which it was similar to (i.e., resembled) the mental representation that the participant was using to make decisions. A positive correlation indicated that the pattern of dark and light pixels in a specific image are similar to the pattern of dark and light pixels in the target image; a negative correlation indicated that dark regions in the target image tend to correspond with light regions in the trial image, and vice versa.

Figure 4 shows the proportion of 'target' responses (i.e., the proportion of images in each bin in the data in which a false alarm occurred) plotted against the average IWC between each image and the CIs shown in Figure 2. The data in their entirety consisted of nearly 32,000 points (8 participants x 4000 trials, minus a few dozen missing trials for some participants). The 32,000 points were grouped into various numbers of equally sized bins. This was both to smooth the data, and to convert the binary response variable into a pseudo continuous variable, interpretable as the proportion of images selected as containing a target. To understand the effect of the binning procedure on the observed relationship, we examined the relationship between 'target' responses and the IWCs at bin numbers of 50, 400, and 1000. The results are shown in Table 1.

These correlations were obtained after fitting a sigmoid function to the data, yielding a slope of 93.76 (SE = 1.34), and a centre constant of 0.008 (SE = 0.0002). The maximum of the sigmoid function was fixed at 1, as participants cannot identify more than 100% of the images shown as containing a target. Using these constants, a sigmoid transformation was performed on the IWCs between the target images and CIs. There was a strong linear relationship between sigmoid-transformed IWCs and proportion of 'yes' responses by participants, $r(398) = 0.99$, $p < .001$. It is important to note that this relationship is a strong over-estimate of the true predictive power of the CI, as we are making predictions of responses using information generated from the decisions we are trying to predict. The crucial point here is that the slope of the sigmoid function is very steep, suggesting that participants could easily discriminate between which images matched their criterion, and which images did not.

Table 1: The effect of bin size on the correlation between the proportion of images selected as containing a target and the IWC between the trial image and the CI.

Number of Bins	Observed Correlation	Statisical Significance
50	$r(48)= .99$	$p < .001$
400	$r(398) = .99$	$p < .001$
1000	$r(998) = .98$	$p < .001$

These results suggest that participants are quite consistent in their internal representations, as they've been measured by the superstitious perception technique. The next question we wanted to answer is whether our technique of estimating their mental representations is actually externally valid, given that they were assigned a target to try to identify in the images. In other words, given that we've told them what to look for in the images, we have some degree of certainty of what their mental representation should be. To estimate the degree of match between these

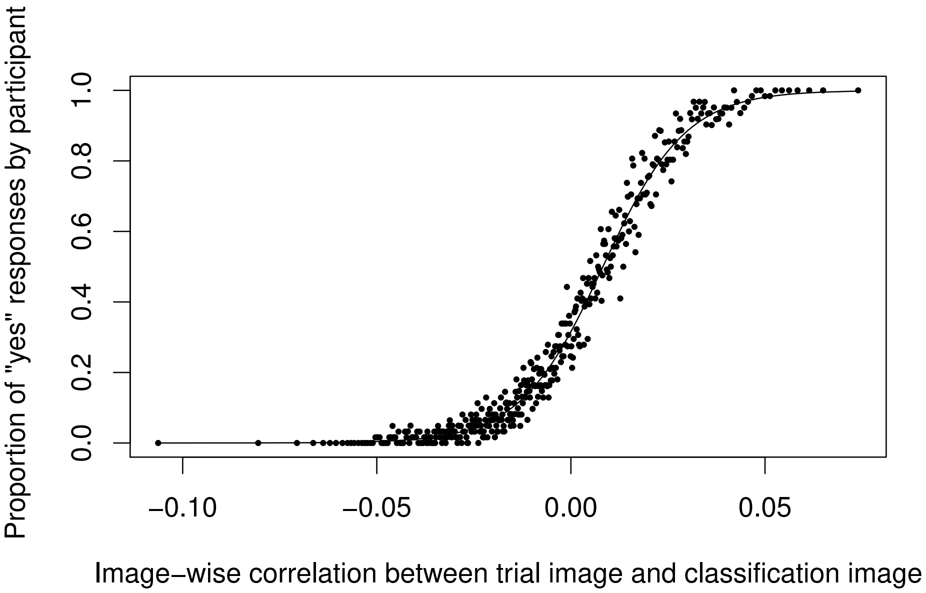

Figure 4: A scatter plot showing the relationship between the IWC of the CI and each trial image, and the proportion of images in each bin selected as containing a target. The relationship is sigmoid in nature, and the line drawn in is the sigmoid of best fit, with slope of 93.76 (SE = 1.34) and a center constant of 0.008 (SE = 0.0002).

two concepts, we will attempt to predict participants' responses to trial images using the image of the target they were assigned. The degree to which we can predict participants' responses using the targets they were assigned will help quantify how accurate our mental representations are.

Evaluating image selection using external prediction

With the target image that was shown to participants to depict what they were trying to detect, a second set of IWCs were calculated. The IWCs were calculated by computing the correlation between the two images at the pixel-by-pixel level.

The IWCs between the target image and each trial were binned into 400 bins.

The proportion of images identified as containing a target was calculated for each bin. This allowed for a continuous measure of the participants' collective likelihood of perceiving a target within an image that has a given IWC score with the target that they were trying to detect. A positive IWC indicated that the respective pixels in the given trial image and the target image are positively correlated, and dark regions in the target image tend to correspond to dark regions in the trial image. In contrast, negative IWC indicated that the respective pixels are negatively correlated, and dark regions in the target image tend to correspond with light regions in the trial image. In other words, the more positive the IWC becomes, the more "target-like" the image at a pixel-by-pixel level. The more negative the IWC becomes, the less 'target-like' the image, again at the pixel-by-pixel level.

Figure Figure 5 shows the proportion of 'yes' responses (i.e. the proportion of images in each bin in which a target was identified) plotted against the average IWC in each bin.

The data shown in Figure 5 are from binning the data into 400 discrete categories. At this level of resolution, there was a moderate linear relationship between the target-generated IWCs and the proportion of 'yes' responses among participants, r(397) = .34, p <.001. One outlier was identified in this dataset, shown in grey. This outlier was removed for the subsequent analyses. However, its removal had no influence on the conclusions drawn. In its most detailed form, the data consisted of 32,000 points (8 participants x 4000 trials). In the non-aggregated form, responses are binary in nature with participants having to choose either to identify a target or not, even if their confidence is approximately 50/50. In order to smooth the noise in the data, and to better estimate participants' confidence in their response, we examined the relationship between 'target' responses and the IWCs at bins ranging from 50 to 1000. The results are shown in Table 2. The apparent magnitude of the relation increases as the number of bins is decreased, though the reliability of the relationship is approximately equal across bin size. This is presumably because the number of data points (degrees of freedom) trades off with the increase in measurement error that accompanies increasing sample sizes. The important point is that there is a reliable relation between false alarms rates and the

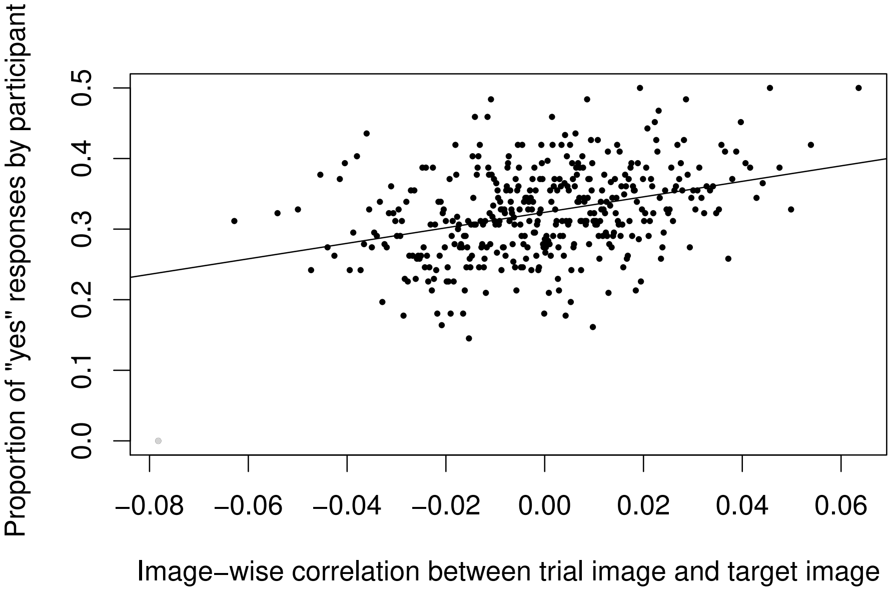

Figure 5: The relationship between the likelihood of an image being selected as a target, and the IWC between that image and the target image.

resemblance of trial images to the target template defined by the experimenter.

Table 2: The magnitude and reliability of relation between the proportion of images selected as target and the IWCs between the trial image and the target image, for the data binned at various levels of resolution.

Number of Bins	Observed Correlation	Statisical Significance
50	*r*(47)= .65	*p* <.001
400	*r*(397) = .34	*p* <.001
1000	*r*(997) = .23	*p* <.001

2.3 Discussion

In Experiment 1, we had three primary goals. First, we aimed to reproduce the results of Gosselin and Schyns (2003) and their superstitious perception technique. We also hoped to estimate the resolution of the technique through the use of targets with varying degrees of similarity. Second, we introduced the concept of internal prediction in order to estimate the degree of participants' confidence in their responses, and thus the strength of their internal representation of the target. This was quantified using the slope of the sigmoid function of their responses. Finally, we validated the concept of external prediction for predicting participants' responses to stimuli using a known representation of their internal template that was not derived from their responses.

Classification images

The CIs generated from the responses in the testing phase were less clear than previous work (Gosselin & Schyns, 2003; Rieth et al., 2011, , for example). This was not surprising, however, since we used less than half the number of images per target. Attempts to evaluate which target was associated with which CI yielded mixed results. On one hand, the CIs were able to distinguish between high- and low-similarity targets. However, they were unable to distinguish between high-similarity targets and the correct target. This could be explained by participants not mentally distinguishing between these targets - participants remembered both the "x" and italic "x" as an "x". On the other hand, the failure to distinguish between high similarities could indicate that the technique itself is not sensitive enough to make such distinctions. Regardless, these results are in line with those of previous studies.

Internal prediction

Using internal prediction, or predicting responses associated with the images used to create the CI, was useful in this case to estimate participants' consistency in response. Aggregating data into bins allowed for the conversion of the binary yes/no responses into a continuous proportion of false alarms. The false alarm could be interpreted as an indirect measure of participants' confidence in a target presence. Since it is unlikely that participants' confidence in their perceptions is binary, the binning procedure allowed us to estimate their aggregate confidence that a target is present. This aggregate confidence was measured as a false alarm rate for a given IWC range. Analyzing the relationship between this proportion of responses and the IWC between targets and trial images allowed us to get a measure of participants' consistency in responses. For example, a steeper slope of the resulting sigmoid curve between the two variables suggests that the participants had a clearer, more distinct decision boundary between their decisions. On the other hand, a shallower slope would have indicated more uncertainty. In this case, there was a slope of 93.76. It remains to be seen if this is a comparatively steep or shallow slope. Furthermore, the internal prediction process verified that the IWC technique is valid in estimating the likelihood of image selection in the superstitious perception task. It suggests that the pixel-wise similarity between the trial image and the CI contains information about the confidence of participants' decisions.

External prediction

Meanwhile, external prediction, or predicting participants' responses based on the image of the target they were assigned, is a more rigorous validation of the superstitious perception technique. Since participants' targets were explicitly assigned to them, we knew the ground truth of the target that participants were using in their decisions. As such, we can estimate the degree to which the participant's responses could be predicted by the image of the target they were trying to detect in the images. Not surprisingly, the external prediction method proved much less effective in predicting participants' responses than the internal prediction method. This is likely because the images we were using to calculate the IWC were not directly generated from the participants' responses in the external prediction method, and thus there can be no over-fitting of the data. The moderate relationship between the proportion of images identified as containing a target and the IWC for the target image suggests that the technique does, in fact, measure the internal template to some extent. It is important to note that the experimenter will not always have access to the exact image that participants were trying to detect. In Gosselin & Schyns' experiment, 2003, for example, participants were told to look for an $'S'$ in the image, but the exact qualities of the "S" that participants were trying to detect was unknown. In such CI experiments, where participants' internal templates for a concept are unknown, a CI for each participant should be used to predict their responses to novel white noise stimuli.

