Conscious access and complexity of visual features

How parallel processing of natural images interacts at different stages of computations

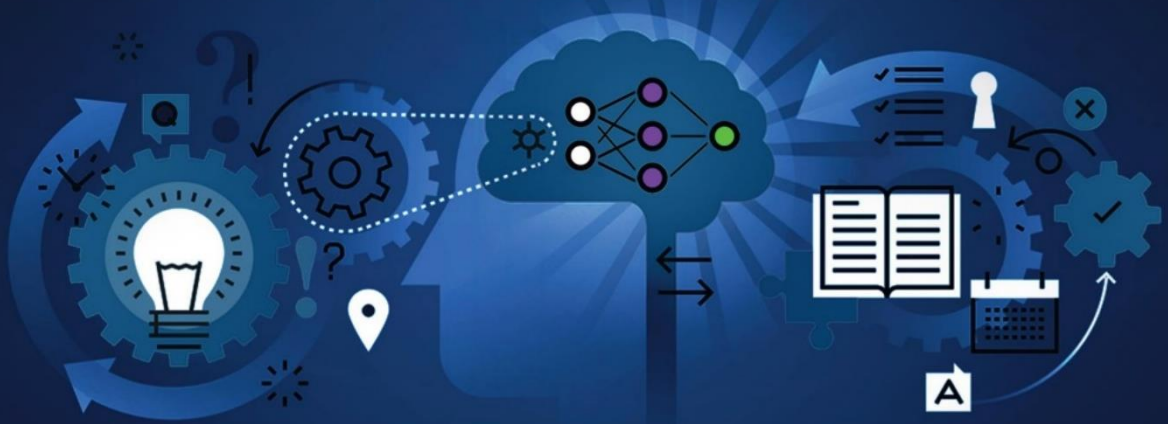
Lindh, P.J.D.

Publication date
2022

Document Version
Final published version

CONSCIOUS ACCESS AND COMPLEXITY OF VISUAL FEATURES

HOW PARALLEL PROCESSING OF NATURAL IMAGES INTERACTS AT DIFFERENT STAGES OF COMPUTATION



PER JOHAN DANIEL LINDH

CONSCIOUS ACCESS AND COMPLEXITY OF VISUAL FEATURES

PER JOHAN DANIEL LINDH

Conscious access and complexity of visual features

How parallel processing of natural images interacts at different stages of computations

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College
voor Promoties ingestelde commissie,
in het openbaar te verdedigen
in de Agnietenkapel op Vrijdag 1 Juli 2022, om 16.00 uur

door Per Johan Daniel Lindh
geboren te Forsa

Cover: https://www.fiverr.com/xee_designs1
Layout: Daniel Lindh
Printing: Proefschriften.nl

Promotiecommissie

Promotores:	prof. dr. V.A.F. Lamme	Universiteit van Amsterdam
	prof. dr. K.L. Shapiro	University of Birmingham
Copromotores:	dr. I.G. Sligte	Universiteit van Amsterdam
	dr. I. Charest	University of Birmingham
Overige leden:	prof. dr. E.H.F. de Haan	Universiteit van Amsterdam
	dr. H.S. Scholte	Universiteit van Amsterdam
	prof. dr. H.A. Slagter Vrije	Universiteit Amsterdam
	prof. dr. C.N.L. Olivers	Vrije Universiteit Amsterdam
	prof. dr. N.G. Kanwisher	Massachusetts Institute of Technology
	prof. dr. B.U. Forstmann	Universiteit van Amsterdam

Faculteit der Maatschappij- en Gedragswetenschappen

Dit proefschrift is tot stand gekomen binnen een samenwerkingsverband tussen de Universiteit van Amsterdam en de University of Birmingham met als doel het behalen van een gezamenlijk doctoraat. Het proefschrift is voorbereid aan de Faculteit der Maatschappij- en Gedragswetenschappen van de Universiteit van Amsterdam en aan de School of Psychology van de University of Birmingham.

This thesis was prepared within the partnership between the University of Amsterdam and the University of Birmingham with the purpose of obtaining a joint doctorate degree. The thesis was prepared in the Faculty of Social and Behavioural Sciences of the University of Amsterdam and in the School of Psychology of the University of Birmingham.

For LovCov

Table of Contents

Introduction

Imagine you are in your car driving to meet a friend at a restaurant you have never been to before. As an experienced driver, you don't need to deliberately direct your gaze. Instead, your attention is automatically drawn to crossings far off in the distance, other moving vehicles, and relevant road signs. Without having to assert effort, your brain suppresses details in your immediate surroundings to enhance relevant information. When you arrive at the restaurant, you swiftly search through the crowd of strangers, assessing whether everyone is your friend within a fraction of a second. Your brain effortlessly evaluates each person with templates in your memory, first on crude features such as hair colour or height, and for anyone who fits these criteria, assessment is carried out on finer facial features. With our ability to use logical inferences based on experience we build templates of a target, which we use to efficiently scan through our environment. Regardless of how mundane this everyday task might seem; its completion requires several fundamental computational problems to be overcome. When driving a car and when searching a crowded room, you need to selectively enhance and suppress visual information, as processing all information equally is an inefficient use of resources. It can take several hundred milliseconds to fully process a complex natural scene (Kar et al., 2019), meaning that the processing of several visual objects must be happening in parallel. To add to this complexity, humans are continuously updating their goals (first, search for the bar across the whole room, then search for a person at the bar) based on information we are gaining within each moment. In this dissertation, I will address how the brain organizes information into categories, how items that are processed in parallel can interfere with each other, and at what levels of processing these interferences occur.

Even though between 20% and 30% of the cortex is dedicated to vision (Essen, 2003), our visual perception is not a veridical representation of the objective world. We fail to observe most of the enormous information flow present in our environment. While this may seem like a shortcoming, this failure to process information can be considered advantageous. Humans have evolved an impressive set of tools that enable us to quickly sift through the massive amount of data surrounding us, picking out and acting on what is most important. We can rapidly isolate important regions in our visual field and allocate additional resources, i.e., attend to that specific area (Posner et al., 1980). In our daily life, we take this ability for granted and it is easy to forget how impressive this achievement is. Combined, the hypothetical situation above illustrates a grand performance, which encompasses object recognition, saccade planning, working memory updating, goal definition, and integration with memories, held together by a sophisticated attentional system. Our environment constantly bombards us with information, and one of the most daunting tasks of the brain is to select the relevant information and filter out noise (i.e., signals that are non-informative for our task goals). This selection process is not without biases. The Baader-Meinhof phenomenon (or the frequency illusion) nicely illustrates how selective attention biases our perception in daily life. For example, have you ever learned a new word and then suddenly seen this word everywhere you look? In reality, you have been walking around all your life with this word frequently reaching your retina but not reaching your conscious awareness. What are the factors that lead us to become consciously aware of a specific stimulus? Are these factors all just related to our task-goals or are there stimuli that we inherently treat differently? Is it possible to manipulate our perceptual system in such a way that we are more likely to perceive certain objects? These are pertinent questions we need to answer to build a comprehensive theoretical framework of perception.

Object recognition and representation

Humans can classify and act upon objects within a fraction of a second (Kirchner & Thorpe, 2006). In fact, a mere 13-millisecond exposure of a visual scene is enough for us to retain the information in neural constellations so we can process the semantic meaning (Broers et al., 2018). Visual information is first received and processed in the retina, located at the innermost part of the eye. Here, spatial resolution is highest in the fovea and strongly declines toward the periphery of the visual field (Daniel & Whitteridge, 1961). The retina consists of rods and cones, two types of photoreceptors that provide vision in dim light and photopic (colour) vision, respectively (Bowmaker & Dartnall, 1980). The simplicity of this retinal setup comes with a few computational problems. For example, since the retina lies like a flat sheet in the back of the eye, the information reaching our brain is by nature two-dimensional, the brain, therefore, needs to infer depth. Another problem is that the same object can produce an immense variation in appearance depending on viewing angle (Logothetis & Steinberg, 1996). Therefore, to perceive the environment coherently it is necessary for the brain to construct view-invariant object representations without causing a combinatorial explosion in the number of cells required. To solve these problems, neurons with the same response properties are not randomly distributed throughout the visual system. Instead, neurons in the early visual cortex are organised retinotopically, which means that neurons' spatial organisation in the cortex corresponds to the locations in the visual field. Following the retinotopic organisation, bundled neurons in the early visual cortex respond to low-level stimulus features such as orientation (Swindale et al., 1987), spatial frequency (Tootell et al., 1981), and colour (Tootell et al., 2004). These features are later combined into more complex representations further along the visual stream, which runs from the most posterior to the anterior part of the brain (Figure 1). An influential idea of the visual system is the two-streams hypothesis (Goodale & Milner, 1992; Ungerleider & Mishkin, 1982) which posits that as information exists the occipital pole there is a partitioning of processes into the dorsal and ventral visual stream. A simplified description is that the dorsal stream encodes spatial properties such as size and location

(known as the “where”-stream) and the ventral stream encodes the identity of an object (known as the “what”-stream). However, the two streams are not independent, but characterized by many reciprocal connections (Budisavljevic et al., 2018; Cloutman, 2013; Zhong & Rockland, 2003) and informational integration between the streams is thought to happen at several levels, including continuous cross-talk between the streams (Budisavljevic et al., 2018; van Polanen & Davare, 2015), and shared target regions in frontal areas (Rauschecker & Scott, 2009), which in turn may facilitate integration through recurrent feedback loops (Cloutman, 2013). Despite this interconnectedness between the two streams, the predominant interest in the field of visual cognition has been with the ventral visual stream. This is presumably since within the ventral visual stream, specifically the human inferior temporal cortex (ITC, Figure 1), researchers have found patches of category-specific areas. For example, there are patches within the ITC that respond to places (Epstein et al., 1999), faces (N. Kanwisher et al., 1997), bodies (Downing et al., 2001), and 3D (Janssen et al., 2000); and the more posterior the patch, the more invariant the responses are to viewing angle (Bao et al., 2020). One recently proposed mechanism for these patches is that they are an emergent property of the evolutionary pressure of metabolic constraints, where it is more energy efficient to organise neurons responding to similar objects adjacent to each other to shorten the length of the axons in the lateral connections between them (Lee et al., 2020). Nevertheless, the behavioural relevance of this category-dependent organisation is a topic of active debate in the field of neuroscience, and answers to questions regarding the relationship between object representations and behaviour could help researchers understand the underlying mechanisms for these emergent structures.

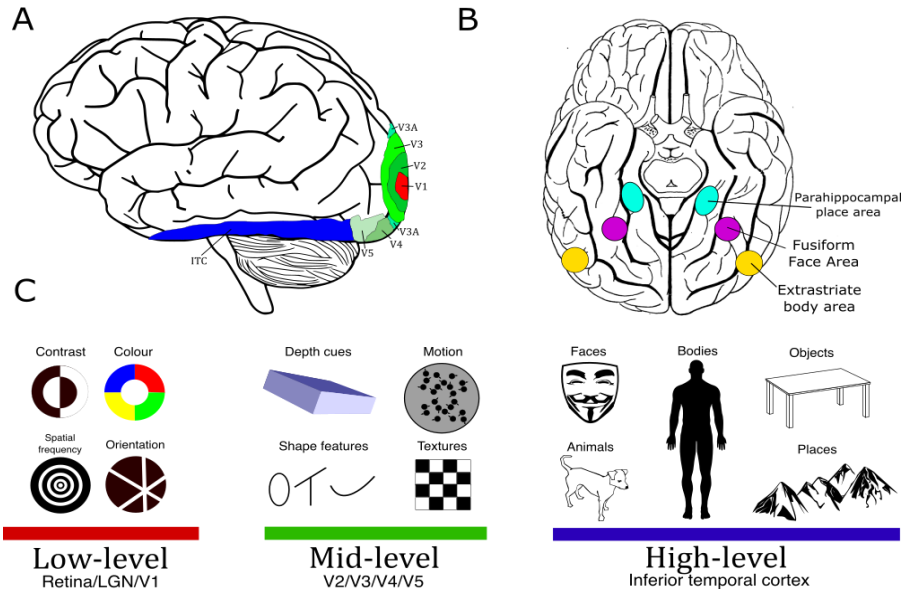


Figure 1. ***A)*** *Approximate anatomical divisions within the ventral visual stream. V1 (red) is the first cortical area to receive visual input. Information is then passed on to V2-V3-V4-V5 (green) before it reaches ITC (blue).* ***B)*** *Ventral view of the brain. Known areas that respond selectively to certain categories of visual stimuli. More patches are known to respond to faces, scenes and bodies, however, these locations are the most studied.* ***C)*** *Representational depiction of visual features processed at different stages. Early in the visual processing, simple features such as contrast and orientation are processed (left), later, more complex features are processed (middle) and eventually category-specific responses are found in the ITC (right). Colours of the underlying bars are reflecting areas as shown in A.*

Animacy and behaviour

One of the most notable functional divisions in ITC is between animate and inanimate objects. Here, researchers have found a continuum from the medial to the lateral, where the more animate an object is, the more lateral in the ITC is its representation encoded (Sha et al., 2015). Behavioural studies have shown that humans performing visual search are quicker to find animals compared to non-animals (Jackson

& Calvillo, 2013), better at remembering words describing animals (Nairne et al., 2013), and that animals seem to have privileged access to our conscious perception (Guerrero & Calvillo, 2016; Lindh et al., 2019). This division makes sense from an evolutionary perspective given how important it is to monitor potential threats such as predators in our surroundings. New et al. (2007) showed that subjects were quicker at detecting non-human animals than vehicles in a change detection task. The authors argued that if visual expertise was driving performance, we should predict the opposite - that is, that vehicles would be more quickly detected due to their prevalence in everyday life compared to exotic animals. This led New and colleagues to propose the "animate monitoring hypothesis", emphasising the evolutionary relevance of detecting animals for ancestral hunter-gatherer societies. This idea of an innate predisposition for detecting animals has been further corroborated by the existence of animate/inanimate distinct regions in the ventral visual stream in both sighted and congenitally blind subjects (Mahon et al., 2009).

The specific representational relationship between animals and non-animals in late visual areas seem to dictate how efficiently they are being processed. For example, (Carlson et al., 2014) trained a support vector machine (SVM) on voxels within the ITC to classify whether a presented stimulus was an animal or not. The resulting "confidence" of the classifier for each image was used as a measure of how much of an animal an object was according to the voxel wise activation in ITC. This "confidence" was later shown to correlate with reaction time in a speeded judgement task, with animals higher on the animal spectrum leading to faster response times. This study supports the notion that representations in visual areas are meaningful for behaviour and the representations we extract through multivariate approaches reflect information used in decision making (Grootswagers et al., 2018). One of the key questions in this dissertation is regarding how different semantic categories, eliciting grouped patches of activation throughout the visual stream, affect conscious awareness. I address this question directly in chapter 2 and further explore a more general mechanism for this in chapter 3 and chapter 4.

Convolutional neural networks

In the field of computational vision, the difficulty of turning pixels into view-invariant categorical representations has not been eluded. Computational scientists have attempted for decades to develop algorithms that can detect and classify objects in natural images, with a wide range of biologically inspired and more statistically based methods. This has been such a prominent problem in the field that it has inspired the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015), which consists of a large open data set of manually annotated natural scenes. The challenge is to construct computational models that classify objects within natural scenes into 1000 different categories. In 2012, (Krizhevsky et al., 2012) entered the competition using a deep convolutional neural network (DCNN, Figure 2) leading to a turning point for large-scale object recognition (this network has later been named AlexNet as a homage to the first author Alex Krizhevsky). Despite the fact that neural networks have been present since the 1980s (Fukushima, 1980), they have been deemed computationally intractable for many decades due to the large amount of parameters (AlexNet has 61 million parameters!), giving preference to less complex models such as Fisher vectors (Sanchez & Perronnin, 2011) and Support Vector Machines (SVM) (Anthony et al., 2007) for object classification. However, with the recent advances in GPUs, optimised for fast matrix calculations, Krizhevsky and colleagues were the undisputed winners of the 2012 competition, inspiring DCNNs to completely dominate the competition the following years. The architecture of DCNNs (Figure 2) is biologically inspired in that they consist of several hierarchical layers, equipped with “neural” units that either are activated or not depending on their input. Information is fed through each layer before reaching the final output layer, reminiscent of the brains’ ventral visual stream in which information propagates through visual areas V1/V2/V3/V4 to ITC (Figure 1). The main revelation underpinning these models is that the progressive build-up in invariance properties of neural responses along the ventral visual

stream could be approximated by a series of convolutions (multiplying areas of the image with learned filters) and local pooling operations (non-linearly combining the output of these filters). In the case of AlexNet (Krizhevsky et al., 2012), this network consists of eight layers where the first five layers are convolutional layers preserving retinotopical information (Figure 2). In each layer, each convolution is pooled into smaller representations, successively decreasing the retinotopic information. The last three layers are fully connected layers where retinotopical information is lost, giving way to view-invariant representations of high-level visual features. Even though AlexNet now has been surpassed in terms of classification accuracy by more complex models, it strikes a good balance between architecture-complexity and performance and is still widely used as a model of the human visual system.

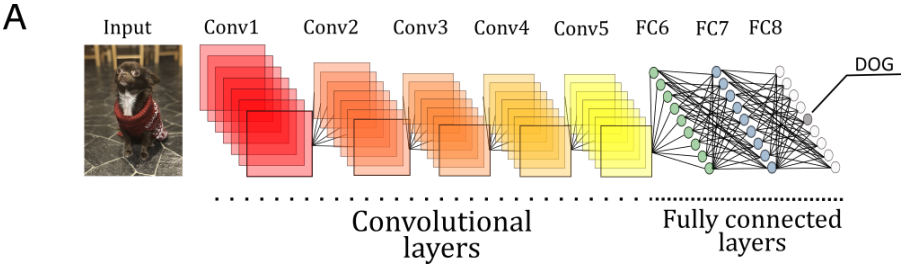


Figure 2. *Pictorial representation of the AlexNet architecture. AlexNet consists of five convolutional layers and three fully connected layers. The input image is fed into layer 1 (Conv1), and after a series of operations, information is fed forward to the next layer. Each convolutional layer consists of a bank of learned filters, that iteratively convolve separate parts of the image to estimate the presence of progressively more abstract visual features. In the fully connected layers, "retinotopical" information is lost and general view-invariant features are processed. Eventually, information reaches the response layer (FC8) where each node represents a category. The node with the highest activity becomes the network's best guess of what object is present in the image.*

The emergence of DCNNs sparked interest in comparing the internal representations produced within these models with the representations of stimuli at different stages of the ventral visual

stream. Despite the fact that the main engineering purpose of DCNNs was to solve object recognition, initial computational neuroscience work showed that AlexNet fit activity in the ITC significantly better than any other commonly used vision model (Khaligh-Razavi & Kriegeskorte, 2014). A series of consecutive studies further showed that the hierarchical order of the layers in DCNNs corresponds progressively to brain data along the ventral visual stream (Cichy et al., 2016; Eickenberg et al., 2017; Güçlü & van Gerven, 2014; Yamins et al., 2014) and the evolving representations over time (Cichy et al., 2016; Greene & Hansen, 2018). With the development of DCNNs came an increase in depth, with DCNNs winning the ILSVRC using >150 layers (He et al., 2016). Interestingly, even though increased depth has led to a lower error rate in image classification, to such a degree that the interest for ILSVRC has stagnated, initial studies did not show that the VGG-net (Simonyan & Zisserman, 2014), with its 19 layers, exhibited a better goodness-of-fit to the brain compared to the relatively parsimonious AlexNet (Abbasi-Asl et al., 2018). However, later studies have confirmed that increased depth (Kar et al., 2019) and recurrent connections (Kar et al., 2019; Kietzmann et al., 2019), increasing the depth of processing without adding more layers, substantially improves the model's ability to explain variance in the ventral visual stream. The current challenges have been to train models on semantics rather than image classification with, so far, promising results (Devereux et al., 2018). Just like Carlson et al. (2014) who showed that voxel activity in the late ventral visual stream is relevant for decision making, DCNNs can potentially be utilised to predict behaviour contingent on how successfully they model the visual processing hierarchy in the brain. Showing that DCNNs not only predict brain activity but also predict behaviour in a similar way as fMRI and EEG, is a crucial step to validate the models. Additionally, these models can eventually be used to probe the human perceptual system non-intrusively by treating them as never-ending variations of lab animals (Scholte, 2018). For these reasons, another main topic of this dissertation is not only to compare DCNNs with brain activation but also to evaluate how well DCNNs can be used to predict behaviour.

Temporal Object Recognition

Despite the computational difficulties of view-invariant object recognition, our ability to detect objects is remarkably fast. Broers et al. (2018) showed that a presentation of 13 ms per image was enough for their subjects to process semantic information. This impressive ability requires the brain to selectively process sensory inputs using a variety of mechanisms, collectively referred to as selective attention (for a recent review, see Fiebelkorn & Kastner, 2019). In early accounts from one of the first psychologists William James (James, 1890), attention "is the taking possession by the mind, in clear and vivid form, of one out of what may seem several simultaneously possible objects or trains of thought. It implies withdrawal from some things to deal effectively with others". At its core, selective attention refers to both enhancing relevant information as well as filtering out distracting information over both space and time. Researchers have developed many tools designed to test selective attention at the edge of our abilities. One of the most prevalent tools at our disposal is the rapid serial visual presentation (RSVP) display, where stimuli are presented in a quick fashion on a screen and subjects are asked to detect targets within the stream. By varying the speed of presentation, and what type of targets are presented at what time, researchers can examine processes related to attention, working memory, and conscious perception. Two of the most common findings using this paradigm are known as Attentional Blink (AB) (Raymond et al., 1992) and Repetition Blindness (RB) (Kanwisher, 1987; Kanwisher & Potter, 1990). These phenomena are closely related but each has important distinctions. Understanding these distinctions will give researchers a better understanding of how to optimally utilise these phenomena to elucidate fundamental mechanisms of human perception.

In a typical AB experiment, a series of distractors are presented at a rate of about 10 images per second. Within the stream of distractors, two targets (named T1 and T2) are presented at different temporal positions in the stream. Each position is often referred to as lags in relation to T1, where the items that are immediately following T1 are defined as lag-1, lag-2, etcetera (Figure 3A). In trials where T2 is

shown between 200-500 ms after T1 (lag-2 and lag-5, respectively when the speed of presentation is 10 items/s), participants show a clear reduction in performance (Figure 3B). This effect is abolished (Raymond et al., 1992), or in some cases partially (Folk et al., 2002; Maki & Mebane, 2006), when subjects are asked to ignore T1. This implies that the reason participants fail to report T2 is because of the attention required for T1. This phenomenon is like a blink of the mind caused by a lapse of attentional resources instead of a physical blink. One of the first, and most influential, models of the AB is the two-stage model (Chun & Potter, 1995). The two-stage model posits that pre-attentive initial processing can be done in parallel, such as processing the visual features of both targets. However, the second stage, where targets are encoded into working memory and become reportable, is constrained by a processing bottleneck to protect stimuli from being overwritten. Thus, the two-stage model proposes that the first target in the stream needs to be encoded fully into working memory before the second target can be processed. In support of this model, Vogel and Luck (2002) showed that, while both T1 and T2 are followed by a P3, an event-related potential (ERP) component strongly linked to working memory consolidation (Başar-Eroglu et al., 2001; Dolu et al., 2005), the P3 following T2 was *delayed*, suggesting that the brain was still consolidating T1 at the time of T2's presentation (Vogel & Luck, 2002).

Interestingly, studies have shown that when T2 is presented immediately after T1, i.e., when T2 is presented at lag-1, the effect of the AB is eradicated (n.b., this is not always the case, for discussion see Visser, 2015; Visser et al., 2009). This finding has been dubbed "lag-1 sparing", and although the T2 performance is almost as high on lag-1 as on late lags (lag-6 and upwards), it comes with order-inversion errors. Chun and Potter (1995) presented participants with a stream of symbols and asked participants to report two digits embedded within the stream. Participants were not instructed to report the targets in the correct order, however, they noticed that at lag-1, where T2 performance was high, participants often reported T2 first, implying a reversal of order of the targets. Later studies have not only corroborated this finding but also showed extended lag-1 sparing where several targets in a row, without intervening masks, can be

detected (Olivers et al., 2007). This extended sparing of multiple target detection and identification is direct evidence against the notion that the AB is due to a bottleneck in processing capacity as indicated by earlier models of the AB such as the two-stage model (Chun & Potter, 1995). Instead, newer models of AB assume that attention is chunked into separate episodic events. The closing of these events can be prolonged as long as new targets are being presented (Wyble et al., 2009), allowing up to 4-5 targets in a standard RSVP (Olivers et al., 2007; Wyble et al., 2009). The idea of our perceptual reality being organised into separate sections of incidents is contrary to our introspective notion of a continuous, coherent conscious experience of the world, however, this idea is neither new nor without substantial supporting evidence, for example, see (di Lollo, 1980).

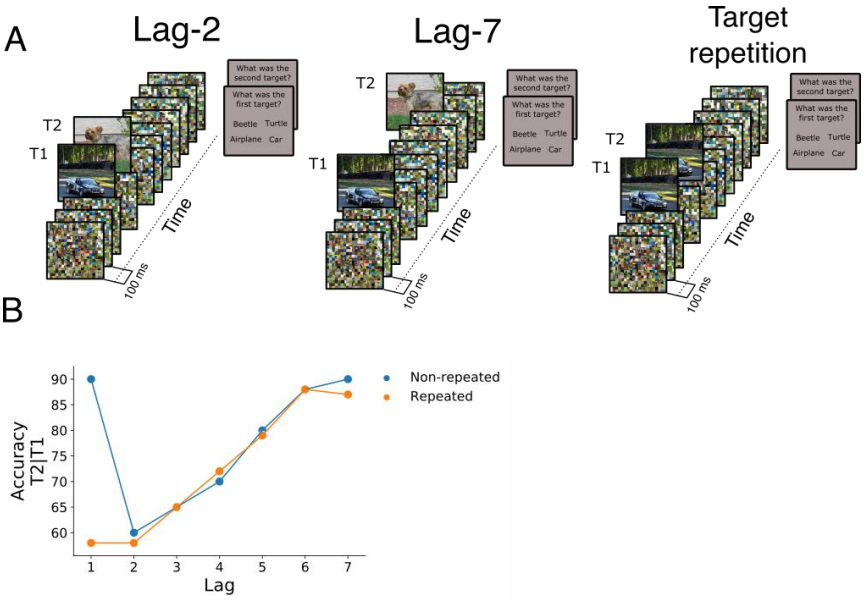


Figure 3. Attentional Blink and Repetition Blindness. ***A)*** *Typical RSVP setup for AB and RB. A stream of masks is shown for about 100 ms each. Within the stream two targets are presented, in all experiments in this thesis, targets are defined by being non-scrambled natural images. Subjects report which was the first and second target after the stream. During lag-2 trials (Left, one intervening distractor) subjects typically show difficulties reporting the second target compared to lag-7 (middle, six intervening distractors) trials. When targets are repeated, subjects often show an*

additional deficiency in reporting T2. ***B)*** *A cartoon plot illustrating the typical AB and RB results (see for example Chun, 1997). During trials when targets are not repeated subjects often show a lag-1 sparing, however, this is typically not as prominent when targets are repeated.*

In normal memory tasks, the repetition of stimuli usually enhances memory consolidation (Gathercole, 2006). However, early on in the history of RSVP, a series of experiments (Kanwisher, 1987; Kanwisher & Potter, 1990) showed that the second occurrence of an item in a stream is often omitted from the report at the end of the stream, dubbed Repetition Blindness (RB, Figure 3A). This effect was even present when a word omitted from a sentence by the subject led to a grammatically incorrect sentence. While the original study used letters, digits, and symbols, RB is also found when the stimuli consist of objects and natural images (Buffat et al., 2013; Coltheart et al., 2005; Harris & Dux, 2005). Interestingly, RB does not require the repetition of the exact same item (Bavelier, 1994; Bavelier & Potter, 1992). Sy and Giesbrecht (2009) showed that when participants were asked to identify the emotional expression of faces, the repetition in the task-relevant domain (i.e., two different angry faces) led to a decrease in T2 performance, but a repetition in gender (two female targets) did not. This finding was reversed when participants were asked to identify the sex of the target faces. Similarly, Stein et al. (2009) showed that emotional faces, which generally affect performance, only affected performance when participants were required to report on emotional content, i.e., when emotion was the task-relevant domain. This provides strong evidence for the importance of task goals in producing the RB effect.

RB is similar to another repetition deficit, the Ranschburg effect (Jahnke, 1969). However, RB and the Ranschburg effect are differentiated on their time scale. RB is only observed when items are presented at a fast rate (100-180 ms per item). The Ranschburg effect, on the other hand, is found with a presentation rate of 1 second per item indicating that there is a time interval distinguishing these two

paradigms. The RB and the Ranschburg effects are intriguing findings that reveal a systematic memory failure (Fagot & Pashler, 1995) at different processing levels. Likewise, AB and RB share many similarities but are also differentiated in their time scale (Arnell & Shapiro, 2011). While AB leads to impaired performance of T2 reportability around 200-500 milliseconds after T1, usually with a sparing of the first target (i.e. lag-1 sparing), the detrimental effect of RB is most pronounced in the first few lags with the largest effect on lag-1 (Chun, 1997). However, despite the overlaps between AB and RB and how the AB paradigm is used in the literature to study attention, working memory, and conscious experience, very few researchers try to control for these confounds. Considering how much parallel information we are tasked with processing while our eyes quickly move around, investigating our complex surroundings, understanding how representations overlap at different stages of processing is crucial for interpreting the processes underlying our perceptual content.

Embracing the complexity of natural images

Historically, studies of AB and RB have mostly concerned themselves with simple stimuli such as digits and letters. This has its own advantages of maximising control over the stimuli; however, it misses out on the complexity offered by natural images. Recent developments in machine learning, such as DCNNs, have facilitated research that embraces the complexity of natural stimuli. Another important development is the increased popularity in using multivariate analysis tools, including representational similarity analysis (RSA; Kriegeskorte et al., 2008), which allows for comparing the representational geometries (Figure 4A) between different modalities. Put simply, RSA allows us to measure how distant the representation of two stimuli is in a certain brain region in the high-dimensional space offered by the voxels within the region. By doing a pairwise comparison of all possible stimuli combinations we achieve a representational dissimilarity matrix (RDM). This can be done for the different layers of a particular CNN architecture, which transforms the idiosyncratic organisation of

features from one modality to a general “representational geometry space”, i.e., the specific pairwise relationship between stimuli. This transformation not only allows for different modalities (such as brain representations and CNNs) to be compared directly but also allows them both to be used to predict behaviour in a comparable way. Previous studies of RB with natural images often use category as a proxy for similarity, where for example two faces are more similar than a house and a face. While this assumption is oftentimes entirely valid, it misses out on the nuance between and within categories, for example, a strong association between a picture of a golf club and a golf ball in a certain brain area despite a lack of any shared visual features. By measuring the distance (or similarity) in representation using multivariate distance/similarity metrics from different brain areas and different layers of a CNN, it is possible to appreciate the complex relationship between each pair of images.

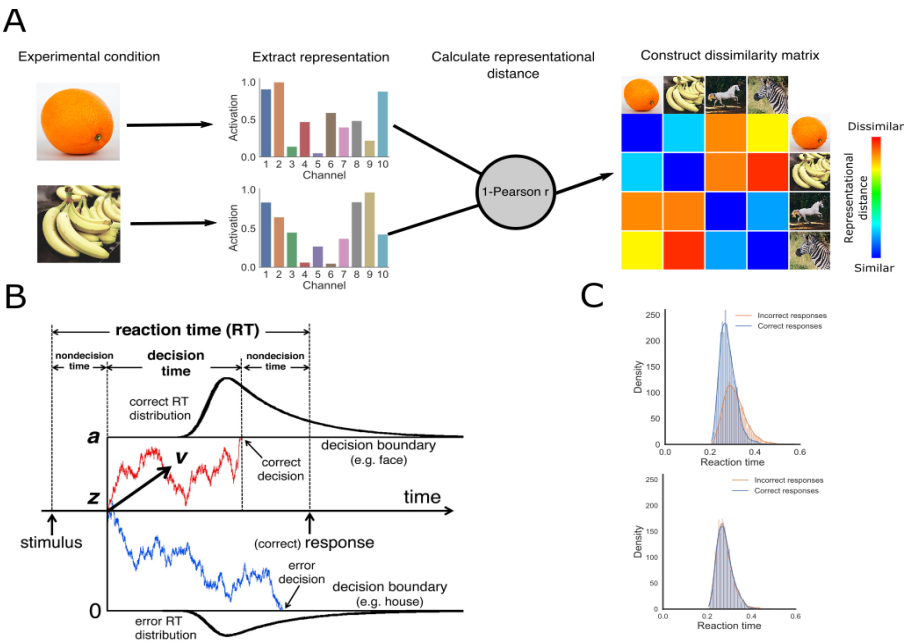


Figure 4. Representational similarity analysis and drift diffusion modelling. ***A)*** *Pictorial representation of the creation of a representational dissimilarity matrix (RDM). First, images are presented to a participant or a vision model. Second, multivariate representations are extracted for each*

image. Here different channels refer to dependent variables from the measured modality, for example voxels recorded with fMRI, electrodes from EEG or units from DNN. Third, a pair-wise comparison of how far away these representations are in high-dimensional space. One common measure of distance is the 1 - Pearson correlation, however, many other distance metrics exist. ***B)*** *Illustration of what different fitted parameters of the drift diffusion model (DDM) correspond to. In the DDM framework, several latent parameters are estimated which are believed to be evident in the reaction time distribution of correct and incorrect trials. Evidence is accumulated over time for two alternatives (for example a house or a face), when evidence for one of the alternatives reaches a boundary, a perceptual decision is made. The a-parameter corresponds to the distance between the starting point and the decision boundary, colloquially describing the decision criterion. The v-parameter refers to the drift rate, the steepness of the evidence accumulation and describes how efficiently a participant accumulates information over time. Other parameters are the t-parameter for non-decision time (for example motor responses and the time it takes a stimulus to reach cortical processing areas) and the z-parameter for bias (for example if a participant is more inclined to respond "house" over "face").* ***C)*** *Cartoon distributions of reaction distribution if the drift rate (v-parameter) is high (top plot) or low (bottom plot). When drift rate is high, the reaction time distribution will shift with a higher and earlier mode for correct (blue) trials compared to incorrect (red) trials. In comparison when drift rate is low the two distributions are indistinguishable.*

In all three following chapters, we combine well-known behavioural RSVP paradigms, such as AB and RB, with state-of-the-art brain analyses and machine learning to answer three main questions. 1) What is the relationship between target categories and their propensity to be consciously accessed. 2) How does the relationship between targets affect performance at different levels of processing. 3) In addition to explaining neural data, can CNNs also be used to explain behaviour? In Chapter 2 we specifically ask whether categories grouped together in multivariate high-dimensional space in ITC are also differentially affected by the AB time window. Using natural images depicting everyday objects from several distinct categories known to be grouped together in high-dimensional space (Charest et al., 2014; Kriegeskorte et al., 2008), we show that there is an extensive variance between semantic categories in the AB. We further

demonstrate that the variance between images in AB can be predicted using high-level visual features, as opposed to low-level visual features. Finally, we show that similarities between targets in terms of visual features, which are not the dimension used by participants to report targets, increase the probability of correctly reporting T2.

