THE TEMPORAL STRUCTURE OF ATTENTION IN MULTI-PART TASKS

Lydia Joy Barnes

Clare Hall
MRC Cognition and Brain Sciences Unit
School of Clinical Medicine
University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy
September 2021

PREFACE

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification. It does not exceed the prescribed word limit for the Faculties of Clinical Medicine and Clinical Veterinary Medicine Degree Committee at the University of Cambridge.

The study in Chapter 2 is under review for publication. This project was done in collaboration with Erin Goddard, who shared initial scripts for a behavioural task and gave advice on design, analysis, and writing.

The study in Chapter 3 was done in collaboration with Erin Goddard and Tim Kietzmann, who gave advice on the analysis. Tim Kietzmann also shared initial scripts for the analysis.

The study in Chapter 4 was done in collaboration with Jason Mattingley and Dragan Rangelov, who shared initial scripts for a behavioural task, and gave advice on design and analysis.

SUMMARY

Lydia Joy Barnes

The Temporal Structure of Attention in Multi-Part Tasks

Human behaviour is extraordinarily flexible. Task fMRI and patient studies highlight a network of frontoparietal brain regions, called "multiple-demand" regions, that serve as a hub for flexible cognition. Neurons within this network adapt to code what is relevant for the current task, across many task types. Under the attentional episodes view, prioritising immediately relevant information in this way allows us to break complex tasks into simple parts and drive neural resources toward the problem at hand. Many studies demonstrate that relevant information is preferentially encoded, such that we can read out features more accurately when they are task-relevant. Yet we do not know whether the preferential coding that we see on slow timescales and in simple tasks supports moments of narrow focus, or "temporal modules", in more complex, multi-part tasks. This is central to the attentional episodes account: that selection of immediately relevant information, through preferential coding, *dynamically shifts* to give us the information that we need for each part of a task.

Chapter 2 begins by asking how preferential coding emerges in a multi-part task. Using MEG data from a dual-epoch task with single object (Experiment 1) and dual-object (Experiment 2) displays, I show that what is relevant can be preferentially encoded in sequential task epochs with similar rapidity. Preferential coding in either epoch was only detected with dual-object displays, mirroring dominant theories of attention as a spatial spotlight, or a filter to reduce complexity.

Chapter 3 builds on this by tracing how ventral visual and multiple-demand regions contribute to, and communicate, coding of relevant stimulus information throughout a task. I resolve MEG data from a multi-epoch visual attention task

(Experiment 2) to source space, to pull apart how the stimulus coding in Chapter 2 arises in visual and domain-general regions. Using Granger causality, I probe the timecourse of top-down and bottom-up information flow as the relevant feature shifts. I show feedback from prefrontal to visual regions emerging in both task epochs, again highlighting how flexibly we are able to direct focus to multiple parts of a task in turn.

Chapter 4 extends Chapters 2 and 3 to a situation like those we face often in daily life, where what is relevant for each part of a task can be present throughout. These distractors with some task-relevance are also common in classical tests of fluid ability, and could be preferentially attended if selection is not strictly directed to the immediate task part. I use a behavioural task with two sequential displays, each containing a relevant- and an irrelevant-coloured moving dot cloud. Despite being cued to attend to two colours in sequence, participants were not more distracted by the second target colour when it appeared as a distractor in the first task epoch. That is, we can effectively direct our focus to what is immediately relevant, even when presented with a future-relevant feature.

Attending to what is currently relevant as we move through parts of a task is a central aspect of flexible behaviour. These studies probe the limits of this temporal modularity in attention. They show that we can preferentially encode distinct stimulus features as what is relevant changes; and that we can overtly respond to what is relevant in each task part, even when a feature relevant for one task part is visible throughout. Together, they emphasise our extraordinary capacity to direct our focus toward what is relevant.

ACKNOWLEDGEMENTS

To Alex, thank you for your brilliance, your commitment, and your inexhaustible enthusiasm. You give so much to your students and there's always more I can learn from you.

To my collaborators, Erin, Tim, Jason, and Dragan, thank you for getting involved, for the impromptu stats lessons, and for being genuinely curious about what we can discover.

To the people who offered advice, encouragement, code, and coffee breaks – Nic, John, Kanad, Tijl, Lina, Amanda, Ilona, the CBU grad students – thank you for making that generosity seem so normal.

To my Cambridge family – the Woolgar lab, the "Tai Chi" crew, my wonderful walking and picnicking buddies of 2020 – thank you for making these last few years special.

To Robert, thank you for being the best friend I could ask for.

I am grateful for the MRC CBU for their financial support.

TABLE OF CONTENTS

Chapter 1

Fluid Ability in Daily Life

Human behaviour is extraordinarily flexible. We can plan and prioritise, navigate a route we have never followed, or take elements from recipes we have cooked to invent something new. At each moment, we select what is relevant in our memories and environment, so that our actions are more than re-actions. Our everyday actions rest on this ability to draw on what we know and what is around us to coherently move towards a goal, though we may not give it much thought.

Flexible cognition is central to our lives, from routine activities through to mentally challenging tasks. Scores on fluid intelligence tests, which use challenging abstract tasks to probe the limits of mental flexibility, powerfully predict educational achievement, and even health (Primi et al., 2010; Wray et al., 2020; Wrulich et al., 2014). When this skill breaks down, everyday actions become disorganised. Damage to the brain systems that support planning and learning can turn simple tasks into an unmanageable challenge. Acquired brain injury to the frontal lobes can produce disordered and risky behaviours (Chevignard et al., 2008). People whose learning and reward systems are disrupted in negative-symptom schizophrenia can struggle to piece together simple daily tasks (Josman et al., 2009). Many of us will not experience such a dramatic drop in our flexibility, but we may know what it is like to be overwhelmed by a mental challenge when we are tired and feel our focus slipping away.

Recently, artificial intelligence research has highlighted the practical value of understanding flexible cognition. Computational power has rocketed forward, along with our ability to build multi-layer neural networks that learn to distinguish subtle features of the environment. Yet human-like flexibility – our capacity to learn about and implement many different tasks – is still out of reach. Teaching

multiple tasks to a neural network can lead to some task information being lost (Flesch et al., 2018), or simply take an unmanageable amount of time. Understanding how the human brain extracts and manipulates task information could give us useful insight into how a system with limited energy can efficiently adapt to a range of challenges.

Studying flexible brain processes, then, has implications for how we educate, how we design, and how we develop artificial systems, as well as giving us a window into how our own minds work.

1.1. Flexibility across the brain

We can begin by understanding how the brain adapts to our task and context. Over the past decades, neuroscience has gained a lot by localising function to specialised brain regions, marking out what information each region responds to consistently over diverse settings. For instance, we can see that neurons in area V4 preferentially respond to colour, and that a portion of the fusiform gyrus preferentially responds to faces. In parallel, we have seen that this specialisation co-exists with adaptation: receptive fields, functional connections, and dynamic population coding combine to push forward the information that is most important for our goal.

In the following sections, I will review our current understanding of brain mechanisms for goal-oriented behaviour, and discuss the unique role of the frontoparietal cortex in directing neural resources towards the current task. From this, I will consider a theory of how coding task-relevant information in highly-adaptive frontoparietal brain regions could support flexible cognition. I will highlight the implications of this theory for multi-step, complex behaviour. Lastly, I will give an overview of three empirical studies that test how we prioritise relevant information in brain and behaviour as we move through parts of a task.

1.1.1. Adaptive sensory coding, receptive fields, and local connections

Even within sensory cortex, multiple features unite to prioritise rewarding or task-relevant information. Rapid processing for information at the fovea (Schira et al., 2009) allows gaze to drive selection within a scene. As we fixate repeatedly on the same stimuli, correlations between neurons in early visual cortex become smaller and more stable (Gutnisky & Dragoi, 2008). This decreased redundancy across the local network could allow us to efficiently encode the stimuli to which we dedicate more time. Our deeply rooted preference to perceive whole objects (see for example Yeari & Goldsmith, 2010) prioritises all features of an item that has captured our attention (Baldauf & Desimone, 2014; Z. Chen, 2012a), meaning that we can search for an object by one feature (for example, its colour) and quickly obtain information about shape and size that could be important when we interact with it. Recordings from mouse somatosensory cortex show that simple inputs, which could be coded by neurons tuned to stimulus feature A or B, are coded across neurons with widely varying preferences (Nogueira et al., 2021), in a high-dimensional format associated with complex behaviours (Fusi et al., 2016; Rigotti et al., 2013). Thus, flexible representations could be ingrained in early sensory cortical processing even when they are not strictly required.

Moving further out, widening receptive fields along the ventral visual stream mean that stimuli compete to feed forward, creating an opportunity for attentional bias, such as top-down communication of the task goal, to select the task-relevant item (Duncan, 2006; Humphreys et al., 1998; Reddy et al., 2009; J. H. Reynolds et al., 1999; Scalf et al., 2013). Connectivity in the visual ventral stream also shows task-adaptive modulations, as primary visual and ventral temporal regions become more connected as we process task-relevant information, particularly under high task demand (Hwang et al., 2018). Together, these mechanisms for adaptation allow

relevant information to take priority from the moment cortical neurons begin to transform the input from our sense.

1.1.2. Rapid reward learning

Beyond sensory brain regions, a critical source of flexibility arises from the way we code reward. With each action we make, dopamine neurons in the midbrain code the reward that we expect based on previous experiences, and the true reward we experience. At first glance, this seems to primarily support flexibility in the long term, by biasing us towards actions that have benefited us. Rewards reinforce specific actions, and anticipating those rewards drives our future choices. But reward coding is remarkably sensitive to our immediate task context, in important ways. Single-unit recordings from non-human primates show that midbrain dopamine neurons adjust their gain according to variance in reward (Tobler et al., 2005). If rewards are very similar, dopamine neurons tune in to be more sensitive. This kind of context-sensitive coding provides a way to scale reward learning from simple settings through to complex situations with subtly different choices. Further afield in the reward learning system, neural populations in the human rostral cingulate cortex track reward probability over time, responding more to reversal learning when learning trials have been informative, so that we adapt our learning rate to the information content of the environment (Jocham et al., 2009).

Reward also plays a role in complex problem solving. Fluid intelligence problems – an extreme measure of flexibility – require that we direct our attention to related features of a problem, then voluntarily disengage and shift focus to the next part. Computational modelling of reward learning in the basal ganglia predicts disengagement, as does decreased basal ganglia BOLD response (Stocco et al., 2021). This ramping BOLD activity makes sense if basal ganglia activity increases or drops as our estimates of reward are exceeded (positive reward prediction error) or not reached (negative reward prediction error). Reward learning could adapt our

actions to the task context, and support flexible cognition, by enabling us to track when our approach to a problem needs to change.

1.1.3. Connectivity hubs

Thirdly, we prioritise task and context through highly-connected, domain-general brain "hubs". The term "hub" comes from graph theory, which defines a set of tools for understanding complex networks by describing their nodes and connections. In graph theory terms, hubs are nodes (regions) where many connections converge (Sporns et al., 2004). In the brain, hubs typically communicate long-range, contrasting with the short-range connections that characterise sensory and motor brain regions (Bullmore & Sporns, 2009). Although the hub terminology comes from graph theory, many neuroscientific approaches identify a similar set of multi-purpose regions that connect widely across the brain. We can see the same core networks emerge when we trace functional connectivity (Yeo et al., 2011), actively contrast tasks (Fedorenko et al., 2013), or combine anatomical and functional boundaries (Crossley et al., 2014).