The near-ceiling performance in the practice phase of the experiment suggests that it was too easy - even in the hard condition - and suggests that participants never struggled to detect the target. This means that participants were never encouraged to make false alarms due to increasing task difficulty, as intended in the hard phase. If participants were not making a considerable number of false alarms during the experiment, this would suggest that they were not seeing superstitiously. If this were the case, participants false alarms could have been the result of a few trials in which the noise looked extremely target-like. However, participants were still reporting false alarms in the testing phase at an average rate of 32%. Participants were reporting these superstitious perceptions at a rate comparable to those reported in previous work (Gosselin & Schyns, 2003; Rieth et al., 2011). For these reasons, the ease of the second practice phase is unlikely to have had a significant impact on the results of the testing phase.

Despite the promising results further validating the superstitious perception technique, there is still a considerable amount of unexplained variance in the participants' responses that is not explained by their assigned template. Some of this variance could be explained by slight differences in the participants' memory of the target (Holden, Toner, Pirogovsky, Kirwan, & Gilbert, 2013, for example). However, given the amount of unexplained variance in responses, it is likely that there is another mechanism at play that is being indirectly measured by this technique.

One possible mechanism that could be impacting the responses of participants is their decision-making process. Over the past decade, it has come to light that decision-making processes and response strategies can dramatically alter observed response patterns (e.g., Lifshitz, Bonn, Fischer, Kashem, & Raz, 2013; Smilek et al., 2006). This is particularly likely, given the reported strategy of a participant in Gosselin & Schyns' original paper: "[She] simply waited to see if the [target] jumped out at [her]." In chapter 3, we examine the effects of encouraging a particular task strategy for participants on the outcome of a superstitious perception task.

Chapter 3: Experiment 2

In Experiment 2, we aimed to employ the techniques we developed, in order to evaluate the effect of task strategy on the measurement error associated with the superstitious perception task. This allowed us to ask pointed questions about the accuracy of associated experimental outcomes when performed under different circumstances.

The purpose of the superstitious perception task is to study the latent representations of objects or concepts within the visual system. Therefore, one key assumption is that the task is dependent predominantly upon the human visual system. However, research areas investigating behavioural measures in both the visual and attention domains have shown the influence of task strategy on experimental outcomes (e.g., Seli, Jonker, Solman, Cheyne, & Smilek, 2013; Smilek et al., 2006). The mechanism behind these differences is thought to be differential recruitment of secondary brain networks to the task. Unfortunately, there is limited information available with respect to the effect of task strategy on the outcomes of the superstitious perception task.

Testimony reported by Gosselin and Schyns (2003) suggests that participants seem to naturally adopt a passive strategy to the task, claiming to check if the target of interest “jumps out at them” in each trial image during the task. Such a finding suggests that participants adopted a strategy of trusting their initial impression of each trial image throughout the task. Unfortunately, subsequent studies using the superstitious perception technique neglect to specify the type of strategy adopted by participants while performing the task, and thus this hypothesis remains unconfirmed. However, a task strategy that is dependent on initial impressions could be induced by reducing the exposure time to each trial image, forcing a reliance on first impressions to guide responses (Kheradpisheh et al., 2016).

The goal of this experiment was therefore to probe the effect of task strategy, as influenced by image presentation time, on the extent to which the CIs generated by the superstitious perception task estimate participants' internal representations. By testing the influence of task strategy on participants' response patterns and generated CI accuracy, we can gain a more thorough understanding of how strategy influences the ability of the superstitious perception task to achieve its stated purpose: to measure and make participants' latent representations accessible. The previous literature, including the testimony of participants from Gosselin & Schyns work 2003, suggests that adopting a passive strategy will yield the most accurate CIs. If this is the case, it suggests that a passive task strategy should be used in order to make the measurements of latent representations as accurate as possible.

3.1 Methods

3.1.1 Participants

8 participants (seven females) between the ages of 20 and 26 were recruited from the paid subjects pool at the University of British Columbia. Only participants with normal or corrected-to-normal vision were recruited, verified by self report upon participants' arrival in the lab. All participants were tested one at a time on the same computer.

3.1.2 Stimuli and procedures

All stimuli were generated in a manner identical to that described in chapter 2. Similarly, all procedures were identical except for the following deviations. First, the trial images were limited in presentation time, meaning that the trial image would appear on screen for only 500ms upon trial onset, after which the image would be replaced with a blank screen. No visual mask was used after the stimulus presentation during this experiment due to the fact that the stimuli were noise, and presenting a noise mask may lead to serious confounds (Eriksen, 1980). The trial would not end until the participant responded, but participants were encouraged to respond as quickly and as accurately as possible. This change in procedure was made to encourage participants to respond based on intuition so as to minimize non-visual decision cues (e.g., deciding to identify a target because they haven't reported seeing one in a while). In other words, the change in procedure was an attempt to force participants to follow a similar decision making strategy as reported by the participants in work by Gosselin and Schyns (2003).

3.1.3 Data analysis

Data analysis procedures were identical to those described in chapter 2.

3.2 Results

3.2.1 Practice phase

Figure 6A shows that there was a large difference in average sensitivity (d') between two conditions of the practice phase, with the easy condition (mean d' = 3.36, *SD* = 0.89) being significantly larger than the hard condition (mean d' = 0.38, *SD* = 0.44), $t(7) = 10.14$, $p < .001$. Figure 6B shows that there was also a difference in average decision bias, where participants tended to be more liberal (more likely to identify a target as being present when uncertain) in the easy condition (*M* = -0.71, *SD* = 0.88) than in the hard condition (*M* = 0.22, *SD* = 0.16), $t(7) = -3.18$, p = .016.

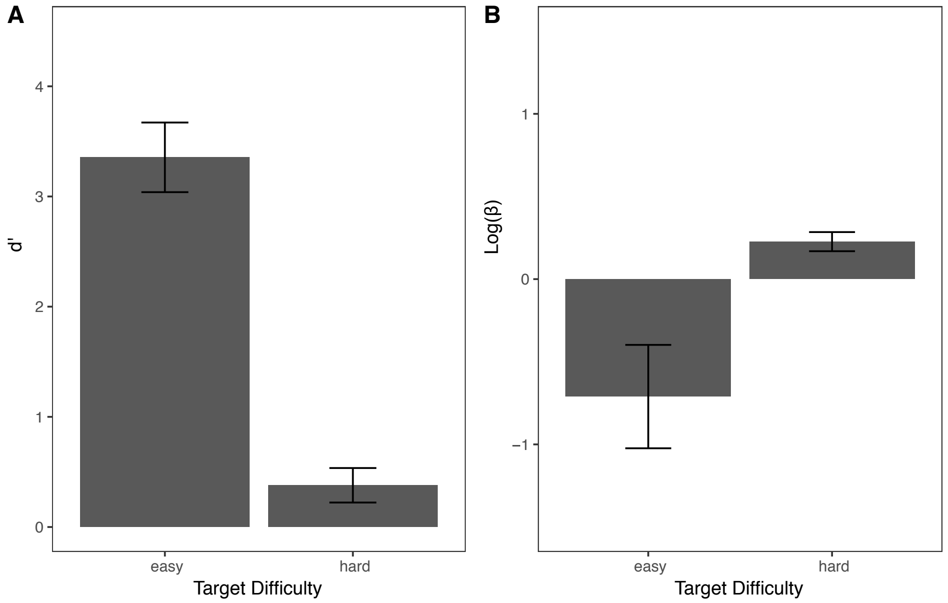

Figure 6: Summary statistics for the training phase of the experiment. A) Mean sensitivity score across the easy and hard practice conditions. B) Decision bias, lnβ, across the easy and hard practice conditions. Error bars represent the standard error of the mean.

3.2.2 Test phase

Similar to chapter 2, participants mistakenly identified an average of 36.4% (*SD* = 6.4%) of trials as containing a target, despite no trial containing a true target, in contrast to an average false alarm rate of 10.7% (*SD* = 11.5%) in the easy condition, and 26.9% (*SD* = 14%) in the hard condition. However, contrary to chapter 2, participants took less time to respond in the test phase (*M* = 503ms, *SD* = 197ms) than in the practice phase (*M* = 855ms, *SD* = 254ms). Overall, participants took much less time to respond in experiment 2 compared to chapter 2. This suggests that the manipulation worked, forcing participants to think less about the reasoning behind their response. No sensitivity or bias scores could be calculated as no targets were presented in the test phase, making all responses either a correct rejection, or a false alarm.

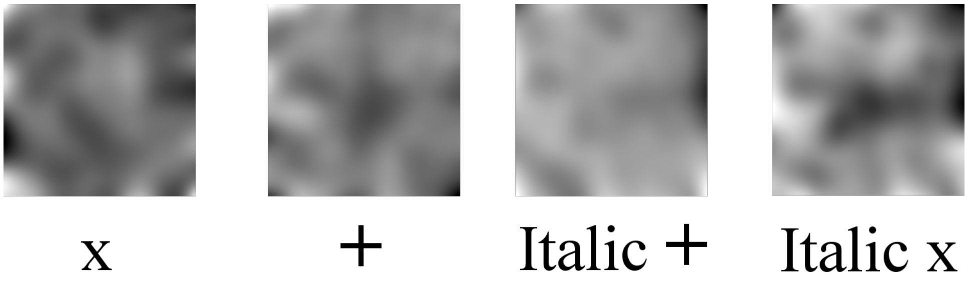

Figure 7: Classification images for each target assigned to participants. These CIs are generated by aggregating across the participants that were assigned to each target. The label under each image denotes which target participants were assigned.

Figure 7 shows the CIs generated for each target in Experiment 2. These images were generated by adding together all of the trial images that were selected as containing a target, and then subtracting all of the trial images that were not selected as containing a target. Visually, these CIs have much less visual similarity to the targets that they were trying to detect.

Target CIs were correlated with each of the 4 possible targets that were assigned to participants. The goal of this procedure was to assess the ability of the CIs to identify the target that participants were assigned. These correlations were grouped into 3 groups. First, the correlation between the target CI and the image of the correct target were put into a "Correct Target" group. Next, the correlation between the CI and the target with high similarity (X and italic X, or + and italic +) were put into a "High Similarity" group. Finally, the remaining correlations between the CI and the two low similarity targets were put into the "Low Similarity" group. Overall, the "Correct Target" group had an average correlation of *r* = .15 (*SD* = .22), the "High Similarity" group had an average correlation of *r* = .12 (*SD* = .19), and the "Low Similarity" group had an average correlation of *r* = .15, (*SD* = .15).

Evaluation of reverse correlation sensitivity

Using the participant's classification images, the same technique as described in chapter 2 was applied to try to predict participants' response to each trial based on their CI. An image-wise correlation was calculated between each trial image shown to participants and that participants' classification image. This image-wise correlation serves as an index of how similar each trial image is to the participants' CI. This approach yielded nearly 32,000 data points (8 participants x 4000 trials, with some participants missing a few dozen responses due to time constraints during data collection). In order to convert the binary yes/no responses into a continuous proportion of targets identified, trials were grouped into equally sized bins based upon similar image-wise correlation values. For each bin, a proportion of 'target-present' responses was calculated. A few bin numbers, 50, 400, and 1000 were calculated with the intention of understanding the effect of bin number on the observed relationship.

Figure 8 shows a very strong relationship between the proportion of 'yes' responses by participants and the image-wise correlation between each trial image and the participants' respective CIs, grouped into 400 bins. Due to the sigmoidal nature of the curve in chapter 2, a sigmoid was fit to the data. The resulting sigmoid had a slope of 17.58 (SE=0.2), and a center constant of 0.008 (SE = 0.0005). The relationship was equally reliable across bin size, though the magnitude of the observed effect decreased slightly as the number of bins increased. The relationship's reliability likely stays consistent due to the trade-off between sample-size and effect size. A summary of the relationship observed at the various bin sizes is shown in Table 3.

Table 3: How the number of bins effects the strength of the relationship between the IWCs for participants' CI and their trial images.