The finding in Chapter 2 that target similarity leads to better T2 performance was a surprising contrast to RB which led to the experiment in Chapter 3, wherein we extended our analysis and included individual brain representations from both functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). The main difference between fMRI and EEG is resolution in space and time. fMRI measures oxygenated blood flow in small voxels (volumetric pixels) and can reach millimetre precision in identifying which brain area is active. However, the drawback is that blood flow is slow, and the presentation of stimuli needs to be separated by several seconds to get a reasonable signal-to-noise ratio. Meanwhile, EEG measures electrical activity outside the scalp and is often recorded with a temporal precision of ~500-1000 Hz but with the caveat, the electrodes on top of the scalp produce an imprecise estimate for the origin of the signal. However, by combining both fMRI and EEG and designing different behavioural paradigms that let us achieve high signal-to-noise in both, as well as using representations in AlexNet, we show that when targets are similar in high-level semantic space there is a decrease in T2 performance. This reflects previous RB findings but with the important extension that the effect of similarity between targets is a gradient, and an exact repetition is not necessary to produce this behaviour. Furthermore, we show that when targets are similar in low-level visual features (such as in V1, the first cortical area to receive visual input), there is an increase in T2 performance. This replicates our findings from Chapter 2 and shows that target similarity can lead to both increased and decreased T2 performance, depending on the level of process in which the targets interact. To our knowledge, this is the first demonstration of such an RSVP effect.

In the AB literature, it is well-known that some participants do not seem to “blink”, referred to as “non-blinkers" (Martens & Valchev, 2009).

Seeing that several clinical conditions, such as ADHD (Amador-Campos et al., 2015; Armstrong & Munoz, 2003) and schizophrenia (Goddard, 2004; Wynn et al., 2006), also lead to individual differences in the attentional blink, it is pertinent that we understand the underlying mechanism behind “non-blinkers". We investigated individual differences by looking at overall performance for each participant and isolated which brain areas correlated with their performance in terms of similarity between image pairs. We found that participants that have large representational distances between images in the right temporoparietal junction (rTPJ) and the right inferior frontal gyrus (rIFG) perform significantly better at the task. These areas have been proposed to constitute a bottom saliency network (Corbetta et al., 2008), and our finding corroborates this network as an important target for investigating idiosyncratic perceptual processing.

While RB is believed to impair memory-related functions (Fagot & Pashler, 1995), we reasoned that the effect of V1-similarity found in Chapter 3 is related to processes prior to working memory updating. Specifically, in Chapter 4 we hypothesised that the T1-evoked activation in V1 would facilitate evidence accumulation rate for T2 if both shared similar representations, regardless of the semantic content of the two natural scenes. Furthermore, one of the key concepts of AB is the role of attending or ignoring T1. In all models of AB, attending T1 has a central role, and to show that target-target similarity to be pertinent for AB it also needs to be modulated by attention. To investigate this, we created a modified RSVP task where participants were presented with two targets and instructed to make a speeded judgment on whether T2 (i.e., the second target) contained an animal. Participants completed two blocks, one where they were asked to ignore the first target and one where they were instructed to memorise and report the first target after the stream. When using reaction time and accuracy for T2 as dependent variables it is common to look at each variable separately. However, the caveat is that there is a trade-off between speed and accuracy such that when participants are faster, they often make more mistakes. Another problem is that reaction time distributions are rarely normally distributed, so a point-estimate (such as the mean) is rarely a good description of the

distribution. Instead, to test our idea, we used Drift Diffusion Modelling (DDM) (Ratcliff & McKoon, 2008; Voss et al., 2004). DDM allows researchers to infer latent variables associated with the decision process in two-alternative forced-choice tasks. Assuming two boundaries (one for each alternative) with a decision value placed somewhere in between. This decision value will vary over time depending on the incoming information. The subject performing the task will reach a decision when the accumulated information reaches one of the two boundaries. By looking at the performance and the reaction time distribution (Figure 4B), DDM infers several latent variables that are associated with the decision process, such as drift-rate (the rate of evidence accumulation over time), bias (if participants have a preference towards one of the two alternatives), criterion (how much evidence does the subject need before making a decision), and non-decision time (length of motor responses and encoding to working memory). The shape of the reaction time distribution for the two alternatives can be described with different values of the latent variables (Figure 4B). We, therefore, fit the variables in such a way that it describes the RT distributions in the best way where the most interesting variable is drift-rate. We show that V1-similarity between T1 and T2 increases drift-rate for detecting T2 targets, lending support to the notion that this facilitation of T2 performance is driven by pre-attentive processes. Importantly, we show that attending T1 is necessary for the effect of target-target similarity in V1 to affect T2 drift rate.

In combination, this series of studies make use of cutting-edge technological advances combined with well-established paradigms to answer a set of questions impossible to answer just a few years ago. We show that the representational geometry present throughout visual cortices has important behavioural relevance. Semantic categories that are grouped together in the high-dimensional space within the ventral visual stream are differentially processed, where mainly animate objects have a higher propensity for conscious access. The inter-stimuli differences measured with brain imaging tools and CNNs provide explanations for contradictory behaviour. While the processing of two stimuli that share high-level, task-relevant similarities have been

argued to be related to memory encoding failures (Fagot & Pashler, 1995), our findings suggest that this is related to neurally overlapping representations in late processing stages. In contrast, the similarity in low-level visual features boosts the processing speed of objects in the evidence accumulation stage. Furthermore, not only can CNNs predict neural activation, but they also successfully predict behaviour. The implication of this is that the representational overlap between CNNs and the brain is not only relevant in direct terms, but they are applicable in a behavioural sense, corroborating CNNs as a promising model of the visual system.

References

Abbasi-Asl, R., Chen, Y., Bloniarz, A., & Oliver, M. (2018). The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. *bioRxiv*. https://www.biorxiv.org/content/10.1101/465534v1.abstract

Amador-Campos, J. A., Aznar-Casanova, J. A., Bezerra, I., Torro-Alves, N., & Sànchez, M. M. (2015). Attentional blink in children with attention deficit hyperactivity disorder. *Revista Brasileira de Psiquiatria* , *37*(2), 133–138.

Anthony, G., Greg, H., & Tshilidzi, M. (2007). Classification of Images Using Support Vector Machines. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/0709.3967

Armstrong, I. T., & Munoz, D. P. (2003). Attentional blink in adults with attention-deficit hyperactivity disorder: Influence of eye movements. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *152*(2), 243–250.

Arnell, K. M., & Shapiro, K. L. (2011). Attentional blink and repetition blindness. *Wiley Interdisciplinary Reviews. Cognitive Science*, *2*(3), 336–344.

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103–108.

Başar-Eroglu, C., Demiralp, T., Schürmann, M., & Başar, E. (2001). Topological distribution of oddball "P300" responses. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *39*(2-3), 213–220.

Bavelier, D. (1994). Repetition blindness between visually different items: the case of pictures and words. *Cognition*, *51*(3), 199–236.

Bavelier, D., & Potter, M. C. (1992). Visual and phonological codes in repetition blindness. *Journal of Experimental Psychology. Human Perception and Performance*, *18*(1), 134–147.

Bowmaker, J. K., & Dartnall, H. J. (1980). Visual pigments of rods and cones in a human retina. *The Journal of Physiology*, *298*, 501–511.

Broers, N., Potter, M. C., & Nieuwenstein, M. R. (2018). Enhanced recognition of memorable pictures in ultra-fast RSVP. *Psychonomic Bulletin & Review*, *25*(3), 1080–1086.

Budisavljevic, S., Dell'Acqua, F., & Castiello, U. (2018). Cross-talk connections underlying dorsal and ventral stream integration during hand actions. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *103*, 224–239.

Buffat, S., Plantier, J., Roumes, C., & Lorenceau, J. (2013). Repetition blindness for natural images of objects with viewpoint changes. *Frontiers in Psychology*, *3*(January), 1–11.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, *26*(1), 132–142.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the*

National Academy of Sciences, *111*(40), 14565–14570.
Chun, M. M. (1997). Types and Tokens in Visual Processing: A Double Dissociation between the Attentional Blink and Repetition Blindness. *Journal of Experimental Psychology. Human Perception and Performance*, *23*(3), 738–755.
Chun, M. M., & Potter, M. C. (1995). A Two-Stage Model for Multiple Target Detection in Rapid Serial Visual Presentation. In *Journal of Experimental Psychology: Human Perception and Performance* (Vol. 21, Issue 1, pp. 109–127). https://doi.org/10.1037/0096-1523.21.1.109
Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(January), 1–13.
Cloutman, L. L. (2013). Interaction between dorsal and ventral processing streams: where, when and how? *Brain and Language*, *127*(2), 251–263.
Coltheart, V., Mondy, S., & Coltheart, M. (2005). Repetition blindness for novel objects. *Visual Cognition*, *12*(3), 519–540.
Corbetta, M., Patel, G., & Shulman, G. L. (2008). The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron*, *58*(3), 306–324.
Daniel, P. M., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *The Journal of Physiology*, *159*, 203–221.
Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, *8*(1), 1–12.
di Lollo, V. (1980). Temporal integration in visual memory. *Journal of Experimental Psychology. General*, *109*(1), 75–97.
Dolu, N., Başar-Eroğlu, C., Özesmi, Ç., & Süer, C. (2005). An assessment of working memory using P300 wave in healthy subjects. *International Congress Series / Excerpta Medica*, *1278*, 7–10.
Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science* , *293*(5539), 2470–2473.
Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*(January 2016), 184–194.
Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). *The Parahippocampal Place Area : Recognition , Navigation , or Encoding ? 23*, 115–125.
Essen, V. (2003). Organization of visual areas in macaque and human cerebral cortex. *The Visual Neurosciences*, *1*, 507–521.
Fagot, C., & Pashler, H. (1995). Repetition Blindness: Perception or Memory Failure? *Journal of Experimental Psychology. Human Perception and Performance*, *21*(2), 275–292.
Fiebelkorn, I. C., & Kastner, S. (2019). Functional Specialization in the Attention Network. *Annual Review of Psychology*, *70*, 77–110.
Folk, C. L., Leber, A. B., & Egeth, H. E. (2002). Made you blink! Contingent attentional capture produces a spatial blink. *Perception and Psychophysics*, *64*(5), 741–753.
Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. In *Biological Cybernetics* (Vol. 36, Issue 4, pp. 193–202). https://doi.org/10.1007/bf00344251
Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*(04), 513–543.
Goddard, K. M. (2004). *The contribution of selective attention as measured by attentional blink (AB) tasks to performance on sustained attention and memory tests in schizophrenia*. https://elibrary.ru/item.asp?id=9362137
Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.
Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS Computational Biology*, *14*(7). https://doi.org/10.1371/journal.pcbi.1006327
Grootswagers, T., Cichy, R. M., & Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage*, *179*(March), 252–262.
Güçlü, U., & van Gerven, M. A. J. (2014). *Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Brain's Ventral Visual Pathway*. *35*(27),

10005–10014.
Guerrero, G., & Calvillo, D. P. (2016). Animacy increases second target reporting in a rapid serial visual presentation task. *Psychonomic Bulletin & Review*, *23*(6), 1832–1838.
Harris, I. M., & Dux, P. E. (2005). Orientation-invariant object recognition: Evidence from repetition blindness. *Cognition*, *95*(1), 73–93.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
Jackson, R. E., & Calvillo, D. P. (2013). Evolutionary relevance facilitates visual information processing. *Evolutionary Psychology: An International Journal of Evolutionary Approaches to Psychology and Behavior*, *11*(5), 1011–1026.
Jahnke, J. C. (1969). The Ranschburg effect. *Psychological Review*, *76*(6), 592–605.
James, W. (1890). The principles of psychology. *Henry Holt and Company*. https://books.google.com/books?hl=en&lr=&id=11gUsvvfrYUC&oi=fnd&pg=PA1&dq=The+Principles+of+Psychology&ots=EyZycVJFbZ&sig=qsbrzGucAwwCZS6kT-EHRfQlME4
Janssen, P., Vogels, R., & Orban, G. A. (2000). Three-dimensional shape coding in inferior temporal cortex. *Neuron*, *27*(2), 385–397.
Kanwisher, N. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, *27*, 117–143.
Kanwisher, N. G., & Potter, M. C. (1990). Repetition Blindness: Levels of Processing. *Journal of Experimental Psychology. Human Perception and Performance*, *16*(1), 30–47.
Kanwisher, N., McDermott, J., & Chun, M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*. https://www.jneurosci.org/content/17/11/4302.short
Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*. https://doi.org/10.1038/s41593-019-0392-5
Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11). https://doi.org/10.1371/journal.pcbi.1003915
Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., & Hauk, O. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1905544116
Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762–1776.
Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 1–28.
Kriegeskorte, N., Mur, M., Ruff, D. A., & Kiani, R. (2008). Matching categorical object representations in inferior temporal cortex of mand and monkey. *Neuron*, *60*(6), 1126–1141.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1–9.
Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., & DiCarlo, J. J. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. In *Cold Spring Harbor Laboratory* (p. 2020.07.09.185116). https://doi.org/10.1101/2020.07.09.185116
Lindh, D., Sligte, I. G., Assecondi, S., Shapiro, K. L., & Charest, I. (2019). Conscious perception of natural images is constrained by category-related visual features. *Nature Communications*, *10*(1), 4106.
Logothetis, N., & Steinberg, D. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621.
Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-Specific Organization in the Human Brain Does Not Require Visual Experience. *Neuron*, *63*(3), 397–405.
Maki, W. S., & Mebane, M. W. (2006). Attentional capture triggers an attentional blink.

Psychonomic Bulletin & Review, *13*(1), 125–131.
Martens, S., & Valchev, N. (2009). Individual differences in the attentional blink: The important role of irrelevant information. *Experimental Psychology*, *56*(1), 18–26.
Nairne, J. S., VanArsdall, J. E., Pandeirada, J. N. S., Cogdill, M., & LeBreton, J. M. (2013). Adaptive Memory: The Mnemonic Value of Animacy. *Psychological Science*, *24*(10), 2099–2105.
New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, *104*(42), 16598–16603.
Olivers, C. N. L., Van Der Stigchel, S., & Hulleman, J. (2007). Spreading the sparing: Against a limited-capacity account of the attentional blink. *Psychological Research*, *71*(2), 126–139.
Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology*, *109*(2), 160–174.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.
Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724.
Raymond, J. D., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in a RSVP task: an attentional blink? *Journal of Experimental Psychology*, *18*(3), 849–860.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.
Sanchez, J., & Perronnin, F. (2011). High-dimensional signature compression for large-scale image classification. *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 1665–1672.
Scholte, H. S. (2018). Fantastic DNimals and where to find them. *NeuroImage*, *180*(Pt A), 112–113.
Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O., & Connolly, A. C. (2015). The animacy continuum in the human ventral vision pathway. *Journal of Cognitive Neuroscience*, *27*(4), 665–678.
Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1409.1556
Stein, T., Zwickel, J., Ritter, J., Kitzmantel, M., & Schneider, W. X. (2009). The effect of fearful faces on the attentional blink is task dependent. *Psychonomic Bulletin & Review*, *16*(1), 104–109.
Swindale, N. V., Matsubara, J. A., & Cynader, M. S. (1987). Surface organization of orientation and direction selectivity in cat area 18. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *7*(5), 1414–1427.
Sy, J. L., & Giesbrecht, B. (2009). Target-target similarity on the attentional blink: Task-relevance matters! *Visual Cognition*, *17*(3), 1–10.
Tootell, R. B. H., Nelissen, K., Vanduffel, W., & Orban, G. A. (2004). Search for color "center (s)"in macaque visual cortex. *Cerebral Cortex* , *14*(4), 353–363.
Tootell, R. B., Silverman, M. S., & De Valois, R. L. (1981). Spatial frequency columns in primary visual cortex. *Science*, *214*(4522), 813–815.
Ungerleider, L., & Mishkin, M. (1982). Two cortical visual systems. *Analysis of Visual Behavior*, 549–586.
van Polanen, V., & Davare, M. (2015). Interactions between dorsal and ventral streams for controlling skilled grasp. *Neuropsychologia*, *79*(Pt B), 186–191.
Visser, T. A. W. (2015). Expectancy-based modulations of lag-1 sparing and extended sparing during the attentional blink. *Journal of Experimental Psychology. Human Perception and Performance*, *41*(2), 462–478.
Visser, T. A. W., Davis, C., & Ohan, J. L. (2009). When similarity leads to sparing: Probing mechanisms underlying the attentional blink. *Psychological Research*, *73*(3), 327–335.
Vogel, E. K., & Luck, S. J. (2002). Delayed working memory consolidation during the attentional blink. *Psychonomic Bulletin & Review*, *9*(4), 739–743.
Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206–1220.

Wyble, B., Bowman, H., & Nieuwenstein, M. (2009). The Attentional Blink Provides Episodic Distinctiveness: Sparing at a Cost. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(3), 787–807.

Wynn, J. K., Breitmeyer, B., Nuechterlein, K. H., & Green, M. F. (2006). Exploring the short term visual store in schizophrenia using the attentional blink. *Journal of Psychiatric Research*, *40*(7), 599–605.

Yamins, D. L. K., 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*, *111*(23), 8619–8624.

Zhong, Y.-M., & Rockland, K. S. (2003). Inferior Parietal Lobule Projections to Anterior Inferotemporal Cortex (Area TE) in Macaque Monkey. *Cerebral Cortex* , *13*, 527–540.

Chapter 2

Conscious perception of natural images is constrained by category-related visual features

Authors:
Daniel Lindh[1,2,3]
Ilja G. Sligte[3,4]
Sara Assecondi[1,2]
Kimron L. Shapiro[1,2]
Ian Charest[1,2]

Affiliations:
[1]School of Psychology, Hills Building, University of Birmingham, B152TT Birmingham, United Kingdom

[2]Centre for Human Brain Health, University of Birmingham, United Kingdom

[3]Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands
[4]Amsterdam Brain and Cognition, University of Amsterdam, The Netherlands

Corresponding author: dnllndh@gmail.com

Published as:
Lindh, D., Sligte, I. G., Assecondi, S., Shapiro, K. L., & Charest, I. (2019). Conscious perception of natural images is constrained by category-related visual features. Nature communications, 10(1), 1-9.

Keywords:
Attention, Deep Convolutional Neural Networks, Consciousness, Object recognition

Abstract

Conscious perception is crucial for adaptive behaviour yet access to consciousness varies for different types of objects. The visual system comprises regions with widely distributed category information and exemplar-level representations that cluster according to category. Does this categorical organisation in the brain provide insight into object-specific access to consciousness? We address this question using the Attentional Blink (AB) approach with visual objects as targets. We find large differences across categories in the AB. We then employ activation patterns extracted from a deep convolutional neural network (DCNN) to reveal that these differences depend on mid- to high-level, rather than low-level, visual features. We further show that these visual features can be used to explain variance in performance across trials. Taken together, our results suggest that the specific organisation of the higher-tier visual system underlies important functions relevant for conscious perception of differing natural images.

Introduction

A long-standing question in cognitive neuroscience is how visual information is transformed from segregated low-level features to fully conscious and coherent representations. Prevailing object recognition models propose that rapid object identification is accomplished by extracting increasingly complex visual features at various stages/locations of the visual stream [1–3]. Objects are first processed through a hierarchy of ventral visual areas where computations evolve from image feature detection, shape and part segmentation, before more invariant, semantic representations of the objects are established [4–6]. Previous research has shown that animate objects are preferably processed in a broad range of perceptual tasks[7]. This led us to question whether or not animacy also has a preferential access to consciousness, and furthermore, if this could also be true for sub-categories within the animate/inanimate distinction.

Animate versus non-animate object processing has been extensively studied, showing distinct processing pathways throughout the visual stream[8]. Behavioural studies have shown that animate objects are more often consciously perceived in rapid serial visual presentations (RSVP)[9–11], more quickly found in visual search[7], elicit faster responses in discrimination tasks[12,13], and animate words are better retained in working memory[14]. Aggregated, these findings point to a preferential visual processing of animate objects, most likely also reflected in the representational organisation of the visual stream[12,13]. However, the animate categorical division contains several sub-categories also known to cluster together, such as scenes in the parahippocampal place area[15], faces in the fusiform face area[16] and body parts in the extrastriate body area[17] (for review see Martin, 2007[18]). It remains unclear how such sub-categories also might differ in visual processing. We address this question by testing differences across several categories (i.e., fruits and vegetables, processed foods, objects, scenes, animal bodies and faces, human bodies, and faces), known to cluster together throughout the visual stream, in their propensity to conscious access using the Attentional Blink paradigm (AB)[19].

In the AB paradigm, two targets (T1 and T2) are embedded in a rapidly presented stimulus stream (RSVP). The frequently replicated finding is a reduced ability to report T2 when it is presented in a temporal window of 200-500 ms after a correctly identified T1. This effect disappears when subjects are asked to ignore T1[19], indicating that the fundamental explanation for this effect is attentional rather than perceptual. Most theoretical accounts of the AB suggest a two-stage information-processing model[20,21]. First, both targets are rapidly and automatically processed to a high-level representational stage. This is followed by a capacity-limited second stage, where the percept is transformed into a reportable state (i.e., working memory). Neural findings[22–26] have suggested that the AB arises at the second stage, after semantic processing of the object. This is in contrast to backwards masking, which is known to interrupt feedback loops in early processing[27–29]. Since feedback loops between visual areas are thought to be intact in the AB[26], combined with a behavioural outcome that typically yields a significant number of both correct and incorrect trials, this paradigm is an ideal approach to investigate the bifurcation between conscious and unconscious visual processing.

One potential problem of studying categorical differences is that many categories share low-level scene statistics[30], which also are known to explain behaviour[31]. Consequently, an issue that must be taken into account is how to control for low-level scene statistics in a neurally plausible way. We address this issue by using a Deep Convolutional Neural Network (DCNN)[32] which is designed in a hierarchy encompassing feature representations of increasing complexity, similar to the visual system. Recent studies using DCNNs trained to classify a large corpus of natural images have revealed a significant correspondence between DCNN layers and the visual hierarchical organisation in the brain both using fMRI[33–36], and MEG[5,37]. This makes DCNNs attractive for modelling visual features rather than relying on manually labelling image features without knowing their relevant correspondence to the visual system.

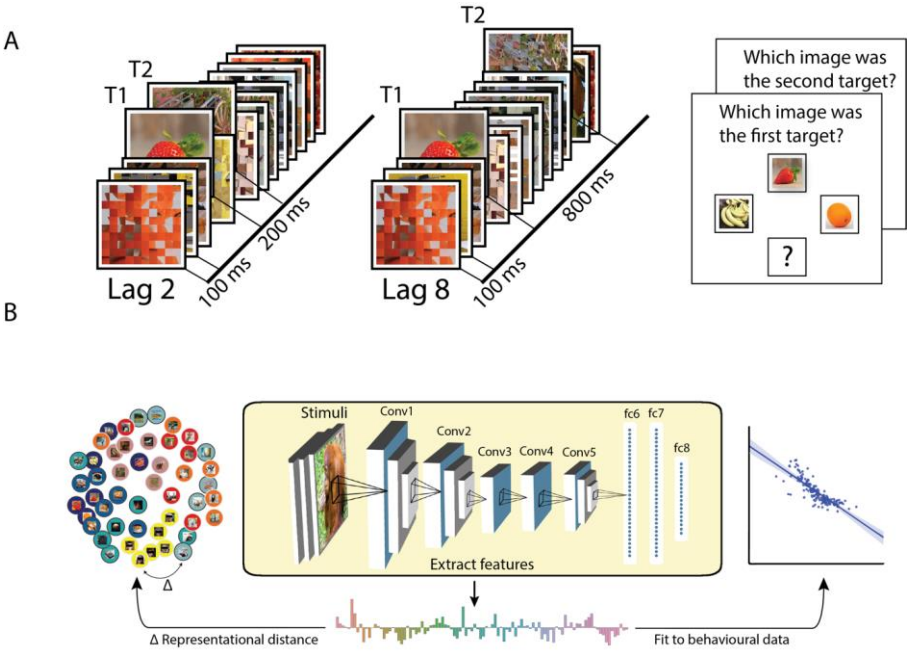


Figure 1. *Modulating conscious access using the Attentional Blink Paradigm. Due to copyright reasons, all photos except for the faces (which were photographed by one of the authors but have been anonymised) have been replaced by representational images. Eye regions are occluded above in the images to protect privacy but were not occluded in the experiment. A) We presented a rapid serial visual presentation to participants, with two targets (T1 and T2) following each other within a stream of distractors. On the left, the second target (T2) is shown 200ms after the first target (T1), and on the right, 800ms after the T1. In every trial, participants had to detect and later recall both T1 and T2 targets. B) We used a deep convolutional neural network (DCNN; yellow insert; 5 convolutional layers and 3 fully connected layers) to model the stimulus representational geometries (left) and predict our participants' behaviour (right). The visual stimuli were fed into the DCNN, providing a hierarchical representation for each image. These unit activations were then analysed layer-by-layer and used to predict behaviour.*

The main question of the current study is if the organisation of the visual system promotes conscious access to certain objects more than to others. A priori, we had two related hypotheses: first we hypothesised that categories will differ in their access to consciousness. Our second hypothesis was that variance in conscious

access between image exemplars could be predicted using high-, as opposed to low-, level features derived from the DCNN. These two predictions are consistent with our current understanding of the categorical organisation of the ventral visual stream[4,6,38,39], the high resemblance in representational geometry between the brain and DCNNs[33–37], and theoretical models positing the AB as a disruption of late selection[20,21]. In addition, we explore whether trial-by-trial variance in performance is related to the similarity between the two targets in terms of visual features. We asked whether this relationship has any impact on conscious access and, if so, at what stage of processing do the two targets interact? To test this formally, we used a method called representational sampling, where trials of the AB are constructed with stimuli selected according to their location in DCNN representational geometries. To foreshadow, we show that there are categorical differences in the probability of conscious access. Differences across images are predicted using mid- to high-level visual features. Furthermore, we find a facilitating interaction effect between targets, increasing the probability to recover T2.

Results

Experiment 1

Differences in AB magnitude as a function of category

Participants were presented with Rapid Serial Visual Presentations (RSVP), consisting of scrambled masks, and two embedded targets. The targets were selected from a stimulus set of 48 images derived from 8 different categories – fruits and vegetables, processed foods, objects, scenes, animal bodies, animal faces, human bodies, and human faces. At the end of each trial, participants were requested to recall the first and the second target (see Figure 1A). First, we observed a significant AB effect using a two-tailed dependent t-test in T2 performance (T2 performance is always conditional on T1 correct trials; T2|T1) between lags (Lag 2; accuracy $M = 0.704$, $SD = 0.041$, Lag 8; $M = 0.847$, $SD = 0.129$, $t(18) = -6.427$, $p < 0.001$, see Fig 2A). We first pooled the images according to animate and inanimate

(excluding scenes) objects (see Table 1). Animate and inanimate objects have previously been shown to be differentially affected during the AB[9–11]. Similarly here, a repeated measures 2x2 ANOVA with lag and animacy as factors showed a main effect of lag ($F(1,18) = 34.09$, $p < 0.001$, $\eta^2 = 0.654$) and animacy ($F(1,18) = 27.72$, $p < 0.001$, $\eta^2 = 0.606$) as well as a significant interaction effect ($F(1,18) = 45.63$, $p < 0.001$, $\eta^2 = 0.606$; see Fig 2B). Thus, in accordance with previous studies, the AB was less pronounced for animate images. For each sub-category (Table 2), using a repeated measures ANOVA, we observed a main effect of T2-lag ($F(1,18)=42.87$, $p < 0.001$, $\eta^2 = 0.704$) and category ($F(7,126) = 45.49$, $p < 0.001$, $\eta^2 = 0.716$), along with an interaction between category and T2-lag ($F(7, 126) = 23.99$, $p < 0.001$, $\eta^2 = 0.571$). Beyond the expected AB effect, the interaction effects reveal that different categories exhibit different attentional blink magnitudes (ABM; difference in performance between lag 8 and lag 2). Separate AB effects were tested by contrasting lag 8 and lag 2 performance within each category using a two-tailed dependent t-test (Fig 2C) – Fruits and Vegetables ((18) = 6.912, p < .001), Processed foods ($t(18) = 6.748$, $p < .001$), Objects ($t(18) = 3.003$, $p = .004$), Scenes ($t(18) = 8.073$, $p < .001$), Animal bodies ($t(18) = 5.259$, $p < .001$), Animal faces ($t(18) = 2.712$, $p = .007$), Human bodies ($t(18) = 1.162$, $p = .13$), Human faces ($t(18) = 2.632$, $p = 0.008$).

Table 1: Mean and SDs for T2 performance for animacy.

Animacy	Mean (Lag 2)	SD (Lag 2)	Mean (Lag 8)	SD (Lag 8)	N
Animate	0.792	0.171	0.872	0.126	19
Inanimate	0.683	0.185	0.871	0.097	19

Table 2: Mean and SDs for T2 performance for each category.

Category	Mean (Lag 2)	SD (Lag 2)	Mean (Lag 8)	SD (Lag 8)	N
Fruits Vegetables	0.651	0.199	0.867	0.139	19
Processed Foods	0.595	0.214	0.853	0.110	19
Objects	0.806	0.173	0.893	0.079	19
Scenes	0.406	0.234	0.695	0.237	19
Animal bodies	0.642	0.232	0.822	0.159	19
Animals faces	0.782	0.197	0.858	0.134	19
Human bodies	0.859	0.179	0.879	0.153	19
Human faces	0.886	0.133	0.927	0.085	19

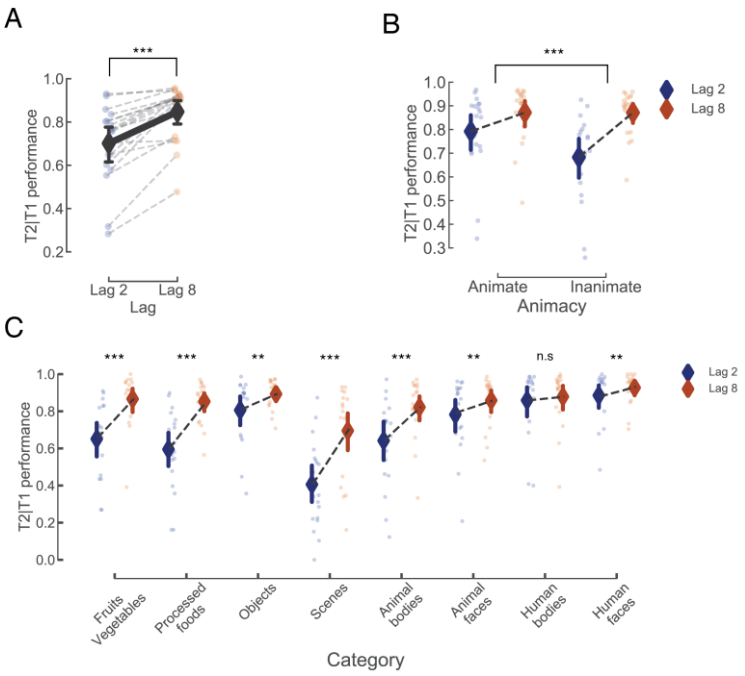


Figure 2. ***Animate objects elicit weaker attentional blink***. ***A)*** *The accuracy in detecting the second target conditional on having detected the first target for lag 2 and lag 8. Individual dots represent the mean performance for each subject, bold dots represent the mean performance across subjects, and error bars indicate 95% confidence interval around the mean in all plots.* ***B)*** *Performance plotted separately for animate and inanimate T2 targets. Attentional Blink Magnitude (ABM) is defined as the difference in performance between lag 8 and lag 2. Asterisk indicate significant difference in ABM between animate and inanimate.* ***C)*** *T2 performance for each category separately. Asterisks indicate p-values significant difference in ABM from zero. Two-tailed dependent t-test * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.*

Mid and high-level image features explain ABM variance

For each image we extracted unit activations from all the layers throughout an AlexNet DCNN (see methods). For the convolutional layers, we averaged over the spatial domain to obtain feature activations. It is important to note that this DCNN was trained on classifying objects into categories from a different set of images than those presented in our experiment, and at no point was trained on the AB. To increase the generalization of the model fits to the test data, we selected informational features through a variance thresholding

approach. The feature selection was done by calculating the variance across samples in the training data (important to note that the test data was never part of the feature selection) and removing features with near-zero variance from both training and test data. The remaining feature activations were then applied to a cross-validated linear regression model aimed at predicting each image's ABM. From these predicted ABMs, we can compute in each participant the mean absolute error (MAE). For significance testing, we permuted the image labels, repeated the cross-validated linear regression model, and computed the average MAE across subjects. We repeated this permutation procedure 3000 times to estimate the distribution of MAE under the null hypothesis that our image labels are interchangeable. We then compared our observed MAE (averaged across subjects) to this null distribution and obtained p-values. We were able to significantly (Bonferroni corrected alpha = 0.00625) predict the ABM using features derived from layer conv4 (MAE M = 0.19, STD = 0.04, $p = 0.003$), conv5 (M = 0.179, STD = 0.04, $p < 0.001$), fc6 (M = 0.159, STD = 0.033, $p < 0.001$), fc7 (M = 0. 1593, STD = 0.033, $p < 0.001$), and fc8 (M = 0.191, STD = 0.048, $p < 0.001$). To see if one layer had significantly lower error than any other layer, we tested the MAE for each pair-wise comparison of layers across subjects with a two-sided dependent t-test. In Fig 3B we show a summary of this result, where we find that Layer 7 (Fig 3C) has a significantly lower error than layer 1 (mean difference = -0.21, $t(17) = -6.14$, $p < 0.001$), layer 2 (mean difference = -0.15, $t(17) = -7.8$, $p < 0.001$), layer 3 (mean difference = -0.16, $t(17) = -10.83$, $p < 0.001$), layer 4 (mean difference = -0.18, $t(17) = -5.8$, $p < 0.001$) and layer 8 (mean difference = -0.18, $t(17) = -5.17$, $p < 0.001$).