Domain-general hubs appear to play a special role in adapting our behaviour to our context. While long-range connections can be metabolically costly, they allow our brains to communicate and integrate across specialised functions (Crossley et al., 2014), orienting these functions toward a shared goal. Thus, dedicated hubs could implement a level of coordinated flexibility that would be difficult to get from local mechanisms. For example, a centralised hub for selection can help artificial neural networks to learn. Embedding a central "attention" module into an artificial neural network can make training more efficient (T. Chen et al., 2021) and allow the network to self-distil relevant and contextual information, such as emotion in speech (Zhang et al., 2021). Connections between hubs and sensory and motor regions may explain how diverse executive functions (like visual or phonological working memory, attention, and inhibition) historically have been linked to overlapping domain-general brain regions (Zink et al., 2021). Empirical work shows

that clearly formed and segregated domain-general hubs characterise healthy development. Children who struggle at school show less distinct hubs (Jones et al., n.d.; Siugzdaite et al., 2020), highlighting the importance of centralised brain networks for everyday functions. These hubs could be key for directing our actions towards what is most rewarding or relevant in our rich sensory environment and mental life.

1.2. Adaptive brain networks for fluid ability

1.2.1. Pre-frontal cortex for task representations

The brain region most consistently associated with flexible cognition is the pre-frontal cortex, or PFC. The PFC is active across a wide range of tasks, and shows both high-dimensional neural coding and hub-like connectivity (Cole et al., 2012; Rigotti et al., 2013). While receptive fields and selectivity vary across the PFC (Riley et al., 2017), the region as a whole appears to be important for tasks that are extended over time (Wilson et al., 2010). Thus, the PFC could support flexible task performance by maintaining our goal across many small actions.

The PFC's function may be best understood through two networks that claim its medial and lateral parts: the default mode network, and the multiple-demand network. These networks are prime examples of domain-general hubs, connecting widely and long-range across the brain. The two networks appear to work in complement, for example, becoming decorrelated as healthy development progresses (DeSerisy et al., 2021). Importantly for our search for brain bases of adaptive behaviour, both appear to be sensitive to information that is currently relevant, from relevant concept knowledge to task identity and task phase (Wang et al., 2021; Wen et al., 2020). In the next sections, we will delve into what these two networks may contribute to flexible cognition.

1.2.2. The default-mode network

The default mode network (DMN) includes the medial PFC, as well as the posterior cingulate cortex and angular gyrus. It is best known for its increased activity during wakeful rest relative to active task (Raichle et al., 2001). However, links between the DMN and current task features suggest that it is also actively involved in externally-directed tasks. For example, activity in the DMN, measured by fMRI, tracks task completion (Farooqui & Manly, 2019) and increases during context-based decisions (V. Smith et al., 2021). The DMN appears to be particularly sensitive to the overall task identity (compared to immediate task phase in the frontoparietal network, Wang et al., 2021) and to decisions about naturalistic stimuli (compared to symbolic stimuli in the frontoparietal network, Smith et al., 2021). Within the DMN, neural codes represent task-relevant information when the task context is naturalistic (V. Smith et al., 2021). The DMN could support flexible cognition by tracking progress toward a goal and prioritising information that is relevant for each step.

1.2.3. The ultra-task-sensitive "multiple-demand" network

In parallel with the DMN, the multiple demand network (MDN) adapts to code key task features. The network can be identified by tracing the brain regions in which BOLD signal increases during high demand across a wide range of tasks – hence the name "multiple-demand" (Assem et al., 2020; Fedorenko et al., 2013; Jung et al., 2021, p. 20). Broadly, the network spans precentral, inferior frontal, and medial frontal gyri; anterior insula/frontal operculum, supplementary motor and pre-supplementary motor areas, and anterior cingulate cortex; and the intraparietal sulcus (Fedorenko et al., 2013). However, detailed single-subject anatomical and functional brain parcellation, paired with harsher thresholding at the group level, reveals a core subset of MD regions that are highly connected and respond strongly to task demand (Assem et al., 2020). These regions (Glasser areas i6-8, p9-46v, a9-

46v, 8BM-SCEF, AVI, IP1, IP2, IFJp, 8C, and PFm, 10 in total; Glasser et al., 2016) map closely to a network otherwise known as the lateral frontoparietal network, frontoparietal control network, task-positive network, central executive network, or executive control network, which has been identified through both task activation and resting state functional connectivity (Fox et al., 2005; Hwang et al., 2018; Ji et al., 2019; Power et al., 2011; S. M. Smith et al., 2009; Uddin et al., 2019). Peripheral or "penumbra" MD regions (Glasser areas p10p, a10p, 11l, a47r, p47r, FOP5, 6r, s6-8, AIP, LIPd, MIP, PGs, TE1m, TE1p, d32, a32pr, and POS2, 17 in total) fall within a network known as cingulo-opercular, ventral frontoparietal, or salience; the dorsal frontoparietal or dorsal attention network; and the default mode network (Assem et al., 2020). The blurred boundary between the core and penumbra MDN could reflect, as other studies have demonstrated, that networks overlapping the lateral PFC flexibly adjust to operate together or independently depending on the task (Camilleri et al., 2018; Fox et al., 2006).

Though its borders vary slightly with method and task, the MDN does appear to broadly operate as a whole. Evidence comes from studies of patients with an acquired brain injury to the MDN, whose unaffected regions of the network appear to compensate for damage by becoming more active (Woolgar et al., 2013). Neurons within frontal and parietal cortex in non-human primates have been shown to code the same task-relevant features (Hall et al., 2020; Quintana & Fuster, 1992), with inactivation of prefrontal cortex disrupting parietal neuronal activity and vice versa suggesting that similar responses in each region arises from co-dependence between them (Chafee & Goldman-Rakic, 2000). In parallel, disrupting activity in the dorsolateral PFC within the human MDN using transcranial magnetic stimulation (TMS) disrupts preferential coding of task-relevant information across the network (Jackson et al., 2021).

So, what does this core task-positive network do? Like its neighbouring dorsal and ventral attention networks, the MDN (or frontoparietal control network) encodes stimulus information and enhances coding of the task-relevant stimulus

dimension (Long & Kuhl, 2018). However, where two other domain-general networks, the dorsal attention and default mode networks, are minimally correlated, the MDN contains connections to both (Spreng et al., 2012). Functional MRI studies with human participants also suggest that the MDN in particular engages more when a task is difficult, compared to when it is simple. Thus, the MDN appears to enhance task-relevant stimulus features, with potentially a special role in directing neural resources towards relevant information under high cognitive demand.

Two features of the MDN appear to be uniquely task-sensitive. First, the MDN's task-adaptive nature is clear in its flexible connectivity. Multiple-demand network connectivity changes flexibly with the task, more than we typically see in other brain networks (Cole et al., 2013). Thus, the MDN is uniquely placed to reconfigure connections between disparate brain regions, driving mental resources towards the task at hand.

This uniquely flexible connectivity is paired with remarkable flexibility in the information that the MDN encodes. Within the MDN, different task-relevant features are coded in close proximity. Distinct object features are coded by overlapping voxels within the MDN, demonstrating the network's versatility (Jackson & Woolgar, 2018). Information maintained in working memory, and planned motor responses, are both held in overlapping but independent traces within non-human primate lateral PFC (Tang et al., 2020), highlighting the multifunctionality of this region. In human fMRI, the pattern of activity across the MDN codes task rules, responses, and stimulus features (Woolgar, Thompson, et al., 2011). Features in working memory can be read out from these regions with a linear classifier (M. G. Stokes et al., 2013). Stimulus features coded in the MDN include colour, form, length, and orientation, as well as auditory tones and vibro-tactile information (Assem et al., 2021; Jackson et al., 2016; Woolgar, Afshar, et al., 2015; Woolgar et al., 2016; Woolgar, Hampshire, et al., 2011; Woolgar, Thompson, et al., 2011; Woolgar, Williams, et al., 2015; Woolgar & Zopf, 2017). Brain regions that

respond to auditory or visual tasks are interdigitated alongside cortical MD regions (Assem et al., 2021). Together, these findings show that diverse stimulus features are coded in task-oriented ways within the MDN.

Time-resolved data suggest that the MDN may be flexible over shorter timescales than we are able to observe with fMRI. Neurons in the non-human primate homologue of lateral PFC initially code the visual features of a display, then quickly adapt to preferentially code the relevant features (Kadohisa et al., 2013; M. G. Stokes et al., 2013). Even more extraordinary are non-human primate data showing that the same lateral PFC neurons can rapidly adapt to code different task features, as what is relevant changes within a trial (Rao et al., 1997). Similarly, the same information can be held in orthogonal population codes during display and delay phases of a task (Sigala et al., 2008). Whereas functional data from MR imaging in humans shows at best that MDN neural populations reconfigure from trial to trial, these invasive recordings raise the possibility that preferential coding of what is relevant in the lateral PFC could rapidly reconfigure even within a task.

Each of these characteristics emphasises the MDN's task-sensitivity. Long-range connections position the network as a hub to integrate diverse cognitive functions. Task-driven changes in connectivity allow this integration to adjust to the current goal. Coding of multiple stimulus modalities within the network reinforces this role in integration, while preferential coding prioritises information within attentional focus. In the following sections, I will explore how two defining features of the MDN – task-sensitive connectivity and rapid coding of relevant information – could contribute to flexible implementation of many different tasks.

1.3. Information flow for top-down control

Many traditional theories of goal-directed behaviour describe a central executive or supervisory attention system (Baddeley, 1996; D'Esposito et al., 1995; Norman & Shallice, 1986; Shallice et al., 1996). The role of this central component

is to bias sensory and motor processing towards goal-relevant information and action. That is, where sensations compete for higher-level processing or motor commands conflict, a central component that encodes the task goal resolves competition or conflict in favour of the task-relevant item. As we have seen in the previous sections, neuroscience research has brought nuance to this story. Goal-directed behaviour can arise from adaptive neural processes on many levels, from primary sensory cortices to domain-general networks.

However, a brain network dedicated to preferentially coding what is immediately relevant could offer something more than we can achieve through local connectivity or reward learning alone, by providing a global source of bias to radically reorient perception and action towards our goal. Such an attentional bias does not negate the importance of adaptive mechanisms within sensory regions or in reward-sensitive neurons, but could work with them. For example, widening receptive fields along the ventral visual stream mean that similar stimuli compete for identification, categorisation, and memory. This provides an opportunity for goal-directed behaviour if attentional bias selects task-relevant over task-relevant competing features (Duncan, 2006; Humphreys et al., 1998; Reddy et al., 2009; Scalf et al., 2013). Task-sensitive information in the MDN could be communicated, through task-sensitive connectivity, to provide this bias.