Number of Bins	Observed Correlation	Statistical Significance
50	$r(48)= .997$	$p <.001$
400	$r(398) = .98$	$p <.001$
1000	$r(998) = .96$	$p <.001$

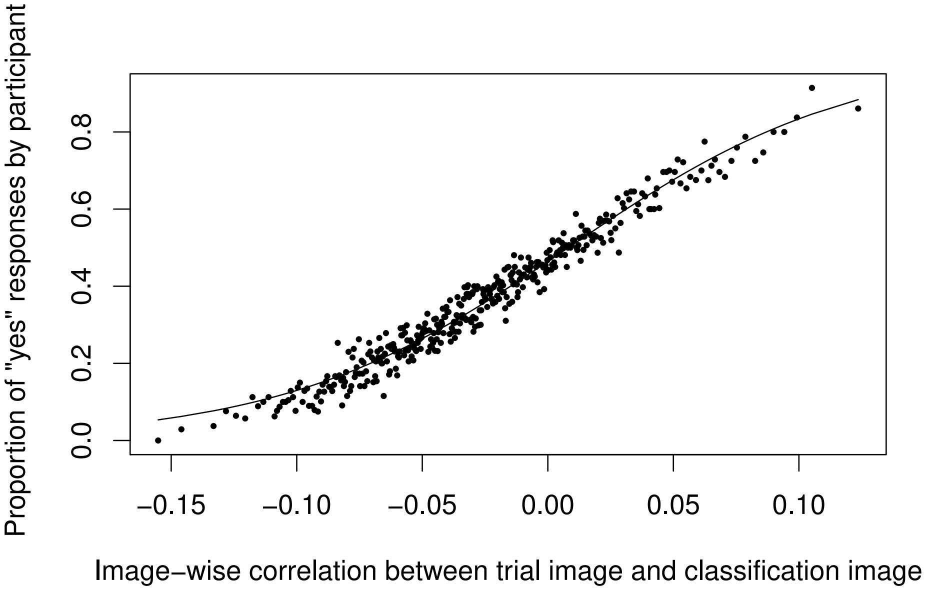

Figure 8: Scatter plot showing the relationship between the IWC for each trial image and the participants' CI with 400 bins. Solid line shows the sigmoid function that was fit to the data. The slope for the sigmoid was 17.6 (SE=0.2), with a center constant of 0.008 (SE = 0.0005).

Evaluating image selection using forward correlation

Using the same technique as described in chapter 2, an image-wise correlation was calculated between each trial image shown to participants and the target that they were assigned to identify (x, +, italic x, or italic +). This image-wise correlation serves as an index of how similar each trial image is to the participants' assigned target. In order to convert the binary responses into a continuous proportion, trials were binned into 400 bins based upon similar image-wise correlation values. For each bin, a proportion of 'target-present' responses was calculated.

Figure 9 shows a weak relationship between the proportion of 'target-present' responses in each bin, with the mean image-wise correlation for each bin. The positive linear relationship between proportion of 'target-present' responses and mean image-wise correlation in each bin for 400 bins was significant, $r(398) = .10$, $p =$

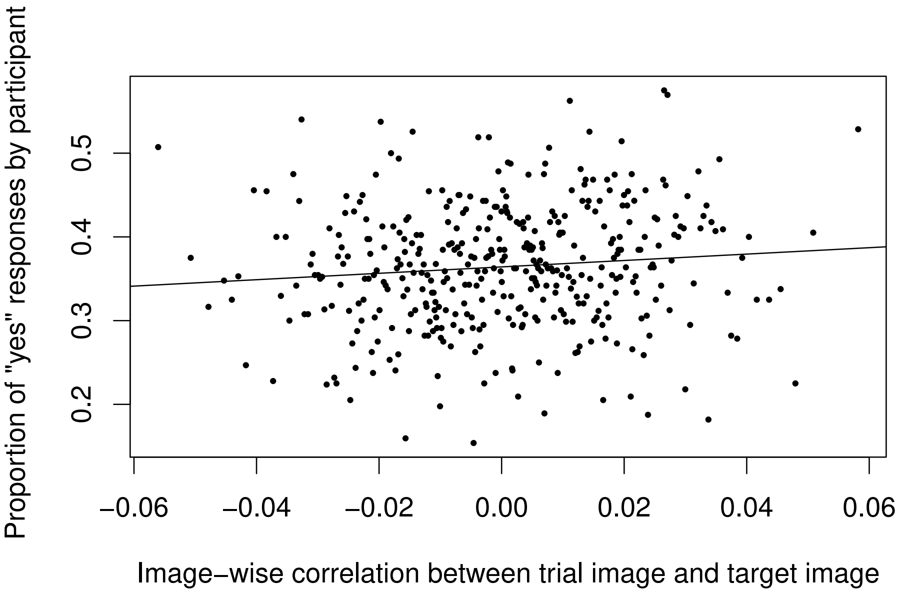

Figure 9: Scatter plot showing the relationship between the IWC for each trial image and the target participants were assigned. Data is aggregated into 400 bins.

.042, 95% CI = [.003, .198]. Again, as in chapter 2, the chosen number of bins was somewhat arbitrary. A range of bin sizes were chosen to map the effect of bin size on the magnitude of the relationship, with 50 bins and 1000 bins being calculated in addition to the 400 bins mentioned above. The effect appeared to be similar to that observed in chapter 2, with the magnitude of the relationship increasing as we decreased the number of bins. However, the reliability remains approximately equal. The relationship, in this case, seems to be only somewhat reliable as it hovers around the decision boundary for statistical significance. In addition, it is much smaller than the relationship observed in chapter 2. A summary of the observed relationship for the different numbers of bins can be found in Table 4.

Table 4: How the number of bins effects the strength of the relationship between the IWCs for participants' assigned target and their trial images.

Number of Bins	Observed Correlation	Statistical Significance
50	$r(47) = .25$	$p = .080$
400	$r(397) = .10$	$p = .042$
1000	$r(997) = .06$	$p = .045$

3.3 Discussion

The changes to the experimental procedure in Experiment 2 were intended to force participants to make decisions based upon their first impressions. This was achieved by limiting the time for which each trial image was presented on the screen. Specifically, participants saw each image for 500ms before the image disappeared, and participants were asked to make a decision as to the presence of a target. This manipulation seems to have decreased deliberation as the mean response time per trial decreased considerably between chapter 2 (at 1661ms) and this study (at 503ms). Additionally, the standard deviation of response times decreased as well, from 1022ms in chapter 2 to 197ms in this experiment. This supports the hypothesis that a variety of strategies were employed in experiment 1, whereas here we have successfully encouraged participants to use a specific strategy — namely one of intuition.

The observed differences in sensitivity between the two target difficulty sessions of the practice phase suggest that the difficulty between the two sessions was likely to have been more effective at drawing participants to see superstitiously, as observed by Rieth et al. (2011). However, there seems to have been only a negligible increase in rates of superstitious perception, with the false alarm rates in the test phase (36.4%) being comparable to that observed in chapter 2 (32.8%). This suggests that such a method is not necessary for inducing superstitious perceptions in participants.

There was very little signal detected in the target CIs, shown in Figure 7. The correlations between these images and the target images confirmed this, with negligible difference between the average IWC for “correct targets” and for “low similarity”. This suggests that the task lost its utility in detecting participants’ internal templates when they were asked to make fast decisions about the presence of the target in the image. One CI, however, seems to bear some similarity with its associated target image: the ’+’ target CI. This could be due to the possibility that the manipulation did not universally have the effect of increasing passive engagement in searching for evidence of a target. Nonetheless, as a group, this study showed a marked loss in the ability to measure participants’ mental representations of the target compared to chapter 2.

This decrease in performance is corroborated by the noticeable decrease in slope observed in the backward prediction of participants’ responses based on the IWCs between their CIs and each trial image. The slope, 17.58, was much shallower than that observed in chapter 2, at 93.76. This suggests that participants were less consistent with their responses, and had a less rigid decision boundary between deciding whether an image contains a target or not. We would expect that these participants would require far more trials in order for a template of their internal representation would become visible, as their internal consistency in responses has decreased.

When evaluating participants’ performance using the forward prediction, a similar picture becomes evident. Participants’ responses were predicted less accurately using their assigned target image in this study than in chapter 2. In the previous study, participants’ responses could be predicted with moderate accuracy, $r(397) = .34$, when the 400 bins were used. However, in this study, we see that this accuracy has dropped to a very low level of accuracy, $r(397) = .10$, suggesting that their responses were guided less by the target that was assigned to participants. On the other hand, participants’ decision criteria in the easy section of the practice phase remained largely unchanged, $\ln(\beta) = -0.77$, when compared to chapter 2, $\ln(\beta) = -0.71$. These results could be explained by an increase in the internal noise of participants’ decision making (Neri & Levi, 2006). In other words, by reducing participants’ exposure time to the trial images, their information about each image is decreased and the amount of uncertainty that participants experience is increased.

This would lead to a noisier decision process despite the criterion for the decision remaining unchanged. In chapter 4, we will address this question directly.

Chapter 4: Experiment 3

The goal of experiment 3 was to build upon the assessment of task strategy on superstitious perception performed in experiment 2. One weakness in experiment 2 was the lack of direct manipulation of task strategy, opening the possibility for confounds. To strengthen our design, and build a better understanding of the effect of task strategy on the accuracy of generated classification images estimating internal representations, a direct manipulation of strategy was necessary. In this experiment, a direct manipulation was employed by comparing and contrasting superstitious perception performance between two groups who were assigned different task instructions, wherein differing task strategies were suggested.

To perform a direct manipulation of task strategy, we borrowed from a previous manipulation performed for the same purpose (Smilek et al., 2006). This experimental manipulation had two groups: (1) In the passive group, participants were asked to follow their intuition — intended to minimize the engagement of executive functions when performing the task; (2) In the active group, participants were asked to critically examine each trial in the task — intended to maximize the engagement of their executive functions during the task. Thus, at a theoretical level, this experiment was examining the effect of executive function engagement on the efficacy of the superstitious perception task in estimating internal representations.

4.1 Methods

4.1.1 Participants

14 participants (nine females) between the ages of 18 and 66 (mean age = 26.57, SD = 12.02) were recruited from the paid subjects pool at the University of British Columbia. Only participants with normal or corrected-to-normal vision were recruited. This was verified by self report upon participants' arrival in the lab. All participants were tested on at a time on the same computer, in the same room.

4.1.2 Stimuli and procedures

Stimuli were generated in the same manner as described in chapter 2. Procedures were also the same, except an additional manipulation was added, and the number of targets was reduced from 4 to 2, using only the standard "x" and "+" from chapters 2 and 3. Specifically, in order to more precisely manipulate the task strategy employed by participants, the participants were assigned into either an Active task group, or a Passive task group. The task was identical for both groups except for a difference in instructions at the beginning of the experiment.

The instructions for the two groups were adapted from Smilek et al. (2006). They used a similar set of instructions to influence search strategy in a visual search task. The only changes that were made to the instructions were changing the context to be relevant to the superstitious perception task, and not a visual search task. In the *Active* group, participants were told:

> *The best strategy for this task, and the one that we want you to use in this study, is to be as active as possible and to "search" for the target in the image as you look at the screen. The idea is to deliberately direct your attention to determine your response. Sometimes people find it difficult or strange to "direct their attention" but we would like you to try your best. Try to respond as quickly and accurately as you*

> *can while using this strategy. Remember, it is very critical for this experiment that you actively search for the target in the image. If you cannot find any reason to suspect that the target is present, respond that the target is not present.*

In the *Passive* group, participants were told:

> *The best strategy for this task, and the one that we want you to use in this study, is to be as receptive as possible and see if the target pops into your mind as you look at the image. The idea is to let the display and your intuition determine your response. Sometimes people find it difficult or strange to tune into their gut feelings but we would like you to try your best. Try to respond as quickly and accurately as you can while using this strategy. Remember, it is very critical for this experiment that you allow the target to just pop into your mind. If this does not happen, respond that the target is not present.*

Participants from both groups completed identical tasks, and differed only in their task instructions.

4.1.3 Data analysis

Data analysis procedures were identical those performed in chapter 2 and chapter 3, but were adapted slightly to accommodate the two instruction groups. Namely, all analyses were performed on the two groups separately, and then the results statistically compared.

Analysis of the practice phase was performed using a mixed effects ANOVA, with instruction group as a between-subject factor with two levels (Active and Passive), and target difficulty as a within-subject factor with two levels (easy and hard). One mixed effects ANOVA was used to compare sensitivity (d’) across factors, and a second was used to compare decision bias, ln(β).

The analysis of the test phase was identical to those of chapter 2 and chapter 3, but was split across instruction group, so the passive condition was analyzed separately from the active condition. When comparisons between groups was necessary, an appropriate statistical test was employed. When comparing the mean proportion of false alarms across groups, a t-test was performed. When comparing the slopes of the two sigmoid curves in the backward prediction analysis, a z-test was performed to test the statistical significance of the difference between the two slopes. Finally, a Fischer's Z-test was used to compare the strength of the relationships between target image IWCs and proportion of false alarms for the two groups.

4.2 Results

4.2.1 Practice phase

Figure 10A shows that there was no interaction between condition and target difficulty for sensivity (d'), $F(1,22) = 0.58$, $p = .46$. Nor was there an effect of condition on sensitivity, $F(1,22) = 1.52$, $p = .23$. There was, however, an effect of target difficulty, with participants being more sensitive to targets in the easy condition than in the hard condition, $F(1,22) = 16.90$, $p < .001$. Similarly, Figure 10B shows that there was no interaction between condition and target difficulty in decision bias ($\ln\beta$), $F(1,22) = 0.05$, $p = .82$. There was no main effect of condition on decision bias, $F(1,22) = 0.68$, $p = .42$, nor was there a main effect of target difficulty, $F(1,22) = 2.23$, $p = .15$.