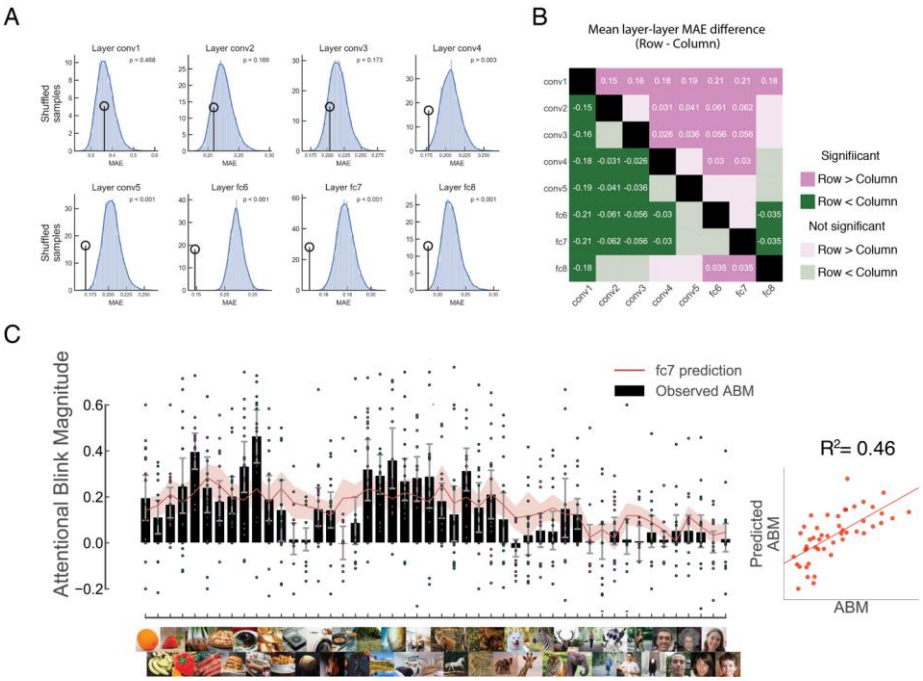


Figure 3. *DCNN activation units predict attentional blink magnitude.* ***A)*** *Permutation test distributions. Histograms show the mean absolute error (MAE) after averaging the prediction across participants with randomised image labels. Circles point to the observed MAE. The Bonferroni corrected alpha value for 8 tests is p < 0.00625.* ***B)*** *Layer by layer comparisons of MAE. Comparisons are done row-wise, where green indicates a lower MAE, or better fit, in comparison to the corresponding column. Only significant (Bonferroni corrected) comparisons are denoted with mean differences in MAE between comparisons.* ***C)*** *ABM per image. Due to copyright reasons, all photos except for the faces (which were photographed by one of the authors) have been replaced by representational images. Eye regions are occluded above in the images to protect privacy but were not occluded in the experiment. Black bars indicate the observed Attentional blink magnitude (ABM), red line is the average predicted ABM based on features from Layer fc7 (which outperformed all other layers, see panel B). Individual dots represent individual participants and error bars indicate the 95% confidence interval. Layer fc7 explained 46% of the variance observed. The insert panel shows the average predicted ABM on the y axis, and the average observed ABM per image, on the x axis.*

Shared image features between targets predicts performance

In addition to predicting the ABM for each image, we sought to better understand the trial-by-trial differences in the AB. For each trial, we correlated the two targets (T1 and T2) based on their features (Pearson correlation, Fig 3B) to obtain a T1-T2 similarity measure within each layer. We then averaged the similarity for all hit and miss trials for each participant and tested the difference for each layer using a two-tailed dependent t-test. Our test revealed a significantly higher representational similarity between targets in hit-trials compared to miss-trials for layer conv2 (Hit; similarity $M = 0.375$, $SD = 0.008$, Miss; $M = 0.354$, $SD = 0.014$, $t(18) = 4.967$, $p < 0.001$, Cohen's $d = 1.761$), conv3 (Hit; $M = 0.329$, $SD = 0.010$, Miss; $M = 0.299$, $SD = 0.016$, $t(18) = 6.273$, $p < 0.001$, Cohen's $d = 2.130$), conv4 (Hit; $M = 0.257$, $SD = 0.009$, Miss; $M = 0.244$, $SD = 0.012$, $t(18) = 3.505$, $p = 0.003$, Cohen's $d = 1.258$), conv5 (Hit; $M = 0.131$, $SD = 0.007$, Miss; $M = 0.119$, $SD = 0.011$, $t(18) = 3.311$, $p = 0.004$, Cohen's $d = 1.233$), fc6 (Hit; $M = 0.023$, $SD = 0.002$, Miss; $M = 0.018$, $SD = 0.004$, $t(18) = 4.009$, $p = 0.001$, Cohen's $d = 1.520$), fc7 (Hit; $M = 0.026$, $SD = 0.003$, Miss; $M = 0.021$, $SD = 0.005$, $t(18) = 3.189$, $p = 0.005$, Cohen's $d = 1.093$), fc8 (Hit; $M = 0.139$, $SD = 0.013$, Miss; $M = 0.104$, $SD = 0.022$, $t(18) = 6.134$, $p < 0.001$, Cohen's $d = 1.864$; Fig 4B). This suggests that the ongoing visual processing of T1 can lower the conscious access threshold for T2, if T2 shares visual features with T1. This was true for all layers except for layer 1.

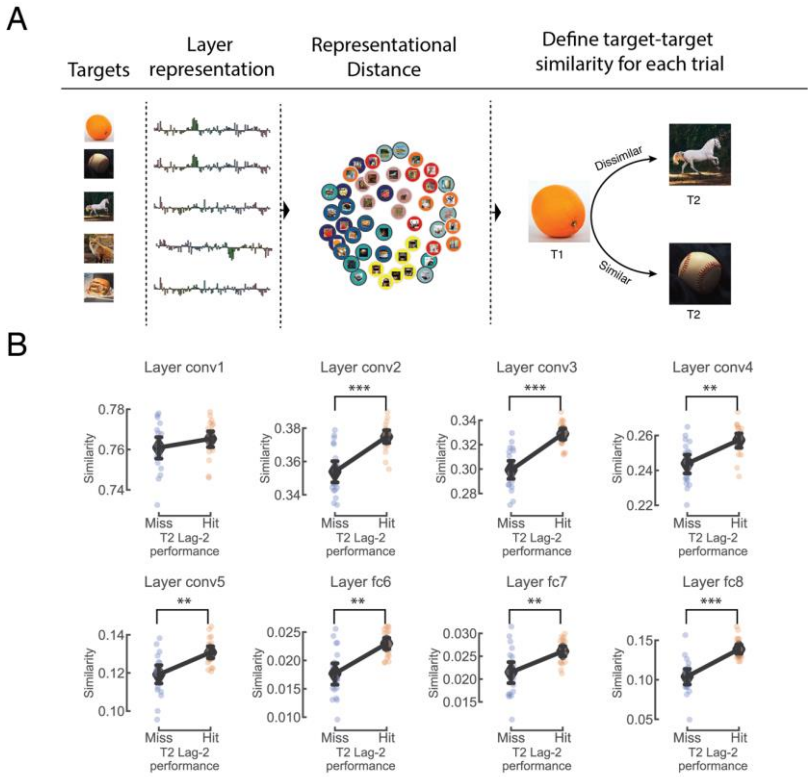


Figure 4. DCNN representational distance and target similarity explain trials of the AB. ***A)*** *Depiction of analysis procedure. For each layer, DCNN representations are extracted for each image. These feature activations were then compared for all image pairs (Pearson correlation), to estimate the similarity between pairs. Due to copyright reasons, all photos except for the faces (which were photographed by one of the authors) have been replaced by representational images.* ***B)*** *Mean similarity between T1 and T2, based on feature activation of each layer, for lag-2 missed and hit trials separately. Separate dots represent single subjects. The mean similarity across subjects is represented by a large black diamond and black bars denote 95% confidence interval. Two-tailed dependent t-test * = p < 0.05, ** = p < 0.01, *** = p < 0.001.*

Experiment 2

Constructing AB trials using representational sampling

The finding that T1-T2 similarity influences T2 performance prompted us to design a follow-up study. We sought to investigate the causal effect of target-target similarity by manipulating the targets' category and feature similarity. We developed a procedure called representational sampling, which first characterises a variety of stimulus response profiles, and samples a subset of stimuli tailored for our experiment. We used unit activations from layer 5 (see methods for rationale) of the DCNN as stimulus response profiles. We measured these unit activations on 250 images, derived from ImageNet[40], to yield 16 images as our T2s; in turn chosen to represent four categorical groups equally (mammals, insects, vehicles, and furniture). For each image we then selected two T1s based on category (same or different) and similarity within layer 5 (similar or dissimilar), resulting in eight T1s per T2. This allowed us to examine the specific contribution of high-level feature similarity and category membership separately. We presented these four conditions to 24 new participants in an AB task similar to that of Experiment 1.

Table 3: Mean and SDs for T2 performance in experiment 2.

T2|T1

Category	Similarity	Mean	SD	N
Same	Similar	0.85	0.10	24
	Dissimilar	0.81	0.09	24
Different	Similar	0.82	0.08	24
	Dissimilar	0.74	0.12	24

Table 3 shows the group means of T2 performance for each of the four conditions. The probability of correctly reporting T2 was the highest when T1 came from the same category and had similar visual feature activation in layer 5 of the DCNN (M = 0.849, SD = 0.097). In contrast, the lowest probability of correctly reporting T2 was observed when T1 came from a different category and was dissimilar (M = 0.741, SD = 0.123). A 2x2 (Category by Similarity) repeated measure ANOVA showed a significant main effect for both category ($F(1,23) = 20.68$, $p = <.001$, $\eta^2 = 0.473$) and similarity ($F(1,23) = 45.468$, $p = <.001$, $\eta^2 = 0.664$), as well as an interaction effect ($F(1,23) = 5.413$, $p = 0.029$, $\eta^2 = 0.191$). The larger effect size for the similarity factor indicates that visual features over semantic relevance determine behaviour.

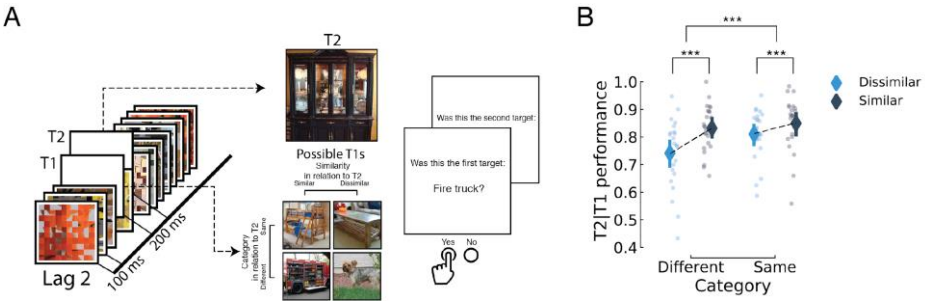


Figure 5. Target similarity between T1 and T2 explains T2 performance. ***A)*** *Representational sampling was used to construct trials of experiment 2. Each of the sixteen T2s were either preceded by a T1 from the same/different category and similar/dissimilar in representational space within layer 5 of the DCNN. To ensure that participants did not use low-level statistics (such as colour) when reporting the targets, we switched the response menu to a semantic task. B) Behavioural results from experiment 2. Our results show that features similarity explain a significant portion of T2 performance. Individual dots correspond to individual subjects. Error bars indicate the 95% confidence interval. Statistics were performed using a repeated measures ANOVA (see results). * = p < 0.05, ** = p < 0.01, *** = p < 0.001.*

Discussion

We investigated the effect of category membership and image features on conscious access using natural images in the Attentional Blink[19] paradigm (Fig 1A and B). By testing images spanning several categories we first show a clear division in performance between animate and inanimate objects, where animate objects reveal a reduced AB caused by the processing of the T1 (Fig 2B), in line with previous reports[9,10]. We further show that this bias is not only expressed between this super-ordinate division, but also extends to various sub-categories. Using a DCNN to model the stimulus visual features, we show that mid- and high-level features in natural images (Fig 3) regulate the AB magnitude. In addition, we show that target-target similarity (Fig 4 and 5) interacts with target selection, providing a mechanistic explanation of the AB phenomenon and of conscious access in object recognition.

Previous studies have shown differences between categories in the AB, most extensively between animate and inanimate objects[9–11,41]. The animacy bias in visual processing has been attributed to evolutionary relevance, as opposed to visual expertise, reflected in its importance for ancestral hunter-gatherer societies (The animate monitoring hypothesis)[42]. Evidence for this hypothesis comes from a wealth of behavioural studies showing that animate objects are more quickly and more often detected in different types of attentional tasks[7,42]. Likewise, animate and inanimate objects are distinctly represented throughout the ventral visual stream[8,43], which has been argued to be an evolutionary phenomenon and not contingent on visual experience[44]. In our current study, we find that the AB magnitude (ABM – performance difference between Lag-8 and Lag-2) is larger for inanimate objects, similar to Guerrero and Calvillo (2016)[10]. The finding by Guerrero and Calvillo has been contested by Hagen and Laeng (2017)[11] who showed that animate objects are simply reported more often, but that the ABM is unaffected. Our results argue against the findings of Hagen and Laeng and, more importantly, reveal that differences in AB magnitude exist in a myriad of sub-categories. Here we examine a significant number of categories, which

are known to cluster throughout the visual cortex. We show a high variance in the effect of the AB across categories (Fig 2C), implying that distinctive sub-categories have special privilege in the path to conscious access. One possible mechanism for categorical differences in conscious access can be related to the findings of Carlson et al. (2014)[12], who showed that animate objects that are neurally coded as more animate (as assessed by a decoding scheme) in the human analogous of inferior temporal cortex (hIT) are more quickly categorised as animate in a speeded discrimination task. Translated to our task, this would mean that certain categories are more distinctly represented, with less representational overlap to other images, leading to more robust processing of these categories. It is important to note that by looking at the differences between Lag-8 and Lag-2, effectively baselining each image with its own Lag-8 performance, our results cannot be explained by differential effects of masking. Importantly, this implies a dissociation between attentional relevance and conscious access, since it would be reasonable to assume that attentional relevance would affect Lag-2 and Lag-8 equally.

The finding that the ABM varies across categories (Fig 2C) is hard to interpret without properly examining image features of different complexities. Many semantic categories share low-level statistics[30,31] and, without delving further than categorical membership, one cannot disentangle at which level of processing the differences occur. The prediction of ABM across visual objects achieved by modelling DCNN unit activations from the mid to late layers explained a large proportion of AB variance across images (~46% of the variance in layer fc7, Fig 3C). This implies that the bottleneck produced by the AB is due to late visual processing and probably reflects the particular categorical organisations within higher-tier visual areas. This relationship between neural representation of images and behavioural outcomes is supported by recent work showing that the particular representational organisation in late visual areas predicts certain behavioural measures, such as reaction time[12,13,45]. This ‘conceptual’ approach to conscious access promotes a more fundamental view to how visual

consciousness might operate by focusing on the organisation of the visual system rather than on top-down mechanisms.

Our experiments further enabled us to explore the importance of T1-T2 similarity. Only a handful of studies have investigated target-target similarity in the context of AB[9,41,46–48]. In one of the earliest attempts to study target-target similarity and its effect on T2 performance, Awh et al. (2004)[46] concluded that similarity between targets is detrimental to T2 reportability. This led to the multiple-resource channel hypothesis (MRCH)[46]. According to the MRCH, two targets (T1 and T2) can be processed in parallel, but only if their visual features are different enough to be processed through distinct feature channels. While a few following studies have corroborated this notion[41,47,48], our study reveals that similarity is beneficial for performance. The difference in results might be explained by the way we define similarity by image features. Previous studies used categorical membership as a proxy to similarity, and thus it is possible that our findings reflect a facilitation effect not found in the previous studies (but see[9]). Importantly, while visual features function as stepping stones toward semantic meaning, it is unclear that such visual features would be maintained in working memory in our paradigm. Task-relevant similarity (i.e. the semantic content stored in working memory necessary to successfully carry out the task) between targets has been shown to be key for inducing a larger blink[48]. We would argue that the visual features within the DCNN models processes that precede working memory representations. As such, the target-target similarity rather enhances visual processing of T2, leading to a more probable recovery. The combined findings of all these studies highlight a relatively unexplored aspect of AB, where the relationship between the targets might play a significant role in explaining many AB phenomena. Further questions could be explored using a combination of brain measures to determine representational similarity within subjects, which might potentially also explain individual differences in performance.

In conclusion, we present compelling evidence that there are categorical differences in conscious access in object recognition. Specifically, we present findings that attribute differences in conscious

access between image exemplars to difficulties in representational readouts of features in higher-tier visual areas. This visual feature-related bias is reflected in a stable functional organisation, where fine-grained category distinctions have a larger impact on conscious access than previously believed. Moreover, we point to a more dynamic way in which the context (i.e. the similarity between T1 and T2) biases the probability for a target to be consciously perceived. In summary, our findings suggest that object categories and high-level visual features constrain conscious perception of natural images.

Methods

Experiment 1

Participants

Twenty participants (19 females; age range: 19-22; mean = 20.1 ± 1.2) were recruited for the study. We excluded two participants due to incomplete data. One additional participant was excluded for the image-by-image analyses due to lack of trials where T2 was correct for one image after filtering for T1 correct. All participants provided and signed informed consent and were rewarded for their time via course credits or financial compensation (at the standard rate of £7/h). All participants had normal or corrected-to-normal vision, and no known history of neurological disorders. The Ethical Review Committee of the University of Birmingham approved the experiment.

Procedure

Participants viewed visual objects in a rapid serial visual presentation (RSVP), and were asked to detect two targets (T1 and T2) embedded into a stream of distractors (Fig 1A). Following the stream, a response menu was presented for T1, which included the T1 and two foils, and the participant had to identify the target with a button press. A similar response menu was then presented to identify the T2. The foils in the menu always belonged to the same category as the targets (Fig 1A, right panel).

Design and Stimuli

Participants were seated 60 cm away from a Stone monitor (60Hz refresh rate), and stimuli covered 5 degrees of visual angle centrally on a grey background. Stimulus presentation was achieved using the Psychtoolbox extension (version 3; Brainard, 1997) in MATLAB 2016b (MathWorks Inc., Natick, USA). Stimuli consisted of 48 images, derived from eight different categories: fruits and vegetables, processed foods, objects, scenes, animal bodies, animal faces, human bodies, and human faces (Fig 1B). It's important to note that images were displayed in greyscale to reduce performance for human observers. To generate the items used as distractors in the stream, each image was divided into 5 x 5 (25 total) squares. Each square was then inverted and randomly assigned to a new square position. Following a standard Attentional Blink (AB) paradigm[19], each trial started with 300 ms of fixation, followed by a rapid serial visual presentation (RSVP) consisting of 19 images. Each image was presented for 16.7 ms with a stimulus-onset asynchrony (SOA) of 100 ms (Fig 1A). Embedded into the stream of distractors, two non-scrambled targets (T1 and T2) were presented at two different lag conditions (Lag-2: 200ms and Lag-8: 800ms). The T1 was always presented as item 5 in the stream, while T2 was either presented as item 7 (Lag 2) or item 13 (Lag 8). Each participant completed 12 runs (excluding one practice run of 5 trials). Across all runs each image was presented 12 times as T2 for both lags, for a total of 24 repetitions per image, and a total of 1152 trials. All 48 images were presented on an equal number of trials either as T1 or as T2, randomized within blocks with no trial having the T1 and T2 coming from the same superordinate category. Importantly, the same pair of T1 and T2 was always presented in both the Lag-2 and Lag-8 conditions, within the exact same stream of distractor masks in the RSVP trial. Participants had to press one out of three buttons to identify the correct target from the foils, or a fourth button when they missed the target. The two foils came from the same category as the target.

Deep Convolutional Neural Network (DCNN)

We employed a DCNN (AlexNet, see Fig 1C)[32], implemented through Python and Caffe[49], as a model of the visual cortex for extracting hierarchical visual features from our stimuli (we don't intend the use of model here to mean an exact biological model, but merely to approximate the hierarchical architecture that is known to exist in both). We chose AlexNet due to its relative simplicity, compared to more recent DCNNs, and its well-studied relation to the human visual system[33–36,43]. AlexNet consists of eight layers of artificial neurons stacked into a hierarchical architecture, where preceding layers feed-forward information to the next layer (Fig 1B). The first five layers are convolutional layers, whereas the last three are fully connected layers. While the fully connected layers (fc6, fc7, and fc8) consist of one-dimensional arrays (sizes of 4096, 4096, and 1000 units respectively), the convolutional layers have the dimensionalities of: layer 1 (conv1) - 96x55x55 (96 features, over 55 x 55 retinotopic units), layer 2 (conv2) – 256x27x27, layer 3 (conv3) – 384x13x13, layer 4 (conv4) – 384x13x13, and layer 5 (conv5) – 256x13x13. For all analyses we averaged the values in the convolutional layers for each image over the spatial dimension, leaving them with the vector length of 96, 256, 384, 384, and 256 respectively. This network was pre-trained on 1.3 million hand-labelled, natural images (ImageNet; Russakovsky et al., 2015) for classification into 1000 different categories (available at http://caffe.berkeleyvision.org/model_zoo.html), reaching near-human performance on image classification (Krizhevsky et al., 2012). Our test set of 48 images were analysed through the network, and we used the last processing stage of each layer as model output for further analyses. To keep the images as close to the training data as possible, and to avoid distortions of all levels of feature representations, the colour versions of the images were used.

Analyses of behaviour and image features

For each image, we calculated mean T2 accuracy at both Lag-2 and Lag-8 across subjects. We then computed attentional blink magnitudes (ABM) by subtracting Lag 2 mean accuracy from Lag 8 mean accuracy. ABM then becomes a measure of how much the AB

time window affects the recall of each image separately. In the interest of quantifying image features, within our DCNN, we extracted unit (neuron) activation patterns for each image from all the layers. For the first five convolutional layers, we averaged the activation over the spatial dimension. These activation patterns were incorporated into a multivariate linear regression model, with the activation patterns from each layer as features in the model to predict each image's ABM within subjects. The prediction pipeline followed a leave-one-image-out procedure (i.e., train on forty-seven images and test on one left out image) – where, based on the training data, the features were thresholded to have a larger variance than 0.15, to remove near-zero-varying features, and later standardised to unit variance with a mean of zero. Our choice of a threshold of 0.15 was arbitrary and had little to no effect when compared to only removing zero variance features. It's important to note that the test data was never part of any feature selection, as this would constitute double dipping. All pre-processing and fitting procedures were implemented using Sci-kit learn[50], for python code see [https://github.com/Charestlab/abdcnn].

Target-target similarity

We further tested the effect of target-target similarity on conscious access. Here, we go beyond using predetermined categories as a proxy for feature similarity and examine the representational distance between images within a given layer of the DCNN. For each layer we calculated the Pearson correlation between all possible T1-T2 pairs (Fig 4A), we then averaged the similarity for hit and miss trials separately. This allowed us to test the difference between hit and misses in terms of the relationship between the targets.

Experiment 2

Participants

We recruited 24 participants (Age - M= 19.38, SD = 0.95, females = 19, males = 5) with normal, or corrected-to-normal, vision. All participants provided and signed informed consent and were rewarded for their time via course credits or financial compensation (at the

standard rate of £7/h). The experiment was approved by the ethics committee at the University of Birmingham.

Procedure and stimuli

Unless stated otherwise, all procedure and visual presentations were identical to Experiment 1 (see fig 5A). Sixteen images, a subset of 250 labelled and processed images from the ImageNet database[40], were selected as T2s. The T2s derived from four different categories (mammals, insects, vehicles, and furniture), and each category was uniformly represented in the T2 selections. Similarity between images was determined by their Pearson correlation coefficient within layer 5 of the DCNN. The layer 5 was chosen because it was a high-performing layer in the first study and to still maintain the retinotopic information for an additional analysis not used in this study. To model the layer-wise unit activations for this new set of images, we used the same pre-trained network (AlexNet)[32] as in Experiment 1. For each T2, we selected two similar and two dissimilar images from the same category and any of the other categories as T1. This resulted in eight potential T1s for each T2 in a 2-by-2 factorial design (Similarity X Category) (Fig 5A). Each condition had the following mean Pearson correlation between T1 and T2, Same category/Similar layer 5 representation (Pearson r M = 0.43, SD = 0.114), Same category/Dissimilar (M = 0.136, SD = 0.113), Different category/Similar (M = 0.337, SD = 0.114) and Different category/Dissimilar (M = -0.056, SD = 0.099). T1 was always placed at position 11, and T2 at position 13 (in a RSVP of 19 items for each trial). Each block consisted of a presentation of each T2 paired with every possible T1, for a total of 128 trials per block divided into 4 runs (32 trials per run). Each participant completed 2 blocks for a total number of 256 trials per session (64 trials per condition).

Data availability

Supplementary data associated with this article can be found, in the online version, at [https://github.com/Charestlab/abdcnn/].

Code availability

Code associated with the manuscript is available at [https://github.com/Charestlab/abdcnn/].

References

1. DiCarlo, J. J., Yoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* 73, 415–434 (2012).
2. Felleman, D. J. & Van Essen, D. C. Distributed hierachical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47 (1991).
3. Ungerleider, L. G. & Haxby, J. V. 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* 4, 157–165 (1994).
4. Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D. & Kriegeskorte, N. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc. Natl. Acad. Sci.* 111, 14565–14570 (2014).
5. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat. Publ. Gr.* 17, 455–462 (2014).
6. Clarke, A. & Tyler, L. K. Object-Specific Semantic Coding in Human Perirhinal Cortex. *J. Neurosci.* 34, 4766–4775 (2014).
7. Jackson, R. E. & Calvillo, D. P. Evolutionary relevance facilitates visual information processing. *Evol. Psychol.* 11, 1011–1026 (2013).
8. Sha, L. *et al.* The Animacy Continuum in the Human Ventral Vision Pathway. *J. co* 27, 665–678 (2015).
9. Evans, K. K. & Treisman, A. Perception of objects in natural scenes: Is it really attention free? *J. Exp. Psychol. Hum. Percept. Perform.* 31, 1476–1492 (2005).
10. Guerrero, G. & Calvillo, D. P. Animacy increases second target reporting in a rapid serial visual presentation task. *Psychon. Bull. Rev.* 23, 1832–1838 (2016).
11. Hagen, T. & Laeng, B. Animals do not induce or reduce attentional blinking, but they are reported more accurately in a rapid serial visual presentation task. *Iperception.* 8, (2017).
12. Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S. & Ma, J. Reaction Time for Object Categorization Is Predicted by Representational Distance. *J. Cogn. Neurosci.* 26, 132–142 (2014).
13. Ritchie, J. B., Tovar, D. A. & Carlson, T. A. Emerging Object Representations in the Visual System Predict Reaction Times for Categorization. *PLoS Comput. Biol.* 11, 1–18 (2015).
14. Nairne, J. S., VanArsdall, J. E., Pandeirada, J. N. S., Cogdill, M. & LeBreton, J. M. Adaptive Memory: The Mnemonic Value of Animacy. *Psychol. Sci.* 24, 2099–2105 (2013).
15. Epstein, R., Harris, A., Stanley, D. & Kanwisher, N. The Parahippocampal Place Area : Recognition , Navigation , or Encoding ? 23, 115–125 (1999).
16. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17, 4302–11 (1997).
17. Downing, P. E., Jiang, Y., Shuman, M. & Kanwisher, N. A cortical area selective for visual processing of the human body. *Science* 293, 2470–3 (2001).
18. Martin, A. The Representation of Object Concepts in the Brain. *Annu. Rev. Psychol.* 58, 25–45 (2007).
19. Raymond, J. D., Shapiro, K. L. & Arnell, K. M. Temporary suppression of visual processing in a RSVP task: an attentional blink? *J. Exp. Psychol.* 18, 849–860 (1992).
20. Chun, M. M. & Potter, M. C. A Two-Stage Model for Multiple Target Detection in Rapid Serial Visual Presentation. *Journal of Experimental Psychology: Human Perception and Performance* 21, 109–127 (1995).
21. Dux, P. E. The attentional blink: A review of data and theory. *Atten. Percept.*

Psychophys. 71, 481–489 (2009).
22. Shapiro, K. L., Johnston, S. J., Vogels, W., Zaman, A. & Roberts, N. Increased functional magnetic resonance imaging activity during nonconscious perception in the attentional blink. *Neuroreport* 18, 341–345 (2007).
23. Luck, S. J., Vogel, E. K. & Shapiro, K. L. Word meanings can be accessed but not reported during the attentional blink. *Nature* 383, 616–618 (1996).
24. Marois, R., Yi, D. J. & Chun, M. M. The Neural Fate of Consciously Perceived and Missed Events in the Attentional Blink. *Neuron* 41, 465–472 (2004).
25. Sergent, C., Baillet, S. & Dehaene, S. Timing of the brain events underlying access to consciousness during the attentional blink. *Nat. Neurosci.* 8, 1391–1400 (2005).
26. Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J. & Sergent, C. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn. Sci.* 10, 204–211 (2006).
27. Fahrenfort, J. J., Scholte, H. S. & Lamme, V. A. F. Masking disrupts recurrent processing in human visual cortex. *J. Cogn. Neurosci.* 19, 1488–1497 (2009).
28. Harris, J. J., Schwarzkopf, D. S., Song, C., Bahrami, B. & Rees, G. Contextual illusions reveal the limit of unconscious visual processing. *Psychol. Sci.* 22, 399–405 (2011).
29. Kovács, G., Vogels, R. & Orban, G. A. Cortical correlate of pattern backward masking. *Proc. Natl. Acad. Sci. United States Am.* 92, 5587–5591 (1995).
30. Torralba, A. & Oliva, A. Statistics of natural image categories. *Netw. Comput. Neural Syst.* 14, 391–412 (2003).
31. Groen, I. I. A., Ghebreab, S., Lamme, V. A. F. & Scholte, H. S. Spatially Pooled Contrast Responses Predict Neural and Perceptual Similarity of Naturalistic Image Categories. *PLoS Comput. Biol.* 8, (2012).
32. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 1–9 (2012). doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007
33. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6, 1–13 (2016).
34. Eickenberg, M., Gramfort, A., Varoquaux, G. & Thirion, B. Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage* 152, 184–194 (2017).
35. Güçlü, U. & van Gerven, M. A. J. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Brain's Ventral Visual Pathway. 35, 10005–10014 (2014).
36. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput. Biol.* 10, (2014).
37. Greene, M. R. & Hansen, B. C. Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS Comput. Biol.* 14, (2018).
38. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron* 76, 1210–1224 (2012).
39. Kriegeskorte, N., Mur, M., Ruff, D. & Kiani, R. Matching categorical object representations in inferior temporal cortex of mand and monkey. *Neuron* 60, 1126–1141 (2008).
40. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 211–252 (2015).
41. Einhäuser, W., Koch, C. & Makeig, S. The duration of the attentional blink in natural scenes depends on stimulus category. *Vision Res.* 47, 597–607 (2007).
42. New, J., Cosmides, L. & Tooby, J. Category-specific attention for animals reflects ancestral priorities, not expertise. *Proc. Natl. Acad. Sci.* 104, 16598–16603 (2007).
43. Wen, H. *et al.* Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision. *Cereb. Cortex* 1–25 (2017). doi:10.1093/cercor/bhx268
44. Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M. & Caramazza, A. Category-Specific Organization in the Human Brain Does Not Require Visual Experience. *Neuron* 63, 397–405 (2009).
45. Grootswagers, T., Cichy, R. M. & Carlson, T. A. Finding decodable information that can

be read out in behaviour. *Neuroimage* 179, 252–262 (2018).
46. Awh, E. *et al.* Evidence against a central bottleneck during the attentional blink: Multiple channels for configural and featural processing. *Cogn. Psychol.* 48, 95–126 (2004).
47. Serences, J., Scolari, M. & Awh, E. Online response-selection and the attentional blink: Multiple-processing channels. *Vis. cogn.* 17, 531–554 (2009).
48. Sy, J. L. & Giesbrecht, B. Target-target similarity on the attentional blink: Task-relevance matters! *Vis. cogn.* 17, 1–10 (2009).
49. Jia, Y. *et al.* Caffe: Convolutional Architecture for Fast Feature Embedding. (2014). doi:10.1145/2647868.2654889
50. Buitinck, L. *et al.* API design for machine learning software: experiences from the scikit-learn project. 1–15 (2013).

Acknowledgements

This work was supported by a European Research Council (ERC) Starting Grant ERC-2017-StG 759432 (to I.C.). We would like to thank Sara Binks and Alfie Brown for their help in collecting the behavioural data from experiments 1 and 2 (respectively), Jasper van den Bosch, and Howard Bowman for comments on the manuscript.

Author contributions

S.A., D.L., and I.C. contributed to the design of the experiments. D.L. analysed the data. D.L., I.S., S.A., K.S., and I.C. contributed to the writing of the manuscript.
The authors report no conflict of interest.

Chapter 3

Sensory and semantic target-target interaction have opposite effect on performance in Attentional Blink

Authors:
Daniel Lindh[1,2,3]
Ilja G. Sligte[3,4]
Kimron L. Shapiro[1,2]
Ian Charest[1,2]

Affiliations:
[1]School of Psychology, Hills Building, University of Birmingham, B152TT Birmingham, United Kingdom

[2]Centre for Human Brain Health, University of Birmingham, United Kingdom

[3]Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands

[4]Amsterdam Brain and Cognition, University of Amsterdam, The Netherlands

Corresponding author: p.j.d.lindh@uva.nl

In preparation

Keywords:
Attention, Deep Convolutional Neural Networks, Consciousness, Object recognition

Abstract

Attention is a crucial component for our survival. By selectively attending to objects in our environment, we can allocate cognitive resources where they are needed. One way to investigate our attentional processing abilities is to push our senses to the limit by presenting consecutive stimuli at a fast pace, known as a rapid serial

visual presentation (RSVP). One common finding in RSVPs, is that when a stream of distractors has two targets embedded (T1 and T2, respectively), T2 is often omitted when placed 200-500 ms after T1, known as the Attentional Blink (AB). In a previous study we showed that when both targets share visual features, T2 performance is enhanced. This finding contrasts with repetition blindness (RB), a phenomenon where a direct repetition, or two targets that are similar in the task-relevant domain, often leads to additional impairments of T2 performance. One explanation to this incongruence could be related to how similarity between two targets is defined. The visual system follows a hierarchical structure, by extracting low-level features first in the early visual cortex. These features are later combined and aid the processing of more complex features until semantic properties emerge. This implies that targets can be similar at many stages of processing and investigating how similarities of targets at different levels of processing affect performance can provide novel insights into AB and RB. In a previous study we found that similarity in visual features between two targets increases T2 performance, in direct contrast to RB. Here, we investigate this apparent conflict between our findings and the literature by defining similarities between targets using functional magnetic resonance imaging, electroencephalography, and a convolutional neural network. We show that target similarity in low-level visual features, such as in V1, decreases the AB magnitude which corroborates our previous findings. We also find that similarities in late stages of processing increase the AB magnitude, in line with RB findings. Furthermore, we also show that individual differences in performance can be explained by a wider representational space in the right temporoparietal junction and inferior frontal gyrus. These findings elucidate how object recognition and conscious access is shaped by attention, concurrent processes, and the context in which objects are presented in. We discuss implications and further studies.

Keywords: Attention; Working memory; CNN; fMRI; EEG; Consciousness

Introduction

Every second our brain is flooded with an overwhelming amount of visual information from our environment. Our eyes move quickly over the visual field, sampling information several times per second. Therefore, humans have evolved to be exceptionally quick at recognizing objects in natural scenes. For example, studies have shown that we can make a saccade towards an animal in 120 ms[1] and extract semantic information from only 13 ms of exposure[2]. This rapid processing of objects is believed first to follow a feedforward hierarchical organisation[3,4]. Low-level image statistics are processed in posterior visual areas (e.g. in the primary visual cortex; V1) and progressively more complex visual features are then abstracted in multiple anterior brain areas with a distributed neural population that encodes semantic categories[5–7]. This initial forward stream is then followed by recurrent information from higher-tier visual areas re-entering lower visual areas[8–10] making it likely that subsequent fixations lead to overlapping object processing. There are several outstanding questions as to how the brain processes temporally adjacent information, especially when similar neural representations are evoked. Here we address several of these questions with a well-studied task to measure temporal attention using state-of-art methods to extract representational similarities between image pairs from functional magnetic resonance imaging (fMRI), electroencephalogram (EEG), and a deep convolutional neural network (DCNN)[11] trained on object categorisation.