Neuroimaging data support the idea that the MDN could drive goal-directed responses in sensory brain regions. For example, patients with dorsolateral PFC lesions show increased mid-latency auditory evoked responses to distracting sounds during auditory working memory delay, suggesting that the dorsolateral PFC is critical for top-down control of early auditory processing (Chao & Knight, 1998). Connectivity between early and late stages of the ventral visual stream increases with task difficulty, in synchrony with increased frontoparietal activity (Hwang et al., 2018), possibly reflecting top-down control of recurrent processing in visual perception. Brain stimulation data further reinforce the proposal that the MDN drives goal-directed responses in visual cortex, showing that preferential coding of

relevant stimulus information in early visual cortex disappears under concurrent transcranial magnetic stimulation to the dorsolateral PFC (Jackson et al., 2021).

Similarly, connections between the MDN and primary visual cortex provide a possible mechanism for maintaining goal-relevant information over time in visual working memory. Researchers have proposed that information held in working memory could be maintained within sensory cortices (Pasternak & Greenlee, 2005; Postle, 2006). One benefit of this is that sensory cortices already have the tools to process stimulus features, and thus could be best placed to represent visual features in working memory, in the sense that they could "re-present" those features. However, maintaining stimulus information in sensory cortices could leave them susceptible to interference from distractors during working memory delay. Functional imaging data show that, while early visual cortex information coding was disrupted by distractors during working memory maintenance, information coding within the intraparietal sulcus (IPS) is stable (Lorenc et al., 2018). Stable coding of relevant information within the MDN, then, could be important to restore information coding in sensory cortex after distraction. Diverse executive functions, such as working memory, could arise from the combination of goal-oriented coding in the MDN and localised stimulus or motor representations.

1.3.1. "Information" flow

The studies above present a compelling case for goal-directed attention and working memory arising from communication between the task-sensitive MDN and sensory cortices – but two crucial features are missing from these sources. Multiple methods converge to show altered MDN activity and connectivity alongside task-specific selectivity in sensory cortices. Yet these findings do not address what information is transmitted between brain regions, nor what part of this information is critical for biasing perception. To do this, we need to trace information flow in time and space.

Information-based connectivity measures (Goddard et al., 2021; Kietzmann et al., 2019) offer new insights into what shared frontal and visual cortex codes reflect top-down communication, or feedback. These data show that the representational geometry of the stimulus space in frontoparietal cortex – how stimuli are represented as similar or dissimilar – can predict later goal-oriented representation in the visual system. Selection emerges in the visual cortex when frontal feedback of stimulus information dominates information flow (Goddard et al., 2021). Feedback information flow is stronger for low coherence relative to high coherence stimuli, suggesting that the type of information conveyed could be biased to support processing of complex inputs (Karimi-Rouzbahani et al., 2021). These findings highlight that task-sensitive information is passed to sensory brain regions when the task requires careful focus.

1.4. Using adaptive codes for complex behaviour

Beyond localising and tracking flexible neural activity, the ultimate goal of cognitive neuroscience is to understand how the brain supports complex real-world function. We know that adaptive neural codes strongly link to real-world function, especially in the task-positive MDN. Scores on novel problem solving tests drop after damage to the frontal lobe (Duncan et al., 1995), with the degree of damage across the MDN after brain injury linearly predicting fluid intelligence scores (Woolgar et al., 2010, 2018). We can calculate fluid intelligence scores from performance on abstract problem-solving, or by performing factor analysis on agglomerated cognitive tasks, but what the scores represent is much more important: our ability to use our knowledge to learn new skills.

But how are we using this adaptability to support complex behaviours? To really understand what flexible neural activity does for behaviour, we need to understand what complex behaviour looks like. Complex behaviours are varied and organised. Varied, in that we can recombine words and actions in a myriad of ways to produce new behaviour sequences. Organised, in that we can construct these

sequences to coherently work towards a goal. This means that our brains need to flexibly assemble plans for focus and action as new situations or demands arise, while clearly maintaining our goal.

1.4.1. Temporal modularity, attentional episodes

We can think about these complex behaviours as a series of steps. On one level, it is simply true that everything we do is a series of steps. You could think of leaving the house as a series of steps, from gathering your belongings, putting on your shoes, walking to the door, leaving, and locking it behind you. But is this step-by-step structure important?

A step-by-step approach could be critical for understanding complex tasks. Fluid intelligence tests often use matrix reasoning problems, in which a grid of images are constructed of elements that repeat systematically across the columns and rows. One image in the grid is left out, and subjects select an image from multiple options that best completes the grid. These problems reliably show individual differences. However, segregating the problems, so that each systematically repeating element appears alone, removes the gap between high and low performers (Duncan et al., 2017). The benefit of segregating complex problems has been replicated in children (aged 7-10), though some children still struggled with the segregated problems (O'Brien et al., 2020). Intuitively, this makes sense: when we segregate the problem, we make it trivial to solve, so that it no longer tests fluid reasoning. What is interesting is that subjects could choose to mentally segregate the problems if they liked. In fact, instructions for matrix reasoning problems often encourage people to look at how elements change across rows and columns. Despite this, high scores on segregated problems and low scores on integrated problems suggest that we struggle to solve each element in turn when we see them all together.

An influential "attentional episodes" theory proposes that this is the MDN's role in flexible cognition: simplifying complex problems by segmenting them and

driving our attention towards each task component in turn (Duncan, 2013).[1] As I highlighted earlier, adaptive coding of task-relevant information within the MDN could provide a global source of bias, directing perception, memory, and action towards our current goal, as required within each episode. Beyond this, the MDN also appears to be sensitive to goal and sub-goal structure within tasks, commensurate with a role in segmenting and sequencing these episodes. MDN BOLD activity ramps up as we complete a task step, with a greater increase in activity as we complete the full task (Farooqui et al., 2012). Moreover, subnetworks within the MDN have been associated with different timescales within a task, with frontoparietal regions rapidly adapting to what is immediately relevant and cingulo-opercular regions maintaining information that is relevant across the task (Dosenbach et al., 2006, 2007). Functional imaging data show that lateral PFC BOLD activation is transient when goals (or rules) change from trial to trial, and sustained when the rule is stable (J. R. Reynolds et al., 2012). Thus, the MDN could drive momentary focus toward each task step while maintaining the overarching goal.

1.5. Research questions

In this thesis, I will consider whether focused attention to simple task parts, implemented by adaptive coding and flexible connectivity, is a defining feature of how we think. First, I will investigate whether we can rapidly reconfigure adaptive coding of relevant information when what is relevant changes mid-trial. Next, I will extend this to communication across the brain, asking what top-down information is

[1] "Attentional episodes" has also been used to describe phenomena like those observed in the attentional blink, in which items that fall in a brief window around a target are bound together without our intention or control. Here, an "episode" refers to a period in which we deliberately direct attention to a series of task-relevant features, and can be long or short depending on our goal.

conveyed, and when, as the focus of attention shifts. Last, I will look at behavioural choices when information about multiple task parts is available, asking whether, in practice, we focus our attention narrowly on the immediate task part or distribute our focus over future task steps. I will argue that adaptive neural resources can enable precise focus on each step of a multi-part task, and that we do approach multi-part tasks in a temporally modular way. I will close with reflections on how this modularity might extend to loosely-defined and self-directed problems that we face outside the lab.

1.5.1. Chapter 2, Rapid reconfiguration for focused task steps

Multiple lines of evidence converge to suggest that adaptive neural populations within the MDN could support momentary focus on what is relevant for each part of a task. Attentional episodes theory proposes that this momentary focus supports complex cognition by allowing us to direct mental resources to simple task steps. Thus, we must reconfigure adaptive codes to smoothly move through parts of a task. Non-human primate studies show that neural populations in lateral PFC can rapidly reconfigure what they encode to prioritise newly relevant features. However, neuroimaging data in humans primarily shows reconfiguration over longer timescales, by tracking what is coded across trials and tasks. In Chapter 2, I will extend non-human primate work to humans, using a multistep task in two magnetoencephalography (MEG) experiments to ask whether we can quickly reconfigure focus within a task to support attentional episodes. I will show that relevant visual features can be preferentially coded even as what is relevant shifts mid-trial. However, I will also show that this rapid preferential coding does not occur in all circumstances, adding an important nuance to the theory: it may be more adaptive to avoid focusing on individual task features where it is computationally possible to do so.

Other lines of evidence suggest that we may not always benefit from approaching tasks as a series of distinct steps. For one thing, we cannot always

identify task parts. Knowing that a matrix reasoning problem can be solved as a series of parts does not get us past the hurdle of finding and focusing on each of those parts. For another, a modular approach could incur some costs. Many behavioural studies show that responses are slow and error prone after a task switch (Meiran et al., 2000; Monsell, 2003; Rogers & Monsell, 1995). Neural network modelling suggests that the depth of focus on one task predicts the degree of switch cost when task change (Musslick et al., 2018). If these same costs apply to attention shifts between steps within a task, closely focusing on each step in turn could be suboptimal. Deep focus on the current task step could also create an opportunity cost, as we overlook the broader task goal and other choices for behaviour. Based on the findings in this chapter, and on the wider literature, I will suggest that understanding behaviour in multi-step tasks is key if we wish to accurately characterise flexible cognition.

1.5.2. Chapter 3, Temporal modularity and connectivity

In Chapter 3, I will investigate the timecourse of information flow to ask whether long-range feedback triggers rapid reconfiguration of stimulus preferences. Rapid changes in adaptive coding suggests rapid implementation in specialised brain regions. Previous work demonstrates goal-oriented coding in sensory cortices. For example, in fMRI of human visual cortex, visual features are coded in independent patterns according to whether they are prospective or current targets (van Loon et al., 2018). Many theories propose or assume that this goal-oriented coding in the visual system arises in part from communication with frontoparietal cortex (for example, D'Esposito, 2007; Duncan, 2010; Humphreys et al., 1998; Lorenc et al., 2018; Scalf et al., 2013). Information-based connectivity offers new insight into this relationship by tracing when the representational structure in one region predicts (or "flows" to) another. Using magnetoencephalography (MEG), this approach has demonstrated that preferential coding of relevant features in visual cortex emerges when stimulus representations are fed back from frontal brain

regions (Goddard et al., 2021). In this chapter, I will build on the findings in Chapter 2 to examine the extent to which preferential coding of relevant stimulus information reflects feedback information flow.

1.5.3. Chapter 4, Characterising everyday behaviour: Is temporal modularity normal?

Lastly, I will move from neural data to behaviour. Asking how we reconfigure adaptive codes, or flexibly redirect connectivity, gives us detailed information about what our brains *can* do, but does not necessarily capture what we choose to do. The results of Chapter 2 demonstrate that neural coding of task-relevant information can be highly dynamic, but that we do not always choose to selectively encode what is relevant. Together with task-switching research, this raises the possibility that focusing deeply on what is immediately relevant could have negative consequences for subsequent focus. We can be blind to this possibility when we observe behaviour in single-step, focused attention tasks. In Chapter 4, I will present results from four behavioural experiments using a multi-step behavioural paradigm to ask whether we choose to direct our focus towards the current task step when other task information is available to us.