4.2.2 Test phase

Figure 11 shows that the difference in proportion of false alarms between conditions was trending towards significance, with the active condition (mean P_{yes} = 44.7%, SD = 28.8%) reporting more false alarms than the passive condition (mean P_{yes} = 20.3%, SD = 11.7%), $t(9.74) = 2.17$, $p = .056$. In addition, participants in

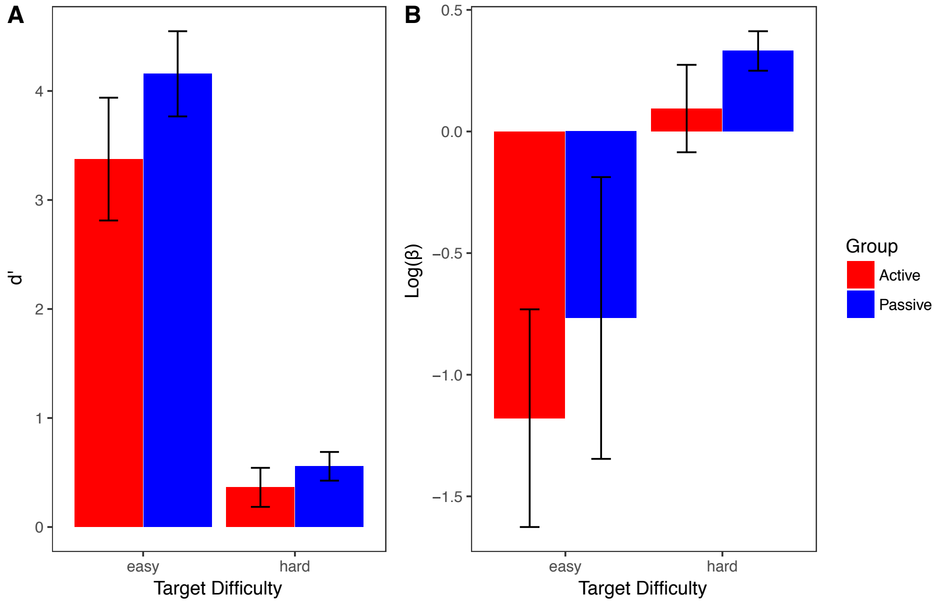

Figure 10: A summary of the practice phase of the experiment. A) Mean sensitivity (d') for both the easy and hard target difficulties, across groups. B) Mean decision bias (lnβ) for both the easy and hard target difficulties across groups. Error bars represent the standard error of the mean.

the two conditions responded about equally quickly, with participants in the active condition (mean RT = 1438ms, SD = 942ms) being non-significantly slower than participants in the passive condition (mean RT = 855ms, SD = 364ms) in response, $t(9.54) = 1.60$, $p = 0.14$.

Figure 12 shows that there were much clearer patterns in the CIs from the passive condition than in those from the active condition. These CIs were generated the same way as they were in experiments 1 and 2, by adding together all of the images that were false alarms, and then subtracting all of the images that were correction rejections. Some linear transformations were applied to each image to ensure that the minimum pixel luminance in the image was 0, and the maximum was 1.

In order to estimate how well the generated CIs were able to discriminate be-

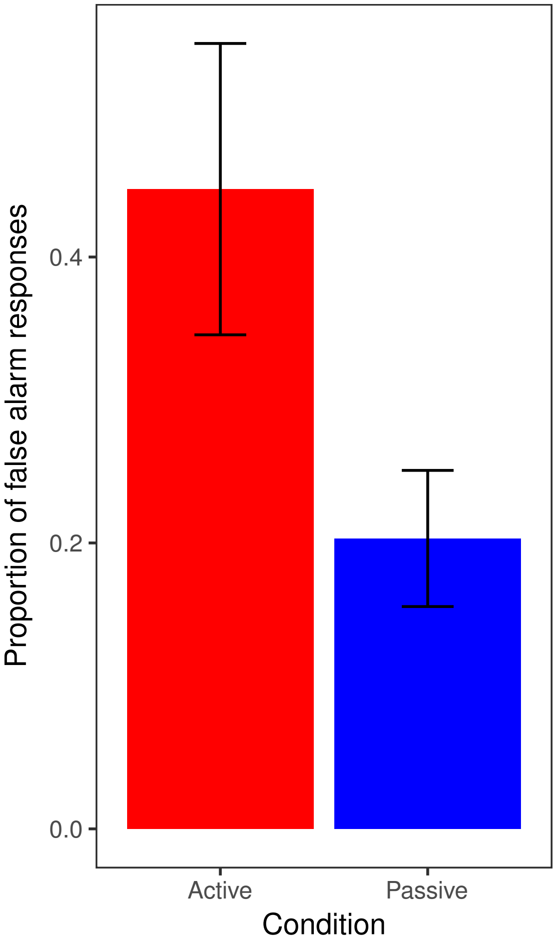

Figure 11: The proportion of false alarms in the test phase, across groups. Error bars represent the standard error of the mean.

tween their respective target and other potential targets, an IWC was calculated between the CI and each of the 4 possible targets from experiments 1 & 2 (x, italic x, plus, italic plus). Discriminability ratings were averaged across assigned targets for each condition to estimate how well the overall condition was able to generate discriminating CIs. In both conditions, discriminability was very low. In the active condition, CIs had an average similarity score of r = 0.003 with their respective targets, a similarity score of r = - 0.001 for the high similarity target, and r = -.002

for the low similarity score. In the passive condition, CIs had an average similarity score of r = 0.045 for their respective targets, a score of r = 0.040 for high similarity targets, and r = 0.027 for low similarity targets. The CIs from the passive condition, while having a low degree of similarity with any targets, were able to discriminate more effectively between the possible targets than were CIs from the active condition.

Evaluation of reverse correlation sensitivity

An IWC was calculated between each participant's CI and each of their trial images. The IWC was interpreted as being a measure of the similarity between each trial image and the CI. Larger, positive IWCs indicate higher similarity between the participant's CI and a trial image. Negative IWCs indicate lower similarity between the participant's CI and the trial image. These IWCs were used to predict the likelihood of a participant identifying a target in a trial image.

Figure 13 shows that the sigmoidal relationship between the IWC for a participants' CI and each trial image, and the proportion of false alarms for images with that IWC value, grouped into 400 bins, is steeper in the active condition than in the passive condition. This was verified by fitting a sigmoid curve to each condition, showing that the active condition data, slope = 48.7 (SE = 0.7), did indeed have a steeper slope than the passive condition, slope = 36.7 (SE = 0.6). A test of significance of the difference between two independently observed slopes was conducted as suggested by Cohen, Cohen, West, and Aiken (2003), $z = 13.22$, $p < .001$. However, the relationship between bin size and the strength of the relationship observed in chapter 2 and chapter 3 was still present in both conditions, with a very strong linear relationship between the proportion of false alarms in each bin, and sigmoid transformed IWC values in both the passive and active conditions. These values are shown in Table 5.

The aforementioned results and relationships were also tested with 50 and 1000 bins to examine how they are affected by bin size. The observed slope for the two conditions remained the same for all 3 bins, suggesting that bin size does not

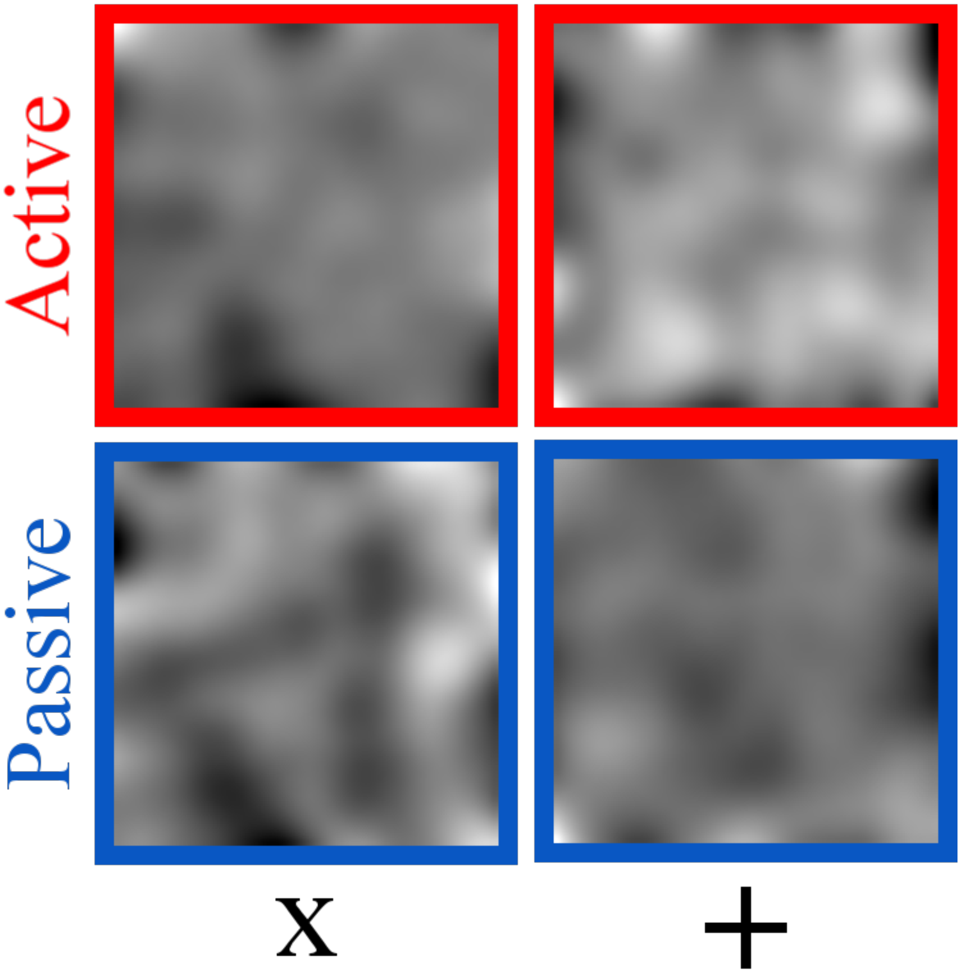

Figure 12: The classification images generated across participants for each target, in each condition. Blue shows the CIs generated for the passive condition, red shows the CIs generated for the active condition.

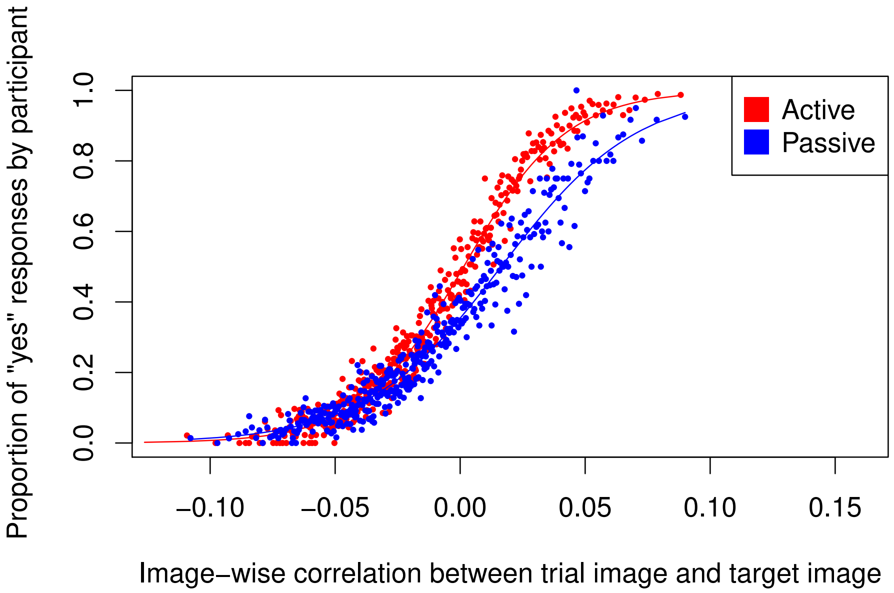

Figure 13: The relationship between IWC for participants' CI and each individual trial image, and the proportion of images in each bin identified as containing a target. Data from the active condition are plotted in red, and data from the passive condition are plotted in blue. All data were aggregated into 400 equally sized bins for each condition. Lines show the sigmoid functions that were fit to the data for each condition.

affect these values. In addition, the relationship between the proportion of false alarms and the sigmoid-transformed mean IWC for each bin decreased slightly as the number of bins increased. The decrease in the strength of the relationship is likely due to less aggregation of the data as the number of bins increases, and thus a weaker signal-to-noise ratio is obtained. Nonetheless, the reliability of these relationships remained relatively constant, likely due to an increase in degrees of freedom as the bin number increases, that strongly suggests that the relationship is present. These results are shown in Table 5

Table 5: How the number of bins effects the strength of the relationship between the IWCs for participants' assigned target and their trial images.