A common method to investigate temporal processing is rapid serial visual presentation (RSVP), where one or more targets are embedded within a stream of distractors. Two well-known phenomena were discovered using this approach, the attentional blink (AB)[12] and repetition blindness (RB)[13,14]. In the AB, a lapse of attention is generated when the second target (T2) is presented 200-500 ms after the first target (T1). Due to the ongoing processing of T1, participants

are often unable to report T2[15]. Specifically, when subjects are asked to ignore T1, the T2 performance improves significantly[12]. This implies that deliberate processing of T1 interferes with T2, however, the exact neural mechanism by which this interference occurs is unknown. Similar to AB, RB describes a phenomenon where T2 is missed due to a repetition of the target, where theoretical accounts suggest that task-relevant dimensions need to be repeated for RB to occur[13,16,17]. For example, Sy and Giesbrecht showed that when gender was the task-relevant feature, the feature to be reported, repeating the emotional content (for example two happy faces) did not negatively affect performance. However, performance was impaired when both T1 and T2 were of the same gender. This implies that RB is not a low-level repetition suppression effect evoked by an exact repetition of the stimulus, but rather is due to repetition in higher-level representations that are related to the task goals and a failure of separating targets into separate working memory representations.

Interestingly, in contrast to findings from RB, Lindh et al. (2019)[18] found that targets that share low-level visual features can be beneficial for reportability of T2. Previous studies that have examined target-target similarity have opted to use simple heuristics such as categorical adherence as a proxy for similarity[16,19]. For example, two target images containing horses can be similar at the level of semantics but depending on viewpoint they are not necessarily similar in low-level visual features. Therefore, these methods are likely to be limited in how well they can describe actual overlap in representational space in the brain. In contrast, Lindh et al. (2019)[18] defined similarities between targets using different layers of a deep convolutional neural network (DCNN)[11]. This DCNN was trained on object categorisation within natural images, which have been shown to lead to a remarkable similarity in representational geometry to the human inferior temporal cortex[20], with early and late layers of the DCNN corresponding to the posterior-anterior hierarchical ordering found in the visual system[21–24]. The usage of the DCNN and the complexity of the scenes could provide a different continuum of image similarity, which had not been considered before. In a multi-target RSVP where several objects are processed simultaneously, it is inevitable that the processing of these

targets interact at some level within the visual hierarchy. Given that AB seems to be driven by deliberate processing of T1[12], one contributing process could be an interaction between the two targets. Studying target-target similarity at different processing levels in the human brain could thus provide insights in how overlapping target representations affect conscious access to visual stimuli.

To fill this gap in our knowledge, we examine to what extent representational similarity between targets affects reportability. We define similarity using a broad spectrum of modalities, including fMRI, EEG, and DCNN. Critically, all three modalities provide a hierarchical account of object processing: brain activity patterns can reflect increasing complexity in space (fMRI; with a posterior-to-anterior gradient) and in time (EEG), and in units (DCNN) as the latter respond to increasingly complex features with increasing network depth, with a gradient from low-level features to high-level features across layers. This allows an examination of the full breadth of possible target-target representational relationships to an unprecedented degree. In the present research we ask how (1) target-target similarity correlates with behavioural performance trial-by-trial, (2) how overall similarity affects AB magnitude for an image and (3) how representational similarity can be used to explain individual differences in AB. To foreshadow, our results merge previous incongruent findings in regards to RB[13,16,25], where target-target similarity leads to an impairment of performance, and to the conclusions drawn by Lindh et al. (2019)[18], where target-target similarity improves performance. We show that similarity in low-level visual features enhances the probability of successful report, while representational overlap in late processing stages leads to an attenuation of conscious access to a second target occurring 200 ms after the first. Importantly, this analysis explains significant variance in both trial-by-trial performance and differences in AB magnitude between images. These results are stable to a large degree regardless of if we use fMRI, EEG or the DCNN to define target-target similarities. We also find that individuals who have larger distances in the representational space of images in the right temporoparietal junction (rTPJ) and right inferior frontal gyrus (rIFG), two brain areas commonly

associated with working memory updating[26], show less of an AB impairment.

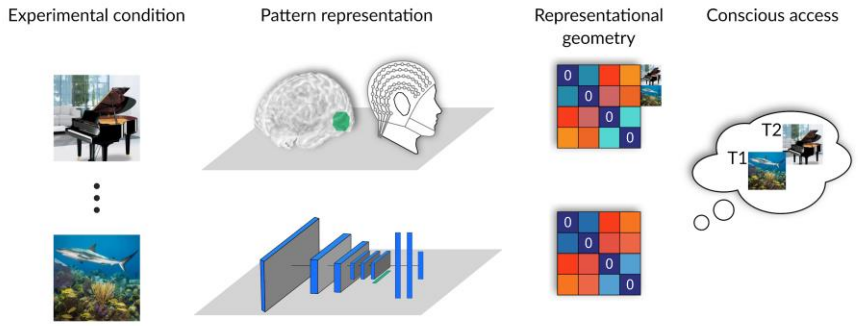


Figure 1. *Predicting conscious access from pattern-similarity estimates. Different stages of processing for each image were captured using fMRI, EEG and a deep convolutional neural network. We extracted the pairwise comparison for all images and used these similarity values to relate to conscious access in three different levels of analysis: lag-2 performance, attentional blink magnitude per image, and individual differences in conscious access performance.*

Results

Previous studies have shown that similarities between T1 and T2 can both facilitate[18] and impede conscious access [13,14,16,19,25]. An important difference between these studies involves the level of complexity used in defining similarity. For example, the improved performance of the T2 report was shown using visual features derived from a DCNN while (visual features in contrast to semantic information). To investigate the role of target-target similarity in conscious access we used multivariate representations of images in fMRI, EEG and DCNNs, providing a precise account of similarity at various levels of description (Figure 1). Specifically, we test whether similarity in early- (V1, early EEG time points and early DCNN layers) vs late-object processing stages

(ventral stream, late EEG time points and late DCNN layers) yields different accounts of the role of target-target similarity for conscious access. Participants (n=16) viewed natural scenes depicting visual objects (animate and inanimate objects) in four sessions of EEG while performing an attentional blink task (Figure 2A). Whole-brain fMRI (3T, $3mm^3$; TR=0.764s; multi-band 4) data were collected in two separate sessions, while participants performed a simple working memory task (Figure 2B). The DCNN was a convolutional neuronal network with 5 convolutional layers and 2 fully connected layers trained on object recognition[11]. Critically, the same natural scenes were used in the fMRI and EEG experiments, and modelled through the DCNN, enabling the use of RSA for comparisons to the attentional blink behavioural data.

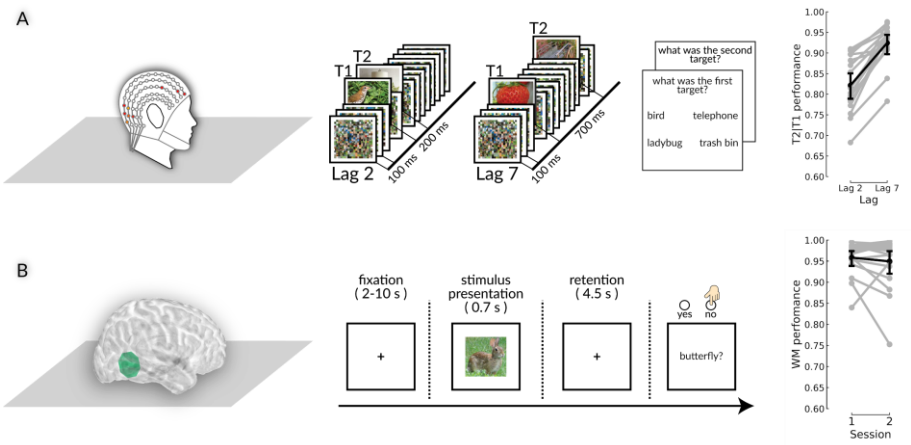


Figure 2. ***A)*** *Attentional Blink (AB) paradigm. Targets (T1 and T2) were embedded in a rapid serial visual presentation of scrambled mask distractors. We manipulated two lags of the AB. For Lag 2 trials, the T2 was presented 200ms following T1 and the two targets were separated by 1 distractor. For Lag 7 Trials, T2 was presented 700ms after T1 (with 6 distractors in between). Participants then had to report the identity of both targets. The behavioural performance for lag-2 and lag-7 shows a significant difference in T2 performance between the two lags, indicative of an AB effect.* ***B)*** *The working*

memory task in the scanner. Participants performed a simple memory task while we collected brain activity patterns using fMRI. Each trial started with 2-10 seconds of fixation, followed by a brief presentation of an image (0.7 seconds), a 4.5-second-long retention period and finally a response menu where participants were asked whether or not the word shown corresponded with the centrally positioned object in the image. We observed no significant difference between the memory performance between the scanning sessions indicating that participants could reliably do the task in both sessions.

Attentional Blink behaviour

The attentional blink task consisted of a rapid serial visual presentation, where two target images were embedded in a stream of scrambled distractors. Each session consisted of 8 blocks of 120 trials. Each trial started with 1.25 seconds of fixation, followed by a stream of 19 images in rapid succession (one frame every 16.7 ms). Within the stream, the two targets (T1 and T2) were presented at either 200 ms (Lag-2) or 700 ms (Lag-7) apart. The residual 17 images were mask distractors constructed by combining random images (see online methods). Participants showed a higher T2 performance at lag 7 ($M = 0.93$, $SD = 0.068$) than at lag 2 ($M = 0.823$, $SD = 0.05$, $t(15) = -7.79$, $p < 0.001$, Figure 2A), indicating the commonly found attentional blink effect (Figure 2). In the scanner, participants completed a working memory task (Figure 2B), specifically designed for the low temporal resolution of fMRI. We used this task to characterise object representations from early visual perception to conscious access, emulating the stages of processing known in the attentional blink task but with a temporal resolution that leads to better SNR in fMRI. To avoid fatigue and discomfort, the ethical committee of University of Amsterdam allows for a maximum of 90 minutes in the scanner per session. Therefore, we divided up our working memory task in the scanner into two 1-hour sessions with 1 hour rest in between. We observed no difference in performance between session 1 ($M = 0.96$, $SD = 0.04$) and session 2 ($M = 0.95$, $SD = 0.06$, $t(16) = 0.75$, $p = 0.46$, Figure 2B).

EEG decoding of attentional blink targets

Before measuring target-target similarity based on the scalp activity in the EEG trials, we trained a shrinkage linear discriminant classifier[27] to decode the identity of the targets presented in the attentional blink. We used a k-fold cross-validation procedure applied to each EEG time-point separately. The classifier was trained on T1 trials, where there is the least disruption of the EEG signals, and tested on either the T1, T2 presented at lag 2, or T2 presented at lag 7 (see online methods). We trained and tested the linear classifier across all possible pairs of conditions that we presented as targets in the AB, and here we report the average decoding accuracy across all pairs (see Supplementary Figure 1). The topographies elicited by the attentional blink targets provided enough representational detail to decode the identity of the targets in all conditions of the AB in a time-window starting at around 80 ms until around 650 ms post-target onset. Moreover, we observed significantly greater decoding accuracies for the T2 presented at lag 7 (in contrast to the T2 presented at lag 2) between ~160 ms and 620 ms post-target onset.

Target-target similarity explains intertrial differences in T2 performance

For each trial, we calculated the representational pattern similarity between T1 and T2 in the spatial (fMRI), hierarchical (DCNN) and temporal (EEG) domain (Figure 3A). Using 40 natural images that could either be the T1 or the T2, but without ever repeating the same image twice, we end up with 780 different T1-T2 combinations. For each participant, we binned the trials into 20 bins based on the similarity between T1 and T2. Then we averaged the T1-T2 similarity and the hit-rate within each bin and correlated the two measures for each participant independently using Pearson's correlation. We tested the resulting correlation coefficients against 0 using a two-tailed one-sample T-test. For fMRI, we found that trial performance and T1-T2 similarity in V1 showed a significant positive correlation ($M = 0.039$, $SD = 0.037$, $t(15) = 3.927$, $p = 0.002$), while Ventral Stream ($M = -0.076$, $SD = 0.054$, $t(15) = -5.342$, $p < 0.001$) showed a significant

negative correlation. In EEG, a cluster-level permutation test (5000 permutations, t-threshold for defining clusters = 3) revealed a significant negative correlation around ~200 ms, peak correlation was at 196 ms, with M = -0.503, SD = 0.215, in line with late processing and confirming the finding in late visual areas in fMRI. Similar to V1 in fMRI, in early layers of the CNN we find that target-target representational similarity correlated positively for layer 1 (M = 0.103, SD = 0.064, $t(15) = 6.042$, $p < 0.001$). However, we do not find that later layers show an opposite effect, as with high-level processing areas of the brain. All p-values are FDR corrected for multiple comparisons. In regard to our fMRI results, we recognize a discrepancy between the EEG (showing similar results as the Ventral Stream ROI using fMRI) and the DCNN (analogous to the V1 result using fMRI). In combination, our fMRI results suggest that low-level interaction of the targets increases the probability of T2 to be perceived while high-level representational overlap interferes with the processing of T2. While the EEG data confirmed the late processing interference, the CNN data confirmed the early process facilitation of T2.

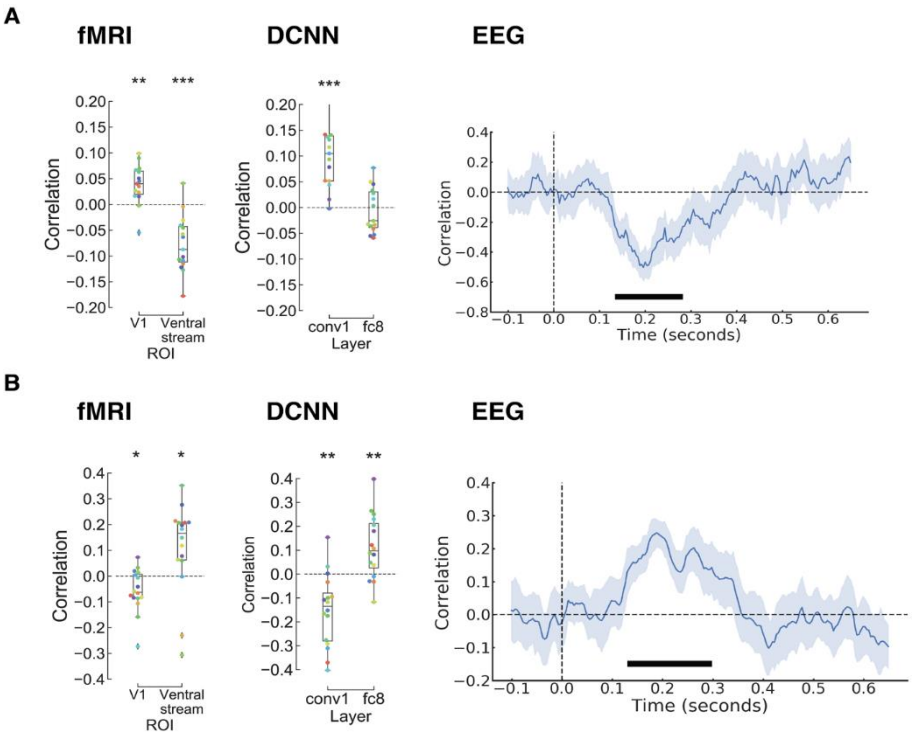


Figure 3. fMRI, DCNN and EEG results. *Early and late processing of objects is presented for all modalities - V1, layer 1 and early EEG time points reflect early processes while ventral stream, layer 8 and late EEG time points reflect late processes. A). Lag-2 T2 performance and T1-T2 similarity correlations. Left, V1 and ventral stream, capturing low-level and high-level visual processing respectively. When T1 and T2 are similar in V1, we see an increased ability to recover T2. This is contrasted when T1-T2 are similar in higher visual areas, i.e. ventral stream. Middle, layer 1 of the DCNN show the same results as V1 with the fMRI data. However, layer 8 does not replicate the reverse effect as seen in the ventral stream. Right, EEG time series. T2 performance has a negative correlation with target-target similarity during 120ms-300ms, mirroring the results of the negative behavioural effect of target-target similarity found in the ventral stream. B) Image specific Attentional Blink Magnitude (ABM: lag 7 performance - lag 2 performance) correlates with an image's overall similarity to all other images. Left, fMRI results show that images that are in general similar to other images in V1 show a lower ABM, while images that are in general similar to other images in the ventral stream have larger ABM. Middle, DCNN results show identical patterns for layer 1 and layer 8 as with fMRI. Right, EEG results mirror the*

*result of the ventral stream in fMRI and layer 8 in the DCNN with images that are generally like other images at around 120-300 ms show a larger ABM. Shaded areas indicate 95% confidence interval. The black bar denotes time point clusters significant from zero. * = p < 0.05, ** = p < 0.01, *** = p < 0.001*

Overall image distinctiveness as an explanation for differences in ABM

Building on the finding that target-target similarity affects T2 processing, we set out to investigate if the representational distinctiveness of an image can explain why some images seem to be more sensitive to the AB window (i.e., the temporal window at which conscious access to the T2 is impaired). To test this, we calculated the Attentional Blink Magnitude (ABM) for each image by comparing the performance at lag-7 to the performance at lag-2. This way we are baselining each image with its performance outside the AB window, which sets this measurement apart from simply looking at lag-2 T2 performance as in the previous section. Based on pattern representations for each ROI in fMRI, a time point in EEG, and a layer of CNN, we calculated the average similarity of one image in respect to all other images (Figure 3B). This yielded one value per image, indicating how similar this image is overall to the rest of the data set. We then correlated the average similarity with the ABM for each image within participants and tested the correlation coefficient against zero using a one-sample T-test. In accordance with the previous analysis, we find a positive correlation between overall distinctiveness and ABM in V1 ($M = 0.056$, $SD = 0.081$, $t(15) = 2.683$, $p = 0.026$), indicating that images that are generally similar to other images in V1 show less of an ABM. Conversely, we find robust positive correlations in the ventral stream ($M = -0.111$, $SD = 0.167$, $t(15) = -2.570$, $p = 0.027$). In EEG, a cluster permutation test (5000 permutations, t-threshold for cluster definition = 3) showed that images that are generally similar to other images in a time window between ~150 – 210 ms (peak time = 188 ms, $M = 0.248$, $SD = 0.091$) exhibit a larger ABM (Figure 3B), consistent with when late visual processing is predicted in higher-order processing brain areas. We tested the layers 1 and 8 (representing initial low-level and later more category related visual features) in the

CNN with the same methods as the ROIs in fMRI. Identically to fMRI, we find a negative correlation between average similarity and ABM in layer 1 ($M = -0.150$, $SD = 0.147$, $t(15) = -3.963$, $p = 0.003$), while layer 8 ($M = 0.114$, $SD = 0.130$, $t(15) = 3.398$, $p = 0.008$) showed a positive correlation. All p-values are corrected for multiple comparisons using FDR. Even though ABM is a different measure than T2 performance in lag-2, these results mirror each other since a low ABM score is equivalent to better performance while a low T2 performance in lag-2 is indicative of a bad performance.

Individual differences in conscious access

One question that has intrigued researchers in attention literature is the finding that some participants don't exhibit an AB[28–30] (Figure 4A). Importantly, looking at individual differences has been argued to be a promising method to understand the processes underlying the AB[28]. Here we investigated the notion of *representational richness* as a factor in explaining the large variability between subjects in the typical AB task. Specifically, we tested the overall similarity between all objects for a given region of the brain related to participants ABM. Here, a large representational richness would be reflected in large differences in the neural representations between objects. Using a searchlight procedure on the fMRI data, iterating over each brain voxel as a centre for a sphere, we averaged the representational similarities (measured using Pearson's correlation) for all pairwise comparisons of the activity patterns elicited by our object stimuli. This average representational similarity score provided a representational richness index in each spherical searchlight for each participant. Across participants, for each sphere, we correlated the representational richness indices to the participants' ABM. The searchlight revealed five main clusters positioned on the right hemisphere of the brain (Figure 4, MAX R = 0.78; MNI 1, 36, 30; see supplementary table 1 for all regions and MNI coordinates, False Discovery Rate cluster forming threshold = 0.01, cluster threshold = 20). Specifically, we found that participants with more representational richness in the right temporoparietal junction (rTPJ) and right inferior frontal gyrus (rIFG)

are more vulnerable to the attentional blink, and conversely, participants with rich and decodable representations in this network perform better in the AB (Figure 4). This network is primarily known for its putative role in bottom-up saliency[26], but has also been specifically noted in the AB[31–34].

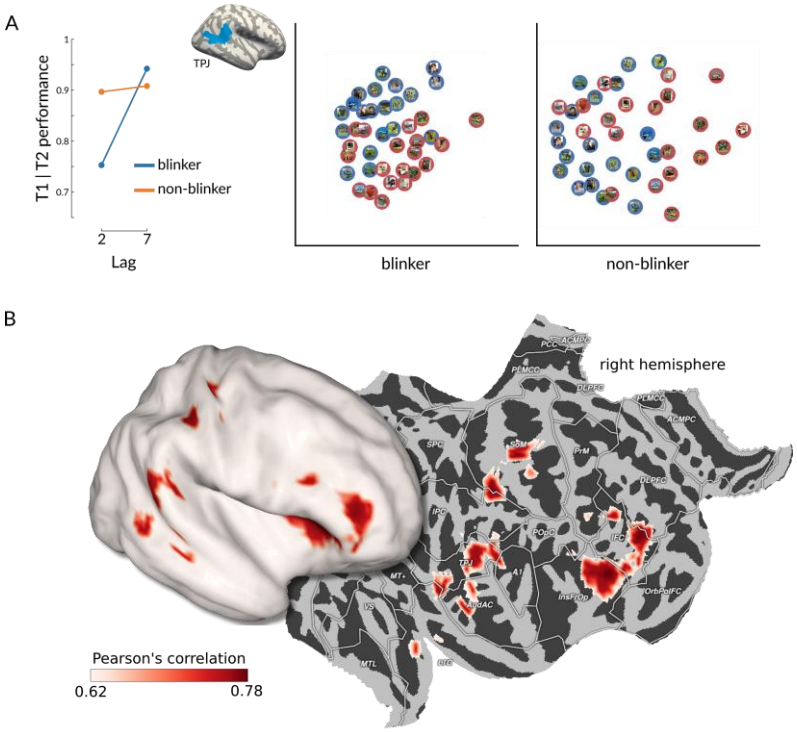


Figure 4. Individual differences. ***A)*** *(left) Example participants in the attentional blink task. The blue line shows the performance of a typical "blinker" participant, and the orange line shows the performance of a typical "non-blinker" participant. The right panel shows the multidimensional scaling projection on a two-dimensional plane of the representational dissimilarity matrix measured in a temporoparietal junction region of interest in each participant. Blue circles indicate animal and red circles indicate a non-animal. B) We performed a searchlight analysis to relate individual differences in representational richness to participants' performance in the attentional blink. In each volumetric spherical searchlight, we measured Pearson's correlation between the average representational distance and the attentional blink magnitude measured from each participant. The resulting correlation map*

was corrected for multiple comparisons using a False Discovery Rate procedure (cluster forming threshold = 0.01, cluster threshold = 20). The searchlight analysis revealed a right-lateralized distributed network of cortical areas, including anterior inferotemporal cortex, temporoparietal junction, supramarginal gyrus, and inferior frontal cortex where the representational richness of a participant predicts that participant's performance at the attentional blink task.

Discussion

In the current study we measured brain activity from functional magnetic resonance imaging and electroencephalography to investigate conscious access in object recognition. Specifically, we measured representational geometries at various levels of processing using representational similarity analysis (RSA) applied to fMRI data (using V1 and the ventral stream ROIs as representative of "early" and "late" visual processing respectively) and EEG data (at different moments in time following target presentation). We further completed this analysis framework with RSA applied to layers of a deep convolutional neural network. We tested the hypothesis of a relationship between target similarities and conscious access in the attentional blink. Previous work has established that similarity between T1 and T2 in a RSVP can lead to both increased[18] or decreased[13,14,16,19,25,35] likelihood of consciously perceiving the second target. We show that these two effects are dependent on where in the hierarchical stage of processing the targets are interacting. Critically, we show that our representational similarity framework can be used to explain three core components of conscious access in object recognition. First, we show that low-level and high-level similarities between T1 and T2 across trials predicts the likelihood of detecting the second target in opposite directions, where similarity in V1 increases T2 performance while similarity in the late ventral stream decreases T2 performance (Figure 3A). Second, image-specific brain activity patterns account for the attentional blink variability across stimulus conditions (Figure 3B). Third, representational richness measured in a core network involved in bottom-up attention explains individual

differences in conscious access. Altogether, our results provide a comprehensive view of the underlying mechanisms supporting conscious access in object recognition at the levels of processing in the brain.

In the decades-long history of rapid serial visual presentation (RSVP) research, two main phenomena have been established: the attentional blink (AB) and repetition blindness (RB). The AB[12] is by far the most prominent of these two, with hundreds of papers being published every year either using the AB as a method to induce failures to report T2 or to understand the mechanism underlying AB. The AB is defined as the inability to perceive a second target (T2) in a stream of distractors when the first target (T1) precedes it by 200-500 ms. On the other hand, RB[14] is defined as when subjects are unable to recollect the second target if it is a repetition of the first target. This effect has later been extended to not necessarily be a repetition of the exact same stimuli, but has also been shown to be present when two words are homophones (e.g. allowed/aloud), when two words from different languages describe the same concept (e.g. Caballo/Horse)[36], two visual objects from different angles[25] and rotated images[37]. These studies indicate that it is not the perceptual similarities between targets that impede performance, but rather semantic relationships related to the task. In fact, task-relevance has been shown to be imperative for the repetition blindness effect[16,17], implying that this is a working memory related phenomenon and not perceptual. Conversely, we have shown in a previous study[18] that similarity between targets can also be beneficial for performance. In this current study, we explain the discrepancies between earlier findings by first showing that trials where T1 and T2 have a similar multivariate representation in V1 and the first layer of the DCNN also lead to a higher probability of successful T2 report (Figure 4A). This effect could not be found in the temporal EEG data, which might reflect difficulties in differentiating these early signals from other perceptual processes. However, we would argue that this is a reliable finding considering that we find this effect in two out of three modalities (fMRI and DCNN) and that we

have found a similar effect in a previous study using a DCNN[18]. Second, we find that similarities between T1 and T2 in the ventral stream and EEG time-points between 120-290 ms has a detrimental effect on T2 performance (Figure 4A). This is in line with earlier findings of RB[13,14,35], however, to the best of our knowledge we are the first to connect similarities of brain representations to this specific effect. Sy and Giesbrecht (2009) showed that this repetition effect is dependent on task-relevance by demonstrating that subjects are less likely to remember that T2 was a male if T1 also was a male face, but only when this was the to-be-reported dimension. Since our experiment has semantic task-demands, it corroborates the notion that T1-T2 similarities in later stages interfere with T2 working memory updating. While RB has been described as an inability to individualize a separate episodic token for T2 when it is similar to T1[14] and computational models of AB[38,39] have proposed that T2 is being attenuated to protect the target to be reported (i.e. the working memory representation of T1) it is unclear as to why T1-T2 similarity in low-level visual features would lead to larger probability of perceiving T2. Previous studies have shown a delayed working memory engagement of T2 at short lags[40,41], indicating that the T2 representation lies dormant until resources are freed up and working memory can be engaged to encode T2. If the T2 neural trace must be maintained within lower-level visual areas until WM resources are freed up, recurrent information from higher-order processing of T1 might interfere with the T2 pattern. Considering the fragility of the percept, this interference would logically be more severe if the two targets are very different at this stage of representation. In other words, there might be a cost associated with having ambiguous information represented in lower-level visual areas, however, further studies are needed to answer this. Aggregated, our data support the notion of repetition blindness, where target-target similarities in representational space during late stages of processing affect T2 performance negatively, while a shared representational space between targets in early visual processing is beneficial for T2 performance.

Previous studies have shown that certain types of stimuli are less

affected by the AB window. For example, the AB can be modulated by animacy[18], emotional content[17], and by attentional biases such as gambling-related stimuli in gamblers[42]. Following our finding that T1-T2 similarity at different levels of processing affect T2 performance we set out the test if the Attentional Blink Magnitude (ABM: T2|T1 Lag-7 performance - T2|T1 Lag-2 performance) for one image is related to how similar a particular image is to all other images. By using the ABM, we are effectively baselining each image's performance during lag-2 with its performance at lag-7. This way, when comparing between images, we can assure that the difference between images is not due to variability in how our specific choice of masks interfered with the processing of one image. We find that the ABM of an image is modulated by its overall similarity to all other images (Figure 4B). Images that in general share a lot of low-level (V1 in fMRI and Layer 1 of the DCNN) similarities with other images are less affected and images that are like other images in high-level (ventral stream in fMRI, 120-290 ms time points in EEG and layer 8 of the DCNN) representations are more affected by the AB window. In addition, we also conducted a searchlight[43] procedure which revealed an extensive language/reading-related network focused on the dorsal visual stream, the left temporal cortex, and the left inferior frontal gyrus (Supplementary Figure 3). Many of the areas that showed a positive correlation with ABM have a history in the language processing literature such as the Visual Word Form Area[44], Wernicke's Area, and Broca's Area[45]. Given the semantic nature of our task, these are areas where you would expect relevant processing for working memory functions and where interference of tokenization[14] for T2 would occur. The left oriented network we found also carries resemblances to the semantic control/episodic network [46] and the top-down control network as described by Corbetta and colleagues[26,47], which is proposed to enable selection of goal-driven stimulus processing. The overlap between the top-down control and semantic network implies that representational similarities are malleable depending on task-demands, which is in line with previous studies showing the importance of task-relevance when investigating the different effects of target repetition[16] and emotional processing[17] in RSVPs. These findings expand on previous studies and underline the effect of

concurrent processing of T1 on conscious perception of T2 and how it can explain variability in performance between different types of stimulus inputs.

Another method that has been suggested to shed light on the mechanism of the AB is to evaluate individual differences in performance [28,29,48]. Throughout the three decades of investigating the AB, studies have found that a significant proportion of the population seem almost unaffected by the AB and have thus been termed "non-blinkers" [28,29]. Investigating how these individuals differ from others is crucial to understand to explain the processes that underlie the AB phenomenon. Earlier reports on individual differences have suggested that non-blinkers show a faster peak of the P3, indicating that they are quicker to consolidate information into working memory [29]. However, some previous research has indicated that cognitive processing speed is not what best describes individual differences[49]. For example, vocal naming tasks[50] and fluid intelligence[51] do not predict individual differences in ABM. On the contrary, executive control functions[51,52] and being able to filter out irrelevant stimuli[53] can significantly predict individual differences. A larger ABM has also been observed in patients with lesions in the right inferior parietal lobe, overlapping with the right temporoparietal junction (TPJ) [54]. Several studies have implicated the right TPJ in AB performance, where grey matter density in the right TPJ[55], connectivity between right TPJ and inferior frontal gyrus[31,55] and transcranial magnetic stimulation (TMS) on the right TPJ all modulate performance in the AB[33,56]. Other lesion studies of the right TPJ show that some patients develop visual extinction, a phenomenon that describes the unsuccessful perception of contralesionally events during competition between the two visual hemifields (in contrast to temporal competition) [57]. That is, when one item is being shown in each visual hemifield, patients report no awareness of the item presented in the left hemifield, an effect that is amplified when the two targets are the same in the task-relevant domain [58]. This post-perceptual role of the right TPJ is also noted in the change detection literature where semantically incongruent changes in a change detection task are more often detected than

when scene congruent items are added. Importantly, this difference between congruent and incongruent change detection was eradicated with TMS to the right TPJ[59]. In line with the literature, we find that individual differences in ABM were related to subjects' average representational similarity in right TPJ and right IFG (Figure 4, Table S2). These two areas are closely associated with the stimulus-driven control network as proposed by Corbetta and colleagues[26,47]. Rather than being activated by expectations, this network responds to the detection of task-relevant stimuli, see Corbetta et al., 2008 for review[26]. This is very similar to how the P3 has been described in the literature, which is one of the reasons why it has been argued that the P3 originates from the TPJ[60]. Thus, our results corroborate earlier findings that non-blinkers show faster working memory updating[29] and attribute this to the separation of image representations, and therefore unambiguous target separation during working memory encoding, in the ventral attentional network[26]. Specifically, we argue that subjects who have a larger representational space for objects in this key brain network for working memory updating are quicker at resolving object identity and can consequently evade the attentional blink window. This argument is further supported by findings showing that the slope of the P3 is related to evidence accumulation[61], however, future studies should also investigate if representational space in the ventral attentional network explains individual differences in evidence accumulation.

In conclusion, we show that not only can representational overlaps explain trial-by-trial variance but also explain why some objects are more probable to reach conscious processing in AB. We conclude that target-target similarity can both have a positive and a negative effect on performance and this depends on the stage of processing in which the targets are interacting. While repetition blindness related effects have been studied to some extent, more research is needed in situations when target-target interaction leads to positive performance. It is unclear as to what type of processes are affected by these target interactions, if they are perceptual or non-perceptual by nature, and future studies could potentially investigate this using speeded judgment tasks rather than a report after the RSVP. Also, we also

propose object separation in the right ventral attentional network[26] as a viable explanation for individual differences in the attentional blink. In line with recent work showing that the distances captured with multivariate methods using fMRI is indeed related to the decision-function used when humans do animate/inanimate speeded categorisation[62], we show that the representational similarities captured by all three of our modalities have substantial power in explaining variance in performance at multiple levels.

Methods

20 participants (mean age = 23, range = 18 to 44, 13 females) participated in the study. Participants completed 4 sessions of the attentional blink task (while we recorded EEG, see below) and two sessions of functional magnetic resonance imaging (fMRI). Three participants did not complete all conditions and were thus excluded from further data analyses. All participants provided informed consent and were compensated for their time (at the rate of €10 per hour for EEG, €20 per hour for fMRI, and €50 for completion for a total of €210). The experiment was approved by the ethics committee at the University of Amsterdam.

Stimuli

The visual objects presented in both tasks consisted of forty natural scene images depicting objects positioned in the centre of the image (twenty animals, twenty non-animals) from the ImageNet database[63]. The final set of 40 stimuli was chosen to have a proclivity for high blink rate in a pilot experiment (see supplementary materials). The experiment was programmed using Psychtoolbox Version 3 (PTB-3; MATLAB and Statistics Toolbox Release 2016, The MathWorks, Inc, Natick, Massachusetts, United States). The distractors were created by dividing an empty image up into a 10x10 grid (each grid cell containing XX pixels, equating to roughly 0.5 visual degrees) and then sample image information from the corresponding grid cell from a random image of our set.

Attentional Blink task

Participants were comfortably sitting in front of a 19" monitor positioned at 60 cm. Targets and distractors were displayed in the centre of the screen subtending 5 degrees of visual angle on a constant grey background. At the beginning of each trial, participants attended a white fixation cross for 1.25s. This was followed by a stream of 19 images (17 distractors and 2 targets). Images were shown for 16.67 ms with a stimulus onset asynchrony (SOA) of 100 ms. The first target (T1) was randomly presented at position 4, 5 or 6 in the stream and the second target (T2) was presented either two (lag 2) or seven (lag 7) items further away (Figure 1A).