The studies presented below demonstrate how dynamically we are able to direct our focus over time. They show that we can preferentially encode distinct stimulus features, even as what is relevant quickly changes; and that we can overtly respond to what is immediately relevant for each task part, even in the presence of future-relevant information. Together, they raise important questions about the role of modular processing in the tasks we perform every day.

Chapter 2

Rapid Reconfiguration for Focused Task Steps

Every day, we respond to the dynamic world around us by choosing actions to meet our goals. Flexible neural populations are thought to support this process by adapting to prioritise task-relevant information, driving coding in specialised brain regions toward stimuli and actions that are currently most important. Accordingly, human fMRI shows that activity patterns in frontoparietal cortex contain more information about visual features when they are task-relevant. However, if this preferential coding drives momentary focus, for example to solve each part of a task in turn, it must reconfigure more quickly than we can observe with fMRI. Here I used multivariate pattern analysis (MVPA) of MEG data to test for rapid reconfiguration of stimulus information when a new feature becomes relevant within a trial. Participants saw two displays on each trial. They attended to the shape of a first target then the colour of a second, or vice versa, and reported the attended features at a choice display. I found evidence of preferential coding for the relevant features in both trial phases, even as participants shifted attention mid-trial, commensurate with fast sub-trial reconfiguration. However, I only found this pattern of results when the stimulus displays contained multiple objects, and not in a simpler task with the same structure. The data suggest that adaptive coding in humans can operate on a fast, sub-trial timescale, suitable for supporting periods of momentary focus when complex tasks are broken down into simpler ones, but may not always do so.

2.1. Introduction

Human cognition is remarkably flexible. We can fluidly direct our focus towards what we need for our current goal, seamlessly adapt to changes in our environment, and generalise from what we know to solve new problems. Several lines of research suggest that this flexibility emerges from activity in frontoparietal cortex. Cognitively challenging tasks elicit robust activity in the 'multiple demand' (MD) system—a distributed network of frontal and parietal cortex recruited by a wide range of tasks (Assem et al., 2020; Duncan, 2010; Fedorenko et al., 2013). Damage to this system linearly predicts fluid intelligence scores (Woolgar et al., 2010, 2018), which in turn powerfully predict how well we are able to acquire new skills.

The characteristic adaptability of frontoparietal regions means that they are ideally suited to supporting flexible cognition. For example, patterns of activity in the MD system, measured with fMRI, adapt to code information that is relevant for the current task. MD patterns can encode many different aspects of a task (for example, visual: Jackson et al., 2016; vibrotactile: Woolgar & Zopf, 2017; for a review see Woolgar et al 2016), commensurate with a high degree of mixed selectivity in these regions (Fusi et al., 2016; Rigotti et al., 2013). Moreover, MD coding for task-relevant stimuli is enhanced when stimuli are more difficult to discriminate (Woolgar, Hampshire, et al., 2011; Woolgar, Williams, et al., 2015) and changes to prioritise information that is at the focus of attention (Jackson & Woolgar, 2018; Woolgar, Williams, et al., 2015). Activity in at least one MD region appears to be causal for facilitating task-relevant information processing elsewhere in the MD system (Jackson et al., 2021). This may provide a source of bias to more specialised brain regions, for example through task-dependent connectivity (Cole et al., 2013; see for example Baldauf & Desimone, 2014). Consequently, adaptive coding has been proposed as a central component of goal-directed attention, biasing sensory and motor brain regions to perceive and respond to information that is relevant to our current task.

A key outstanding question concerns the temporal scale of this process. Here, I explore the 'attentional episodes' account of flexible behaviour (Duncan, 2013) which predicts a fast temporal scale. This account draws on studies of human and artificial intelligence to propose that flexible behaviour rests on our ability to break a complex task down into a series of simpler parts, and to focus, moment-to-moment, on the information needed for each part (Duncan, 2013; Duncan et al., 2012, 2017). Indeed, there is some evidence that this ability may underpin performance on novel problem solving tasks. For example, explicitly breaking a complex task into simple parts removes the performance gap between people with high and low fluid intelligence scores (Duncan et al., 2017; see also O'Brien et al., 2020). In this matrix reasoning study, participants viewed a 2x2 grid with three of the four squares filled with an image. They had to abstract relationships between the images to fill in the remaining square. Images consisted of multiple features. In another condition, each feature was presented separately. These segmented problems were trivial for participants to solve, regardless of whether they struggled or performed well on the difficult, unsegmented problems. This led the authors to propose that participants who were able to solve the unsegmented problems were better able to mentally break them down into their relevant parts. Adaptive coding could be a key component of this segmentation by driving momentary focus toward subsets of the available information in turn.

From these studies, it seems intuitive that flexible cognition involves identifying simple problems that we can solve, and addressing them in an ordered sequence. However, we do not have clear insight into whether codes reconfigure quickly enough to prioritise relevant information throughout a task. The bulk of research on adaptive coding in humans uses fMRI. While these studies show trial-to-trial shifts in what information can be discriminated from activity patterns (for example, Woolgar et al., 2011, 2015), the coarse temporal resolution of fMRI does not support precise, sub-second measurement of changes in task information.

Time-resolved methods, such as electrophysiology, EEG and MEG, offer promising evidence for sub-second or sub-trial changes in task representation. Non-human primate studies show that the same frontal neurons can encode object identity, then location, within a single trial, as monkeys attended to what, then where, an object was (Rao et al., 1997). These data demonstrate that the neural population can systematically change its activity pattern in synchrony with the task. However, they are taken from highly trained monkeys and could rely on a learned response rather than instantaneous shifts in a flexible brain system. More recent work by Spaak et al. (2017) demonstrates that, even when the same information is encoded across phases of a task, neurons in primate lateral prefrontal cortex dynamically update what they encode. This dynamic reallocation of selectivity within a trial makes it plausible that the information represented in these adaptive brain regions could indeed rapidly shift.

In humans, stronger coding for visual features when they are task-relevant compared to task-irrelevant emerges in MEG data as early as 100 ms from stimulus onset (Battistoni et al., 2020; Goddard et al., 2021; Moerel et al., 2021; Wen et al., 2019), with sustained coding of the relevant feature emerging around 200-400 ms in the MEG/EEG signal (Goddard et al., 2021; Grootswagers et al., 2021; Moerel et al., 2021; Yip et al., 2021). This provides preliminary evidence that relevance effects on population coding for visual features emerge quickly.

However, this previous time-resolved human neuroimaging work did not require participants to shift their attention within trials, so we do not know how rapidly information codes update to redirect attention in each part of a task. Rapid reorganisation of information coding within a task has been proposed as key component of how we solve complex tasks. In fact, the attentional episodes account crucially depends on our being able to fluidly reconfigure our attention as we piece together the information that we need for a decision. Thus, we should expect to see sub-trial shifts in what is preferentially encoded, and any limits on how quickly we can reconfigure our attention within a task could have widespread effects on how

we segregate and integrate complex information. Despite the likely role that shifting attention between components of a task plays in how we perform complex actions, it has been difficult to isolate fully the effects of shifting attention from concurrent shifts in stimuli or stimulus-response mapping in behaviour alone; while neurophysiological studies of non-human primates that have been able to isolate the neural correlates of shifting attention without requiring an overt response mid-trial have not yet been translated to the human brain.

Here, I explored the limits of sub-trial reconfiguration of task-relevant information in humans. I tested the dynamic adaptation of task representations when what is relevant changes within single trials. I used MEG to track shifts in adaptive coding with sub-second precision across fragments of two rapidly changing tasks. Considering the strong association between task difficulty and the brain regions implicated in adaptive coding (Crittenden & Duncan, 2014; Fedorenko et al., 2013), I tested whether preferential coding of task-relevant information reconfigured mid-trial at low and high levels of attentional demand. In Experiment 1, I used simple stimuli to track preferential coding of relevant information under low attentional demands. In Experiment 2, I used a complex stimulus space, abstracted decisions, and the presence of distractors to track preferential coding of relevant information under high attentional demands. Across both experiments, I asked whether neural codes for relevant stimulus information rapidly reconfigure when what is relevant changes mid-trial.

2.2. Methods

2.2.1. Participants

Participants were selected to (a) have normal or corrected-to-normal visual acuity and normal colour vision; (b) be right handed; (c) have no exposure to fMRI in the previous week; (d) have no non-removable metal objects; and (e) have no history of neurological damage or current psychoactive medication. Prospective participants

were informed of the study's selection criteria, aims, and procedure, through a research participation site.

For Experiment 1, 20 participants (17 female, 3 male, mean age 25±6 years) were recruited from the paid participant pool at Macquarie University (Sydney). They gave written informed consent before participating, and were paid AUD$30 for their time. Ethical approval was obtained from the Human Research Ethics Committee at Macquarie University (5201300602).

For Experiment 2, 20 participants (16 female, 4 male, mean age 31±12 years) were recruited from the volunteer panel at the MRC Cognition and Brain Sciences Unit (Cambridge). They gave written informed consent prior to each testing session, and were paid GBP£40 for their time. Participants were additionally asked to only volunteer if they had existing structural MRI scans on the panel database. Two participants took part prior to completing a structural scan; one obtained a scan through another study conducted at the MRC Cognition and Brain Sciences Unit, while the other returned for a separate MRI session as part of this study. This participant gave written informed consent before completing the structural scan and was paid an additional GBP£20 for this component of their time. Ethical approval was obtained from the Psychology Research Ethics Committee at the University of Cambridge (PRE.2018.101).

2.2.2. Stimuli

Stimuli were created in MATLAB and presented with Psychtoolbox (Brainard, 1997; Kleiner et al., 2007). In Experiment 1, they were displayed with an InFocus IN5108 LCD back projection monitor (InFocus, Portland, Oregon, USA) at a viewing distance of 113 cm. In Experiment 2, they were displayed with a Panasonic PT-D7700 projector at a viewing distance of 150 cm.

Experiment 1 stimuli consistent of four novel objects (Op de Beeck et al., 2006; see Figure 1) that were either 'cubie' or 'smoothie' shaped, and green or red

(RGB 0-194-155 and 224-0-98). Colours were chosen for high chromatic variation and strong contrast against the dark grey background (RGB 30-30-30).

Experiment 2 stimuli consisted of 16 novel "spiky" objects, adapted from the Op de Beeck et al (2006) 'spiky' stimuli, selected at four points on a spectrum of red to green, and upright to flat (Goddard et al., 2021). Colour values were numerically equally spaced in u'v' colour space between [0.35,0.53] and [0.16,0.56]. Shapes were also equally spaced to create four steps in orientation from upright to flat. Each step included 100 shape exemplars, with different spikes indicating the orientation, to discourage participants from judging orientation based on a single spike.