Condition	Number of Bins	Observed Correlation	Statistical Significance
Active	50	$r(48) = .998$	$p < .001$
	400	$r(398) = .0.99$	$p < .001$
	1000	$r(998) = .97$	$p < .001$
Passive	50	$r(48) = .997$	$p < .001$
	400	$r(398) = .98$	$p < .001$
	1000	$r(998) = .94$	$p < .001$

Evaluating image selection using forward correlation

A second set of IWCs were calculated between the target assigned to each participant, and their respective trial images. As was the case for the IWCs involving the CIs, these can be interpreted as the similarity between a given trial image and the target assigned to the participant. As in chapter 2 and chapter 3, the IWC scores can be used to predict the likelihood with which participants will identify a target in a given trial image.

Figure 14 shows the relationship between the IWC with participants' target images, and the proportion of false alarms in each of 400 equally sized bins of IWC observations. In the Passive condition, there was a moderate relationship between the observed IWC and the proportion of false alarms, $r(397) = .25$, $p < .001$. However, the active condition shows no relationship between the IWC with participants' target images and the proportion of false alarms, $r(397) = .01$, $p = .87$. These relationships were significantly different from one another, $z = 3.45$, $p < .001$, but the difference is in the opposite direction than expected. To verify that this difference is not a result of the number of bins, the two correlations were re-calculated for 50 and 1000 bins. The results were similar, showing the same trends as observed in the previous experiments, but with reliability remaining consistent. A summary of these results can be found in Table 6.

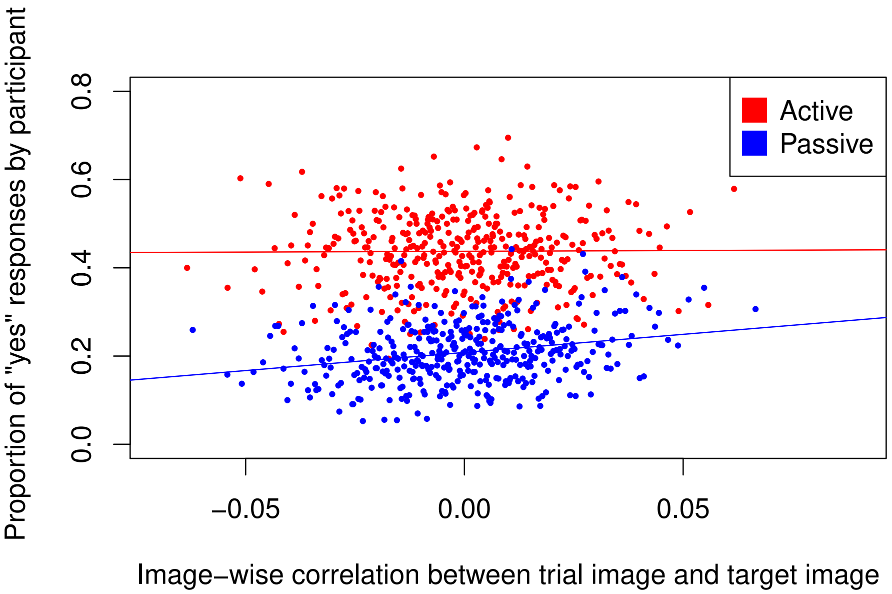

Figure 14: The relationship between the IWC for participants' assigned target and each individual trial image, and the proportion of images in each bin identified as containing a target. Data for the active condition are shown in red, and data for the passive condition are shown in blue. All data were aggregated into 400 bins of equal size per condition.

Table 6: How the number of bins effects the strength of the relationship between the IWCs for participants' assigned target and their trial images.

Condition	Number of Bins	Observed Correlation	Statistical Significance
Active	50	$r(47) = .04$	$p = .76$
	400	$r(397) = .01$	$p = .87$
	1000	$r(997) = .00$	$p = .95$
Passive	50	$r(47) = .57$	$p < .001$
	400	$r(397) = .25$	$p < .001$
	1000	$r(997) = .15$	$p < .001$

4.3 Discussion

The goal of experiment 3 was to directly compare the accuracy of the classification images with respect to estimating participants' internal templates across task strategies. To do so, we used task instructions similar to those of Smilek et al. (2006) to directly compare two task strategies using the techniques developed in chapter 2. The two task strategies included a passive strategy, which is presumed to reduce the recruitment of executive functions to the task, and an active strategy, presumed to increase the recruitment of executive functions to the task. Investigations of this nature in other domains has suggested that task strategy can have a notable impact on task outcome (Smilek et al., 2006). Furthermore, indirect comparisons between Experiments 1 and 2 suggested that if participants adopt a passive task strategy, classification images would be less accurate. However, the work of Smilek et al. (2006) suggested the opposite outcome; Adopting a passive task strategy yields better outcomes in visual search.

4.3.1 Internal prediction

As observed in chapters 2 and 3, the relationship between the proportion of false alarms and trial image similarity to the classification image was sigmoid in nature in both conditions. Moreover, the slope of the sigmoid curve in the active condition was significantly steeper than that in the passive condition, a result that is in agreement with the results of chapter 3, where we induced a passive task strategy by reducing stimulus presentation time. chapter 3 also yielded a shallower slope in the sigmoid function in comparison to the unfettered responses of participants in chapter 2. This suggests that participants were less consistent in their responses in the passive condition, meaning that there was a more distributed decision boundary between trial images that were target-like, and those that were not. Alternatively, this could be interpreted as participants in the passive condition being less confident in their responses, yielding a noisier decision process than in the active condition.

4.3.2 Classification images

The classification images, shown in Figure 12, showed very little resemblance to their respective targets in either condition, with the highest IWC of the 4 CIs being r = 0.049. Qualitatively, there is very little signal in either set of CIs, but those in the passive condition seem to have more detail than do those from the active condition. However, the analysis of the discriminative ability of the generated CIs, compared across the two conditions suggests that the passive condition yielded CIs that were better able to discriminate between the correct target and other targets of high or low similarity. This suggests that when using the superstitious perception task to estimate internal representations of a specific nature, the passive task strategy is a more reliable method to employ.

4.3.3 External prediction

Figure 14 shows a very large difference in predictive ability of human responses between the active and passive conditions in experiment 3. In the active condition, the non-significant correlation between target-image similarity and false alarm rate suggests that the participants' target images did not describe the signal participants were detecting. This could be caused by participants in the active condition becoming focused on small subsets of the trial image that appear target-like, while disregarding the rest of the image. This may indicate that while participants may focus on one subset of a trial image (e.g., participants focus on the top half of the image which may be target-like while the bottom half of the image, disregarded by participants, may be particularly *un*-target-like). Since our image-wise correlation procedure considers the image as a whole, and not subsets thereof, the image described would appear to have no similarity to the target since the two subsets of the images would be collapsed together for analysis. In contrast, the passive condition shows that as image similarity increases, false alarm rate increases, a result that was comparable in direction and magnitude to chapter 3. This suggests that participants considered the image as a whole when making predictions in the passive condition.

With respect to the practice phase of the experiment, no significant difference was found in sensitivity or decision bias between conditions. This is as expected given the above explanation of results, since target trials in the practice condition genuinely contained a target hidden in noise, and thus all subsets of the image were likely to contain some portion of signal. As a result, whether one looked at the trial image as a whole, or focused on a particular subset, one was likely to come to the same conclusion regarding the target's presence. However, future work could replicate this finding with increased power for between-subject analyses to increase confidence in the null results observed in this experiment.

Notably, the false alarm rate differed significantly between the passive and active conditions, with participants reporting more false alarms in the active condition than in the passive condition. An alternative explanation of the results observed in experiment 3 is that the differences in false alarm reports could have driven the rest of the experimental results. Specifically, that by increasing the false alarm rate, participants' responses became more difficult to predict, and the generated CIs lost their discriminability. The superstitious perception task is based upon the isolation of minute traces of signal that occur spontaneously within randomly generated noise images. It is conceivable that by changing participants decision biases to become more liberal, the underlying statistical processes upon which the task relies became less powerful and thus were unable to isolate the signal participants were detecting within the noise. Experiment 4 addresses this issue through the use of computer simulations in an attempt to show that the active-passive differences are not simply a result of shifting decision biases.

Chapter 5: Experiment 4

In this experiment, we used a neural network that was trained to identify objects in images, and modified it such that it could identify targets within white noise. Given the body of literature that suggests a similarity in representation between HCNNs and the human visual system, we expected that the low-level representations of the targets in the network would approximate those in the human visual system (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016). If these low-level representations in the HCNN truly did approximate those of the human visual system, then they would be sufficient to perform tasks tangential to that for which the network was trained.

To this end, we attempted to induce superstitious perception in the AlexNet HCNN (Krizhevsky et al., 2012). However, since the AlexNet network was trained to classify objects in images contained within the ImageNet database, some secondary training was required in order to teach the network to identify targets in noise. Specifically, the network was trained to identify an "X" in white noise. In order to test the degree to which the low-level representations are generalizable, the low-level layers (layers 1-5) of the AlexNet HCNN were left unchanged, meaning that the weights connecting nodes were not affected during learning. Only two fully connected layers (layers 6 & 7) were trained to perform the new task.

The goals of experiment 4 were two-fold. First, we intended to test the neural network's ability to perceive superstitiously in the manner that humans do without direct training on the task based upon human responses. If the neural network displays a similar pattern of results to the human participants, it would suggest that there is a similarity in their performance. Specifically, we would expect its responses to be comparable to humans who are not actively engaging their executive functions on the task (i.e. the ‘passive’ condition in chapter 4), given the beliefs that the neural network is similar to humans only in visual processing.

Second, if the neural network could see superstitiously, we intended to determine whether the loss of CI quality in the liberal condition in experiment 3 was simply due to increased proportion of false alarms. If this were the case, then selecting a liberal criterion such that 44% of images that were most target-like according to the neural network - like in the liberal condition - would result in a CI of poor similarity to the target. On the other hand, if CI quality is equal between a liberal criterion, the analogue of the liberal condition in experiment 3, and a more conservative criterion, the analogue of the passive condition in experiment 3, then we could conclude that the decrease in CI quality in Experiment 3 was due to interference of executive decision making on superstitious perception.

5.1 Methods

5.1.1 Model description

The HCNN model used in this experiment was a modified, pre-trained form of AlexNet (Krizhevsky et al., 2012), an HCNN which attained state of the art image classification accuracy for the ImageNet benchmark in 2012. This model was chosen due to its relative popularity in comparing HCNNs to the human visual system (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). With the additional information given by the aforementioned comparisons, AlexNet was the model with the highest documented similarity with the human visual system, making it the ideal model to use when comparing the ability of neural networks and humans to perceive superstitiously.

Since the model was not originally trained to identify an $'x'$ in noise, the network needed to be modified in order to perform such a task. To do so, the weights of the nodes in the final 2 fully connected layers of AlexNet were randomly initiated and a new binary classifier replaced the original classification layer of the network. These new layers were trained to perform the novel detection task. The decision to retrain the final 2 fully connected layers was made due to these layers being the point in the network hierarchy in which spatial information is lost. All previous layers of the network were convolutional layers, meaning that spatial information in those previous layers is preserved from one layer to the next (Krizhevsky et al., 2012). Notably, however, the weights in the 5 convolutional layers were not modified from their trained state, as provided by its creators (Krizhevsky et al., 2012).

Training of the 3 randomly initiated layers of the network was performed using stimuli generated in the same fashion as the easy training images used in chapters 2, 3 and 4. These stimuli were generated either by reducing the contrast slightly between the target and the image background (target present trials) or by creating a consistent grey image, luminance matched to the target (target absent trials). All training images were then overlayed with gaussian random noise. Unlike previous chapters, the network was trained exclusively on images containing an ‘x’ as the target image, learning to distinguish between images that contained an ‘x’ hidden in noise from images that contained pure noise. The network was trained on 100,000 training images (50% of them containing a target) for 19 epochs, meaning that the network was shown each training image 19 times during training.

The number of epochs for training was chosen in a very specific manner. After each training epoch, the network was testing on novel “easy training” images and on “hard training” images from chapters 2, 3 and 4. From the networks’ responses, a sensitivity and bias score was calculated. Training continued until the sensitivity and bias at the end of the epoch were approximately matched in sensitivity and bias to the participants observed in chapter 4.