After each trial, participants were prompted with a response menu for T1 and asked to choose which of the four possible words corresponded to the first target (Figure 1A). Following this, a similar menu was displayed for T2. T1 and T2 were never the same image. Participants completed four sessions of approximately three hours of AB (including EEG preparation). Each session comprised 8 runs of 120 trials, where each image was presented 2 times as T2 in lag 2 and 1 time as T2 in lag 7 for a total of 96 repetitions for each image. In total, each participant completed 3840 trials of AB across sessions. T2 performance was computed as a proportion of correct identification. Attentional Blink Magnitudes (ABM) were computed as the difference between T2 performance in the lag 2 and lag 7 conditions for all images separately.

Working memory task

The same natural visual objects were used in the working memory task completed during fMRI scanning. Images were shown with a 5 degrees visual angle through a back-projected screen visible via a head-mounted mirror. The fMRI consisted of two sessions of 1 hour each completed within the same day with 1-hour rest in between. Within each session, participants completed up to 10 runs. Within each run, each of the forty images was displayed two times. The event-related design was created using optseq[64], with the number of time points (ntp): 610, psdwin: 3.056, nsearch: 10000, nkeep: 500. Each trial

started with an image displayed for 500 ms followed by 4084 ms of retention before a word was displayed (500 + 4084 = 6 TRs with a TR of 764 ms). Participants were asked to respond whether the word corresponded to the semantic content of the image or not using the buttons placed under the left or right index finger. This task was designed to accommodate the slow BOLD response but still capture the working memory and conscious access components of the AB task a1fnd this way provides a canonical working memory representation of all images. Trial onsets were timed to TR onset. Fixation time between trials varied based on the output from optseq between 2-16 seconds.

fMRI acquisition

Participants completed two sessions of fMRI on the same day with one hour rest in between. Each session was designed to last for an hour, to ensure that subjects could stay vigilant for the entire period. fMRI data were acquired using a Philips Achieva 3T MRI scanner and a 32-channel SENSE head coil. A survey scan was made for spatial planning of the subsequent scans. After the survey scan, a 3-min structural T1-weighted scan was acquired using 3D fast field echo (R: 82 ms, TE: 38 ms, flip angle: 8, FOV: 240 x 188 mm, 220 slices acquired using single-shot ascending slice order and a voxel size of 1.0 x 1.0 x 1.0 mm). For the working memory task, functional T2*-weighted sequences were acquired using single shot gradient echo, echo planar imaging (EPI; TR: 764 ms, TE: 27.62 ms, flip angle: 60, FOV: 240x240x118.5 mm, number of slices: 36, slice thickness: 3 mm, slice gap: 0.3 mm, voxel size: 3x3x3 mm, multi-band factor: 3), covering the entire brain.

fMRI pre-processing

fMRI data was converted to BIDS[65], before being pre-processed using fMRIPrep[66]. EPI images were corrected for spatial alignment and normalised to the Montreal Neurological Institute (MNI) ICBM template space[67]. Top-up scans were included as an option in fMRIPrep to mitigate field inhomogeneities. No slice-time correction was made

given our sub-second TR (764 ms) and multi-band acquisition parameters.

fMRI analyses

Beta weights for each stimulus condition were obtained using GLMdenoise[68,69], implemented in MATLAB 2016b (MathWorks) and converted into pseudo t-statistics by dividing the betas with the pooled variance obtained from the bootstrapping in GLMdenoise. Regions of interest (ROI) were defined using the Glasser atlas parcellations (Glasser et al., 2017). Each ROI was registered from fsaverage to subject space, then transformed from surface to volume and registered to functional space using the warp-file provided by the fMRIPrep output using Advanced Normalization Tools (ANTs)[70]. Pattern similarity from each ROI was measured using Pearson's correlation across all pairs of condition pseudo t-patterns. For narrative reasons we choose to stick with the non-inverted Pearson correlation which we refer to as "similarity" between two multivariate patterns (the 1-Pearson coefficient is commonly used in RSA and is often referred to as a "distance"). Searchlight procedure was done using a custom Python script (https://github.com/Charestlab/pySearchlight) with a sphere of 6 voxels radius centred around every voxel. For each voxel position, we did a pair-wise comparison of all images using a Pearson correlation based on the voxels contained within the sphere.

EEG acquisition and pre-processing

Electroencephalographic (EEG) activity was collected with 64 scalp electrodes (BioSemi ActiveTwo System). EEG electrodes were arranged according to the International 10–10 system, in addition to two reference electrodes on the left and right mastoids. Eye movements were monitored using two electrodes placed above and below the pupil of the left eye. The EEG signal was recorded with a 1024 Hz sampling rate. The data pre-processing was performed using mne-python [71,72]. For preprocessing, data were first down-sampled to 256 Hz and then a 0.001 Hz high-pass filter was applied. Epochs were defined with T2 presentation as time 0, with each epoch starting at -800 ms and ending at +700 ms. Automatic rejection of trials was done

using AutoReject[73], where peak-to-peak rejection thresholds are determined automatically through 10-fold cross-validation using the built-in Bayesian optimization method for thresholding. Epochs were then baselined using the mean of the time window between 1900 ms – 1800 ms before T2 onset. This window was always within the fixation phase before each trial. Bad channels were defined by manual inspection and later interpolated over based on the nearest neighbour approach.

EEG analyses

In contrast to fMRI, where we are able to remove shared noise between voxels using GLMdenoise[68] before correlating patterns, there is, to the best of our knowledge, no established method of estimating the noise pool for our particular EEG data set. Therefore, in order to obtain representational similarities in pattern representation between images using EEG, we trained a shrinkage (Ledoit-Wolf lemma shrinkage[74]) linear discriminant analysis (LDA, Sci-kit learn)[75] classifier with for each pair of images, separately for each time-point. For each image pair, we trained on trials when these images were presented as T1 and time 0 was centred around the onset of T1. To evaluate an overall distance between images, we evaluated the performance using 10-fold cross-validation. The resulting accuracies were then defined as our representational distances between each image pair. To conform to the similarity measure from the Pearson coefficient in fMRI, we inverted the decoding scores to obtain “similarity” rather than “distance”.

Acknowledgments

This work was supported by a European Research Council (ERC) Starting Grant ERC-2017-StG 759432 (to I.C.).

References

1. Kirchner, H. & Thorpe, S. J. Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Res.* **46**, 1762–1776 (2006).

2. Potter, M. C., Wyble, B., Hagmann, C. E. & McCourt, E. S. Detecting meaning in RSVP at 13 ms per picture. *Atten. Percept. Psychophys.* **76**, 270–279 (2014).
3. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
4. Felleman, D. J. & Van Essen, D. C. Distributed hierachical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
5. Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D. & Kriegeskorte, N. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences* **111**, 14565–14570 (2014).
6. Kriegeskorte, N., Mur, M., Ruff, D. A. & Kiani, R. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).
7. Jozwik, K. M., Kriegeskorte, N. & Mur, M. Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia* **83**, 201–226 (2016).
8. Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M. & Hauk, O. Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U. S. A.* (2019) doi:10.1073/pnas.1905544116.
9. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* (2019) doi:10.1038/s41593-019-0392-5.
10. Lamme, V. A. F. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
11. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 1–9 (2012).
12. Raymond, J. D., Shapiro, K. L. & Arnell, K. M. Temporary suppression of visual processing in a RSVP task: an attentional blink? *J. Exp. Psychol.* **18**, 849–860 (1992).
13. Kanwisher, N. G. & Potter, M. C. Repetition Blindness: Levels of Processing. *J. Exp. Psychol. Hum. Percept. Perform.* **16**, 30–47 (1990).
14. Kanwisher, N. G. Repetition blindness: Type recognition without token individuation. *Cognition* **27**, 117–143 (1987).
15. Dux, P. E. & Marois, R. The attentional blink: a review of data and theory. *Atten. Percept. Psychophys.* **71**, 1683–1700 (2009).
16. Sy, J. L. & Giesbrecht, B. Target-target similarity on the attentional blink: Task-relevance matters! *Vis. cogn.* **17**, 1–10 (2009).
17. Stein, T., Zwickel, J., Ritter, J., Kitzmantel, M. & Schneider, W. X. The effect of fearful faces on the attentional blink is task dependent. *Psychon. Bull. Rev.* **16**, 104–109 (2009).
18. Lindh, D., Sligte, I. G., Assecondi, S., Shapiro, K. L. & Charest, I. Conscious perception of natural images is constrained by category-related visual features. *Nat. Commun.* **10**, 4106 (2019).
19. Coltheart, V., Mondy, S. & Coltheart, M. Repetition blindness for novel objects. *Vis. cogn.* **12**, 519–540 (2005).
20. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput. Biol.* **10**, (2014).
21. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* **6**, 1–13 (2016).
22. Güçlü, U. & van Gerven, M. A. J. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Brain's Ventral Visual Pathway. **35**, 10005–10014 (2014).
23. Greene, M. R. & Hansen, B. C. Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS Comput. Biol.* **14**, (2018).
24. Eickenberg, M., Gramfort, A., Varoquaux, G. & Thirion, B. Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage* **152**, 184–194 (2017).
25. Buffat, S., Plantier, J., Roumes, C. & Lorenceau, J. Repetition blindness for natural images of objects with viewpoint changes. *Front. Psychol.* **3**, 1–11 (2013).
26. Corbetta, M., Patel, G. & Shulman, G. L. The Reorienting System of the Human Brain:

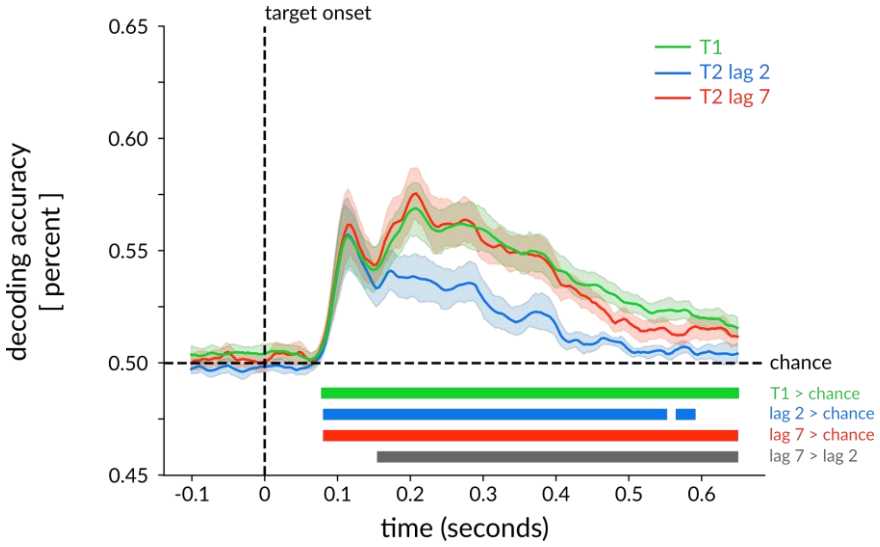
From Environment to Theory of Mind. *Neuron* **58**, 306–324 (2008).
27. Fisher, R. A. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Ann. Eugen.* **7**, 179–188 (1936).
28. Martens, S. & Valchev, N. Individual differences in the attentional blink: The important role of irrelevant information. *Exp. Psychol.* **56**, 18–26 (2009).
29. Martens, S., Munneke, J., Smid, H. & Johnson, A. Quick minds don't blink: Electrophysiological correlates of individual differences in attentional selection. *J. Cogn. Neurosci.* **18**, 1423–1438 (2006).
30. Dale, G. & Arnell, K. M. Individual differences in dispositional focus of attention predict attentional blink magnitude. *Atten. Percept. Psychophys.* **72**, 602–606 (2010).
31. Gross, J. *et al.* Modulation of long-range neural synchrony reflects temporal limitations of visual attention in humans. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13050–13055 (2004).
32. Husain, M., Shapiro, K., Martin, J. & Kennard, C. Abnormal temporal dynamics of visual attention in spatial neglect patients. *Nature* **385**, 154–156 (1997).
33. Cooper, A. C. G., Humphreys, G. W., Hulleman, J., Praamstra, P. & Georgeson, M. Transcranial magnetic stimulation to right parietal cortex modifies the attentional blink. *Exp. Brain Res.* **155**, 24–29 (2004).
34. Yaple, Z. & Vakhrushev, R. Modulation of the frontal-parietal network by low intensity anti-phase 20 Hz transcranial electrical stimulation boosts performance in the attentional blink task. *Int. J. Psychophysiol.* **127**, 11–16 (2018).
35. Bavelier, D. Repetition blindness between visually different items: the case of pictures and words. *Cognition* **51**, 199–236 (1994).
36. MacKay, D. G. & Miller, M. D. Semantic Blindness: Repeated Concepts Are Difficult to Encode and Recall Under Time Pressure. *Psychological science* **5**, 52–55 (1994).
37. Harris, I. M. & Dux, P. E. Orientation-invariant object recognition: Evidence from repetition blindness. *Cognition* **95**, 73–93 (2005).
38. Fragopanagos, N., Kockelkoren, S. & Taylor, J. G. A neurodynamic model of the attentional blink. *Cognitive Brain Research* **24**, 568–586 (2005).
39. Wyble Brad, B., Potter, M. C., Bowman, H. & Nieuwenstein, M. Attentional episodes in visual perception. *J. Exp. Psychol. Gen.* **140**, 488–505 (2011).
40. Dell'Acqua, R. *et al.* The attentional blink impairs detection and delays encoding of visual information: evidence from human electrophysiology. *J. Cogn. Neurosci.* **27**, 720–735 (2015).
41. Vogel, E. K. & Luck, S. J. Delayed working memory consolidation during the attentional blink. *Psychon. Bull. Rev.* **9**, 739–743 (2002).
42. Brevers, D. *et al.* Reduced attentional blink for gambling-related stimuli in problem gamblers. *J. Behav. Ther. Exp. Psychiatry* **42**, 265–269 (2011).
43. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 3863–3868 (2006).
44. Fiez, J. A. & Petersen, S. E. Neuroimaging studies of word reading. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 914–921 (1998).
45. Blank, S. C., Scott, S. K., Murphy, K., Warburton, E. & Wise, R. J. S. Speech production: Wernicke, Broca and beyond. *Brain* **125**, 1829–1838 (2002).
46. Stampacchia, S. *et al.* Shared processes resolve competition within and between episodic and semantic memory: Evidence from patients with LIFG lesions. *Cortex* **108**, 127–143 (2018).
47. Corbetta, M. & Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **3**, 201–215 (2002).
48. Willems, C. & Martens, S. Time to see the bigger picture: Individual differences in the attentional blink. *Psychon. Bull. Rev.* **23**, 1289–1299 (2016).
49. Arnell, K. M. & Shapiro, K. L. Attentional blink and repetition blindness. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 336–344 (2011).
50. Arnell, K. M., Howe, A. E., Joanisse, M. F. & Klein, R. M. Relationships between attentional blink magnitude, RSVP target accuracy, and performance on other cognitive tasks. *Mem. Cognit.* **34**, 1472–1483 (2006).
51. Colzato, L. S., Spapé, M. M. A., Pannebakker, M. M. & Hommel, B. Working memory and the attentional blink: Blink size is predicted by individual differences in operation span. *Psychon. Bull. Rev.* **14**, 1051–1057 (2007).

52. Arnell, K. M., Stokes, K. A., MacLean, M. H. & Gicante, C. Executive control processes of working memory predict attentional blink magnitude over and above storage capacity. *Psychol. Res.* **74**, 1–11 (2009).
53. Arnell, K. M. & Stubitz, S. M. Attentional blink magnitude is predicted by the ability to keep irrelevant material out of working memory. *Psychol. Res.* **74**, 457–467 (2010).
54. Shapiro, K., Hillstrom, A. P. & Husain, M. Control of visuotemporal attention by inferior parietal and superior temporal cortex. *Curr. Biol.* **12**, 1320–1325 (2002).
55. Zhou, L., Zhen, Z., Liu, J. & Zhou, K. Brain Structure and Functional Connectivity Associated with Individual Differences in the Attentional Blink. *Cereb. Cortex* **00**, 1–14 (2020).
56. Kihara, K. *et al.* Differential contributions of the intraparietal sulcus and the inferior parietal lobe to attentional blink: Evidence from transcranial magnetic stimulation. *J. Cogn. Neurosci.* **23**, 247–256 (2011).
57. Molenberghs, P., Sale, M. V. & Mattingley, J. B. Is there a critical lesion site for unilateral spatial neglect? A meta-analysis using activation likelihood estimation. *Front. Hum. Neurosci.* **6**, 1–10 (2012).
58. Ptak, R. & Schnider, A. Visual extinction of similar and dissimilar stimuli: Evidence for level-dependent attentional competition. *Cogn. Neuropsychol.* **22**, 111–127 (2005).
59. Ortiz-tudela, J., Martı, E., Chica, A. B. & Lupi, J. Semantic incongruity attracts attention at a pre-conscious level : Evidence from a TMS study. *Cortex* **102**, 96–106 (2017).
60. Geng, J. J. & Vossel, S. Re-evaluating the role of TPJ in attentional control : Contextual updating ? *Neurosci. Biobehav. Rev.* **37**, 2608–2620 (2013).
61. Twomey, D. M., Murphy, P. R., Kelly, S. P. & O'Connell, R. G. The classic P300 encodes a build-to-threshold decision variable. *Eur. J. Neurosci.* **42**, 1636–1643 (2015).
62. Grootswagers, T., Cichy, R. M. & Carlson, T. A. Finding decodable information that can be read out in behaviour. *Neuroimage* **179**, 252–262 (2018).
63. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
64. Dale, A. M. Optimal experimental design for event-related fMRI. *Hum. Brain Mapp.* **8**, 109–114 (1999).
65. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**, 160044 (2016).
66. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
67. Mazziotta, J. *et al.* A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356**, 1293–1322 (2001).
68. Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F. & Wandell, B. A. GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* **7**, 1–15 (2013).
69. Charest, I., Kriegeskorte, N. & Kay, K. N. GLMdenoise improves multivariate pattern analysis of fMRI data. *Neuroimage* **183**, 606–616 (2018).
70. Avants, B., Tustison, N. & Song, G. Advanced Normalization Tools (ANTS). *Insight J.* 1–35 (2009).
71. Gramfort, A. *et al.* MNE software for processing MEG and EEG data. *Neuroimage* **86**, 446–460 (2014).
72. Gramfort, A. *et al.* MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7**, 267 (2013).
73. Jas, M., Engemann, D. A., Raimondo, F., Bekhti, Y. & Gramfort, A. Automated rejection and repair of bad trials in To cite this version : Automated rejection and repair of bad trials in MEG / EEG. *6th International Workshop on Pattern Recognition in Neuroimaging (PRNI)* (2016).
74. Ledoit, O. & Wolf, M. Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management* **30**, 110–119 (2004).
75. Varoquaux, G. *et al.* Scikit-learn. *GetMobile: Mobile Computing and Communications* **19**, 29–33 (2015).

Supplementary

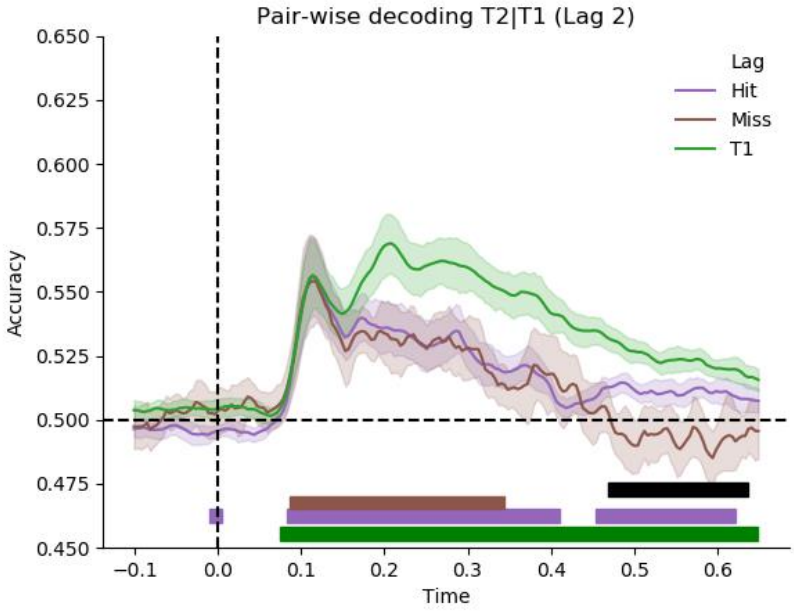
Supplementary Table 1. Brain clusters defined from the searchlight procedures on individual differences.

cluster id	MNI x peak	MNI y peak	MNI z peak	Peak value (Pearson r)	Volume mm	aal
1	36	30	7.8	0.77654	6771.6	Insula R
2	63	-45	7.8	0.763386	5256.9	Temporal Mid R
3	42	-24	54	0.722177	1930.5	Postcentral R
4	51	-30	-8.7	0.695877	415.8	Temporal Mid R
5	45	-27	1.2	0.663015	237.6	Temporal Sup R

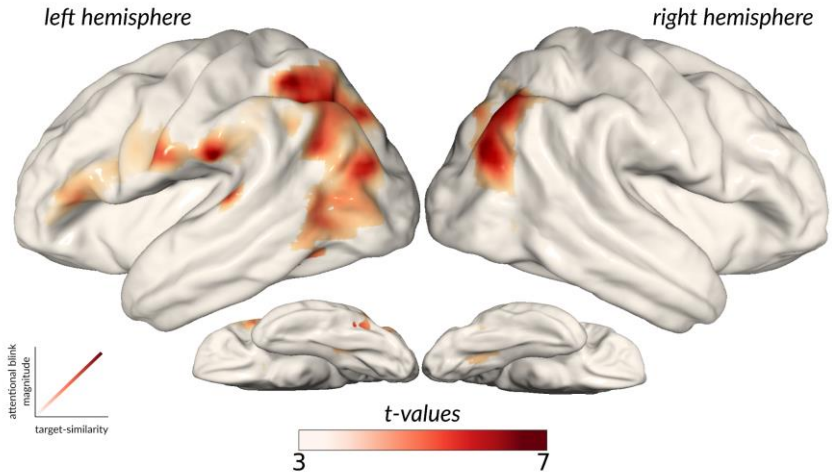


Supplementary Figure 1. *Average decoding accuracy for all pairwise comparisons, separately T1 (green curve), T2 Lag-2 (blue curve) and T2 Lag-7 (red curve) trials. A classifier was trained using T1 presentations when T2*

was at lag-7, to avoid any more than necessary noise, and then tested on lag-2 trials. Green, blue and red bars denote time points where the decoding for T1, Lag-2 and Lag-7 (respectively) is significantly different from chance-level (cluster-permutation test, cluster threshold (t-value) = 3). The gray bar denotes time points where the decoding of lag 7 trials was significantly more accurate than the decoding of lag 2 trials.



Supplementary Figure 2. *Average decoding accuracy for all pairwise comparisons, separately for hit (purple) and miss (brown) trials (Lag-2). A classifier was trained using T1 presentations when T2 was at lag-7, to avoid any more than necessary noise, and then tested on lag-2 trials. In green, T1 decoding using a 10-fold crossfold for comparison. Green, purple and brown bars denote time points where T1, hit and misses (respectively are significant from zero (cluster-permutation test, cluster threshold (t-value) = 3). The black bar denotes time points where hit and miss is significantly different.*



Supplementary Figure 3. *Target-target similarity in late perceptual and semantic brain networks is related to inflated AB magnitude. Using a whole-brain searchlight procedure to correlate behaviour (AB magnitude) with the overall similarity of one image in relation to the rest of the image set, we identified a left hemisphere semantic brain network. Images that are like other images in these areas showed a significantly larger impairment within the AB windo*

Chapter 4

Attention modulates the effect of target-target similarity in opposite ways depending on levels of processing

Authors:
Daniel Lindh[1,2,3]
Ilja G. Sligte[3,4]
Kimron L. Shapiro[1,2]
Ian Charest[1,2]

Affiliations:
[1]School of Psychology, Hills Building, University of Birmingham, B152TT Birmingham, United Kingdom

[2]Centre for Human Brain Health, University of Birmingham, United Kingdom

[3]Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, The Netherlands

[4]Amsterdam Brain and Cognition, University of Amsterdam, The Netherlands

Corresponding author: p.j.d.lindh@uva.nl

In preparation

Keywords:
Attention, Deep Convolutional Neural Networks, Consciousness, Object recognition, fMRI

Abstract

When a complex natural scene is quickly presented, the human visual system is remarkably fast at detecting and identifying characteristics of the image that are diagnostic of the semantic content. To study this ability, researchers often use rapid serial visual presentations (RSVP), where a set of stimuli is presented at a high rate and participants are asked to detect one or several targets within a stream of distractors. One common finding is that when the second target (T2) is presented 200-500 ms after the first target (T1), participants are often unable to report T2 correctly. However, when participants are asked to ignore T1 and only report T2, participants are again remarkably good at reporting T2. This phenomenon where attending T1 attenuates T2 processing is called Attentional Blink. In two previous studies, we have shown that similarity between targets modulates the AB effect in two different ways depending on the level of processing targets share representational overlap. First, in the literature it is well known that repetition of targets often leads to a memory failure where T2 is omitted. We have previously corroborated these findings and have shown that T2's that are similar to T1 in posterior parietal cortex and within a left semantic brain network are less often reported. On the other hand, we also showed that when images are similar early on in processing (V1) T2 performance is elevated. In the current study we sought to test how similarity between targets interacts with attending T1, one of the core components of AB. We test this using a hybrid task where participants were asked to either attend or ignore T1 and to make a speeded judgment if the T2 scene contained an animal or not. By modelling the reaction time distribution using drift diffusion modelling we find several important notions. We find that attending T1 affects both perceptual and non-perceptual processes, undermining theoretical frameworks that propose a bottleneck, such as the two-stage model of AB. We also show that attending T1 is imperative for the beneficial V1 similarity to affect T2 targets. These findings outline a series of behaviours present in humans that can be used to benchmark future models of attention in RSVP settings.

Introduction

While exploring our surroundings, our eyes move and sample information at a high rate (~4-5 saccades per second). Through a collection of processes, often referred to as attention, our brain can selectively filter out noise and efficiently process important information that helps us navigate our complex environment. Investigations into how attention modulates visual representations, and how the processing of distinct items interact at different stages is crucial for our understanding of the human perceptual system. One of the most common paradigms to probe our ability of temporal information-selection is a rapid serial visual presentation (RSVP). In an RSVP, one or several targets are embedded within a stream of distractors, often presented at a rate of ~10 items per second. Interestingly, when two targets (T1 and T2, respectively) are embedded in the RSVP, participants often miss T2 when it is presented 200-500 ms after T1. This phenomenon is known as the Attentional Blink (AB) (Raymond et al., 1992) and is one of the most well-studied attentional paradigms, with thousands of studies done since its discovery 30 years ago. One core aspect of the AB is that when participants are asked to ignore T1, reportability of T2 is high regardless of what position in the stream T2 is presented (Dux & Marois, 2009; Raymond et al., 1992). This indicates that deliberate engagement with T1 is interrupting T2 processing at some critical stage, causing participants to miss the second target. Most theories of AB are so-called two-stage models (Chun & Potter, 1995), which posit that both targets can be processed in parallel in an identification stage (first stage). However, by attending to T1 and encoding the item into working memory (the second stage), which is assumed to be a serial bottleneck, it is effectively interfering with the encoding of T2 causing participants to be unable to report T2. Neural evidence of these two-stage, bottleneck, models consists of studies showing that high-level stimulus information is still present in neural code (Dehaene et al., 2006; Luck et al., 1996; Marois et al., 2004), despite participants' inability to correctly report T2. Other evidence for a bottleneck account comes from (Vogel & Luck, 2002) who showed that masking T2, and thereby impairing T2 performance, leads to delayed working memory engagement. Despite all of this

evidence for T1 disrupting the working memory encoding of T2, there is still no reason to think that T1 cannot also affect perceptual processes as well, however this idea has been proven harder to test. We attempt to test this idea by conceptualizing ideas from reaction time modelling and our previous findings of interaction between targets.

In a previous study (Lindh et al., 2021) we showed that different levels of similarity between targets affect T2 performance in opposite ways. Specifically, we showed that when the two targets are similar in high-level visual/semantic brain areas, participants are less likely to perceive T2. This finding is in line with another RSVP phenomena, known as repetition blindness (RB) (Buffat et al., 2013; Kanwisher, 1987; Kanwisher & Potter, 1990; Park & Kanwisher, 1994). Previous studies have shown that RB can occur for several types of repetitions, for example, with direct repetitions such as “ink/ink” and “3/3” (Kanwisher, 1987; Kanwisher & Potter, 1990), the same objects from different viewpoints (Buffat et al., 2013), phonetically similar words such as “won/one” (Bavelier & Potter, 1992) but also in bilingual participants where the two targets are in different languages but have the same meaning, for example Caballo/Horse (MacKay & Miller, 1994). However, in certain contexts target repetitions can also prime and therefore increase performance. For example, when T2 is missed in a three target RSVP (with T1, T2, and T3 targets), T3 no longer exhibits an AB but is instead primed if the missed T2 was a repetition (Shapiro et al., 1997). Similarly, categorical repetitions (for example T1 and T2 both being animate objects) can also improve T2 performance when participants report identity (for example horse and dog) (Evans & Treisman, 2005). Related to this, an additional important factor is task-relevance - what type of information is to be encoded into working memory. For example, (Bavelier, 1994) showed that a picture of a sun (T1) induced an RB on the word “son” (T2) when the task required phonetic encoding. (Sy & Giesbrecht, 2009) presented participants with faces and asked them to either report gender or emotional expression. When the two target faces had the same gender there was a significant decrease in T2 performance, but contingent on participants reporting gender and not emotional expression. To most

people familiar with priming (Schacter & Buckner, 1998), RB is unintuitive, but decades of research proposes that RB occurs in very specific contexts due to the similarity in neural codes initially used in short-term memory (Bavelier, 1994).

Studies looking at similarities between targets often resort to using categories as a proxy for similarity. However, the visual system has been shown to follow a hierarchical structure (Felleman & Van Essen, 1991), implying that overlap in neural representation between two targets can occur at multiple levels. Our previous study extended these ideas and showed the extent of the brain network at which similarity between targets is detrimental for performance, which included inferior temporal cortex, posterior parietal cortex, and along the left lateral sulcus (Lindh et al., 2021). Interestingly, in contrast to RB, we also showed that when targets were similar in V1, the earliest cortical processing stage for visual information, T2 performance instead increased. This increase in performance, based on low-level visual feature similarities, were in line with our previous study utilising a convolutional neural network (CNN) to define similarities (Lindh et al., 2019). While RB has been investigated extensively (Buffat et al., 2013; Chun, 1997; Fagot & Pashler, 1995; Harris & Dux, 2005; Kanwisher, 1987; Kanwisher & Potter, 1990; Park & Kanwisher, 1994), and seems to be related to late-stage memory functions (Bavelier, 1994; Fagot & Pashler, 1995), our knowledge of the enhancing effect of low-level visual feature similarity is limited to our own two studies. One possible mechanism behind enhanced performance related to target-target similarity in V1 is neural adaptation. Neural adaptation to recent stimulus history can significantly alter perception through neural suppression (Sawamura et al., 2006), neural enhancement (Kasper Vinken et al., 2017) or shifts in tuning functions (Dragoi et al., 2000). One of the main computational roles of neural adaptation that has been proposed is that it facilitates detection by increasing sensitivity to small changes in the environment (Clifford et al., 2007). In a recent study, (K. Vinken et al., 2020) showed that, after an adapter phase with a noise pattern, the following presentation of an object had a higher detection rate when the background noise was the same as during the adapter phase. Importantly, they continued to show that a

CNN with local neural adaptation implemented on a unit-level replicated these results, while a CNN without local adaptation was unable to. It is conceivable that neural representational overlap between two targets in V1 also leads to facilitation of object identification in natural scenes. Therefore, it logically follows that this would increase the speed of evidence accumulation, related to the quality of stimulus information, for target objects embedded in natural scenes. Here, we test this hypothesis by modelling the reaction time distribution of speeded judgments of T2. It is possible that target-target similarities at different levels are unrelated to AB. In order to investigate this we also manipulated attention to T1, one of the core ideas derived from the AB literature.

In the current study, we investigate how target-target similarity affects processing of T2 while manipulating one of the core aspects of the AB, attending or ignoring T1. AB and RB are two phenomena that imply that when targets are presented adjacently in time, the processing of the targets will interact at one or several stages. How this interaction affects different aspects of the decision process (evidence accumulation) and non-decision variables is unknown. Using natural images from the Microsoft Common Objects in Context (COCO) image dataset (Lin et al., 2014) and neural data from the natural scenes data set (NSD) (Allen et al., 2021), we tested how attending T1, target-target similarity, and their interaction modulated the speed of evidence accumulation. Given our experimental approach (Figure 1), we do not expect an RB effect due to the different task-demands for T1 and T2, however, we hypothesized that T1-T2 similarity would prime T2. This is in line with previous passive high-level priming findings using animals and vehicles (Evans & Treisman, 2005), and our own studies showing priming of low-level features (Lindh et al., 2019; Lindh et al., 2021). We further predicted that if target-similarity is related to the AB phenomenon, we expect an interaction effect of attending to T1 and similarity in regards to evidence accumulation. The results reveal a complex relationship between target-target similarities at different stages of processing and attention. These findings elucidate how multiple items are interacting, what type of processes are being affected and present additional challenges for models of attention.