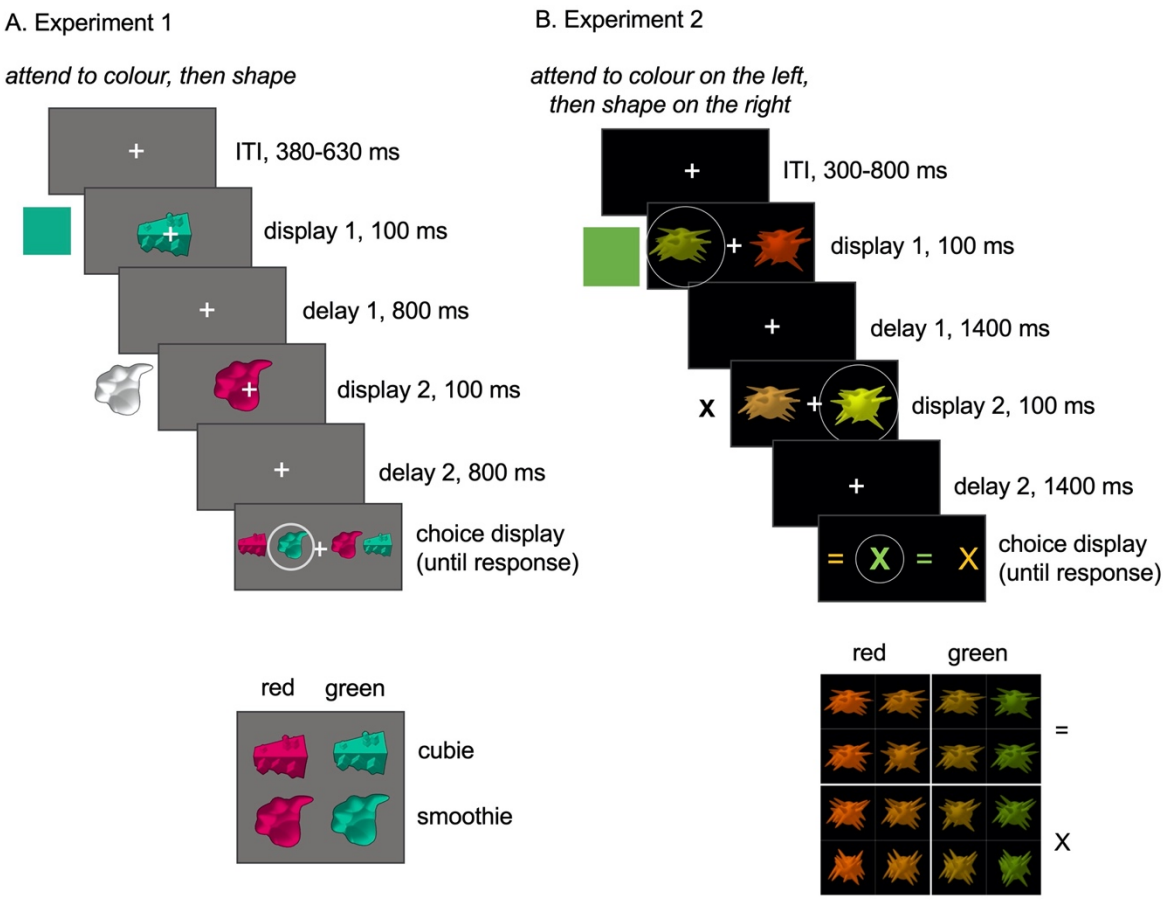


Figure 1. Stimuli and example trials for Experiment 1 and 2. Relevant information for each epoch is shown beside the display. Panel A shows an example trial for Experiment 1, with a single object on each display. In this trial, the relevant features are "green" (Target 1) and "smoothie" (Target 2), resulting in a "green smoothie" response on the choice display. Stimuli could be red, green, "cubie", or "smoothie". Panel B shows an example trial for Experiment 2, in which the participant was cued to attend to colour on the left, then shape on the right. The relevant features were thus green and "X", leading to a response of "green X" on the choice display. Stimuli varied in four steps from red to green, and from X to =, but were assigned to binary red / green, X / = categories. Circles represent the focus of attention and correct choice and were not shown to participants.

2.2.3. Task

Experiment 1 used simple displays and stimuli, optimised for strong visual signals. Each block began with a written cue instructing participants to attend to the colour of the first object, and the shape of the second object, or vice versa. On each trial, participants viewed two brief displays (100 ms), each followed by a delay

(800 ms; see Figure 1). Finally, they were prompted to select an object from a choice display that comprised the combination of the remembered features. All four objects appeared on the choice display, and participants selected the object that matched the colour and shape they had extracted from the preceding displays. For example, under the rule "attend shape, then colour", if the first object was a 'cubie' and the second object was 'red', the target on the choice display was a red cubie. Participants indicated their choice by pressing one of four buttons on a bimanual fibre optic response pad operated with the four fingers of the right hand. The mapping from object location to response button was intuitive (far left button for far left object, etc) and consistent across trials; however, the arrangement of the four objects on the choice display varied to prevent participants preparing a motor response until the display screen was shown. Stimulus arrangements were presented in pseudo-random order and balanced within each rule such that all stimuli on the second display were equally preceded by each stimulus on the first display, and the correct choice pertained equally to all motor responses. If a participant made three consecutive incorrect or slow responses (>3 s), the task was paused and the cue was presented again until the participant verbally confirmed that they understood the rule for that block. Average accuracy and response times were displayed at the end of each block.

Experiment 2 followed the structure of Experiment 1, but used simultaneously presented objects and subtler stimulus discriminations, optimised for high attentional load. For this experiment, each display contained two objects. Participants were cued to both a location and feature, for example, "attend to shape on the right, then colour on the left". Relevant location and feature always changed from display 1 to display 2, creating four possible rules. Delay periods were increased to 1500 ms to allow accurate responses, following piloting of the task. Participants judged the colour and shape *category* of the target objects' features. The choice display contained the symbols X and =, presented in the average of the two 'red' colours and the average of the two 'green' colours, to represent the four

possible answers. These symbols were chosen to encourage participants to make category-level decisions about the objects. As in Experiment 1, the spatial arrangement of the items on the choice display was updated on each trial.

2.2.4. Procedure

2.2.4.1. Experiment 1

Each participant first completed four blocks of 10 practice trials outside the shielded room. These were identical to test trials except that (a) participants received feedback of 'Correct', 'Incorrect', or a red screen signifying a slow response (>3 s), on every trial, (b) display durations in the first two practice blocks were slowed from 100 ms to 500 ms to ease participants into the task, and (c) response key codes were marked on the choice display to train participants in the location-response mapping. Once in the MEG scanner, participants completed eight blocks of 96 trials each, with feedback at the end of each block. Each block lasted approximately seven minutes. Blocks alternated between the two rules, 'attend shape, then colour' and 'attend colour, then shape', with the order counterbalanced across participants.

2.2.4.2. Experiment 2

Participants learned the stimulus categories (red vs green, upright vs flat) and the task in a separate training session. Training could be on the day of or the day before the scanner session. Training consisted of two blocks of 50 category learning trials, in which they saw a single object for 100 ms and pressed a button to indicate its shape or colour category. They then began training on the core task. Within-trial delay periods began at 4 s and reduced to 1.5 s in three steps (3 s, 2 s, 1.5 s). Participants completed a minimum of 10 trials at each of the four speeds for each of the four rules (that is, at least 40 trials per rule). After 10 trials were completed, the speed increased when the participant got 8 trials correct in any 10

consecutive trials. Feedback was given on each trial by a brighter fixation cross for correct responses and a blue fixation cross for incorrect responses, shown for the first 100 ms of the post-trial interval. This procedure trained each participant to the same criterion without penalising them for errors early in the block.

Once in the MEG, participants completed four blocks, each corresponding to a single rule and comprising 258 trials, lasting approximately 20 minutes. Rule order was balanced across participants.

2.2.5. MEG data acquisition

2.2.5.1. Experiment 1

I acquired MEG data in the Macquarie University KIT-MEG lab using a whole-head horizontal dewar with 160 coaxial-type first-order gradiometers with a 50 mm baseline (Model PQ1160R-N2; KIT, Kanazawa, Japan; Kado et al., 1999; Uehara et al., 2003) in a magnetically shielded room (Fujihara Co. Ltd., Tokyo, Japan). First, the tester fit the participant with a cap containing five head position indicator coils. The location of the nasion, left and right pre-auricular, and each of the head position indicators were digitised with a Polhemus Fastrak digitiser (Polhemus, VT, USA). This information was copied to the data acquisition computer to track head position during data collection. Participants lay supine during the scan, and were positioned with the top of the head just touching the top of the MEG helmet. Any change in head position relative to the start of the session was checked and recorded after four blocks. MEG data were recorded at 1000 Hz.

2.2.5.2. Experiment 2

I acquired MEG data with the MRC Cognition and Brain Sciences' Elekta-Neuromag 306-sensor Vectorview system with active shielding. Ground and reference EEG electrodes were placed on the cheek and nose. Bipolar electrodes for eye movements were placed at the outer canthi, above and below the left eye.

Heartbeat electrodes were on the left abdomen and right shoulder. Scalp EEG were also applied for a separate project. Head position indicators were placed on top of the EEG cap. Both head shape and the location of the head position indicators were digitised with a Polhemus Fastrak digitiser. Head position was recorded continuously throughout the scan and viewed after each block to ensure that the top of the participant's head stayed within 6 cm of the top of the helmet in the dewar (mean movement across task 3.94 mm, range 0.5:15 mm). Because targets in this experiment could appear to either side of fixation, I also recorded eye-movements with an EyeLink 1000 eye tracker, which I calibrated before each block. Two participants were incompatible with the eye tracker due to eyewear interfering with the tracker camera's view of their pupil.

2.2.6. Analyses

2.2.6.1. MEG processing

Due to active shielding and artefacts from continuous head position indicators, data from Experiment 2 were first processed with Neuromag's proprietary filtering software (*Maxfilter*, 2010). I applied temporal signal space separation to remove environmental artefacts, used continuous head position information to correct for head movement within each block, and reoriented each block to the subjects' initial head position.

All other processing was the same across experiments. I used a minimal pre-processing pipeline to minimise the chance of removing meaningful data. This was especially appropriate in the present case, as the planned multivariate analyses are typically robust to noise (Grootswagers et al., 2016). MEG data were imported into MATLAB v2018b using Fieldtrip (Oostenveld et al., 2011), and bandpass filtered (0.1-45 Hz). I did not apply a notch filter, as the line noise frequency fell outside the bandpass filtering range. I also did not apply baseline correction, as I planned to compare similar features (i.e. red and orange) within a run, and did not expect

broad changes in mean response to impact decoding. I later ran the analysis with baseline correction and confirmed that this did not change the results. Trials were epoched from a 100 ms pre-stimulus baseline to the maximum possible trial duration (Exp 1: 4800 ms, Exp 2: 5000 ms). All sensors were included in the analysis. I also included all trials in the analyses, both correct and incorrect, as incorrect responses could occur on trials in which participants accurately perceived and attended to one task-relevant feature dimension; and because excluding trials would disrupt counterbalancing, which would be imperfectly solved by subsampling at the decoding stage.

2.2.6.2. MEG decoding

I used multivariate pattern analysis to trace the information about rule, colour, and shape in each task phase. I then compared the information about colour when it was relevant and irrelevant, repeating the comparison for shape. Following previous studies, I expected that rule information, which was known before each trial, would be present throughout the trial and increase briefly after visual displays (Goddard et al., 2021; Hebart et al., 2018). I predicted that preferential coding would be reflected in improved decoding of visual features when they were relevant, compared to irrelevant (Battistoni et al., 2020; Goddard et al., 2021; Grootswagers et al., 2021; Hebart et al., 2018; Moerel et al., 2021; Wen et al., 2019; Yip et al., 2021). Increased colour information when colour was relevant would indicate that information was flexibly coded according to task demands. The critical comparison, then, was how this happened for the two task phases. If information about the relevant feature was prioritised in both task epochs, this would indicate that preferential coding can reconfigure in line with sub-second shifts in what is relevant to the task.