5.1.2 Stimuli and procedures

White noise trial images were generated in the same way as in chapters 2, 3 and 4. That is to say that each image was generated pixel by pixel, randomly sampling a luminance from a gaussian distribution centered at 0.5, which a standard deviation of 0.1 for each of the 2500 pixels. As an analogue to the superstitious perception task completed by humans, the network was fed 16,000 50px x 50px noise images centered in a 227px x 227px grey image exactly as done for the human participants. The size of the overall image, grey border included, was set to match the 227px x 227px input the AlexNet architecture was built to accept (Krizhevsky et al., 2012). The network's reported confidence in the presence of a target was recorded for each of the 16,000 images resulting in a 16,000 response vector.

Once the network confidence was recorded for each trial image, decision criteria were set at two different points. In a liberal condition, the decision criterion was set such that the false alarm rate during the test phase was matched with the mean false alarm rate observed in the active condition of chapter 4. In a conservative condition, the decision criterion was set such that the false alarm during the test phase was matched with the mean false alarm rate observed in the passive condition of chapter 4. Once again, due to the complete lack of any targets among the images presented to the network in the test phase, any positive identification of the presence of a target was a false alarm.

5.1.3 Data analysis

Data analysis procedures were identical to those performed in chapter 2 and chapter 3, but were adapted slightly to accommodate the two criterion settings. Namely, all analyses were performed on the two groups separately, and then the results statistically compared.

Analysis of the practice phase was performed using a mixed effects ANOVA, with criterion setting as a between-subject factor with two levels (liberal and conservative), and target difficulty as a within-subject factor with two levels (easy and hard). One mixed effects ANOVA was used to compare sensitivity (d') across factors, and a second was used to compare decision bias, $\ln(\beta)$.

The analysis of the test phase was identical to that of chapter 4. However, the liberal and conservative criterion groups were analyzed separately. When comparisons between groups was necessary, and appropriate statistical test was employed. When comparing the mean proportion of false alarms across groups, a t-test was performed. When comparing the slopes of the two sigmoid curves in the backward prediction analysis, a z-test was performed to test the statistical significance of the difference between the two slopes. Finally, a Fischer's Z-test was used to compare the strength of the relationships between target image IWCs and proportion of false alarms for the two groups.

5.2 Results

5.2.1 Practice phase - training and validation stimuli

Figure 15 shows a summary of the neural network's performance on trials generated in the same way as the easy and hard trials in Experiments 1-3. For this phase of the experiment, the network was making its own decisions. Specifically, if the network was more than 50% confident that a target was present, it would respond with "Target Present". As a result, the network's response to each trial (either "Target Present" or "no target present") was used to estimate the network's sensitivity to the 'x' target. Figure 15A shows a large difference in the network's sensitivity (d') to targets in the easy (d'= 6.34) and hard target (d'= 0.57) difficulty sessions. Figure 15B shows a large difference between the neural network and human participants. Namely, the neural networks shows little to no bias in its decisions, unlike in the human data observed in previous chapters, in both the easy ($\ln\beta$ = $-1.95 \times 10^{-}14$), and hard ($\ln\beta$ = $9.00 \times 10^{-}5$) target difficulty sessions.

5.2.2 Test phase

Figure 16 shows the distribution of confidence ratings output by the neural network for the presence of a target image in the white noise test trials. The confidence ratings could range from zero to one, with zero denoting a high degree of confidence

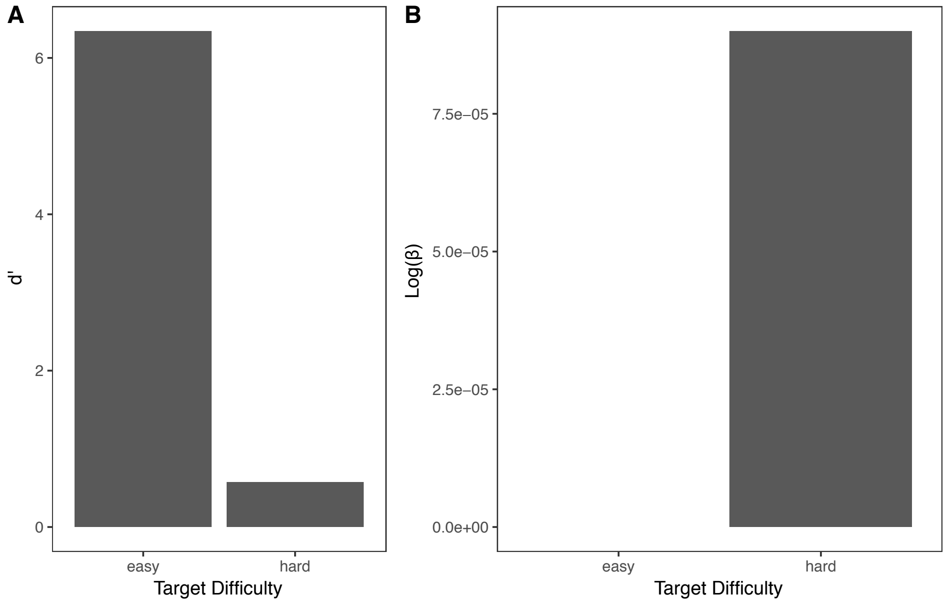

Figure 15: A summary of the practice phase of the experiment. A) Sensitivity (d') of the neural network over 50,000 trials of easy and hard target difficulties. B) Decision bias ($\ln\beta$) of the neural network over 50,000 trials of easy and hard target difficulties. These point estimates have no variability.

that a target is not present, and a one denoting a high degree of confidence that a target is present in the given trial. The standard approach to making decisions based on the neural network output is to consider a target to be present if the confidence is above 0.5, and then consider a target to be absent if the confidence rating is below 0.5. Based on these criteria, the network did not perceive superstitiously at all, with the network confidence ratings ranging from 0.0006, to 0.0117. However, by pushing the network to be more liberal in its decisions, namely by lowering the decision boundary to be a specific percentile point of observed confidence ratings, we can induce superstitious perceptions.

In order to mimic the observed proportions of false alarms in the liberal and conservative conditions of chapter 4, two separate decisions boundaries were calculated. One decision boundary was at the 55.3 percentile point, to mimic the

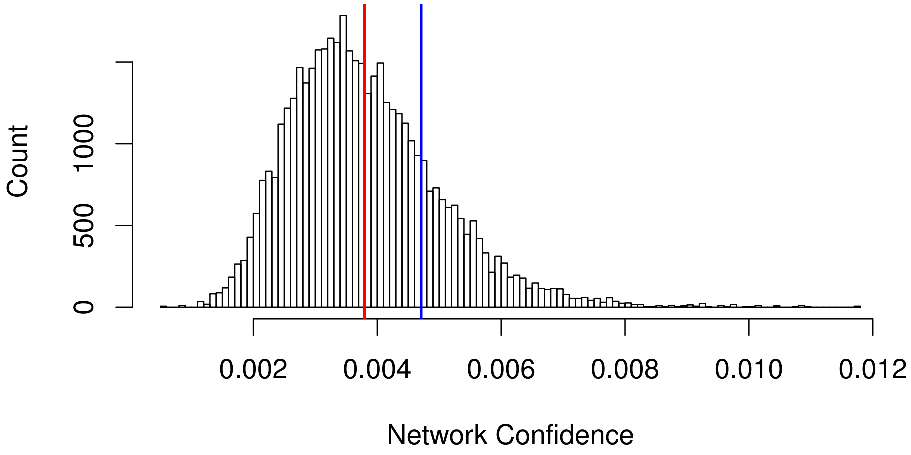

Figure 16: A histogram of the confidence ratings given by the neural network for the presence of a target image in a white noise trial images. The red vertical line denotes the decision cutoff for the liberal condition. The blue vertical line denotes the decision cutoff for the conservative condition.

44.7% false alarm rate observed in the liberal condition. The other boundary was at the 79.7 percentile point to mimic the 20.3% false alarm rate observed in the conservative condition. These decision boundaries are depicted in Figure 16 as vertical bars on the histogram. If the network's confidence was above the decision boundary, the trial was labeled as a false alarm. On the other hand, if the network's confidence was below the decision boundary, the trial was labeled as a correct rejection.

Figure 17 shows the CIs calculated from the responses generated by the two decision boundaries. As in the previous chapters, these CIs were generated by taking the sum of all the images for which there was a false alarm, and subtracting all of the images for which there was a correct rejection. The two CIs are very similar in form. The image-wise correlation between the two CIs was $r = 0.64$.

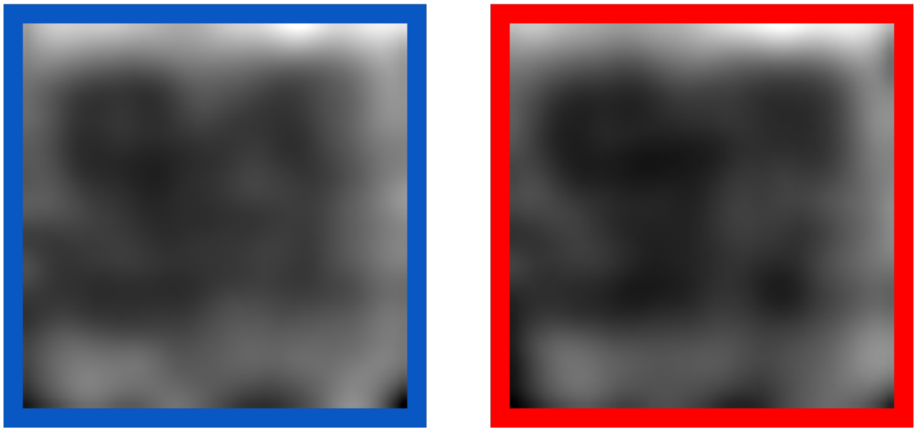

Figure 17: The classification images generated from the responses of the neural network that was searching for an X in each trial. Blue shows the CI generated for the conservative condition, where responses were generated from the conservative decision boundary. Red shows the CI generated for the liberal condition, where responses were generated from the active decision boundary.

An IWC was calculated between each CI and the target image in order to evaluate their similarity to the target. The CI generated from the conservative responses correlated with the target with a strength of $r = .55$, while the CI generated from the liberal responses correlated with the target with a strength of $r = .54$. To test the ability of the generated CIs to discriminate what the target of the network from a low similarity target, the IWCs were calculated between the conservative and liberal CIs and the "+" target. The IWCs were both markedly smaller than for the correct target, $r = .38$ and $r = .33$ respectively.

Evaluation of reverse correlation sensitivity using internal prediction

Figure 18 shows a sigmoidal relationship between the IWC generated for each trial image and the neural network's CIs, and the proportion of false alarms. The figure shows the relationship for both the conservative and liberal conditions. The proportion of false alarms was generated by binning responses into equal groups based on similar IWC scores between the CI and each trial image. A sigmoid function was fit to each set of data, resulting in a slope and center constant for both the conservative and liberal conditions. In the liberal condition, a relatively steep slope of 71.7, (*SE* = 1.9), and a center constant of 0.004, (*SE* = 0.0003). In the conservative condition, there was also a relatively steep slope of 74.2, (*SE* = 2.2), and a center constant of 0.022, (*SE* = 0.004). The two slope coefficients were not significantly different from one another, $z = 0.84$, $p = .40$. Furthermore, in order to test the goodness of fit of the two models, we calculated the correlation between the sigmoid transformed IWC scores and proportion of false alarms per bin. The goodness of fit was very high in both the liberal condition, $r(398) = .96$, $p < .001$, and in the conservative condition, $r(398) = .993$, $p < .001$. A range of bin sizes were calculated in order to estimate the effect of the number of bins on the magnitude of the of the observed goodness of fit. We performed the same analysis on 50, 400, and 1000 bins generated from 16000 trial images fed to the network; the results are summarized in Table 7.

Table 7: How the number of bins effects the strength of the relationship between the IWCs for the neural network's target and the test trial images.