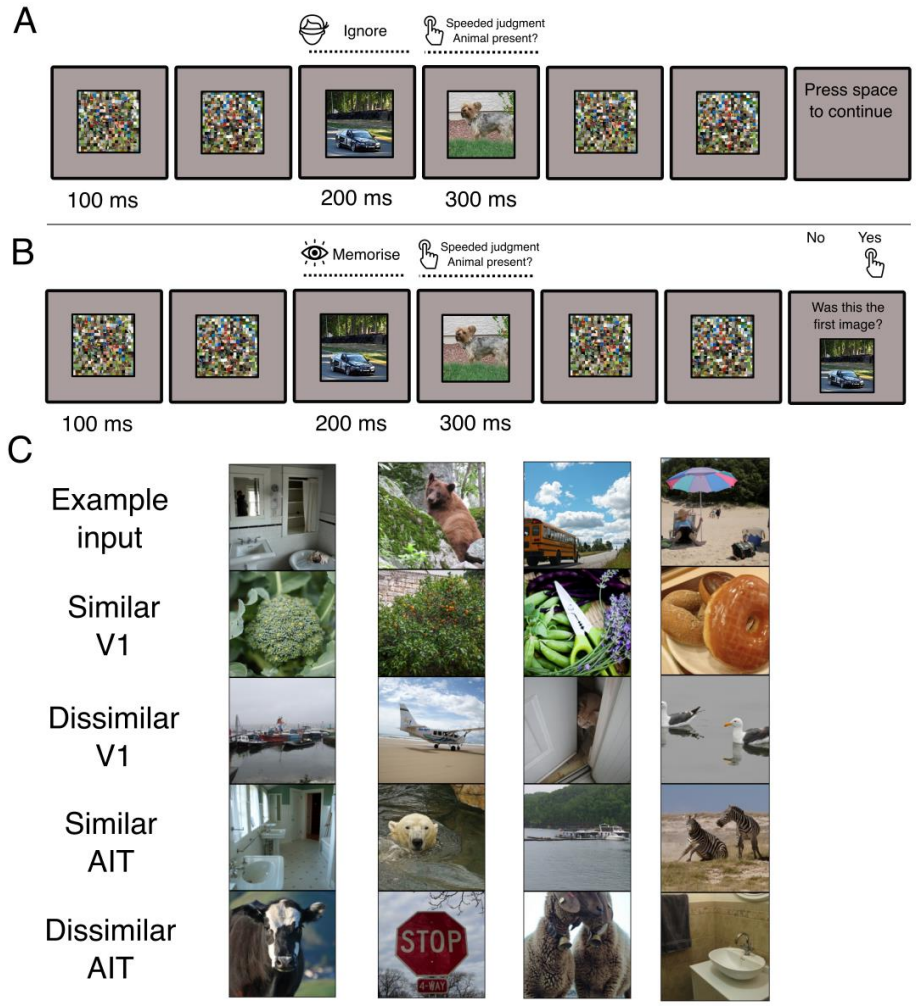


Figure 1. Pictorial description of the dual task. ***A)*** *Participants were presented with a stream of four masks (100 ms SOA) followed by T1 (200 ms), T2 (300 ms) and four more masks. Participants were asked to ignore T1 and to make a speeded judgment if an animal was present in the T2 scene.* ***B)*** *In a different session, participants were instructed to do the same as A, however, with the added instruction to attend T1 for the report after the stream.* ***C)*** *Example images from the NSD / COCOs data set. Top row: randomly selected images from the 1000 seen by all participants in the NSD dataset. 2nd row: The images selected to be maximally similar to the top row in V1. 3rd row: images that are maximally dissimilar in V1 to the top row. 4th and 5th row: same concept but using similarities based on AIT.*

Methods

Participants. 65 participants (60 female, 5 male, age M = 19.5, SD = 2.15, age range = 18-35) were recruited through a participant website hosted by the University of Birmingham. Of these, 13 were excluded due to sub-chance performance (≤50% accuracy on T2 response) or zero performance due to technical difficulties. This resulted in a sample size of 52 participants, all of whom had provided informed consent anonymously, and were rewarded with 1 psychology course credit. All participants reported to have normal- or corrected-to-normal vision. The experiment was approved by the ethical review board of the School of Psychology at the University of Birmingham.

Stimuli. From the 1000 images derived from the COCOs dataset (Lin et al., 2014) and seen by all participants in the Natural Scenes Data set (NSD (Allen et al., 2021)), see *estimating V1 and AIT similarity* below), we removed all images that contained humans resulting in 562 potential target images. For our T2s, we randomly selected 50 images that contained animals and 50 images with no animals. We then selected T1 images such that they would represent a balanced variety of V1 similarities (equal selection of low/mid/high V1 similarity between T1 and T2, see Supplementary Figure 1 for distributions of similarities) for each T2, half of the T1s either contained animals or not. For the total of 100 T2s, each T1-T2 combination (3 (similarity low/mid/high) x 2 (T1 animal/non-animal)) was only shown one trial for a total of 600 trials. Masks were made by subdividing an image into a 10 x 10 matrix, and for each cell we copied the content from the same location from a random image within the data set (see Figure 1A for examples). A total of 200 masks were made and for each trial, masks were randomly selected without replacement. Stimuli were presented on a grey background, and images were sized such that they covered one third of the screen. This data was collected online using the Meadows Research online platform (http://meadows-research.com). Therefore, we could not control the distance between participants and their screen and opted for displaying images in the same size relative to the window size for each device. We argue that

since participants were only allowed to participate using a computer, and that we have a within-subject design, any differences between conditions cannot be explained by the variance of image sizes between participants. The task was programmed using the Python library PsychoPy (Peirce et al., 2019), which was converted into PsychoJS for the online platform.

Procedure. Due to the COVID-19 pandemic of 2020, we administered the task online using the Meadows Research (http://meadows-research.com) online platform. Each participant completed two sessions, one of which participants were asked to ignore T1 and in the other to attend T1 to report after the stream. Each session consisted of 600 trials, divided into 10 blocks. Each trial started with a central fixation cross over a grey screen for 500 ms, then four masks followed, presented for 100 ms immediately following each other. After the masks, T1 was presented for 200 ms followed by the presentation T2 for 300 ms followed by four masks of 100 ms SOA. The absence of intervening masks between T1 and T2 and the SOAs were chosen to optimise the influence of T1 on T2, and to ensure that T2 was presented long enough for it to be solved in the inferotemporal (IT) cortex. Previous monkey studies have shown that object categories within complex visual scenes can be linearly decoded in IT within 120-250 ms (Kar et al., 2019). Participants were instructed to make a speeded judgment to indicate if an animal was present in T2, 'Z' for 'Yes' and 'M' for 'No', as soon as it was presented. In one of the sessions, participants were either instructed to attend or ignore T1 with the order of sessions counterbalanced across participants. After the stream of items, the true T1 image (50% of trials) or a random image was shown, and participants were asked to indicate if this was the T1 or not. In the other session, participants were asked to ignore T1 and after the end of the stream participants were asked to press space to continue. For all participants, the order of session one and two was counterbalanced. Importantly, in our pilot studies we noticed that the Attend T1-condition (which included a task-switch from memorising T1 and to convert the T2 decision into an immediate motor response, see figure 1B) was hard for participants to properly carry out and there was confusion regarding the task instructions. Therefore, before each

session, participants completed two training blocks of 6 trials each. In the first training block, the RSVP stream was slowed down by a factor of ten and instructional text was presented to indicate when to react to T2 and whether to ignore or attend to T1. The second training block had the same presentation speed as the actual experiment. We found that this significantly increased the participants' performance on the real task.

Estimating V1 and AIT similarity

We estimated pairwise similarities in V1 using the Natural Scenes Dataset (NSD, REF). The NSD consists of 8 participants who in total viewed >70 000 images (1000 shared images) from the Microsoft COCOs dataset (Lin et al., 2014), while brain responses were recorded using a 7T Siemens Magnetom 48 passively-shielded scanner and a single-channel-transmit, 32-channel-receive RF head coil. Whole-brain functional data were collected with 84 axial slices, 1.8277 mm slice thickness, 216 mm (FE) and 216 mm (PE) field-of-view, 1600ms TR, 62° flip angle, 0.66 echo spacing, and multiband slice acceleration factor 3. For more details and quality testing of this data set, please see (Allen et al., 2021). Functional data was pre-processed using a novel development of GLMdenoise (Charest et al., 2018; Kay et al., 2013), which allows for single-trial beta estimations (https://github.com/kendrickkay/GLMdenoise). The NSD comes with a collection of regions-of-interests (ROIs) where the visual areas were hand drawn using population receptive field (pRF) data by two cortical surface experts. We selected our ROIs, V1 and anterior inferior temporal cortex (AIT), based on our previous study (Lindh et al., 2021). For each of the 8 participants, we computed the pairwise similarity between each of the 1000 shared images using Pearson correlation on the z-scored beta estimates for both V1 and AIT. These pairwise similarities were then averaged across participants, resulting in one 1000 x 1000 representational similarity matrix. For each T1-T2 combination, we then indexed at the appropriate row and column to identify their similarity coefficient. It is important to note that the

participants in NSD were different from the ones participating in our behavioural study.

Behaviour

In addition to drift diffusion modelling (DDM, see below), to describe the accuracy performance difference between the attend and ignore T1 conditions, we calculated d-prime, and criterion based on the T2 responses**.**

Drift diffusion modelling

Hierarchical Diffusion Decision Modelling (HDDM; (Wiecki et al., 2013)) implemented in Python 2.7 was used to model the reaction time distributions for T2 correct and incorrect responses. A hierarchical model controls the shrinkage of the parameter space by centring the individuals prior to the group mean and can thus be seen as an optimal combination of fixed and random effects. Therefore, HDDM is preferable for small sample sizes (20-100 participants) (Ratcliff & Childers, 2015). Similar to SDT, DDM makes assumptions based on popular computational decision making ideas which posit that sensory evidence for a decision is accumulated over time until it reaches a certain boundary (Gold & Shadlen, 2007; Ratcliff & McKoon, 2008). Translated to our task, when T2 is presented, from starting point *z (bias, however, we modelled correct or incorrect responses, so this variable was not included in our model)* evidence (in favour for either "animal" or "no-animal") is accumulated with drift-rate *v* (evidence accumulation) until it reaches boundary *a* (decision criterion, see Supplementary Figure 2). Another important parameter is *t*, or the non-decision time parameter, which describes ancillary latent variables unrelated to the decision process (such as encoding to working memory or conversion into motor response). Similar to d-prime in SDT, drift rate reflects the quality of the sensory information and is directly related to perceptual processing (Voss et al., 2004). Therefore, our parameter of interest was first and foremost drift-rate v (or evidence accumulation speed) and how our manipulations and trial-by-trial

covariates affected this metric. In a full model using both sessions, we first removed all responses < 100 ms (considered too fast to be a properly evaluated response) and then we added the following regressors of interest on the drift-rate parameter: 1) V1 similarity between T1 and T2. 2) AIT similarity between T1 and T2. 3) Attend or Ignore T1 condition. 4) Interaction between V1 similarity and attending T1. 5) Interaction between AIT similarity and attending T1. In addition, we also added covariates that we reasoned could potentially bias our results: Covariate 1) If T1 and T2 were from the same category (both animal/both non-animal or targets were from different categories, henceforth known as *category congruence*). Since we modelled correct/incorrect responses, estimating the bias (z) was not possible. However, for example, it is possible that the T1 animal / T2 animal pairs would differ from T1 non-animal / T2 animal pairs in both similarity and their semantic relationship and potentially confound the results. Covariate 2) T1 complexity. Upon visual inspection of a previous pilot (see Figure 1C) and further simulations (Supplementary Figure 3) we observed that T1 images that were "dissimilar" in V1 from our T2s regularly had a lower scene complexity. Scene complexity is known to affect performance (Seijdel et al., 2021). In the AB literature, T1 difficulty is well-known to affect T2 performance (Akyürek et al., 2007) and was thus a necessary covariate. In addition, we added one regressor of interest for the non-decision variable *t*: 1) Ignore or attend T1. HDDM uses Markov Chain Monte Carlo (MCMC) sampling to estimate the latent decision parameters associated with DDM (as well as the coefficients for the regressor of interests and covariates for drift rate) by generating samples from the posterior distribution by means of constructing a reversible Markov-chain which is centred around its ground truth posterior distribution. We ran 25000 samples in total. As recommended when sampling with MCMC, for stable estimates, we burnt the first 1000 samples (discarded), resulting in 24000 samples. For each sample, a small step is made in parameter space from the current parameter position and is accepted if the probability of the new parameters (given our data) is higher than the previous. Therefore, the resulting trace for a given coefficient on drift-rate can be used in hypothesis testing by comparing the values against zero, however, note that this is different from the p-value in classical statistics. A priori,

we decided that we would accept the H1 (there is an effect) of any coefficient if > 95% of the accepted posterior estimates were below or above zero. After running the full model, which included both sessions, we ran the two sessions separately using the following regressors of interest: 1) V1 similarity between T1 and T2. 2) AIT similarity between T1 and T2. We also modelled the two same covariates as described for the full model.

Results

Attending T1 does not decrease sensitivity, but impairs reaction time

In the Attend T1 condition, we confirmed that participants executed the task properly with a high accuracy on T1 (M=89.6% correct, SD=6.23%, d-prime M=2.78, SD=0.689). For T2 performance we evaluated both d-primes for both the Attend and Ignore T1 conditions. We found no significant difference between the conditions in terms of d-prime; Ignore T1 (M = 2.50, SD = 0.72), Attend T1 (M=2.57, SD=0.68), t(51)=-0.95, p=0.346). The finding of no difference in d-prime is not surprising considering the 300 ms presentation time of T2, 10 Hz presentation rate (Shapiro et al., 2017) and masking (although see Nieuwenstein et al., (2009)) are crucial components for detecting AB effects. However, there was a significant difference in median reaction time between the two condition; Ignore T1 (M = 611ms, SD = 188ms), Attend T1 (M=669ms, SD=181ms), paired t-test (t(51)=-2.29, p = 0.026, Cohen's d = -0.315). See Figure 2. We further investigated the effect of an animal being present in T1 or not by conducting a repeated measures ANOVA with T1 attention (attend or ignore) and T1 animacy (animal present or not) as factors. For d-prime there were no main effect of attending T1 ($F(1,51) = 0.893$, $p = 0.349$) or T1 animacy ($F(1,51) = 0.087$, $p = 0.769$), and further no interaction effect of attending T1 and T1 animacy ($F(1,51) = 0.025$, $p = 0.874$). For reaction time, the ANOVA confirmed the main effect of attending T1 ($F(1,51) = 5.301$, $p = 0.025$, $\eta^2 = 0.093$), but with no main effect of T1 animacy ($F(1,51) = 0.145$, $p = 0.705$) and no interaction effect ($F(1,51) = 0.243$, $p = 0.624$). We evaluated the convergence of the chains for

the different parameters through visual inspection (Supplementary figure 2).

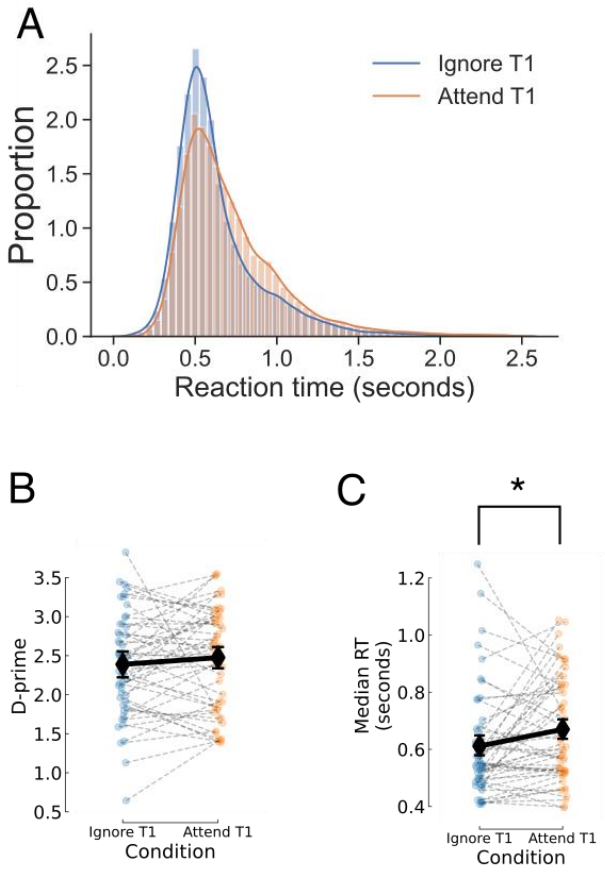


Figure 2. *A) Reaction time distributions on speeded judgment for animal-detection in T2 for the Ignore T1 (blue) and Attend T1 (red) condition. There is a qualitative difference between the distributions, with a larger peak for the Ignore T1 condition. B) D-prime (sensitivity to stimulus) for detecting animals in T2. No significant differsence in the two conditions. C) As implied by the distribution plot in A, there was a significant difference in median reaction time between attending and ignoring T1 conditions. With a Cohen's d of 0.315 (see Attending T1 does not decrease sensitivity, but impairs reaction time under results), this can be considered to be a small to medium sized effect. * = p < 0.05, ** = p < 0.01, *** = p < 0.001.*

Attending to T1 affects drift rate in opposite directions for V1 similarity and AIT similarity

Note that with 24000 samples our precision is 1/24000 = 0.00004 implying that if all the chains end up on either side of zero it can only be described as > 0.99996. For the full model (both sessions, Figure 3A), the estimated decision-related DDM-parameters were the following: parameter **a** (criterion, M=1.74, SD=0.052), **v** (drift-rate/accumulation speed, M=1.55, SD=0.07), **t** (non-decision time, M=0.2, SD=0.012). The estimated coefficients from the five regressors-of-interest on drift-rate (as ordered in *Methods Drift diffusion modelling*): 1) No main effect of V1 similarity between targets on drift rate (M=-0.027, SD=0.067, P(coefficient > 0) = 0.3). 2) A main effect of target-target similarity in AIT on drift rate (M=0.438, SD=0.035, P(coefficient > 0) > 0.99996). 3) Attending T1 led to a decreased drift rate of T2, showing that attending T1 affects perceptual processes (M=-0.223, SD=0.055, P(coefficient < 0) > 0.99996). 4) We found an interaction between attending T1 and V1 similarity, implying that attention to T1 is imperative for the beneficial effect of V1 similarity (M=0.169, SD=0.05, P(coefficient > 0) > 0.9729). 5) We also found a negative interaction between attending T1 and AIT similarity, which corroborates the notion that attending T1 increases the amount of interference between targets in late processing stages (M=-0.15, SD=0.05, P(coefficient < 0) > 0.9996). For our two covariates we got the following estimates: Covariate 1) A positive effect on drift rate for category congruence, meaning that when both targets were from the same category (animal or not) participants were faster at accumulating evidence for the correct target category (M=0.367, SD=0.013, P(coefficient > 0) > 0.99996). Covariate 2) A negative effect of T1 complexity, showing that when T1 is less complex (in terms of number of available visual features) it interferes less with T2 processing (M=-0.0059, SD=0.0007, P(coefficient < 0) > 0.99996). We evaluated the fit using the Deviance Information Criterion (DIC), deviance (the function of the probability density) and pD (DIC - deviance). DIC= 30149.584, deviance = 29987.991, pD = 161.593.

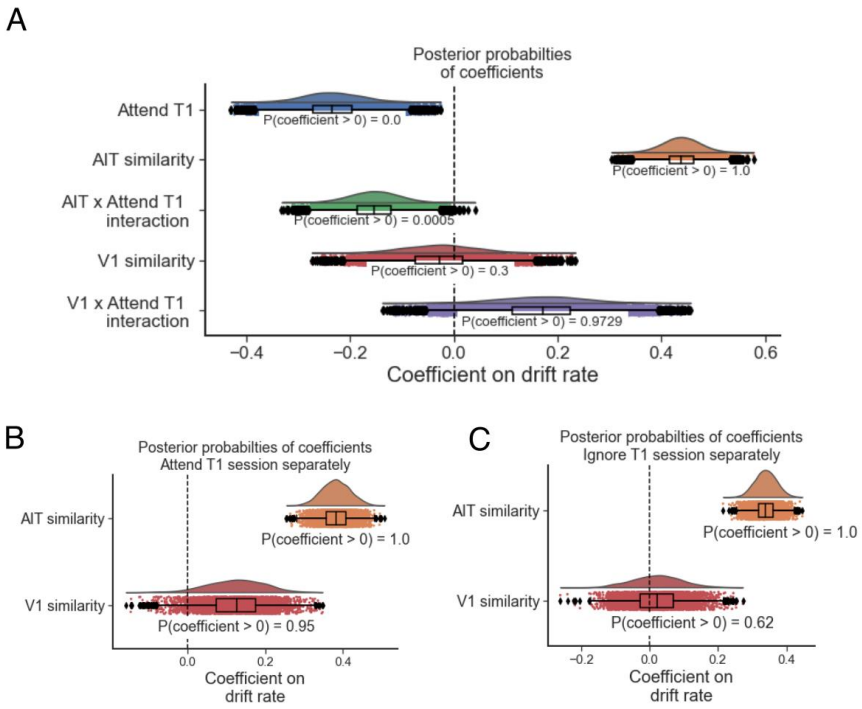


Figure 3. *Posterior probabilities of coefficients on drift rate from 24000 MCMC samples (see methods). The posterior probabilities can be used in hypothesis testing by accepting any outcome that is 95% (0.95 fractional, see (Cavanagh et al., 2011) for similar methods) above or below 0.* ***A)*** *Full model with regression coefficients on drift rate for Attending T1, AIT similarity, AIT x Attending T1 interaction, V1 similarity, and V1 similarity x Attending T1 interaction (see methods for the complementary covariates). Attending T1 was associated with a negative coefficient on drift rate, implying that attending T1 reduces the speed of evidence accumulation. While the main effect AIT similarity increases the drift rate, V1 similarity showed no such effect. However, the interaction effect with attending T1 was significant for both levels of similarity. These interaction effects were interestingly in opposite directions, where attention interferes with the priming effect of AIT similarity but enhances the effect of V1 similarity on drift rate.* ***B)*** *By repeating the analysis on the Attending T1 condition separately, we confirm that there is a main effect of both AIT and V1 similarity.* ***C)*** *The same analysis on the Ignore T1 condition showed no main effect of V1 similarity on drift rate, but an effect of AIT similarity was present.*

AIT and V1 similarity affect drift rate when attending to T1

Separately for the Attend T1 condition (Figure 3B) we estimated the same DDM parameters as in the full model: parameter a (criterion, M=1.663, SD=0.057), v (drift-rate/accumulation speed, M=1.4, SD=0.06), t (non-decision time, M=0.28, SD=0.013). For our regressors of interest we estimated these coefficients on drift-rate: 1) Target-target similarity in V1 (M=0.123, SD=0.075, P(coefficient > 0) = 0.95). 2) Target-target similarity in AIT (M=0.381, SD=0.036, P(coefficient > 0) > 0.99996). Similar to the full model, we also estimated these covariates: Covariate 1) Category congruence (M=0.367, SD=0.013, P(coefficient > 0) > 0.99996). Covariate 2) T1 complexity (M=-0.0056, SD=0.001, P(coefficient < 0) > 0.99996). These results show that attention modulates the effect of similarity differently depending on where the two targets overlap in neural representation. It is important to note that we included condition repetition (if both targets were animate or if both were inanimate), so the effect AIT similarity was not just present regardless of if T1 was attended or not, but it cannot be explained by response bias. We evaluated the convergence of the chains for the different parameters through visual inspection (Supplementary figure 2).

AIT similarity only modulates drift rate when ignoring T1

Since we did not find any main effect of V1 similarity on drift rate, but there was an interaction effect with attention, we decided to analyse both sessions separately. We first analysed the Attending T1 condition (Figure 3B) and estimated the same model parameters as with the full model: parameter *a* (criterion, M=1.663, SD=0.05), *v* (drift-rate/accumulation speed, M=1.407, SD=0.065), *t* (non-decision time, M=0.286, SD=0.013). We also included the same covariates as the full model (condition repetition and T1 complexity) and these regressors of interest: V1 similarity (M=0.123, SD=0.07, P(coefficient > 0 = 0.95)) and AIT similarity (M=0.381, SD=0.036, P(coefficient > 0 = 0.99996)). We then analysed the Ignore T1 condition (Figure 3C): parameter *a* (criterion, M=1.759, SD=0.054), *v* (drift-rate/accumulation speed, M=1.481, SD=0.075), *t* (non-decision time, M=0.211, SD=0.012). For

our regressors of interest we estimated these coefficients on drift-rate: 1) Target-target similarity in V1 (M=0.021, SD=0.071, P(coefficient > 0) = 0.62). 2) Target-target similarity in AIT (M=0.339, SD=0.031, P(coefficient > 0) > 0.99996). Similar to the full model, we also estimated these covariates: Covariate 1) Category congruence (M=0.0496, SD=0.016, P(coefficient > 0) > 0.99996). Covariate 2) T1 complexity (M=-0.0058, SD=0.001, P(coefficient < 0) > 0.99996).

Discussion

The aim of the current study was to evaluate the role of target competition in RSVP, specifically the effect of target similarity on perceptual decisions in RSVPs. We presented participants with two targets (T1 and T2) embedded with distractors and instructed them to make a speeded judgment if an animal was present in T2. Participants in separate conditions were instructed either to ignore or to attend T1 for a subsequent report. This allowed us to investigate how attention, one of the core theoretical elements of AB, interacts with target similarity at either the stage of V1 or AIT. It further allowed us to test one aspect of the most popular models of AB, the two-stage model (Chun & Potter, 1995), which posits that attending T1 has no effect on the perceptual processing of T2. Using HDDM (Wiecki et al., 2013) we estimated coefficients for our conditions on two latent parameters within DDM, i.e., drift rate (speed of evidence accumulation) and the non-decision variable (associated with encoding and motor response). We find that attention exhibits a push and pull relationship with target similarity, whereby attention increases the speed of evidence accumulation for targets that are similar in V1 while decreasing evidence accumulation of T2 when targets are similar in AIT (Figure 3).

In two previous studies (Lindh et al., 2019; Lindh et al., 2021), we have shown that target-target similarity in low-level visual features of natural images can enhance T2 performance in an AB task. These findings have been at odds with another well-known phenomena where repetition of a stimulus lead to impairment of reporting T2 (RB,

(Bavelier, 1994; Buffat et al., 2013; Fagot & Pashler, 1995; Kanwisher, 1987; Kanwisher & Potter, 1990; Park & Kanwisher, 1994)). However, consistent with theories of RB, we have also shown that representational similarity between targets in high-level visual and semantic brain areas is detrimental for T2 performance (Lindh et al., 2021). RB is a counterintuitive notion, considering the robustness of priming phenomena (Monahan et al., 2008; Schacter & Buckner, 1998). RB-like effects seem to depend on task-relevance (Bavelier, 1994; Sy & Giesbrecht, 2009), where it is crucial that the two targets are reported on the same dimension, implicating memory failure, and not perceptual interference, as the underlying cause (Fagot & Pashler, 1995). In our experiment we controlled for target-congruence (if T1 and T2 both contained an animal or if both did not), therefore, any effect of AIT similarity cannot be due to a response bias but an inherent effect of similarity in other high-level visual features. We show that target-target similarity in AIT increases the drift rate for T2, regardless of whether participants were asked to attend or ignore T1. This was an expected effect based on the different task requirements for T1 and T2. However, a clear negative interaction effect between AIT similarity and attending T1 was found, indicating that attention reduces the priming effect of similarity in higher-tier visual areas, suggesting a complementary type of deficiency to RB. Furthermore, we corroborate our previous findings of a facilitating effect through V1 similarity on T2 performance (Lindh et al., 2021) by showing that V1 similarity also increases speed of evidence accumulation. However, this effect is only present when participants are asked to attend T1 (Figure 3B). In the introduction we argued that one potential mechanism for V1 similarity to enhance performance is through neural adaptation. Attention is known to amplify neural activity (Luck et al., 1997; Posner & Gilbert, 1999; Roelfsema et al., 1998), and attention can modify neural adaptation (Alais & Blake, 1999), presumably through recurrent mechanisms (Quiroga et al., 2019). Therefore, it is possible that attending T1 is necessary to increase the adaptation effects, which in turn leads to faster evidence accumulation if T2 shares similar scene statistics with T1.

It has been argued that AB and RB are two distinct phenomena (Arnell & Shapiro, 2011; Chun, 1997), however, our results imply that they might be closer to each other in certain respects. The AB typically requires two preconditions, lag and attention to T1. In our previous paper (Lindh et al., 2021) we showed that AB magnitude (ABM, defined as the difference in lag-7 and lag-2 performance) can be explained in part by the similarity between targets arising at different levels of processing. The task in our current experiment is neither a pure AB nor RB task, in the traditional sense, considering our design has longer presentation times with a speeded judgment on T2. However, this setup allowed us to test the second concept of AB, attention to T1, with a more sensitive measure than pure T2 performance by instead modelling reaction time distributions together with accuracy. We show that the attention to T1 also modulates how similarity between targets affects perceptual processes associated with T2. This interaction with attention and evidence accumulations has clear consequences for many popular theories of AB. Specifically, most AB theories revolve around the two-stage model (Chun & Potter, 1995; Dux & Marois, 2009). Simplified, in a two-stage model, the two targets are first processed, in parallel, up to a semantic level without interference. In the *serial* second stage, T1 is being consolidated into memory and T2 cannot be consolidated until T1 has been fully processed. DDM allows for an important distinction between perceptual decision parameters (the *a*, *v*, and *z* parameters) and the non-decision parameter *t*. The non-decision parameter is associated with auxiliary processes such as motor initiation and memory encoding. A strict two-stage, late bottleneck, model predicts that attending T1 would only affect the t-parameter, and not interfere with the perceptual processing of T2. While we find that attending T1 does affect the t-parameter, corroborating the notion of a bottleneck, we also find that attending T1 also affects drift rate (Figure 3). Not only is the main effect of attention present, but also interaction effects with both AIT and V1 similarity between targets. First, this implies that attending T1 affects both decision-related and unrelated processes, where attention has double negative consequences for T2 processing by increasing the non-decision time as well as slowing down the drift rate. Second, it points to a duality where attention interacts with similarity in

opposite directions depending on where in the visual hierarchy the targets share neural representations.

In this study we elucidate the role of similarity between targets and attention in animacy detection. Looking at similarities between targets at different levels of processing provides a new window into mechanisms underlying phenomenon such as AB and RB. This provides further information into how processing of several targets interacts, depending on where in the visual hierarchy they overlap in representational space. We show that attending T1 interacts with how similarity between targets in V1 and AIT affects the speed of evidence accumulation of animal detection in natural scenes. Interestingly, this interaction goes in opposite ways for V1 and AIT, where attention is needed for V1 to increase evidence accumulation while attention suppresses the effect of AIT. Our data provides evidence that attending T1 disrupts T2 processing by both prolonging the non-decision time as well as slowing down the speed of evidence accumulation, providing specific behaviour which can be utilized to evaluate future models of attentional blink.

References

Akyürek, E. G., Hommel, B., & Jolicoeur, P. (2007). Direct evidence for a role of working memory in the attentional blink. *Memory & Cognition*, *35*(4), 621–627.

Alais, D., & Blake, R. (1999). Neural strength of visual attention gauged by motion adaptation. *Nature Neuroscience*, *2*(11), 1015–1018.

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Dowdle, L. T., Caron, B., Pestilli, F., Charest, I., Benjamin Hutchinson, J., Naselaris, T., & Kay, K. (2021). A massive 7T fMRI dataset to bridge cognitive and computational neuroscience. In *Cold Spring Harbor Laboratory* (p. 2021.02.22.432340). https://doi.org/10.1101/2021.02.22.432340

Arnell, K. M., & Shapiro, K. L. (2011). Attentional blink and repetition blindness. *Wiley Interdisciplinary Reviews. Cognitive Science*, *2*(3), 336–344.

Bavelier, D. (1994). Repetition blindness between visually different items: the case of pictures and words. *Cognition*, *51*(3), 199–236.

Bavelier, D., & Potter, M. C. (1992). Visual and phonological codes in repetition blindness. *Journal of Experimental Psychology. Human Perception and Performance*, *18*(1), 134–147.

Buffat, S., Plantier, J., Roumes, C., & Lorenceau, J. (2013). Repetition blindness for natural images of objects with viewpoint changes. *Frontiers in Psychology*, *3*(January), 1–11.

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*(11), 1462–1467.

Charest, I., Kriegeskorte, N., & Kay, K. N. (2018). GLMdenoise improves multivariate pattern analysis of fMRI data. *NeuroImage*, *183*(May), 606–616.

Chun, M. M. (1997). Types and Tokens in Visual Processing: A Double Dissociation between the Attentional Blink and Repetition Blindness. *Journal of Experimental Psychology. Human Perception and Performance*, *23*(3), 738–755.

Chun, M. M., & Potter, M. C. (1995). A Two-Stage Model for Multiple Target Detection in Rapid Serial Visual Presentation. In *Journal of Experimental Psychology: Human Perception and Performance* (Vol. 21, Issue 1, pp. 109–127). https://doi.org/10.1037/0096-1523.21.1.109

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., & Schwartz, O. (2007). Visual adaptation: neural, psychological and computational aspects. *Vision Research*, *47*(25), 3125–3131.

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

Dragoi, V., Sharma, J., & Sur, M. (2000). Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron*, *28*(1), 287–298.

Dux, P. E., & Marois, R. (2009). The attentional blink: a review of data and theory. *Attention, Perception & Psychophysics*, *71*(8), 1683–1700.

Evans, K. K., & Treisman, A. (2005). Perception of objects in natural scenes: Is it really attention free? *Journal of Experimental Psychology. Human Perception and Performance*, *31*(6), 1476–1492.

Fagot, C., & Pashler, H. (1995). Repetition Blindness: Perception or Memory Failure? *Journal of Experimental Psychology. Human Perception and Performance*, *21*(2), 275–292.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierachical processing in the primate cerebral cortex. *Cerebral Cortex* , *1*(1), 1–47.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Harris, I. M., & Dux, P. E. (2005). Orientation-invariant object recognition: Evidence from repetition blindness. *Cognition*, *95*(1), 73–93.

Kanwisher, N. G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, *27*, 117–143.

Kanwisher, N. G., & Potter, M. C. (1990). Repetition Blindness: Levels of Processing. *Journal of Experimental Psychology. Human Perception and Performance*, *16*(1), 30–47.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*. https://doi.org/10.1038/s41593-019-0392-5

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013). GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Frontiers in Neuroscience*, *7*(7 DEC), 1–15.

Lindh, D., Sligte, I. G., Assecondi, S., Shapiro, K. L., & Charest, I. (2019). Conscious perception of natural images is constrained by category-related visual features. *Nature Communications*, *10*(1), 4106.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*, 740–755.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology*, *77*(1), 24–42.

Luck, S. J., Vogel, E. K., & Shapiro, K. L. (1996). Word meanings can be accessed but not reported during the attentional blink. In *Nature* (Vol. 383, Issue 6601, pp. 616–618). https://doi.org/10.1038/383616a0

MacKay, D. G., & Miller, M. D. (1994). Semantic Blindness: Repeated Concepts Are Difficult to Encode and Recall Under Time Pressure. *Psychological Science*, *5*(1), 52–55.

Marois, R., Yi, D. J., & Chun, M. M. (2004). The Neural Fate of Consciously Perceived and Missed Events in the Attentional Blink. *Neuron*, *41*(3), 465–472.

Monahan, P. J., Fiorentino, R., & Poeppel, D. (2008). Masked repetition priming using magnetoencephalography. *Brain and Language*, *106*(1), 65–71.

Nieuwenstein, M. R., Potter, M. C., & Theeuwes, J. (2009). Unmasking the attentional blink. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 159–169.

Park, J., & Kanwisher, N. (1994). Determinants of Repetition Blindness. In *Journal of Experimental Psychology: Human Perception and Performance* (Vol. 20, Issue 3, pp. 500–519). https://doi.org/10.1037/0096-1523.20.3.500

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203.

Posner, M. I., & Gilbert, C. D. (1999). Attention and primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(6), 2585–2587.

Quiroga, M. D. M., Morris, A. P., & Krekelberg, B. (2019). Short-Term Attractive Tilt Aftereffects Predicted by a Recurrent Network Model of Primary Visual Cortex. *Frontiers in Systems Neuroscience*, *13*, 67.