To implement these analyses, I used a linear classifier (linear discriminant analysis, LDA; see Grootswagers et al., 2016) implemented through CoSMoMVPA. Sensors were normalised so that different sensor types (magnetometers and

gradiometers) could be included in the same analysis. I defined some training data, taking labelled trials from the two feature rules—"attend colour, then shape" and "attend shape, then colour"—and using all but one trial from each category. For speed and to reduce noise, I implemented a principal component analysis on the sensors in the training data (also implemented in CoSMoMVPA), retaining the first components that explained 99% of the variance. The covariance matrix was not regularised. Then, I trained the classifier on the labelled training data. The training step allows the classifier to generate a model of the distributions underlying the data for each class, mapping out what patterns of activation are likely in each condition. I then tested whether the distributions that the classifier had generated to discriminate the classes during training would generalise to the remaining, unobserved trials. I repeated the process, leaving out a different pair of trials each time, until all trials had acted as test data. I then averaged the classification accuracy across all test sets.

For colour and shape classification, I trained a linear classifier on labelled data from two categories—for example, "red" and "green"—using all but one trial from each category, for each feature rule separately. For Experiment 2, I decoded pairs of shape or colour, at a fixed location, for each feature and location rule. For example, I took trials under the rule "attend colour on the left, then shape on the right". For items on the left on the first display, I decoded strong red vs yellow red, yellow red vs yellow green, and so on for all six pairs of colour. I then averaged classifier accuracy across the six pairs into a single measure of colour information coding in the left hemifield under this rule. I repeated this for each rule to obtain four traces of left hemifield colour information coding, representing colour information when that location and feature were relevant or irrelevant. I conducted the same pairwise decoding and averaging for colour in the right hemifield. Conducting the analyses for each hemifield separately minimised the requirement for the classifier to generalise patterns over space. Finally, I averaged the four traces of left hemifield colour information coding with the corresponding right

hemifield traces to produce a single trace for each attention condition: "attended location, attended feature" (the task-relevant trace), "attended location, unattended feature", "unattended location, attended feature", and "unattended location, unattended feature". The two traces for colour (or shape) information at the attended location parallel the two traces for each target in Experiment 1 and form the central part of my analysis.

2.2.6.3. Statistical tests

I tested whether decoding accuracy scores were above chance using a null distribution generated from the data. To generate this, I permuted the predicted class labels so that they were randomly assigned over trials (Bae & Luck, 2019). I calculated decoding accuracy as above and repeated the process 10,000 times to produce a decoding distribution for each participant and each comparison. I then sampled 10,000 times across participants' null distributions to form a group-level null distribution. At each timepoint, I calculated t-scores for classification accuracy relative to the null distribution (Stelzer et al., 2013). I used a threshold-free cluster statistic (threshold step 0.1; Smith & Nichols, 2009), implemented in CoSMoMVPA, to flexibly set a cluster-forming threshold to identify peaks in the t-score time-course that were more strong and/or sustained than expected from the null distribution ($p<.05$). This maximises sensitivity to peaks that are most likely to reflect meaningful change while down-weighting peaks that are small or transient (S. M. Smith & Nichols, 2009). I then corrected for multiple comparisons at the cluster level across the whole trial. Decoding onset was the onset of the first cluster for which decoding accuracy was reliably above chance.

For between-condition comparisons, I contrasted the decoding trace for the target when it was the relevant or irrelevant feature using a two-sided t-test, implemented in CoSMoMVPA (Oosterhof et al., 2016) with threshold-free cluster enhancement and a threshold step of 0.1 ($p<.05$; Smith & Nichols, 2009; Figures 4 & 5).

For Experiment 2, I also conducted secondary analyses to assess the combined effects of spatial- and feature-selective, as reported in Goddard et al. (2021). I conducted 2x2 ANOVAs to test, for each time bin, whether stimulus colour and shape information coding was boosted (1) at the relevant compared to irrelevant location, (2) when that stimulus feature was relevant for the task compared to when it was irrelevant, and (3) when both feature and location were relevant compared to all other attention conditions. I quantified these as main effects of spatial and feature-selective attention, and as a planned comparison between the coding of the reported feature at the attended location and the coding of that feature at that location in the other 3 attention conditions (following our prediction from Goddard et al., 2021). For example, I contrasted decoding for colour on the left when people were attending to colour on the left, with decoding for colour on the left when attending to shape on the left, colour on the right, and shape on the right. I present the results of these secondary analyses in Figure 6.

Lastly, in Experiment 2 I asked whether attentional effects had similar temporal profiles in epoch 1 and epoch 2 of the trial. I epoched the stimulus decoding traces for the target, separately around the first and second stimulus displays (0:1500 ms), using the same pre-trial baseline (-100:0 ms) for all traces. This created four overlaid traces, a relevant and an irrelevant feature trace for Epoch 1 and Epoch 2. I conducted a 2x2 ANOVA with main effects of relevance and epoch. An interaction term tested the hypothesis that preferential coding of relevant information emerges earlier, or is more substantial, in one epoch compared to the other.

2.3. Results

2.3.1. Behavioural performance

In Experiment 1, median accuracy was 93.3% (std 7.5%), with median reaction time 829.2 ms (std 210.7 ms). In Experiment 2, median accuracy was 75.9%

(std 10.9%), with median reaction time 665.2 ms (std 92.1 ms). In both tasks chance accuracy was 25%.

Trials with partly accurate responses, in which participants selected an item with either the correct colour or the correct shape, were treated as incorrect. However, these trials provide an estimate of whether participants were more able to perceive and retain information about one feature. In Experiment 1, participants reported the correct colour or shape on approximately half of the error trials (colour: median 45.6%, std 11.2%; shape: median 39.2%, std 11.0%). In Experiment 2, participants similarly reported the correct colour or shape on approximately half of the error trials (colour: median 55.7%, std 9.1%; shape: median 36.5%, std 6.7%). These values roughly correspond to chance.

Notably, partly correct responses in Experiment 1 did not appear to reliably favour one feature over the other. By contrast, partly correct responses in Experiment 2 were more often due to participants correctly reporting colour, compared to shape. This difference was statistically reliable at the alpha = .05 level, based on a two-sided paired t-test ($t_{1,19} = 4.54$, 95% CI = [.08,.21], $p<.01$). Participants were clearly able to perform both colour and shape categorisation, as overall performance was ~76%. Yet the difference between colour and shape categorisation performance remains when we consider the total percentage of trials in which colour or shape were reported correctly (colour: median 88.7%, std 6.9%; shape: median 84.1%, std 8.1%; two-sided paired t-test: $t_{1,19} = 3.74$, 95% CI = [.02, .06], $p < .01$). Thus, colour information may have been easier to perceive or to split into binary categories. In that case, neural differences between task rules ("attend colour then shape", or "attend shape then colour") could reflect differences in difficulty, effort, and engagement (e.g. more effort in the second epoch when attending to colour then shape) rather than the content of the rule. Easier perception or categorisation of colour could also produce different time-courses for stimulus information and attention effects in the neural data.

We might then predict that decoding accuracy would be higher for colour and create an artificial "attention" effect on relevant feature coding when participants are attending to colour. We would also expect to see higher decoding accuracy for colour when participants are attending to shape. The results instead show higher decoding accuracy for the task-relevant feature in both attend-colour and attend-shape conditions. More effective categorisation for colour does not explain this outcome. However, future studies could reduce unwanted variation by matching the difficulty of colour and shape categorisation with a pilot group, or within each subject.

2.3.2. Eye tracking analysis

Attentional effects on eye movements could in theory enhance decoding of the task-relevant feature. Fixating toward the target object could engage strong foveal processing. Directing gaze toward many points on the target could enhance detection of its outline, though this should be limited by the short stimulus durations. While they were instructed to fixate on the centre of the screen, subjects could have benefitted from looking left when attending to the leftward object, or from looking around the target when trying to judge its shape. Consequently, I report below some key gaze parameters for location and feature rules.

I extracted each subject's horizontal and vertical eye position during each stimulus display. I then averaged over epochs with corresponding location or feature rules (for example, the first epoch when attending to colour then shape averaged with the second epoch when attending to shape then colour). I calculated some group-level statistics to test whether gaze was (a) biased toward the attended location, and (b) more varied when subjects attended to shape compared to colour.

Gaze was, on average, 19.47 pixels further left of centre when subjects were asked to attend to an object on the left, compared to the right. This is roughly 1.5% of the screen's width (19.47/1280 pixels). The difference was not statistically reliable

at the alpha=.05 level, based on a group-level, two-sided, one-sample t-test against zero ($t_{1,17}$=-2.08, p=.053).

Standard deviations were, on average, 4.19 pixels larger horizontally and 1.81 pixels larger vertically when subjects were asked to attend to shape, compared to colour. These differences roughly amount to .33% (4.19/1280 pixels) and .18% (1.81/1024) of the screen's width and height. Again, these numerical differences did not reach statistical significance at the alpha=.05 level, based on two-sided, one-sample t-tests against zero (left-right position: $t_{1,17}$=1.44, p=.167; up-down position: $t_{1,17}$=.58, p=.570).

Overall, these data do not support the idea that gaze parameters substantially changed with location or feature rules. Attentional effects on gaze followed logical predictions, with more leftward bias when attending left and more variation when attending to shape. However, neither effect reached statistical significance. Perhaps more importantly, the impact of the attention conditions on gaze position and variation was very small, around 1% of the screen in either dimension. While the gaze data cannot rule out the possibility that eye movements altered decoding accuracy, the subtlety of these changes in eye movements across attention conditions should limit their impact on MEG signals.

2.3.3. Rule information coding

I trained a classifier to discriminate between feature attention rules ("attend shape, then colour" from "attend colour, then shape") from MEG data to extract a time-course of rule information coding (Figure 2). Since the rule was cued at the start of the block, I expected that participants might prepare their task set in advance of the stimulus display. I anticipated that rule information would be more decodable after each display, when the rule could be applied to extract relevant information (as in Goddard et al., 2021). Indeed, rule information coding emerged early in both experiments, increasing after each stimulus onset, and remaining above chance throughout the trial. In fact, in Experiment 1, rule decoding was

reliably above chance for a component beginning before the stimulus onset. This could reflect differences in head position or univariate responses between "attend shape, then colour" and "attend colour, then shape" blocks. However, it could also reflect true rule information, as rules were cued well in advance of the trial. Rule information coding gradually ramped up after each display in Experiment 1, whereas in Experiment 2 rule information coding was elevated throughout the trial and peaked steeply after each display. For Experiment 2, I collapsed the feature rule analysis over locations to mirror Experiment 1 (Figure 2). I also decoded the location rule (i.e., "attend left, then right" and "attend right, then left"), which I show in Figure 3 for completeness.