Condition	Number of Bins	Observed Correlation	Statistical Significance
liberal	50	$r(48) = .995$	$p < .001$
	400	$r(398) = .96$	$p < .001$
	1000	$r(998) = .90$	$p < .001$
conservative	50	$r(48) = .993$	$p < .001$
	400	$r(398) = .94$	$p < .001$
	1000	$r(998) = .87$	$p < .001$

Evaluating image selection using external prediction

Figure 19 shows a linear relationship between IWC scores for each trial image with the target "x" image, and the proportion of false alarm responses by the neural network at both the conservative and liberal decision points. The proportion of false alarm scores was generated by taking the proportion of false alarms from equal bins

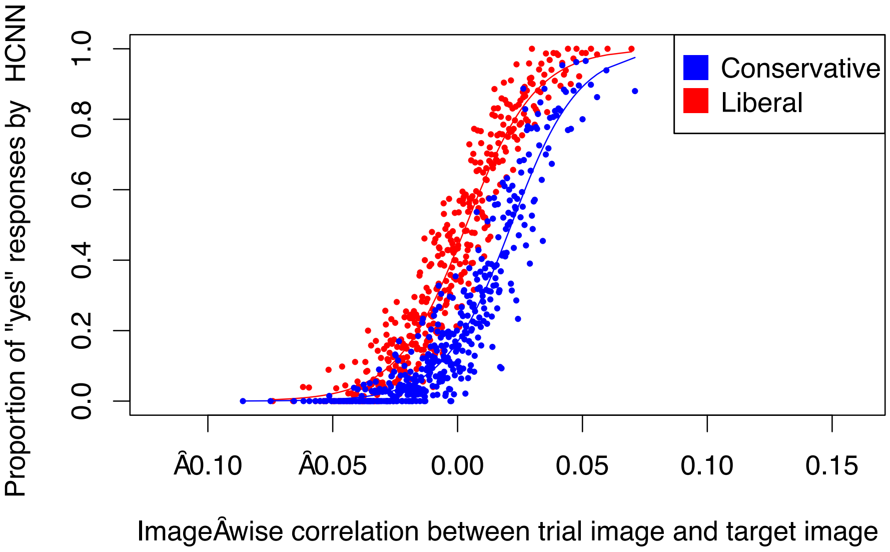

Figure 18: The relationship between IWC for the neural network's CI and each individual trial image, and the proportion of images in each bin identified as containing a target. Aggregated responses from the liberal condition are plotted in red, and aggregated responses from the conservative condition are plotted in blue. All data were aggregated into 400 equally sized bins for each condition. Lines show the sigmoid functions that were fit to the data for each condition.

of responses whose trial images had similar IWC scores with the target. Figure 19 shows the relationship as observed when the data are grouped into 400 equally sized bins. In the liberal condition, there was a relatively strong linear relationship between IWC scores and proportion of false alarms, $r(398) = 0.58$, $p < .001$. In the conservative condition, there was also a relatively strong relationship between IWC scores and the proportion of false alarms, $r(398) = 0.60$, $p < .001$. Most notably, there was no significant difference in the strength of the relationship between the conservative and liberal conditions, $z = 0.43$, $p = 0.67$. In order to verify that the number of bins had no effect on our conclusions, we tested the relationship at bin sizes of 50, 400, and 1000. A summary of the results for the different bin sizes are

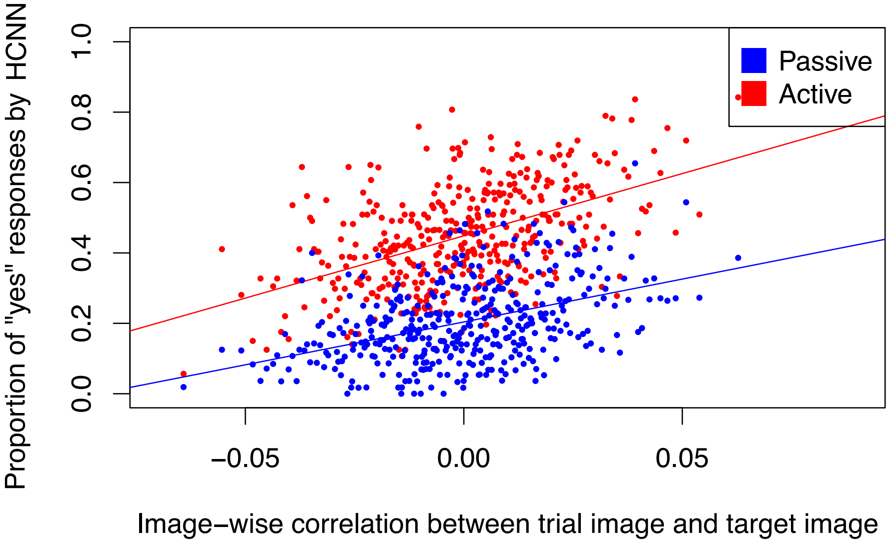

Figure 19: The relationship between the IWC for neural network's target X image and each individual trial image, and the proportion of images in each bin identified as containing a target. Aggregated responses from the liberal condition are shown in red, and aggregated responses from the conservative condition are shown in blue. All data were aggregated into 400 bins of equal size per condition.

found in Table 8 [1].

[1] In order to ensure the stability of the HCNN model, all model weights and biases were truncated to four decimal places as opposed to the unlimited number of significant digits allowed under normal CNN procedures. All analyses covered in this chapter were performed anew using the model with truncated weights. The changes had no influence on the various analyses, but did noticeably reduce the quality of the generated CIs. The shape of the target was still present in both the liberal and conservative conditions. Nonetheless, the CIs lacked the clarity observed in the CIs generated from the model without truncated weights. These observations demonstrate the sensitivity of the HCNN model to restricting the significant digits of the model weights. However, the observations do not change the overall conclusions of this work.

Table 8: How the number of bins effects the strength of the relationship between the IWCs for the neural network's target and the test trial images.

Condition	Number of Bins	Observed Correlation	Statistical Significance
liberal	50	$r(48) = .87$	$p < .001$
	400	$r(398) = .58$	$p < .001$
	1000	$r(998) = .42$	$p < .001$
conservative	50	$r(48) = .90$	$p < .001$
	400	$r(398) = .60$	$p < .001$
	1000	$r(998) = .41$	$p < .001$

5.3 Discussion

Experiment 4 had two goals, with success in the first goal being a requirement for pursuing the second one. The first goal was to compare the behaviour of the neural network's performance on the superstitious perception task with the performance of human observers in the conservative condition of Experiment 3. We chose to compare them on two specific measures: (1) the ability of the generated CI to discriminate between the assigned target, and other possible targets; (2) the strength of the internal and external predictions. If the neural network and human visual system represent visual information in a similar way, we would expect very accurate internal predictions, with a very large correlation between IWC for the CI and proportion of false alarms, and moderate accuracy in the external predictions, with a moderate correlation between IWC for the target image and proportion of false alarms. Furthermore, we would expect that the generated CI would be able to distinguish between the target, and low-similarity targets.

The second goal was to investigate the effect of task strategy on the reliability of the CI. In Experiment 3 we found that adopting a liberal cognitive strategy led to increased response consistency, but decreased external prediction ability of the generated CI. We compared the performance of the neural net under both conservative and liberal decision biasing conditions, to see if the results would mimic those found in human subjects.

To train an HCNN that can see superstitiously, we first needed to ensure that the system could identify the relevant target image at a level comparable to our human observers (i.e., essentially perfect identification and discrimination of noisy 'X's from pure noise). We began by using the 5 pre-trained convolutional layers of the AlexNet HCNN (Krizhevsky et al., 2012).

These 5 layers were then left unchanged for the duration of the training process. In order to allow the neural network to perform the superstitious perception task, its higher-level layers needed to be retrained to identify an 'x' within noise based on the output of the 5th layers of AlexNet. In other words, the final 2 layers of the network that were trained to perform the superstitious perception task had only the representations derived from the 5 pre-trained layers of the AlexNet to use as input. This meant that any success the network had in identifying the targets was due to the information present in the AlexNet convolutional layers.

During training, the modified AlexNet was taught to distinguish noisy images containing an X' from those containing pure noise. 100,000 images were used for training, of which 50% were targets containing an X', and 50% were non-targets containing only noise. Upon training completion, the HCNN was able to achieve a very high standard of performance, correctly labelling near all of the images that were presented to it, with a sensitivity that was near ceiling ($d' = 6.34$) in the easy target condition. Since the only part of the network that was learning to perform the task was the final 2 fully connected layers, one can conclude that sufficient information is preserved about the images in the first 5 layers of AlexNet to correctly identify targets in noise. This finding suggests that the HCNN had a sufficiently complex representation of the image presented to it within the first 5 layers that it can perform simple tasks without any retraining. It also suggests that, much like the human visual system, the low-level representations can be recruited to perform tasks that are tangential to that for which it was trained.

Classification images

The Classification Images generated from the responses of the HCNN both qualitatively and quantitatively approximate the target for which the network was searching. Figure 17 shows that both CIs appear to contain an “x”-like form in the center of the image. Furthermore, the two CIs have high IWC scores with the target image, and lower IWC scores with a low similarity target, the + image, suggesting that the two CIs generated from the neural network’s responses were of adequate quality to distinguish the network’s internal template of an “x” from a target of low similarity, the “+”. These data suggest that the neural network did indeed see superstitiously during the task. However, it is important to note that these classification images could only be generated after the reduction of the decision criterion to match the false alarm rates of participants.

Internal prediction

With regards to the internal prediction of neural network responses, the observed patterns were similar to those generated by human observers in Experiment 1. The shape of the relationship between IWC for trial images and CI, and the proportion of false alarms was sigmoid in nature. The sigmoidal nature of the relationship is likely an artifact of the restricted range of a proportion, and so the similar shapes here should not be used as definitive evidence that the neural network and humans responded similarly.

Nonetheless, it is worth noting that there was similarity in the strength of the net and human’s internal predictions. Furthermore, the response consistency, as measured by the slope of the sigmoid, was within the (admittedly large) range of observed consistencies in humans.

External prediction

There was a moderate-to-strong relationship between the trial-target IWC and the proportion of false alarms, as shown in Figure 19. Such a relationship suggests that the neural network’s responses were guided by the similarity of the trial images to the target. However, the strength of this relationship based on the responses of the neural network was considerably larger than that observed based on the responses of human participants, suggesting that the neural network's responses were governed more by their similarity to the target than were those of the human participants.

5.3.1 The effect of false alarm rate on CI quality

Overall, the neural network's performance on the superstitious perception task, as indicated by the CIs and measures of internal and external prediction, were similar to that observed in humans. Qualitatively, the CIs showed a resemblance to the target for which the neural network was looking. Quantitatively, the same CIs were able to distinguish the network's target from other targets with similar features. The HCNN responses exhibited similar patterns to those observed in human responses. Given these findings, we conclude that the network did indeed see superstitiously. Following this conclusion, we used the HCNN responses to meet our second goal of the experiment: to evaluate the degree to which the rate of false alarm reports throughout the task affects the end CI quality. In short, to manipulate the decision criterion of the neural net, through the increase or decrease of its false alarm rate, has a negligible effect on CI quality, and quantitative target similarity.

Looking at the strength of the CI-target IWCs for the liberal and conservative response conditions shows a negligible difference in image similarity between the two conditions. In addition, the two CIs showed a similar pattern in the degree to which they were able to distinguish the network's target from a low similarity target image. This suggests that the proportion of false alarms has little or no impact on CI quality.

The same can be said of the effect of manipulating the false alarm rate on the quantitative measures used to assess response patterns during the superstitious perception task. With respect to the internal prediction, the two conditions had similar slopes, suggesting that response consistency between the two conditions was comparable. Furthermore, the slopes found for the two conditions in the external prediction procedure were also very similar. Both of these findings are contrary to what would be expected if the difference in false alarm rate between the liberal and conservative conditions in chapter 4 was the key factor leading to the qualitative and quantitative differences in task performance for human observers.

Given the array of results in this experiment, we can conclude that the differences observed in chapter 4 between the liberal and conservative task strategies were *not* due to differences in false alarm rate. More specifically, that modifying neural network false alarm rate yielded no significant changes in the quantitative measures developed in chapter 2 to evaluate how well the generated CI estimates the system's internal representation. If the differences in CI estimate accuracy observed in chapter 4 across task strategy were not due to false alarm rate differences, one can conclude that task strategy influences the underlying mechanism that is being measured.

Chapter 6: General Discussion

In chapter 5, we compared the performance of Hierarchical Convolutional Neural Network (HCNN) with the performance of human observers in a superstitious perception task (chapters 2, 3 and 4). To our knowledge, past comparisons of the behaviour of the human visual system and that of HCNNs have been limited to conditions in which the neural networks were trained to respond based on image labels generated by humans, and then tested on the same task (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016).

In the present work, we took the comparison an important step further, because we first trained HCNNs to identify objects in noisy images, before testing them on images in which there was no objective signal. Any relation between the false alarms of the net and the images that yielded those false alarms were thus an indication of the internal template (i.e., the mental representation) that was learned by the net and used to make best guesses in noisy images.