Ratcliff, R., & Childers, R. (2015). Individual Differences and Fitting Methods for the Two-Choice Diffusion Model of Decision Making. *Decision (Washington, D.C.)*, *2015*. https://doi.org/10.1037/dec0000030

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Raymond, J. D., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in a RSVP task: an attentional blink? *Journal of Experimental Psychology*, *18*(3), 849–860.

Roelfsema, P. R., Lamme, V. A., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*(6700), 376–381.

Sawamura, H., Orban, G. A., & Vogels, R. (2006). Selectivity of neuronal adaptation does not match response selectivity: a single-cell study of the FMRI adaptation paradigm. *Neuron*, *49*(2), 307–318.

Schacter, D. L., & Buckner, R. L. (1998). Priming and the brain. *Neuron*, *20*(2), 185–195.

Seijdel, N., Loke, J., van de Klundert, R., van der Meer, M., Quispel, E., van Gaal, S., de Haan, E. H. F., & Scholte, H. S. (2021). On the necessity of recurrent processing during object recognition: it depends on the need for scene segmentation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2851-20.2021

Shapiro, K., Driver, J., Ward, R., & Sorensen, R. (1997). *Priming from the Attentional Blink : A Failure to Extract Visual Tokens but Not Visual Types*. *8*(2), 95–100.

Shapiro, K. L., Hanslmayr, S., Enns, J. T., & Lleras, A. (2017). Alpha, beta: The rhythm of the attentional blink. *Psychonomic Bulletin & Review*, *24*(6), 1862–1869.

Sy, J. L., & Giesbrecht, B. (2009). Target-target similarity on the attentional blink: Task-relevance matters! *Visual Cognition*, *17*(3), 1–10.

Vinken, K., Boix, X., & Kreiman, G. (2020). Incorporating intrinsic suppression in deep neural networks captures dynamics of adaptation in neurophysiology and perception. *Science Advances*, *6*(42). https://doi.org/10.1126/sciadv.abd4205

Vinken, K., Vogels, R., & Op de Beeck, H. (2017). Recent Visual Experience Shapes Visual Processing in Rats through Stimulus-Specific Adaptation and Response Enhancement. *Current Biology: CB*, *27*(6), 914–919.

Vogel, E. K., & Luck, S. J. (2002). Delayed working memory consolidation during the attentional blink. *Psychonomic Bulletin & Review*, *9*(4), 739–743.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206–1220.

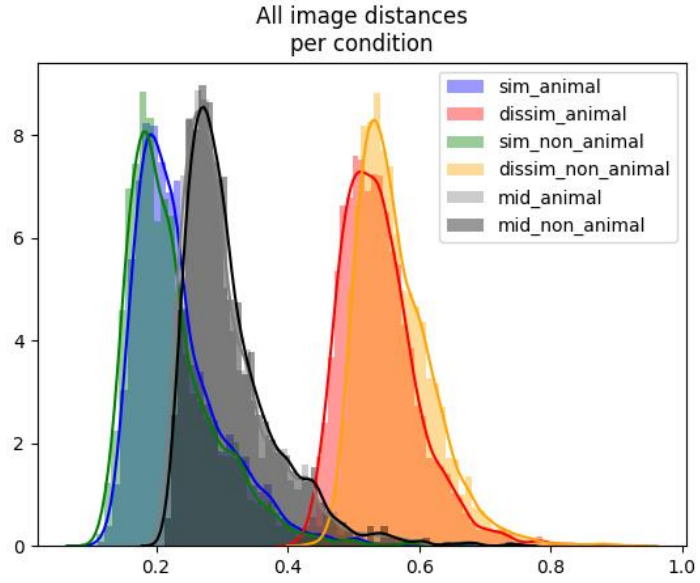
Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 14.

Supplementary

Table S1

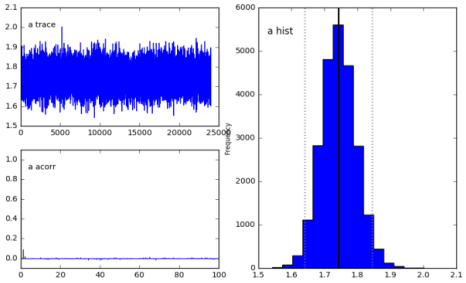
D-prime for attending/ignoring T1 and T1 animacy

Attend T1	T1 animal	Mean	SD	N
Attend	Animal	2.453	0.646	52
	Non-animal	2.499	0.862	52
Ignore	Animal	2.380	0.652	52
	Non-animal	2.397	0.960	52

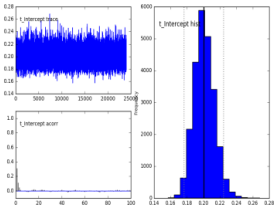


Supplementary Figure 1. Distributions of image similarities within all types of conditions. No visible difference between animal and non-animal distributions, indicating no bias between conditions.

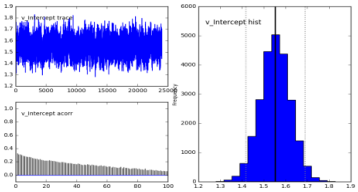
A



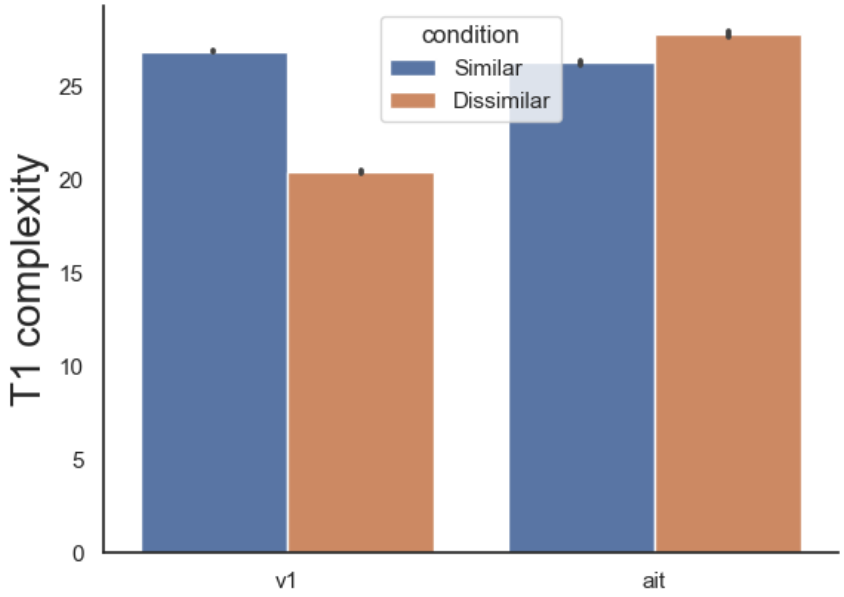
B



C



Supplementary Figure 2. *Posterior plots for DDM parameters. By visually inspecting the normal distribution of the chains one can infer that the chains have converged. While there is no guarantee of convergence for a finite sample set, ensuring that there are no drifts or jumps in the trace (the trace seems to overall be fluctuating over a specific value) is a good heuristic for convergence. Another heuristic is to ensure that the autocorrelation is relatively low. A) Posterior values for parameter a (threshold) indicates how much evidence (criterion) needed for subjects to make a decision. Top-left panel indicates the values for each chain, the right panel shows the distribution of these values. A normal distribution is indicative of converging chains. Bottom-left panel shows the autocorrelation. B and C shows the same for t (the non-decision time) and v (drift-rate).*



Supplementary Figure 3. *In a pilot study we noticed that when selecting image-pairs based on similarity from V1 and AIT that the T1 complexity became biased. As it is known that more complex T1 images (see Figure 1) would be processed quicker (Kar et al., 2019) we argue that this would affect the influence of T1 on T2 processing positively. We defined the complexity of an image by its average activation of the first layer of an AlexNet, where if the image would activate more feature units in layer 1 of AlexNet it would be considered more complex. To simulate the bias, we randomly selected 50 animate and 50 non-animate hypothetical T2s and then selected T1s that were either similar or dissimilar in V1 and AIT. We then saved the average complexity of T1 for each condition. This procedure was repeated 1000 times to obtain a confidence interval. This simulation confirmed that when image pairs are selected based on V1 similarity, dissimilar T1s are less complex and would thus lead to a quicker T2 processing. This effect was reversed for AIT similar*

General discussion

In every waking moment in our lives an immense amount of information reaches our sensory organs. The ability to filter out non-essential information is crucial for preserving computational resources used to recognise the objects which are pertinent for our current goals. Decades of research has given us a good understanding of the visual processing stream, from the retina to early visual areas to the more category-dependent and view-independent representations in inferior temporal cortex (ITC; DiCarlo, Yoccolan, and Rust 2012). The realisation of this hierarchical organisation has inspired a new generation of computational vision models; deep convolutional neural networks (DCNNs; Krizhevsky, Sutskever, & Hinton 2012). These networks mimic the hierarchical structure seen throughout the human visual system and are arguably the most promising models of the brain's visual system to date (Khaligh-Razavi and Kriegeskorte 2014). In both the brain and in DCNNs, the further along the processing stream the more the representations become category-specific and view-invariant. With a large corpus emerging where networks are fine tuned to fit brain data even better, an increased interest is also surfacing to use DCNNs as models for clinical conditions (Bonnen, Yamins, and Wagner 2020) or to manipulate neural populations (Bashivan, Kar, and DiCarlo 2019). To complement this, in this thesis I explore the usage of the intrinsic representations within DCNNs and the brain to predict behaviour in rapid object recognition. However, no sensory stimulus is an island. The perception (and the concomitant neural responses) of a target is strongly dependent on both spatial (what surrounds the object) and temporal (what was observed in the past) context. Processing a stimulus in isolation is challenging enough, however, the complex world we live in does not provide information in discrete, easily distinguishable portions but rather with an abundance of information perpetually reaching our sensory organs. In this thesis, I sought to understand how object recognition, parallel processing of natural images and conscious access relate to each other. I evaluate

the role of the relationship between object representations, both between specific image pairs and between categories of stimuli, and the propensity for conscious access to rapidly presented natural scenes. With this, this thesis attempts to provide a nuanced view of how semantic categories and target interactions during parallel processing affect our ability to perceive the world.

In chapter two we investigated categorical differences regarding their propensity of conscious access during short and challenging presentation rates. Previous research has indicated that animate objects are more efficiently processed compared to inanimate objects (Jackson and Calvillo 2013; Nairne et al. 2013; Guerrero and Calvillo 2016; New, Cosmides, and Tooby 2007), and the specific representational geometry in the human ITC predicts how quickly an object is correctly identified as animate (Carlson et al. 2014). This implies that the categorical organisation within ITC might reflect a bias towards processing certain categories over others. However, beyond the broad category boundaries between animate and inanimate, there is not much knowledge regarding how more specific categorical groups might differ. We narrowed this gap in our knowledge by selecting groups of visual objects known to cluster together in a distributed multivariate representational code within ITC (Charest et al. 2014). By presenting two targets (T1 and T2, respectively) embedded into a stream of distractors, we tested the difference in propensity for conscious access between semantic categories. In a typical Attentional Blink (AB) task, participants are impressively good at identifying a single target within the stream, even at very rapid presentation rates. However, when two targets are presented and T2 follows T1 by 200-500 ms, participants are often unable to correctly report the T2 identity (Raymond, Shapiro, and Arnell 1992). We first show that animate objects are less affected by the AB window, corroborating previous evidence that animate objects are more efficiently processed (New, Cosmides, and Tooby 2007; Guerrero and Calvillo 2016; Jackson and Calvillo 2013). This finding is different from only detection (for example only looking at lag-2 performance) since we baseline each image with its performance at lag-7, meaning that the effect cannot be attributed to the choice of masks and their

influence on performance for each category. Furthermore, we show that there is a significant variance between smaller sub-categories within the animate and inanimate division, extending previous notions of categorical differences in visual processing. Using the hierarchical organisation of a DCNN we tested which type of features (derived from different layers of the DCNN) best predict the variance of AB magnitude (ABM) between images. Here, importantly, the variance in ABM was best predicted by high-level visual features implying that it is the categorical organisation, not the shared low-level visual features within categories, which explain the differences in processing priority. In an exploratory phase, we further tested how similarity between features in the DCNN between targets affects the T2 reportability. We found, contrary to the literature on repetition blindness, that similarity between targets is beneficial for T2 reportability. In a second experiment, we confirm this finding by directly creating trials where T1 and T2 were either similar or dissimilar in terms of mid-level visual features, and thus directly manipulating participants' reportability rather than post hoc correlations. We hypothesised that our finding is since these visual features are not used in working memory consolidation of the actual object, whereas previous studies have shown how important task-relevance is for repetition blindness. It is possible that the proclivity in the literature for using simple stimuli, such as letters and digits, have prohibited researchers from discovering this effect earlier. By embracing the complexity of natural scenes with the usage of multivariate analysis methods and DCNNs, this enabled me to not only to show that categorical differences in ABM are due to high-level visual features but also that there is an effect of similarity that affects T2 reportability.

In the third chapter, inspired by our finding of target-target similarity enhancing performance, we continued to probe a conundrum of target-target similarity. On one hand, we have a large corpus of repetition blindness findings, where a repetition of targets leads to an impediment of the T2 report. On the other hand, we note in chapter two an improvement of T2 reportability when targets have similar visual features. By utilising similarity measures between natural images using functional magnetic resonance imaging (fMRI),

electroencephalogram (EEG), and a DCNN we asked how target-target similarity at different levels of processing affects T2 performance. Here, we replicated repetition blindness findings and showed that when targets are similar in high-level visual and brain areas associated with semantic representations (Binder et al. 2009) the T2 reportability suffers. This finding extends previous experiments of repetition blindness and implies that overlap in neural representation is key to the deficiency. Prior research on repetition blindness has shown that an exact replication of the stimulus is not necessary to obtain the effect. Participants will miss the second target even when the two targets are represented differently, e.g., 7 and "SEVEN" or homophonic word pairs such as rain/reign. Indeed, these previous studies imply that the suppressive mechanism is not related to visual features per se, but to phonological overlap. However, our results suggest that this is only part of the story, with target-target similarity in several brain areas associated with high-level visual features, semantics as well as phonological processing all correlate with behaviour.

We replicate our findings from chapter two to show that similarity in V1 between T1 and T2 leads to increased T2 performance. There are several possible reasons for our opposing findings, where the similarity between targets can both increase and decrease the probability for correct T2 reports depending on which level the processing of targets is interacting. Previous research has shown that when T2 is preceded with a cue of the same colour, T2 performance is enhanced (Nieuwenstein et al. 2005), implying that T2 processing is susceptible to subtle priming by low-level visual features. However, although this finding was robust using similarities from fMRI brain data and DCNN, it was not reflected in similarities derived from EEG. Here, we would expect this to be seen in the early time points (around 100 ms and onwards) where the initial decoding of images is often found. This might reflect a difficulty of extracting similarity measures from rapidly presented stimuli (stimuli was only shown for 16 ms) or possibly the difference in estimating distances (with Pearson correlation used in CNN and fMRI and decoding used in EEG). Furthermore, while the DCNN modelling successfully replicated the results of V1 similarity, it

failed to replicate findings of impaired T2 performance for trials when targets were similar in high-level semantic brain areas and late processing stages from the EEG. This might reflect the fact that AlexNet is trained on object recognition, for example, to locate distinct visual features that represent a dog compared to a cat, rather than semantics, i.e., realising that a dog is related to a leash without displaying any similar visual features. Recent efforts to train DCNNs with semantics have indeed led to better fit to neural data from the late ventral visual stream (Devereux, Clarke, and Tyler 2018). However, one caveat is that many objects that we reason are semantically related are defined by how they are used together in action.

While representational overlap in V1 and high-level visual and semantic brain areas seem to explain inter-stimuli and trial variance of ABM, we also asked if representational distinct representations could explain individual differences. We, therefore, ran a searchlight procedure where we correlated individual performance with the average similarity between targets based on iteratively centring a sphere on each voxel in the brain and including all voxels within the sphere in a pairwise similarity metric. We found that participants who perform well in the task have a larger distance between target representations in the right temporoparietal junction (rTPJ) and right inferior frontal gyrus (rIFG). These areas have previously been associated with a bottom-up saliency network (Corbetta, Patel, and Shulman 2008) and are believed to be crucial for working memory updating. Furthermore, recent research on individual differences has also highlighted rTPJ, with higher grey matter density and connectivity with IFG (Zhou et al. 2020). Connectivity between rTPJ and IFG has also been shown to be a hallmark sign for successful reports of T2 (Gross et al. 2004). Earlier research has argued that the speed of encoding, indicated by an earlier P300 peak, explains why certain participants are so called “non-blinkers” (Martens et al. 2006). Therefore, it is important to note that our task in the fMRI was a slow working memory task, designed to achieve a high signal-to-noise ratio and stable representations for each image, and only several weeks later did participants do the RSVP task. The task in the fMRI was like the RSVP in the regard that participants needed to encode an image

in working memory for several seconds, engaging attention, perception, and memory-related processes. However, an important difference is that the display time of stimuli in the fMRI task was at 700 ms and participants had ample time to encode the natural scene and participants performed at the ceiling. It is then possible that this more distinct representation between objects in the bottom-up saliency network, evident when encoding slowly presented images, facilitates fast processing during RSVP. Further research is needed to evaluate this possibility.

In the fourth chapter, we probed the relationship between image similarities and attention to T1. Within the AB, there are two main concepts: temporal distance between targets (i.e., lag) and attending or ignoring T1. In chapters two and three we focused on AB magnitude, the difference in T2 performance between lag-7 and lag-2. This is a common measure to evaluate how much performance suffers from being within the AB window, that is, when T2 is presented 200-500 ms after T1. However, the reason the phenomenon was named "attentional" blink was due to the fact that when participants were asked to ignore T1, performance on T2 was improved indicating that it was an attention-related depletion that lead to the main effect (Raymond, Shapiro, and Arnell 1992). In fact, most theories of AB have been focused on how attending T1 affects T2 processing. To investigate how our findings of target-target similarity effects on T2 are related to attending T1 we designed a hybrid task with modulation of attention to T1, where participants were asked to make speeded judgments on whether the T2 scene contained an animal. In two different sessions we also asked participants to either ignore T1 or memorise T1 to report after the stream. This allowed us to collect reaction time on T2 as well as accuracy allowing for more informative dependent variables. We modelled several latent decision variables using drift diffusion modelling, where we were primarily interested in drift rate (speed of evidence accumulation) and non-decision time as a way of separating between perceptual and non-perceptual processing. While RB has been associated with memory failure (Fagot and Pashler 1995), we hypothesised that V1 similarity will prime perceptual processing before memory processes are engaged.

Surprisingly, we did not find a main effect of V1 similarity on drift rate, however, we did show a positive interaction effect between attending T1 and V1 similarity on drift rate. Further post-hoc analysis showed that target-target V1 similarity did affect drift rate, but only in the "attending T1"-condition. This implies that attending T1 is a necessary condition for V1 similarity between the targets to affect perceptual processing of T2. Furthermore, many models of the AB are based on strict bottleneck ideas, where attending T1 affects T2 processing in late stages after T2 has been perceptually processed. In corroboration of this we do find that attending T1 does, in fact, affect the non-decision time which is more associated with motor initiation and memory encoding. However, we also show that attending T1 negatively modulates drift rate, indicating that attending T1 both prolongs the non-decision time and slows down drift rate.

Neural mechanisms underlying performance modulation of competing stimuli

There are two related neurophysiological ideas that are interesting candidates for explaining our findings of similarity between targets, repetition suppression and more general, neural adaptation. Adaptation, in the context of neural processing, refers to the idea that neurons that are continuously firing will gradually lower their response over time (Whitmire and Stanley 2016). The proposed main advantage of such a mechanism is that it facilitates detection of changes in the environment by suppressing static information flow, where activation based suppression would decrease the salience of recently seen visual features (Schwartz, Hsu, and Dayan 2007). This neural adaptation is increased along the visual hierarchy (Dhruv and Carandini 2014), implying that there are cumulative contributions at multiple stages. Adaptation is known to change the tuning function of neurons (Whitmire and Stanley 2016), referring to the sensitivity a neuron has for a specific feature dimension. For example, for a neuron that encodes orientation, it will have a preferred angle to which responds the most with a gradual lower response to more distant angles. By presenting an orientation grating stimulus of a larger angle, for example 45 degrees larger than its preferred angle, the tuning

function of the neuron will shift temporarily in the opposite direction (Dragoi, Sharma, and Sur 2000). The timing of the following stimuli can drastically change if its representation is being attracted to or repulsed from the previous orientation (Quiroga, Morris, & Krekelberg, 2019), i.e., biasing the perception of the second orientation away or towards the first orientation. This idea could potentially explain how similarity in low-level visual features might be beneficial when two targets are presented in short succession.

Another interesting notion is how adaptation is used as a mechanism to discount the effects of visual noise. Vinken et al. (2020) presented participants with an adaptor image (a random noise pattern) for an extended time, and then superimposed a target-object using the same background adapter image or another noise pattern. The authors showed that when the object was presented with the same noise pattern as the adaptor image there was a significant increase in detection performance. By implementing a simple local "neuron" adaptation mechanism into an AlexNet DCNN architecture, the network exhibited a similar behaviour as to humans. One of the proposed functions for neural adaptation is that it increases our sensitivity to small changes in the environment, taking advantage of statistical regularities in image structures to optimise sensory coding (Schwartz, Hsu, and Dayan 2007). This idea was the rationale for why evidence accumulation speed might be enhanced when two targets share low-level similarities. By adapting to irrelevant scene statistics, and thus decreasing the neural response, the visual system could potentially be more efficient at evaluating the scene. Future work could attempt to model the findings reported in chapter four, with the interaction effect of attention and similarity, using a similar model as Vinken et al. (2020) together with an intrinsic attention module. At least on the surface, our findings in the fourth chapter reveal a seemingly idiosyncratic behaviour, which could imply a specific architecture and definite processing modules. To successfully recover these results using a simple local neural adaptation model would be a convincing finding and an inspiring start to refine models of the Attentional Blink.

While neural adaptation is best studied using single cell recordings, a related phenomenon from the fMRI literature is repetition suppression. Repetition suppression refers to the fact that the Blood Oxygen Level Dependent (BOLD) signal, or the MRI contrast of blood deoxyhaemoglobin, decreases in certain brain areas when an image is repeatedly shown. This effect distinguishes itself from neural adaptation since fMRI integrates signals from millions of neurons, where a decrease in BOLD does not necessarily relate directly to the lowered firing rate of single neurons. However, a recent study showed that the only model that captures a large variety of second order statistics, within fusiform face area (FFA) and V1, was a local scaling model which outperformed competing models such as neural tuning, repulsion or attraction models (Alink, Abdulrahman, and Henson 2018). This corroborates a link in the literature between local neural adaptation and repetition suppression in fMRI. Like repetition blindness, which is a behavioural phenomenon, repetition suppression does not always occur. For example, in FFA, an area known to respond strongly to faces (Kanwisher, McDermott, and Chun 1997), repeating face stimuli lead to repetition suppression but only when a face or symbol is familiar (Henson, Shallice, and Dolan 2000). Here, a repetition of unfamiliar stimuli instead led to an increase of BOLD response. Interestingly, a similar finding has been reported for repetition blindness, where repetitions of known words induce an impairment of reporting the second target but not a repetition of nonsense words (Coltheart and Langdon 2003). However, due to the complications of capturing fMRI data together with behaviour, to the best of my knowledge, there are no current studies that have successfully shown that repetition suppression and neural adaptation are the main mechanisms behind the behavioural effects of repetition blindness. Nevertheless, they can be argued to be one of the top contenders and further research is needed to establish their role.

Modelling the attentional blink

Throughout my years of investigating object recognition and conscious access using the AB as a tool, I have considered many of the available models that strive to explain all the different findings from the AB

literature. A model that describes the AB behaviour would be of great utility, not only for understanding attention but also to inform researchers using AB as a tool, allowing them to design more precise experiments. In an attempt to make a brief summary of the best known models, I'll start with the inhibition theory (Raymond, Shapiro, and Arnell 1992), which proposes that an attentional "gate" opens when T1 is observed. Immediately following stimuli is suppressed (gate closed) to reduce confusion during feature binding. However, Chun and Potter (1995) showed that both perceptually and categorically defined targets led to a blink, showing that the AB is not due to feature binding problems. Chun and Potter, in their two-stage model, instead proposed that all targets are initially processed perceptually, but need to pass a capacity-limited second stage to be impervious to decay/overwriting. Although, this notion didn't seem to work either. Di Lollo et al. (2005) showed that participants can report 3 consecutive targets (known as the extended lag-1 sparing), which arguably seems inconsistent with a capacity-limited account (Olivers, Van Der Stigchel, and Hulleman 2007). Di Lollo and colleagues instead proposed the temporal loss of control (TLC) model which posits a filter that selects targets and excludes distractors. A T1+1 distractor causes a disruption in the filter's configuration (loss of control) leading to slower processing of the following target. In the boost and bounce model (Olivers and Meeter 2008), T1 ignites an attentional "boost", which allows the T1+1 distractor to be processed. However, the detection of a non-target distractor elicits a "bounce" mechanism that inhibits T2, causing it to be overwritten and forgotten. In both TLC and the boost and bounce model the distractors are a crucial component for AB, however, the effect of AB has been observed even without intermediate distractors (Nieuwenstein, Potter, and Theeuwes 2009). Furthermore, as seen in our fourth chapter, where we don't have any intermediate distractors, a clear deficiency can be found in the drift rate when attending T1. It is possible that solely considering accuracy, where you only have "correct" or "incorrect" responses for each trial, is not sensitive enough to detect the scope of processing deficiencies induced by attending T1.

My personal, probably contentious, opinion of directly modelling AB is that it is a backwards notion at best. The reason for this, is that I argue that AB is an epiphenomenon of an attentional system that has evolved to expect new stimuli to occur at a certain cyclic rate. One of the main contenders for this cyclic expectation is the fact that humans make about 4-5 saccades per second (or once every 200-250 ms). This means that every day, from the time you wake up to the time you go back to bed, you are continuously sampling your environment, processing objects at the focal point of your current fixation and updating your working memory. Or in other words, the precise timings for AB are aligned with the sampling rate to which your brain is adherent to every waking moment. This sampling rate of 4-5 Hertz is known as theta when applied to brain waves and the phase of theta has a crucial role in object detection. For example, when measuring theta in monkeys at the frontal eye fields, lateral intraparietal area, and the mediodorsal pulvinar, studies have found that performance is significantly higher when a target is presented concurrently with the theta phase being at its peak (for review see Fiebelkorn and Kastner 2019). Interestingly, saccades modulate activity in thalamus (Leszczynski et al. 2020), early visual cortex (Purpura, Kalik, and Schiff 2003), as well as hippocampus (Hoffman et al. 2013). In fact, neurons in V1 are particularly responsive to targets presented within 100-150 ms after a saccade (Lowet et al. 2016; Gallant, Connor, and Van Essen 1998), implying a reset of an attentional episode with expectations of incoming stimuli. A similar reset can be argued to relate to theta waves within the hippocampus (Lisman and Jensen 2013), a brain area associated with episodic (or temporal) memory (Umbach et al. 2020). Within the hippocampus, the theta-phase has been theorised to support encoding at the trough and retrieval at the peak (Hasselmo, Bodelón, and Wyble 2002), and saccades reset the theta-phase, such that, at every new fixation the theta phase is at its peak (Hoffman et al. 2013). These studies corroborate the idea that saccades are central to the cyclic expectation that underlies attentional episodes. Attentional episodes refer to the fact that although your experience of the world seems continuous, the brain rather seems to integrate information within ‘‘volleys’’ of activity occurring in a cyclic manner at around 4-5 Hertz. If you hypothesise that target detection in

an RSVP is similar to a saccade, in the respect that it resets an attentional episode, then these animal studies (Lowet et al. 2016; Gallant, Connor, and Van Essen 1998) provide direct evidence for why we have lag-1 sparing and why target detection after 200-500 is impaired.

However, in the AB literature there is little mention of the role of saccades when explaining why a second target presented 200-500 ms after a first target is missed. To the best of my knowledge, there is only one study that investigates the role of saccades in the AB. The authors of this study demonstrated an improvement of T2 performance when participants were asked to make a saccade directly after T1 (Kamienkowski, Navajas, and Sigman 2012), thereby (theoretically) making a hard reset of the attentional episode. Therefore, I argue that we should not model the AB per se. Instead, we should strive to make models that have the same constraints as humans in terms of processing ability. Although our eye muscles are the fastest muscle movement, we are capable of, in a sense, our perception of the world is constrained by our actions (or ability to act). Therefore, a model with similar constraints, such as small focal points and a limitation to how fast it can make saccades to sample new information, and later trained to do object recognition of visual scenes could potentially be a promising model of AB. In fact, the idea of constraining a recurrent neural network with a small fovea, and the ability to sample new information, has been successfully implemented and trained on handwritten digits (Mnih et al. 2014). These restrictions enhanced the model's performance, indicating a computational role for saccades (Mnih et al. 2014). In my view, the most promising model of the AB would not model the epiphenomenon but instead the phenomenon. That is, a model with similar types of constraints as humans together with local neural adaptation to model the interaction potentially also between targets when they share representational geometry at different stages of processing.

The role of DCNNs in neuroscience

Most of the current studies on DCNNs and brain data have been focused on correlating the representations within the brain and DCNNs. These studies have successfully shown that DCNNs are our current most reliable model of the representational geometry at different stages of processing in the visual ventral stream (Khaligh-Razavi and Kriegeskorte 2014; Yamins et al. 2014), following a similar hierarchical structure as the visual stream (Eickenberg et al. 2017; Cichy, Pantazis, and Oliva 2014; Greene and Hansen 2018; Kietzmann et al. 2019). In recent years several studies have shown that representational geometry in high-level visual areas predict behaviour (Charest et al. 2014; Carlson et al. 2014; Ritchie, Tovar, and Carlson 2015). These findings provide crucial evidence for the notion that the information that is being decoded from these areas also have behavioural consequences as opposed to just being epiphenomenal (Grootswagers, Cichy, and Carlson 2018). Almost a decade after DCNNs changed the field of vision science, we are now starting to use DCNNs as models of clinical conditions (Bonnen, Yamins, and Wagner 2020) and researchers have been able to produce images designed to activate only a select population of neurons (Bashivan, Kar, and DiCarlo 2019). Overall, models of the visual system can be infinitely useful, and it is up to researchers to find ways of using them to propel knowledge.

In chapter two we show two creative ways of how DCNNS can be used to explain and manipulate behaviour, both of which are. First, after our finding that there are categorical differences in AB magnitude, we were posed with the conundrum that visual categories not only share high-level visual features but also low-level features (Torralba and Oliva 2003). By taking advantage of the hierarchical structure of AlexNet, where low-level visual features are processed in the first few layers and high-level information emerges in later layers, we showed that the prediction of AB magnitude increased with each layer. We thereby provide evidence that the best explanation for the categorical

differences in AB magnitude is that they are due to the high-level visual features. This implies that some semantic categories, that are known to share a distributed code in ITC (Charest et al. 2014; Bao et al. 2020; DiCarlo, Yoccolan, and Rust 2012; Kriegeskorte, Mur, and Bandettini 2008; Kanwisher, McDermott, and Chun 1997; Downing et al. 2001), are more likely to be consciously accessed and their high-level visual features best explains their differences. Second, we also show that similarity between the two targets can affect behaviour. In order to manipulate behaviour, we used the inner representations of AlexNet to select image pairs that were either similar or dissimilar. We showed that similar image pairs lead to an increased probability of correct T2 report, a notion that seems to contradict the repetition blindness literature (Kanwisher 1987; Coltheart, Mondy, and Coltheart 2005; Fagot and Pashler 1995; Park and Kanwisher 1994; Bavelier and Potter 1992). However, the usage of DCNNs allowed us to define similarity in a different, more objective, way than relying on perceptual intuitions, ratings or similar methods that are subject to our own biases. Since the network we used was trained on strict object classification, the similarities between images are probably related to specific object features. However, one can imagine that training networks on different tasks, from semantics to perceptual qualities, will yield a larger range of similarities which in turn can be used in a more systematic way to investigate the interactive effects of representational overlap.

Problems with DCNNs as models

Despite the success of DCNNs in vision sciences, their presence has not been without criticism. The main concerns can be summarised into three problems: (1) DCNNs do not learn the way humans do, (2) DCNNs make mistakes humans would never do, (3) and we are just substituting one black box with another. These problems have been argued over for many decades, but with the recent upsurge of DCNNs it seems as if more researchers are finding them useful and acceptance in the field is increasing.

First, neural networks were for many decades considered a pipe dream for scientists due the difficulty of determining how modifying the weights one unit would affect the system's overall behaviour. However, this changed in the 1980s with the popularisation of backpropagation (Rumelhart et al. 1985; LeCun et al. 1988), a method where errors of the output layer are propagated backwards throughout the network and therefore solving the problem by iteratively moving the parameters closer to a state that produces a desired output. The brain also learns by adjusting the connection strength between neurons, but feedback connections in the brain seem to have a very different role (Gilbert & Li 2013; Lamme, Supèr, and Spekreijse 1998) and human children seem to learn unsupervised, without the correct labels on every item in their surroundings. Despite these differences, recent proposals have been made arguing that the brain might approximate backpropagation as a learning mechanism using locally computed errors (Lillicrap et al. 2020). This is a drastically different idea than the commonly accepted Hebbian notion of learning, which states that “cells that fire together, wire together” (Hebb 1949). That is, correlated activity between connected neurons leads to a stronger synaptic connection between them, a principle that has been successful in explaining a wide range of plasticity mechanisms (Sumner et al. 2020). However, even if neural networks learn differently from humans, one could argue that the learning process is inconsequential if the final model shares computational characteristics with the brain.

The second problem is related to the fact that DCNNs are well-known to be vulnerable to network adversarial attacks (Goodfellow, Shlens, and Szegedy 2014). This can be done by adding an imperceptibly small amount of designed noise on top of an image of (for example) a panda, and the network would classify it as a gibbon with over 90% confidence (Goodfellow, Shlens, and Szegedy 2014). However, Firestone (2020) argues that these examples of unhuman behaviour do not reflect a meaningful difference in how information is processed. Instead, Firestone argues that humans also differentiate in how they perceive world, where some are scared of spiders, some perceive a blue and black dress to be gold and white (Schlaffke et al. 2015) and some humans make errors when constructing sentences, e.g.,

Spoonerisms (Palmquist 1980). But importantly, neural networks have different constraints than humans that make them vulnerable to different types of errors that are attributable to their constraints rather than their competence (Firestone 2020). This idea is illustrated in a study wherein by constraining a DCNN with a retina model and then perturbing an image of a cat until the network misclassified it to a dog. Importantly, the resulting image also fooled humans (Elsayed et al. 2018). Overall, biologically constrained DCNNs as a way of increasing robustness is gaining more ground in recent years (Girard et al. 2021; Zhang et al. 2019; Evans, Malhotra, and Bowers 2021). Developing neural networks with biological constraints might not only be beneficial in making them more like human cognition but biology can also inspire engineers to make more computationally efficient models.