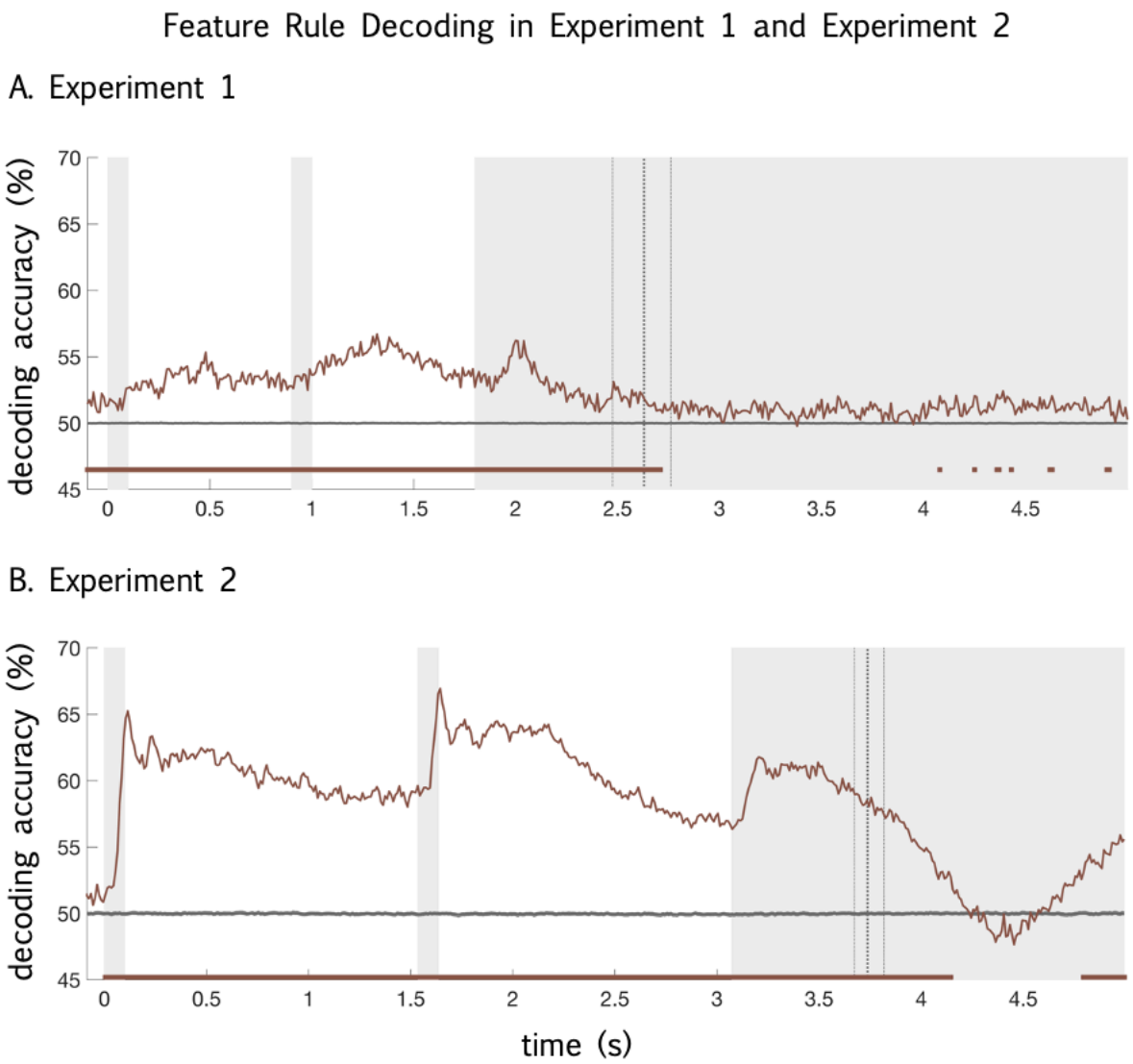


Figure 2. Feature rule decoding ("attend colour then shape" vs "attend shape then colour") for Experiment 1 (A) and Experiment 2 (B). Vertical grey patches mark the stimulus displays and the maximum possible duration of the choice display. Vertical dotted lines mark the median response time with one quartile on either side. Horizontal grey lines show chance (50%) bounded by the 95% confidence interval for the null mean, which I estimated from permutation-based null data. Timepoints at which decoding was reliably different to the null based on threshold-free cluster correction are marked below the trace in brown.

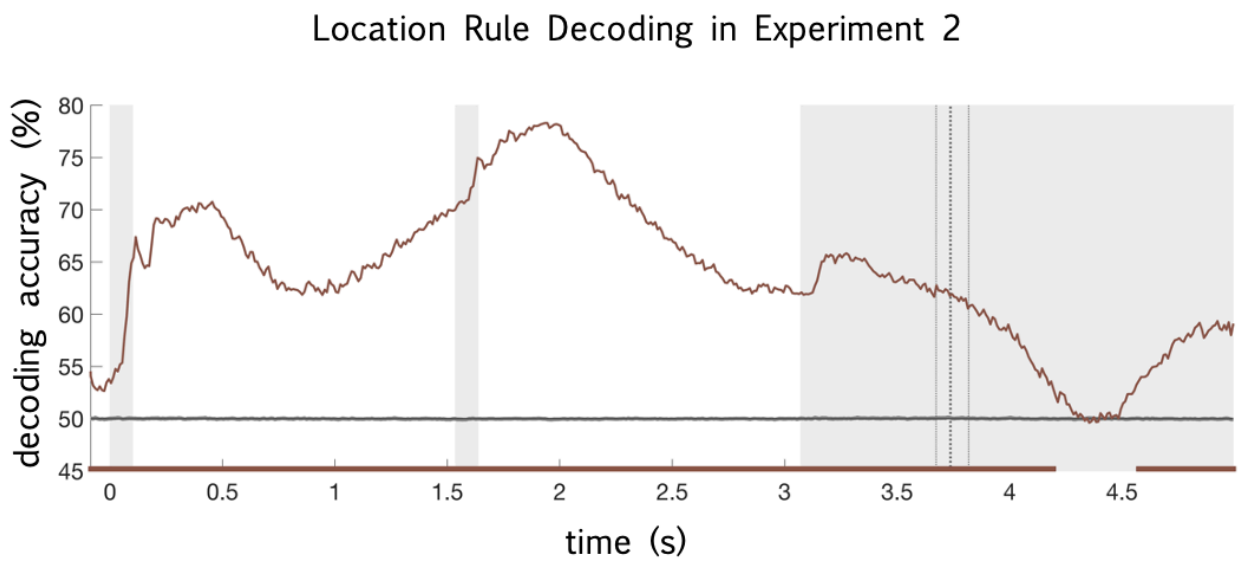


Figure 3. Location rule decoding ("attend left then right" vs "attend right then left") for Experiment 2. Vertical grey patches mark the stimulus displays and the maximum possible duration of the choice display. Vertical dotted lines mark the median response time with one quartile on either side. Horizontal grey lines show chance (50%) bounded by the 95% confidence interval for the null mean, which I estimated from permutation-based null data. Timepoints at which decoding was reliably different to the null based on threshold-free cluster correction are marked below the trace in brown.

2.3.4. Preferential coding of visual features

Next, I examined the time-course with which I could decode stimulus colour and shape from the pattern of MEG activity. I quantified this separately when a feature was relevant or irrelevant for the participant's task so that I could examine the effect of attention on coding of this information. I predicted that both relevant and irrelevant stimulus features would be decodable from the sensor data, but that each feature would be more readily decoded when it was relevant compared to when it was irrelevant, particularly at later timepoints (Goddard et al., 2021; Hebart et al., 2018; Moerel et al., 2021). In Experiment 1, robust decoding of stimulus information emerged rapidly after the onset of each display, remaining through the initial part of the delay phase for each epoch (Figure 4). Contrary to my prediction, however, in Experiment 1 there was no reliable evidence of preferential coding of

the currently relevant information, in either task epoch, for colour or shape information (Figure 4).

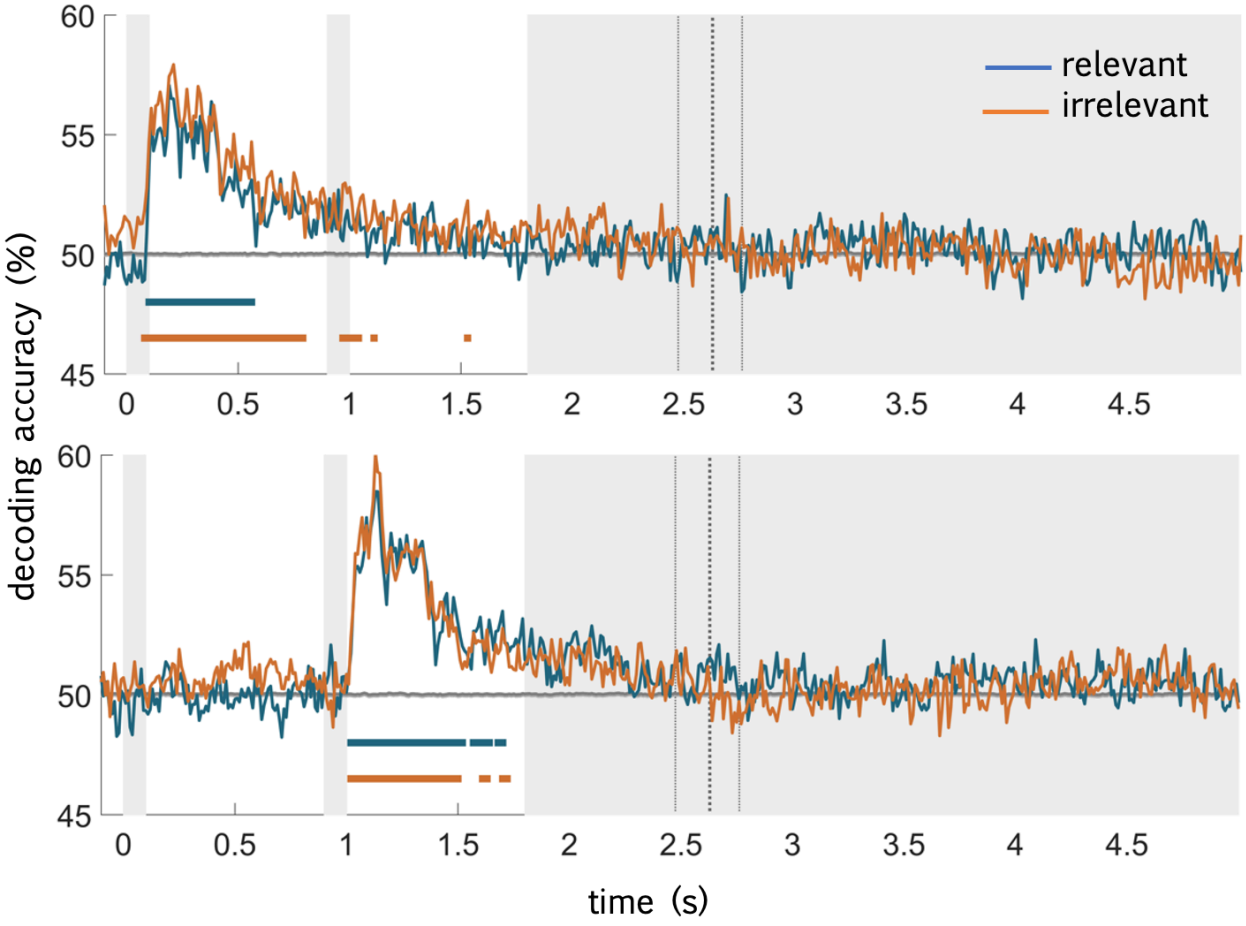
Experiment 2 stimulus decoding was similarly rapid (Figure 5). Although less pronounced (potentially due to the busier displays and more subtle colour and shape differences) initial stimulus decoding peaks followed a similar timecourse to Experiment 1. For coding of colour, there was an initial stimulus-driven response peaking at 100 ms, which was similar when that information was relevant or irrelevant, and which occurred for both epochs, though these peaks did not reach statistical significance. For shape, the pattern was broadly similar and statistically significant, with an initial stimulus-driven response at 100 ms from each display onset. Critically, in contrast to Experiment 1, in Experiment 2 I now saw evidence of additional, sustained, preferential coding of relevant information. Whereas decoding for the target's colour remained close to chance when that feature was irrelevant, coding for the same information when it was relevant was higher and sustained (Figure 5). Coding of relevant colour information was reliably different to chance and to the irrelevant feature trace from approximately 500 ms after stimulus presentation and was sustained into the subsequent trial epoch. I observed the same pattern for shape decoding, with a sustained response only for the relevant information in both epochs, although this was statistically reliable only in the second epoch.