With respect to the process that generates the behaviours, comparisons of the neural processing through which the human visual system and select HCNNs identify objects has revealed that they represent features of an object in similar ways (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins & DiCarlo, 2016). These comparisons were performed using representational dissimilarity analysis (Kriegeskorte et al., 2008), an analysis that compares the spatial relationships in neural activity in response to different stimuli. The logic underlying the technique states that the degree of difference in a neural system's representations between two stimuli can be indexed using the dissimilarity in neural activation in response to the two stimuli. Put differently, imagine an experiment in which a researcher was comparing the neural activation in a brain area in response to letters and houses. One set of neurons in the brain area that were highly active in response to houses became very inactive in response to letters, whereas another set followed the opposite pattern. Since the spatial correlation in activation between these two sets of neurons is very low, we would say that the activation is very dissimilar. From this, we can conclude that the brain area from which the recordings were measured represents the letters and houses very differently. If such an analysis is done comparing activity between a large set of different object classes, one can create a matrix of dissimilarity scores comparing each object class with each of the other object classes.

If one does such an analysis for two separate systems, such as the human visual system (using fMRI) and an HCNN, one can measure the degree to which the two systems share representations by correlating the two associated representational dissimilarity matrices. In short, Yamins and DiCarlo (2016), among others, showed a high degree of similarity between the representation of objects in the human visual system, and in an HCNN trained to identify these objects in images. This means that neurons in the human brain and the nodes of the neural network show similar activation patterns when exposed to the same set of varied stimuli of different object classes, and suggests that the two systems share representations of everyday objects.

Representational similarity between HCNNs and the human visual system has also been observed at different levels of visual complexity (Cichy et al., 2016). Early processing areas of the HCNN seem to share the highest degree of representational similarity with early processing areas of the human visual system, such as V1. Likewise, late processing areas of the HCNN seem to share the highest degree of representational similarity with late processing areas of the human visual system such as IT. In sum, not only do HCNNs show representation similarity with the human visual system at the global level, but representations seem to be similar from the lowest to the highest levels of processing.

To date, the HCNN has only been compared to human performance on tasks that the neural network was specifically trained to perform. Moreover, the neural network was trained based upon the labels created by humans. Unsurprisingly, the neural network showed human-like patterns in response to the task for which it was specifically trained. The novel aspect of the present study was that we compared neural network behaviour on a superstitious perception task — a task that is by definition subjective and therefore has no right answer — to the behavior of humans on the same task.

6.1 Superstitious perceptions in an HCNN and humans

In chapter 5, we used the newly developed techniques to compare a HCNN's behaviour to that of our human participants. The comparison between the HCNN and humans yielded a number of similarities, and some differences. First and foremost, the network, like the human participants, was able to perceive superstitiously. The low level representations in the network trained to identify objects in natural images contained sufficient information about novel images to detect very minute traces of signal in white noise. The flexibility in the task is similar to that of humans, who can use the visual system to perform a plethora of tasks, from negotiating objects in a room in order to sit in front of a computer, to completing an exercise such as that performed in these studies. We can conclude that a neural network trained to identify objects in naturalistic images can be co-opted to perceive superstitiously. Such a conclusions suggests that, to some degree, the procedures through which it identifies objects mimics the human visual system in more than just object recognition.

Beyond the HCNN's ability to perceive superstitiously, its response patterns during the superstitious perception task showed marked similarities to humans as well. Like humans, the relationship between the IWC for classification images and false alarm rate was sigmoid in nature. However, as mentioned earlier, the sigmoidal nature of the relationship may simply be due to the restricted range of the false alarm rate measure, which is a proportion and thus is limited in range from 0 to 1.

When comparing the external prediction capability between the human responses and HCNN responses, the HCNN responses are significantly better predicted by IWCs between the target and trial image than were the human responses. This suggests that the neural network's responses were based on trial images' likeness to the target, but to a greater extent than were the human responses. This difference is highlighted by the fact that the neural network will give identical confidence ratings when shown the same trial image due to the construction of the neural network, and the algorithm upon which it relies. This consistency is likely not present in humans, who are much noisier in their responses.

It was in getting the network to perceive superstitiously, however, that the most notable difference between the network and humans was observed in this experiment. The decision boundary of the network — the point of how target-like an image needs to be before being considered a target — is not flexible in a neural network as it is in humans. Instead, the decision boundary needed to be manually set post-hoc, otherwise the network would not have reported any targets among the noise with the default decision boundary used in most cases (Krizhevsky et al., 2012; LeCun et al., 1989). While the HCNN and human visual system may seem to be similar under certain circumstances, the aforementioned lack of flexibility in response criterion suggests that HCNNs as a model of the human visual system is incomplete.

Overall, the similarities displayed between humans and HCNNs in the superstitious perception task adds to the growing body of evidence pointing towards HCNNs as viable models for human vision. Most notable in the present research is the observed similarities in response patterns, despite the network never being exposed to human response data. All data on which the network was trained were generated algorithmically, so the network could not have learned patterns of human response from its training data. The two systems, HCNN and human visual system, appear to behave similarly on tasks in which the network was trained on human labels of images (Kheradpisheh et al., 2016; Peterson et al., 2016), but these similarities could originate from the networks simply learning to match inputs with human responses through trainning, and not by simulating the computations occurring in the human visual system. However, the present work, where the trained network was co-opted to perform a task in which it was never exposed to human data, suggests that the similarities between HCNN models and the human visual system may be more fundamental in nature.

One area worthy of further exploration is the necessity of manually setting the decision criterion in the HCNN in order for the network to perceive superstitiously. In humans, this seems to occur automatically, with participants reporting very few false alarms during the training phases, but reporting as many as 50% false alarms during the test phase, despite all images being pure noise. On the other hand, the neural network required a manual shift in criterion to match the false alarm rate observed in humans. Without the human data available, such a shift would be completely arbitrary. For this reason, the development of a dynamic decision criterion algorithm may be beneficial to understanding how this process occurs in humans. Specifically, humans seem to take into account high-level information, such as the expected frequency of targets during a task, in order to modify their decision criterion.

6.2 Implications for the study of superstitious perception

In chapters 2, 3 and 4, we developed a method to quantitatively evaluate a classification image generated from responses in the superstitious perception task (Gosselin & Schyns, 2003). Using a combination of internal and external prediction, we evaluated a number of key features of participants' performance on the task. First, we evaluated participants' response consistency by finding the slope of the sigmoidal relationship between the proportion of false alarms and the IWC score between participants' CI and each trial image using internal prediction. Second, we estimated how much the participants' internal representations, estimated by the CI, influences their responses using external prediction. These three measures together give an estimate of how well the CI captured participants' internal representations.

We also demonstrated that the accuracy of such classification images varies based upon the cognitive strategy employed during the task. Namely, we showed that when one engages executive functions to identify targets in the superstitious perception task, we see a marked decrease in CI similarity to one's target. On the other hand, one becomes more consistent in their responses relative to those who do not engage their executive functions. By employing a neural network to see superstitiously, we were able to show that simply changing the decision criterion does not account for the differences in sigmoid slope in internal prediction and external prediction accuracy observed between the passive and active conditions in chapter 4.

Comparing the results of chapter 2 to those of chapters 3 and 4, there is a notable difference across all analyses. Comparing the results of chapter 2 to those of the subsequent experiments, the CIs were much better able to discriminate the targets, the slope of the sigmoid was notably steeper, implying greater response consistency, and the false alarm rate was better explained by the target image. All of these results imply that participants performed better on the task when their task-strategy was not imposed by the experimenter. Such an implication is surprising given that, in chapter 3, and the passive condition of chapter 4, we directed participants to adopt the strategy that Gosselin and Schyns (2003) reported their participants to be using. However, results from our experiments suggest that task strategy has a profound impact on the conclusions that can be drawn from the results of this task. For this reason, we suggest that future experiments attempt to isolate the optimal task strategy for superstitious perception results. Before results generated from the superstitious perception task can be taken at face value, we must generate techniques and measures to ensure that participants adhere to the optimal strategy when performing the task.

References

Bottou, L. (1991). Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nmes*.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of compstat'2010* (pp. 177–186). Springer.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433–436.

Brown-Iannuzzi, J. L., Dotsch, R., Cooley, E., & Payne, B. K. (2017). The Relationship Between Mental Representations of Welfare Recipients and Attitudes Toward Welfare. *Psychological Science*, *28*(1), 92–103. doi: doi:10.1177/0956797616674999

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Deep Neural Networks predict Hierarchical Spatio-temporal Cortical Dynamics of Human Visual Object Recognition. *arXiv*, 15. doi: doi:10.1038/srep27755

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd. ed.). Mahwah, N.J.: L. Erlbaum Associates.

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1993). Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of temporal and spatial summation. *Journal of Neurophysiology*, *69*(4), 1118–1135.

Dyan, P., & Abbott, L. (2001). *Theoretical neuroscience. computational modeling of neural systems.* Cambridge, Mass.: MIT Press.

Eriksen, C. W. (1980). The use of a visual mask may seriously confound your experiment. *Attention, Perception, & Psychophysics*, *28*(1), 89–92.

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.

Fyfe, S., Williams, C., Mason, O. J., & Pickup, G. J. (2008). Apophenia, theory of mind and schizotypy: perceiving meaning and intentionality in randomness. *Cortex*, *44*(10), 1316–1325.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261–2271. doi: doi:10.1016/S0042-6989(01)00097-9

Gosselin, F., & Schyns, P. G. (2003). Superstitious Perceptions Reveal Properties of Internal Representations. *Psychological Science*, *14*(5), 505–509. doi: doi:10.1111/1467-9280.03452

Holden, H. M., Toner, C., Pirogovsky, E., Kirwan, C. B., & Gilbert, P. E. (2013). Visual object pattern separation varies in older adults. *Learning & memory*, *20*(7), 358–362.

Jacoby, L., & Brooks, L. (1984). Nonanalytic cognition: Memory, perception, and concept learning. *Psychology of learning and motivation*.

Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, *58*(6), 1187–211.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915. doi: `doi:10.1371/journal.pcbi.1003915`

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific reports*, *6*, 32672. doi: `doi:10.1038/srep32672`

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). Whats new in psychtoolbox-3. *Perception*, *36*(14), 1.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4. doi: `doi:10.3389/neuro.06.004.2008`

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25*.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

Lifshitz, M., Bonn, N. A., Fischer, A., Kashem, I. F., & Raz, A. (2013). Using suggestion to modulate automatic processes: From stroop to mcgurk and beyond. *Cortex*, *49*(2), 463–473.

Liu, J., Li, J., Feng, L., Li, L., Tian, J., & Lee, K. (2014). Seeing jesus in toast: neural and behavioral correlates of face pareidolia. *Cortex*, *53*, 60–77.

Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology*, *15*(2), 197–237. doi: `doi:10.1016/0010-0285(83)90009-9`

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision research*, *46*(16), 2465–74. doi: `doi:10.1016/j.visres.2006.02.002`

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, *10*(4), 437–442.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. *CoRR*.

Poggio, T., & Riesenhuber, M. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. doi: `doi:10.1038/14819`

Rieth, C. A., Lee, K., Lui, J., Tian, J., & Huber, D. E. (2011). Faces in the mist: illusory face and letter detection. *i-Perception*, *2*(5), 458–76. doi: `doi:10.1068/i0421`

Ringach, D., G., S., & Shapley, R. (1997). A subspace reverse correlation technique for the study of visual neurons. *Vision Research*, *37*(17), 2455–2464.

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*(2), 147–166.

Seli, P., Jonker, T. R., Solman, G. J. F., Cheyne, J. A., & Smilek, D. (2013). A methodological note on evaluating performance in a sustained-attention-to-response task. *Behavior Research Methods*, *45*(2), 355–363. doi: doi:10.3758/s13428-012-0266-1

Shermer, M. (2008). Patternicity: Finding meaningful patterns in meaningless noise. *Scientific American*, *299*(5).

Smilek, D., Enns, J. T., Eastwood, J. D., & Merikle, P. M. (2006). Relax! cognitive strategy influences visual search. *Visual Cognition*, *14*(4-8), 543–564.

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network (Bristol, England)*, *12*(3), 289–316. doi: doi:10.1088/0954-898X/12/3/304

Van Selst, M., & Merikle, P. M. (1993). Perception below the Objective Threshold? *Consciousness and Cognition*, *2*(3), 194–203. doi: doi:10.1006/ccog.1993.1018

Whittlesea, B., & Brooks, L. (1994). Journal of Experimental Psychology: Learning, Memory, and Cognition. *Journal of Experimental*.

Yamins, D., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. doi: doi:10.1038/nn.4244

Yamins, D., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–24. doi:

doi:10.1073/pnas.1403112111