Finally, Kay (2017) contends the idea of DCNNs as useful models of the visual system. He argues that neither the implementation nor the goals of the neural networks are comparable to humans, and therefore their utility to understand the brain is trivial at best. He also iterates a common critique of DCNNs; the enormous parameter space, in combination with non-linearities, makes for an opaque black box which is equally mysterious as the brain. In response to Kay (2017), (Scholte 2018) agrees that on an implementation-level DCNNs and brains are different but this is not a problem if we instead consider DCNNs the same way we do with animal models. In a sense, the complexity argument against DCNNs should also apply to animal models. For example, using rodent brains as models for medical treatments or understanding basic perception has been invaluable for science despite not having a full understanding of how their brains work. Having a potentially unlimited zoo of neural networks with different architectures, trained on different data sets with a variety of task goals could yield new knowledge unreachable with conventional methods. These models can be lesioned and manipulated with precision, enabling us to probe and make predictions about human behaviour and brain function. The discussion for the role of DCNNs in vision science will most certainly continue for a long time (unless they get replaced by newer ideas, such as transformers (Tuli et al. 2021)), but their impact on the field the past ten years is impossible to dismiss.

Relationship to clinical populations

Although all the chapters in this thesis include only healthy participants with no known neurological diseases, some of our findings may shed some light on the understanding of certain clinical populations. Many types of disorders have been associated with a lower AB performance, such as schizophrenia (Wynn et al. 2006), attention deficit hyperactivity disorder (Armstrong and Munoz 2003), and lesions in the parietal lobe (Husain et al. 1997; Shapiro, Hillstrom, and Husain 2002). Two related, but dissociable, neural disorders that connect to our findings are visual extinction and simultanagnosia. In visual extinction, most associated with a lesion around the right TPJ, the patient can attend to an object in the contralesional field (i.e., the left visual field) as long as there is no other salient object in the ipsilateral (to the lesion) visual field. In this case, it seems as if attending to the right visual field interferes with the patient's ability to perceive any object in the left visual field. In contrast, simultanagnosia describes the inability to perceive two objects even when presented within the same visual field. A patient with simultanagnosia being presented with a table with food and cutlery would for example only perceive a spoon. While visual extinction is associated with unilateral damage, simultanagnosia occurs after bilateral damage to the parietal lobe. The main commonality is that in both instances patients have an inability to resolve attentional conflict - an inability similar to that which causes the AB.

Damage to the right TPJ has been associated with a clear reduction of performance in the AB (Husain et al. 1997; Shapiro, Hillstrom, and Husain 2002), where the maximal extinction occurs when the ipsilateral object is presented slightly before the contralateral object (Cate and Behrmann 2002). Similarly, our findings in chapter 3 showed that individual differences in AB can be related to a more differentiable neural code between visual objects in high dimensional representational space within the right TPJ. In our data, there were no clear categorical organisations within the right TPJ, however,

participants who overall had more distinct representations between objects in TPJ were more successful in the task. This implies that this region does not maintain the semantic input itself but is instrumental in transferring perceptual information into working memory, resolving attentional conflict when multiple objects are being processed simultaneously. Furthermore, in both visual extinction and simultanagnosia there is evidence for the notion that similarity between the objects exacerbates the condition (Rafal et al. 2002; Ptak and Schnider 2005; Baylis, Driver, and Rafal 1993; Coslett and Lie 2008). For example, Rafal et al. (2002) asked patients with visual extinction to report the value of digits and numerical words presented simultaneously in each visual field. Patients showed a decreased ability to correctly report the left target when it was paired with a right visual field target that required the same response (e.g., 1 vs 1) regardless of if they were visually similar or not (e.g., 1 vs ONE). Similar to findings in repetition blindness (Bavelier and Potter 1992), Rafal et al. (2002) also found that phonologically similar pairs (e.g., ONE vs WON) led to a direct reduction in performance for reporting the item in the left visual field. One interpretation of this is that the deficits these patient experiences are not perceptual per se, but rather on a response level and related to the current task demands. This is mirrored in RSVPs studies with healthy participants showing that similarity between targets only affects performance when they are similar in the task-relevant domain (Sy and Giesbrecht 2009). Interestingly, even in a healthy population, with intact parietal function, when instructing participants to report two simultaneously presented stimuli (one in each visual field), participants show “pseudoextinction” (Goodbourn and Holcombe 2015), where targets in the right visual field more often omitted. Together, these findings indicate that even healthy individuals exhibit similar deficits (and due to the same competitive mechanism) as patients who suffered damage to the parietal lobe when they are put under highly demanding situations, such as RSVP. By furthering our understanding of the mechanisms behind all types of RSVP phenomena, we might be able to model the experience of certain clinical populations.

Conclusions

The goal for a researcher in any field is to investigate the world around them while attempting to describe an unfathomably complex reality while constrained by the limits of our language and cognitive abilities. We often use terms such as “understanding” to describe the pinnacle of our efforts. However, without a clear definition of what the term “understanding” means, it rather becomes an umbrella term used by researchers to avoid defining their exact desires. In psychology and neuroscience research, the word “understanding” often ends up meaning “is there a difference between these two conditions”, which indeed is an important starting point when there isn’t sufficient knowledge available to make any other predictions. However, I’d argue that modelling, prediction, and precise manipulation based on models are necessary, but perhaps not sufficient, goals for proper understanding.

Throughout this dissertation, these have been the key concepts used for understanding. In chapter 2 our initial question started with the examination of semantic categories and if they are differentially sensitive to the AB window. We found a large variance between categories and continued by predicting the individual images AB magnitude using hierarchical visual features derived from a DCNN. This allowed us to conclude that we can predict how likely an image is to gain conscious access based on its high-level visual features, demonstrating a new understanding at what level processing the AB might be occurring. In an explorative phase we also found that similarity in visual features between targets are beneficial for performance. We replicated this finding by selecting T1-T2 pairs based on their similarity in visual features, thereby using our model to manipulate behaviour. These findings seemed contrary to the current literature where similarity between targets often led to lower performance. In the two following studies (chapter 3 and 4) we show that the word “similarity” has been arguably under-defined in previous studies in the sense that natural images are complex, just like the world around us, and what it means for two natural scenes to be similar depends on where in the brain they are being processed at a given

moment. This denotes one of the main contributions of this thesis. Traditionally, studies of attention, perception, and working memory has made use of simple stimuli (such as simple geometric shapes, letters, digits, etc.) which has allowed a great deal of control over the experimental conditions. However, by embracing the complexity of natural images using DCNNs, fMRI and EEG together with multivariate methods lends utility in exploring the intricacies of the brain. The concept of a dog can be presented in unlimited variations, with different breeds, viewing angles, low-level scene statistics etc. On the other hand, the letter “G” does not allow for the same breath of variation, which thus limits the range of analyses for the researcher. Arguably, our findings of low-level similarity in V1 being beneficial for performance and conflict in high-level vision and semantic brain areas would probably not be possible without the usage of complex stimuli.

In conclusion, the complexity of our environment has shaped our brains, our bodies, and how we interact with our surroundings. The evolutionary relevance of certain categories has not only affected our preferences but also how the multivariate representations of categories in high-level visual areas are related to their proclivity to conscious access. By embracing the complexity of natural images, and utilizing a range of methods from machine learning, deep learning, fMRI/EEG, and cognitive modelling I have shown a nuanced picture of how natural images interact at different levels of processing and emerge into a state underlying conscious report of stimuli. Specifically, V1 similarity in multivariate representational space between images interacts with attention and enhances the speed of accumulating evidence for targets. Meanwhile, high-level similarity between targets has a negative impact on target performance where attention has an opposite effect compared to V1-similarity and decreases this impairment. These are important findings that not only shed new light on how object processing and attention interacts but can also be used by future modelling work as benchmark behaviours the models should exhibit.

References

Alink, Arjen, Hunar Abdulrahman, and Richard N. Henson. 2018. “Forward Models Demonstrate That Repetition Suppression Is Best Modelled by Local Neural Scaling.” *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-05957-0.

Armstrong, I. T., and D. P. Munoz. 2003. “Attentional Blink in Adults with Attention-Deficit Hyperactivity Disorder: Influence of Eye Movements.” *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale* 152 (2): 243–50.

Bao, Pinglei, Liang She, Mason McGill, and Doris Y. Tsao. 2020. “A Map of Object Space in Primate Inferotemporal Cortex.” *Nature* 583 (7814): 103–8.

Bashivan, Pouya, Kohitij Kar, and James J. DiCarlo. 2019. “Neural Population Control via Deep Image Synthesis.” *Science* 364 (6439). https://doi.org/10.1126/science.aav9436.

Bavelier, Daphne, and Mary C. Potter. 1992. “Visual and Phonological Codes in Repetition Blindness.” *Journal of Experimental Psychology. Human Perception and Performance* 18 (1): 134–47.

Baylis, Gordon C., Jon Driver, and Robert D. Rafal. 1993. “Visual Extinction and Stimulus Repetition.” *Journal of Cognitive Neuroscience* 5 (4): 453–66.

Bonnen, Tyler, Daniel L. K. Yamins, and Anthony D. Wagner. 2020. “When the Ventral Visual Stream Is Not Enough: A Deep Learning Account of Medial Temporal Lobe Involvement in Perception.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.10.07.327171.

Carlson, Thomas A., J. Brendan Ritchie, Nikolaus Kriegeskorte, Samir Durvasula, and Junsheng Ma. 2014. “Reaction Time for Object Categorization Is Predicted by Representational Distance.” *Journal of Cognitive Neuroscience* 26 (1): 132–42.

Cate, Anthony, and Marlene Behrmann. 2002. “Spatial and Temporal Influences on Extinction.” *Neuropsychologia* 40 (13): 2206–25.

Charest, Ian, Rogier A. Kievit, Taylor W. Schmitz, Diana Deca, and Nikolaus Kriegeskorte. 2014. “Unique Semantic Space in the Brain of Each Beholder Predicts Perceived Similarity.” *Proceedings of the National Academy of Sciences* 111 (40): 14565–70.

Cichy, Radoslaw Martin, Dimitrios Pantazis, and Aude Oliva. 2014. “Resolving Human Object Recognition in Space and Time.” *Nature Publishing Group* 17 (3): 455–62.

Coltheart, Veronika, and Robyn Langdon. 2003. “Repetition Blindness for Words yet Repetition Advantage for Nonwords.” *Journal of Experimental Psychology. Learning, Memory, and Cognition* 29 (2): 171–85.

Coltheart, Veronika, Stephen Mondy, and Max Coltheart. 2005. “Repetition Blindness for Novel Objects.” *Visual Cognition* 12 (3): 519–40.

Corbetta, Maurizio, Gaurav Patel, and Gordon L. Shulman. 2008. “The Reorienting System of the Human Brain: From Environment to Theory of Mind.” *Neuron* 58 (3): 306–24.

Coslett, H. Branch, and Eunhui Lie. 2008. “Simultanagnosia: Effects of Semantic Category and Repetition Blindness.” *Neuropsychologia* 46 (7): 1853–63.

Devereux, Barry J., Alex Clarke, and Lorraine K. Tyler. 2018. “Integrated Deep Visual and Semantic Attractor Neural Networks Predict fMRI Pattern-Information along the Ventral Object Processing Pathway.” *Scientific Reports* 8 (1): 1–12.

Dhruv, Neel T., and Matteo Carandini. 2014. “Cascaded Effects of Spatial Adaptation in the Early Visual System.” *Neuron* 81 (3): 529–35.

DiCarlo, James J., Davide Yoccolan, and Nicole C. Rust. 2012. “How Does the Brain Solve Visual Object Recognition?” *Neuron* 73 (3): 415–34.

Di Lollo, Vincent, Jun-Ichiro Kawahara, S. M. Shahab Ghorashi, and James T. Enns. 2005. “The Attentional Blink: Resource Depletion or Temporary Loss of Control?” *Psychological Research* 69 (3): 191–200.

Dragoi, V., J. Sharma, and M. Sur. 2000. “Adaptation-Induced Plasticity of Orientation Tuning in Adult Visual Cortex.” *Neuron* 28 (1): 287–98.

Eickenberg, Michael, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. 2017. “Seeing It All: Convolutional Network Layers Map the Function of the Human Visual System.” *NeuroImage* 152 (January 2016): 184–94.

Elsayed, Gamaleldin F., Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian

Goodfellow, and Jascha Sohl-Dickstein. 2018. “Adversarial Examples That Fool Both Computer Vision and Time-Limited Humans.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1802.08195.

Evans, Benjamin D., Gaurav Malhotra, and Jeffrey S. Bowers. 2021. “Biological Convolutions Improve DNN Robustness to Noise and Generalisation.” *bioRxiv*. https://doi.org/10.1101/2021.02.18.431827.

Fagot, Clark, and Harold Pashler. 1995. “Repetition Blindness: Perception or Memory Failure?” *Journal of Experimental Psychology. Human Perception and Performance* 21 (2): 275–92.

Fiebelkorn, Ian C., and Sabine Kastner. 2019. “Functional Specialization in the Attention Network.” *Annual Review of Psychology* 70: 77–110.

Firestone, Chaz. 2020. “Performance vs. Competence in Human-Machine Comparisons.” *Proceedings of the National Academy of Sciences of the United States of America*, 1–10.

Gallant, Jack, Charles E. Connor, and David Van Essen. 1998. “Neural Activity in areasV1, V2 and V4 during Free Viewing of Natural Scenes Compared to Controlled Viewing.” *Neuroreport* 9 (1): 1673–78.

Gilbert, Charles D., and Wu Li. 2013. “Top-down Influences on Visual Processing.” *Nature Reviews. Neuroscience* 14 (5): 350–63.

Girard, Benoît, Jean Lienard, Carlos Enrique Gutierrez, Bruno Delord, and Kenji Doya. 2021. “A Biologically Constrained Spiking Neural Network Model of the Primate Basal Ganglia with Overlapping Pathways Exhibits Action Selection.” *The European Journal of Neuroscience* 53 (7): 2254–77.

Goodbourn, Patrick, and Alex Holcombe. 2015. “‘Pseudoextinction’: Asymmetries in Simultaneous Attentional Selection.” *Journal of Experimental Psychology. Human Perception and Performance* 41 (2): 364–84.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. “Explaining and Harnessing Adversarial Examples.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1412.6572.

Greene, Michelle R., and Bruce C. Hansen. 2018. “Shared Spatiotemporal Category Representations in Biological and Artificial Deep Neural Networks.” *PLoS Computational Biology* 14 (7). https://doi.org/10.1371/journal.pcbi.1006327.

Grootswagers, Tijl, Radoslaw M. Cichy, and Thomas A. Carlson. 2018. “Finding Decodable Information That Can Be Read out in Behaviour.” *NeuroImage* 179 (March): 252–62.

Gross, Joachim, Frank Schmitz, Irmtraud Schnitzler, Klaus Kessler, Kimron Shapiro, Bernhard Hommel, and Alfons Schnitzler. 2004. “Modulation of Long-Range Neural Synchrony Reflects Temporal Limitations of Visual Attention in Humans.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (35): 13050–55.

Guerrero, Guadalupe, and Dustin P. Calvillo. 2016. “Animacy Increases Second Target Reporting in a Rapid Serial Visual Presentation Task.” *Psychonomic Bulletin & Review* 23 (6): 1832–38.

Hasselmo, Michael E., Clara Bodelón, and Bradley P. Wyble. 2002. “A Proposed Function for Hippocampal Theta Rhythm: Separate Phases of Encoding and Retrieval Enhance Reversal of Prior Learning.” *Neural Computation* 14 (4): 793–817.

Hebb, Donald Olding. 1949. “The Organization of Behavior; a Neuropsycholocigal Theory.” *A Wiley Book in Clinical Psychology* 62: 78.

Henson, Ron, Tim Shallice, and Raymond Dolan. 2000. “Neuroimaging Evidence for Dissociable Forms of Repetition Priming.” *Science* 287 (5456): 1269–72.

Hoffman, Kari L., Michelle C. Dragan, Timothy K. Leonard, Cristiano Micheli, Rodrigo Montefusco-Siegmund, and Taufik A. Valiante. 2013. “Saccades during Visual Exploration Align Hippocampal 3–8 Hz Rhythms in Human and Non-Human Primates.” *Frontiers in Systems Neuroscience* 7 (August): 1–10.

Husain, Masud, Kimron Shapiro, Jesse Martin, and Chistopher Kennard. 1997. “Abnormal Temporal Dynamics of Visual Attention in Spatial Neglect Patients.” *Nature* 385 (January): 154–56.

Jackson, Russell E., and Dusti P. Calvillo. 2013. “Evolutionary Relevance Facilitates Visual Information Processing.” *Evolutionary Psychology: An International Journal of Evolutionary Approaches to Psychology and Behavior* 11 (5): 1011–26.

Kamienkowski, Juan E., Joaquín Navajas, and Mariano Sigman. 2012. “Eye Movements Blink the Attentional Blink.” *Journal of Experimental Psychology. Human Perception and Performance* 38 (3): 555–60.

Kanwisher, Nancy. 1987. “Repetition Blindness: Type Recognition without Token Individuation.” *Cognition* 27: 117–43.
Kanwisher, Nancy, Josh McDermott, and Marvin Chun. 1997. “The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception.” *Journal of Neuroscience*. https://www.jneurosci.org/content/17/11/4302.short.
Khaligh-Razavi, Seyed Mahdi, and Nikolaus Kriegeskorte. 2014. “Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation.” *PLoS Computational Biology* 10 (11). https://doi.org/10.1371/journal.pcbi.1003915.
Kietzmann, Tim C., Courtney J. Spoerer, Lynn K. A. Sörensen, Radoslaw M. Cichy, and Olaf Hauk. 2019. “Recurrence Is Required to Capture the Representational Dynamics of the Human Visual System.” *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1905544116.
Kriegeskorte, Nikolaus, Marieke Mur, and Peter Bandettini. 2008. “Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience.” *Frontiers in Systems Neuroscience* 2 (November): 1–28.
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” *Advances in Neural Information Processing Systems*, 1–9.
Lamme, V. A., H. Supèr, and H. Spekreijse. 1998. “Feedforward, Horizontal, and Feedback Processing in the Visual Cortex.” *Current Opinion in Neurobiology* 8 (4): 529–35.
LeCun, Yann, D. Touresky, G. Hinton, and T. Sejnowski. 1988. “A Theoretical Framework for Back-Propagation.” In *Proceedings of the 1988 Connectionist Models Summer School*, 1:21–28.
Leszczynski, Marcin, Tobias Staudigl, Leila Chaieb, and Simon Jonas Enkirch. 2020. “Saccadic Modulation of Neural Activity in the Human Anterior Thalamus during Visual Active Sensing.” *bioRxiv*, 1–26.
Lillicrap, Timothy P., Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. 2020. “Backpropagation and the Brain.” *Nature Reviews. Neuroscience* 21 (6): 335–46.
Lisman, John E., and Ole Jensen. 2013. “The Theta-Gamma Neural Code.” *Neuron* 77 (6): 1002–16.
Lowet, E., M. J. Roberts, C. A. Bosman, P. Fries, and P. de Weerd. 2016. “Areas V1 and V2 Show Microsaccade-Related 3-4-Hz Covariation in Gamma Power and Frequency.” *The European Journal of Neuroscience* 43 (10): 1286–96.
Martens, Sander, Jaap Munneke, Hendrikus Smid, and Addie Johnson. 2006. “Quick Minds Don’t Blink: Electrophysiological Correlates of Individual Differences in Attentional Selection.” *Journal of Cognitive Neuroscience* 18 (9): 1423–38.
Mnih, Volodymyr, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. “Recurrent Models of Visual Attention.” *Advances in Neural Information Processing Systems* 3 (January): 2204–12.
Nairne, James S., Joshua E. VanArsdall, Josefa N. S. Pandeirada, Mindi Cogdill, and James M. LeBreton. 2013. “Adaptive Memory: The Mnemonic Value of Animacy.” *Psychological Science* 24 (10): 2099–2105.
New, J., L. Cosmides, and J. Tooby. 2007. “Category-Specific Attention for Animals Reflects Ancestral Priorities, Not Expertise.” *Proceedings of the National Academy of Sciences* 104 (42): 16598–603.
Nieuwenstein, Mark R., Marvin M. Chun, Rob H. J. Van Der Lubbe, and Ignace T. C. Hooge. 2005. “Delayed Attentional Engagement in the Attentional Blink.” *Journal of Experimental Psychology. Human Perception and Performance* 31 (6): 1463–75.
Nieuwenstein, Mark R., Mary C. Potter, and Jan Theeuwes. 2009. “Unmasking the Attentional Blink.” *Journal of Experimental Psychology. Human Perception and Performance* 35 (1): 159–69.
Olivers, Christian N. L., and Martijn Meeter. 2008. “A Boost and Bounce Theory of Temporal Attention.” *Psychological Review* 115 (4): 836–63.
Olivers, Christian N. L., Stefan Van Der Stigchel, and Johan Hulleman. 2007. “Spreading the Sparing: Against a Limited-Capacity Account of the Attentional Blink.” *Psychological Research* 71 (2): 126–39.
Palmquist, Peter E. 1980. “Spoonerism.” *History of Photography* 4 (1): 38–38.
Park, Jooyong, and Nancy Kanwisher. 1994. “Determinants of Repetition Blindness.” *Journal of*

Experimental Psychology: Human Perception and Performance. https://doi.org/10.1037/0096-1523.20.3.500.
Ptak, Radek, and Armin Schnider. 2005. “Visual Extinction of Similar and Dissimilar Stimuli: Evidence for Level-Dependent Attentional Competition.” *Cognitive Neuropsychology* 22 (1): 111–27.
Purpura, Keith P., Steven F. Kalik, and Nicholas D. Schiff. 2003. “Analysis of Perisaccadic Field Potentials in the Occipitotemporal Pathway during Active Vision.” *Journal of Neurophysiology* 90 (5): 3455–78.
Quiroga, Maria Del Mar, Adam P. Morris, and Bart Krekelberg. 2019. “Short-Term Attractive Tilt Aftereffects Predicted by a Recurrent Network Model of Primary Visual Cortex.” *Frontiers in Systems Neuroscience* 13 (November): 67.
Rafal, Robert, Shai Danziger, Giordana Grossi, Liana Machado, and Robert Ward. 2002. “Visual Detection Is Gated by Attending for Action: Evidence from Hemispatial Neglect Robert.” *Proceedings of the National Academy of Sciences* 99 (25): 16371–75.
Raymond, J. D., K. L. Shapiro, and K. M. Arnell. 1992. “Temporary Suppression of Visual Processing in a RSVP Task: An Attentional Blink?” *Journal of Experimental Psychology* 18 (3): 849–60.
Ritchie, J. Brendan, David A. Tovar, and Thomas A. Carlson. 2015. “Emerging Object Representations in the Visual System Predict Reaction Times for Categorization.” *PLoS Computational Biology* 11 (6): 1–18.
Rumelhart, David E., Geoffrey E. Hinton, Ronald J. Williams, and CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE. 1985. “Learning Internal Representations by Error Propagation.” CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE. https://apps.dtic.mil/sti/citations/ADA164453.
Schlaffke, Lara, Anne Golisch, Lauren M. Haag, Melanie Lenz, Stefanie Heba, Silke Lissek, Tobias Schmidt-Wilcke, Ulf T. Eysel, and Martin Tegenthoff. 2015. “The Brain’s Dress Code: How The Dress Allows to Decode the Neuronal Pathway of an Optical Illusion.” *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior* 73 (December): 271–75.
Scholte, H. Steven. 2018. “Fantastic DNimals and Where to Find Them.” *NeuroImage* 180 (Pt A): 112–13.
Schwartz, Odelia, Anne Hsu, and Peter Dayan. 2007. “Space and Time in Visual Context.” *Nature Reviews. Neuroscience* 8 (7): 522–35.
Shapiro, Kimron, Anne P. Hillstrom, and Masud Husain. 2002. “Control of Visuotemporal Attention by Inferior Parietal and Superior Temporal Cortex.” *Current Biology: CB* 12 (15): 1320–25.
Sumner, Rachael L., Meg J. Spriggs, Suresh D. Muthukumaraswamy, and Ian J. Kirk. 2020. “The Role of Hebbian Learning in Human Perception: A Methodological and Theoretical Review of the Human Visual Long-Term Potentiation Paradigm.” *Neuroscience and Biobehavioral Reviews* 115 (August): 220–37.
Sy, Jocelyn L., and Barry Giesbrecht. 2009. “Target-Target Similarity on the Attentional Blink: Task-Relevance Matters!” *Visual Cognition* 17 (3): 1–10.
Torralba, Antonio, and Aude Oliva. 2003. “Statistics of Natural Image Categories.” *Network: Comput. Neural Syst.* 14 (3): 391–412.
Tuli, Shikhar, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. 2021. “Are Convolutional Neural Networks or Transformers More like Human Vision?” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2105.07197.
Umbach, Gray, Pranish Kantak, Joshua Jacobs, Michael Kahana, Brad E. Pfeiffer, Michael Sperling, and Bradley Lega. 2020. “Time Cells in the Human Hippocampus and Entorhinal Cortex Support Episodic Memory.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (45): 28463–74.
Whitmire, Clarissa J., and Garrett B. Stanley. 2016. “Rapid Sensory Adaptation Redux: A Circuit Perspective.” *Neuron* 92 (2): 298–315.
Wynn, Jonathan K., Bruno Breitmeyer, Keith H. Nuechterlein, and Michael F. Green. 2006. “Exploring the Short Term Visual Store in Schizophrenia Using the Attentional Blink.” *Journal of Psychiatric Research* 40 (7): 599–605.
Yamins, D. L. K., H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. 2014. “Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual

Cortex." *Proceedings of the National Academy of Sciences* 111 (23): 8619–24.
Zhang, Xiaodan, Xinbo Gao, Wen Lu, and Lihuo He. 2019. "A Gated Peripheral-Foveal Convolutional Neural Network for Unified Image Aesthetic Prediction." *IEEE Transactions on Multimedia* 21 (11): 2815–26.

Nederlandse samenvatting

Als we onze ogen snel over een visuele scène heen laten gaan, verwerken we allerlei informatie zonder enige schijnbare moeite en lijkt de informatie continu over de tijd heen geïntegreerd te worden. Mensen zijn zo goed in het verwerken van visuele scènes dat de betekenis van een scène al geduid kan worden als een plaatje heel kort (13 ms) wordt aangeboden (Broers, Potter en Nieuwenstein, 2018) en binnen een fractie van een seconde kunnen mensen al reageren op specifieke informatie uit een scène (Kirchner en Thorpe 2006). In de afgelopen decennia is vastgesteld dat deze buitengewone prestatie bereikt wordt door de inrichting van ons hiërarchische, visuele systeem. Met name visuele kenmerken die laag in de visuele hiërarchie verwerkt worden, zoals randen, oriëntaties en kleur, worden heel snel verwerkt en deze basale visuele kenmerken worden later in de visuele hiërarchie gecombineerd tot complexere visuele kenmerken die vaak semantische eigenschappen hebben (DiCarlo, Yoccolan en Rust 2012).

In een reeks onderzoeken heb ik verschillende varianten van het zogeheten “Rapid Serial Visual Presentation” (RSVP) paradigma gebruikt om te achterhalen waarom specifieke semantische informatie makkelijker in het werkgeheugen terecht komt en hoe scènes die kort na elkaar gepresenteerd worden elkaar beïnvloeden. Twee veelvoorkomende bevindingen bij multi-target RSVP's zijn de zogeheten “Attentional Blink” (AB: Raymond, Shapiro en Arnell 1992) en “Repetition Blindness” (RB: Kanwisher 1987). In het AB paradigma worden twee target-plaatjes (T1 en T2) kort na elkaar gepresenteerd binnen een serie van afleider-plaatjes. Als het tweede target-plaatje (T2) 200-500 ms na het eerste target-plaatje (T1) wordt getoond, kunnen proefpersonen de T2 vaak veel minder goed detecteren/rapporteren. In het RB-paradigma (Kanwisher 1987) presteren proefpersonen veel minder goed als zowel T1 als T2 relevant zijn voor de taakprestatie (Sy en Giesbrecht 2009). Een groot probleem in het al bestaande AB- en RB-onderzoek is dat de gebruikte

stimuli vaak erg simpel en kunstmatig zijn. Daardoor is het moeilijk te begrijpen hoe interacties tussen de visuele targets leiden tot veranderingen in taakprestatie. Recente ontwikkelingen in analytische methoden maken het mogelijk om verschillen tussen complexe, naturalistische stimuli mathematisch te beschrijven. Hierdoor is het mogelijk om interacties tussen targets op verschillende niveaus in de visuele hiërarchie te kwantificeren en daarmee beter te begrijpen waardoor interacties tussen T1 en T2 tot veranderingen in taakprestaties in de AB- en RB-paradigma's leiden.

In mijn onderzoek heb ik verschillende neuroimaging-methoden (EEG, fMRI) en computationele modellen (drift diffusion modelling, DDM; convolutionele neurale netwerken, CNN) gecombineerd om beter te begrijpen, waardoor de "Attentional Blink" en "Repetition Blindness" worden veroorzaakt. Het lijkt erop dat visuele target-plaatjes elkaar op meerdere verschillende visuele verwerkingsniveaus kunnen beïnvloeden en dat afhankelijk van waar deze interactie plaatsvindt (laag in de visuele hiërarchie vs. hoog in de visuele hiërarchie) de veranderingen in taakprestaties tegengesteld kunnen zijn. Met behulp van een rijke dataset aan natuurlijke scènes laat ik zien dat als target-plaatjes op elkaar lijken hoog in de visuele hiërarchie (waar taal en semantiek tot stand komen), dat er dan een verslechtering in taakprestaties plaatsvindt. Dit impliceert dat als de betekenis van de targets te veel overeenkomt, er slechts 1 scène kan worden gerepresenteerd het werkgeheugen (Kanwisher 1987; Wyble et al. 2011). Als target-plaatjes juist op elkaar lijken laag in de visuele hiërarchie (waar basale beeldeigenschappen worden verwerkt), dan worden taakprestaties beter, wat doet denken aan de welbekende priming-effecten. Ook lijken target-plaatjes sneller verwerkt te worden als de beeldeigenschappen overeenkomen laag in de visuele hiërarchie, wat extra evidentie is dat er dan priming plaatsvindt.

Samenvattend lijkt het erop dat deze nieuwe onderzoeksbenadering, waarbij natuurlijke scènes worden gebruikt als stimuli in staat is om vrijwel alle, soms tegenstrijdige bevindingen, in de AB- en RB-literatuur te verklaren.

Referenties

Broers, N., Potter, M. C., & Nieuwenstein, M. R. (2018). Enhanced recognition of memorable pictures in ultra-fast RSVP. *Psychonomic Bulletin & Review*, *25*(3), 1080–1086.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762–1776.

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

Raymond, J. D., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in a RSVP task: an attentional blink? *Journal of Experimental Psychology*, *18*(3), 849–860.

Kanwisher, N. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, *27*, 117–143.

Sy, J. L., & Giesbrecht, B. (2009). Target-target similarity on the attentional blink: Task-relevance matters! *Visual Cognition*, *17*(3), 1–10.

Wyble, B., Bowman, H., & Nieuwenstein, M. (2009). The Attentional Blink Provides Episodic Distinctiveness: Sparing at a Cost. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(3), 787–807.

Acknowledgements

The results and discussions presented in this book represent only the tip of the iceberg of all the projects I have undertaken with my colleagues during my PhD. Countless projects that were started but never quite completed, many projects that are invisible because they were the seed for this book, and many methodological innovations that are now available to the field of neuroscience on GitHub. None of this would have been possible without the people around me. I have been fortunate to be surrounded by world-class researchers throughout my career, and I could not have asked for more kind-hearted, talented, and amazing mentors: Ian, Ilja, and Kimron. Ian was my constant inspiration throughout the process and I would not have survived without him. I have never met anyone so genuinely curious and fascinated by the mysteries of science, and who constantly reminds everyone around him why science is such an incredible privilege to be a part of. Aside from our many travel-related escapades (do not ever lend me your only key again), what I will remember most are our long evenings inside and outside the office developing new, exciting ideas and solving problems in ways no one had thought of before. Ilja was my rock, he was always there when I needed him and supported me in every decision (good or bad) I made. I do not know how I can ever express how grateful I am for all of your expertise, your personal investment, and ability to always help me see the big picture. Kimron, even though I was probably the most difficult graduate student the University of Birmingham has ever dealt with, you were always there for me and helped me. I have lost count of the number of times I have come to you with a problem and you have given me a new perspective. Your expertise and experience in the field have been invaluable to my own growth and understanding of science.

It is easy to read the above and think that is enough. But I also have friends. Michael Rojek Giffin, my constant companion and dearest friend. He's always there to bring me down to earth when needed and

lift me to the heavens when needed. A boy who is as smart as he is funny - a lethal combination that makes him one of the most fascinating people I have ever met. Mary Jo, a person with a spontaneity that rivals my own, and I can count on her support for literally anything, except for my opinion. The first time I met her, she spent the entire evening making fun of my stupid-sounding accent. I'm still working on a comeback. Aaron, my big brown bear! You get to land here. Third place. Not bad for a boy from Brighton. Leon, your ability to make outrageous jokes is impressive. I never know where you get your momentum from, but that's another reason I love you. Sasha, you are by far the least respectful of personal boundaries (in the most loving way possible) and always ask probing questions out of your inexhaustible curiosity. You have my permission to say that this book is about you. Eliska, you have intentionally and accidentally made me laugh more than anyone else. You are my treasure. Luisa, you are an endless source of entertainment and compassion (40 coming up). Josi, my favorite support animal! If it becomes possible to clone humans, I will make a dozen of you. Elio, you clever little button! My favorite human to share far-fetched nerd ideas with! Miriam, you have been by my side since the beginning and I do not know where I'd be without you. May you always lose in the thumb wars though. Lukasz, my sporting nemesis. Don't you dare leave my side. Seb, my nemesis in all other ways, but in a loving way. Auke, the best millennial-man character ever. Anne, there is not a room in the world that would not be brighter with your presence. Logan, you are the person I call when the phone does not work. You are loud, Logan. That's the joke. Josipa, my favorite person to laugh at how crazy we are. Timo, my beautiful German companion! Although I think you are one of the best researchers at UvA, I still love you. Elsa, you beautiful little weirdo. I have never met anyone like you and that is the biggest compliment I can give someone. Lukas, my favorite colleague and pythonista! Whenever I have an unsolvable problem, I come to your office and halfway through my sentence you have already stood up and started sketching the solution on the whiteboard. You are one of the most brilliant, humble and kind people I have ever met. Lynn, constantly curious and full of smart insights. Noor, knows all the lyrics to every song ever written and will always be my first choice for any pub quiz.

Jessica, quirky in the best possible way. Also, I must extend my heartfelt thanks to Steven Scholte, who took me into his lab when I returned from Birmingham.

I also thank my family! My supportive and caring parents, Lennart and Heidie, who are always there for me and always make sure I have something to eat when I am home. My siblings - Mikael, Jennie and Olivier. My niece and nephews: Rebecka, Elliot, Alicia, Marvin, Alma, and Julian! Thank you all for making me always look forward to coming home for Christmas and summer.

Jag har också så klart mina polare i Sverige. Speciellt där har jag Pär, Erik, Pontus och Lars. Trots att jag flyttade ifrån Sverige 2013 så har jag lyckats klänga mig fast vid er än idag, vilket är finare än vad jag kan beskriva här.

Per Johan Daniel Lindh, 2022