As a secondary analysis, I additionally considered coding of the features of the distractor object. All four traces (relevant and irrelevant feature of target and distractor) are shown in Figure 6. Colour and shape information was briefly decodable in all four attention conditions, after which there was a sustained preferential coding of the relevant target feature compared to the average of all other features (Figure 6, black lines). Main effects of spatial or feature-selective attention sometimes emerged before the interaction (Figure 6, epoch 2 colour and epoch 1 shape), suggesting that attention to the relevant location and feature independently drive preferential coding. Spatial attention effects were also

significant for shape in both epochs, overlapping (epoch 1) or interspersed (epoch 2) with the time-course of the interaction. This main effect of spatial attention could be explained by selective enhancement of the relevant target feature, which is best described by the interaction. Alternately, it could reflect a broad spatial bias throughout the trial. The interaction term compared the relevant target feature to the average of all other features, meaning that within-target differences could be small, and the interaction significant due to target-distractor differences. Timepoints showing reliable differences within the target location (Figure 5) were relatively sparse, suggesting that the enhancement of relevant shape information captured by the interaction was partly driven by an overall bias toward the target location. Broadly, though, both primary and secondary analyses showed preferential coding specifically for the information that the participants needed to retain.

Colour and Shape Decoding in Experiment 1

A. Colour Decoding, Epoch 1 and Epoch 2 Stimuli



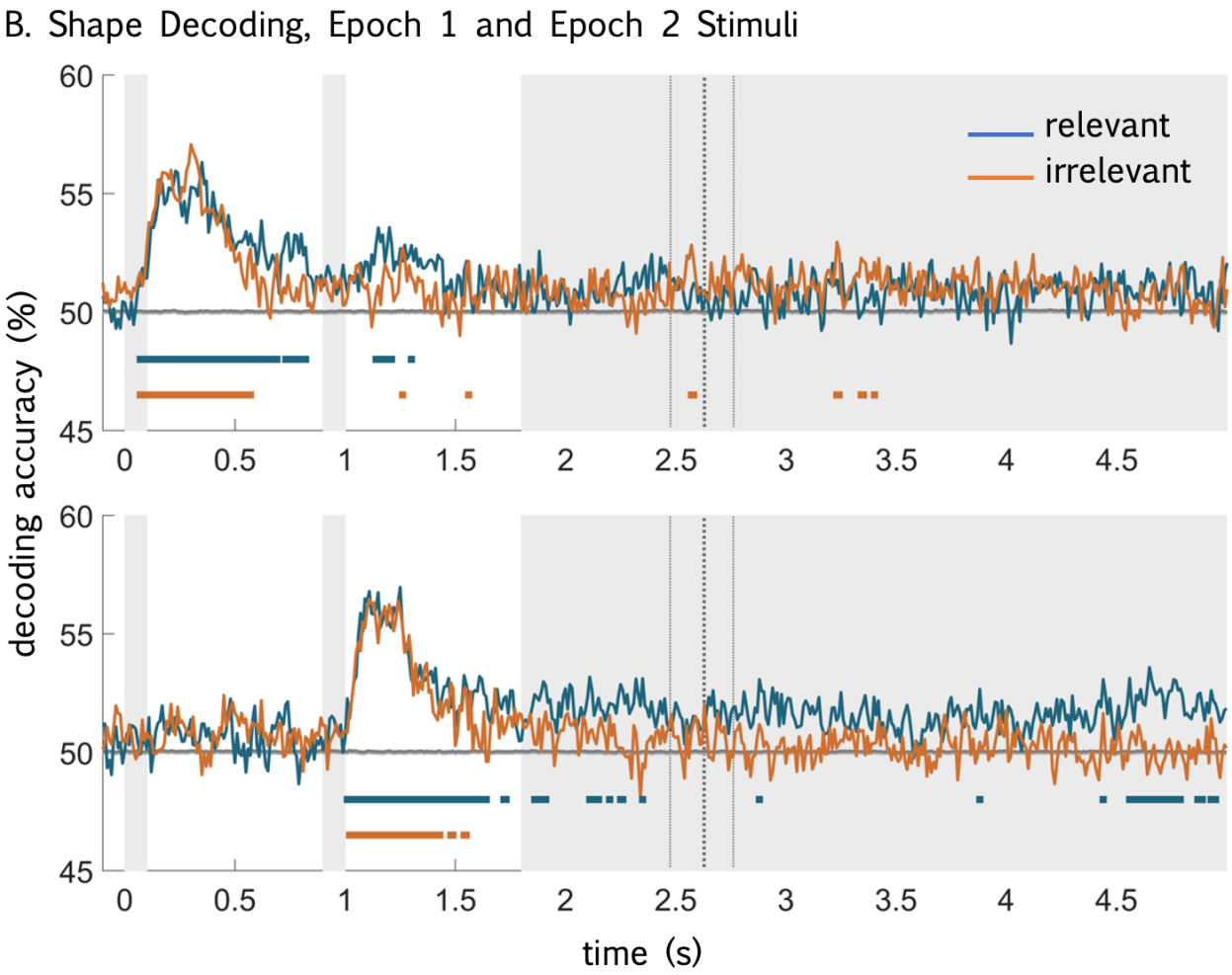
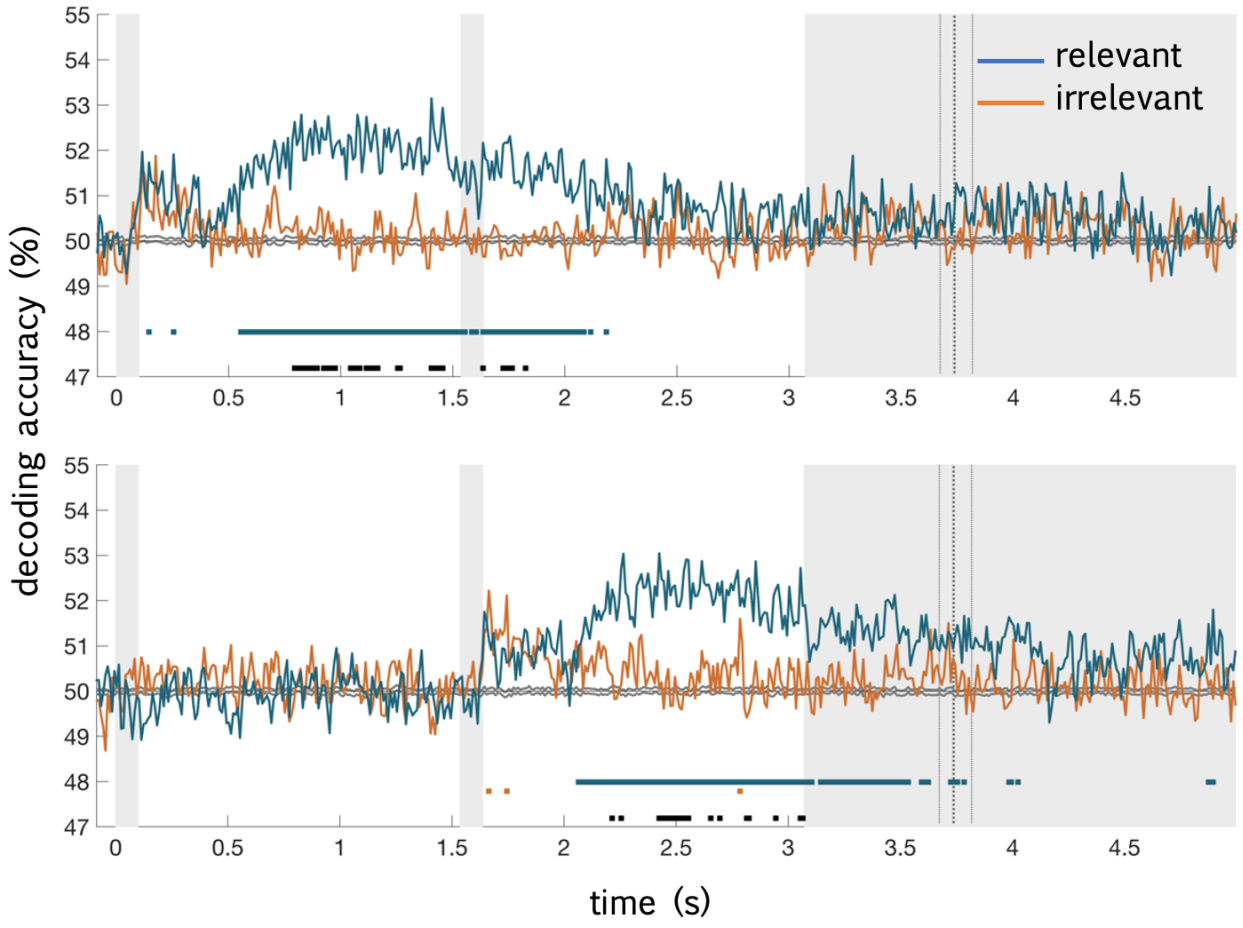


Figure 4. Colour (A) and shape (B) decoding for Experiment 1. A and B show decoding traces for the first and second targets in the upper and lower panels. Decoding accuracies are shown for each feature when it was relevant (blue) or irrelevant (orange) for the task. Grey bars mark the stimulus and response display durations. Vertical lines show the median response time, ± one quartile. Times at which decoding was greater than chance, $p<0.05$ using a cluster-based correction for multiple comparisons, are marked below each trace in the corresponding colour. Relevant information coding did not reliably exceed coding for the irrelevant feature at any timepoint.

Colour and Shape Decoding in Experiment 2

A. Colour Decoding, Epoch 1 and Epoch 2 Stimuli



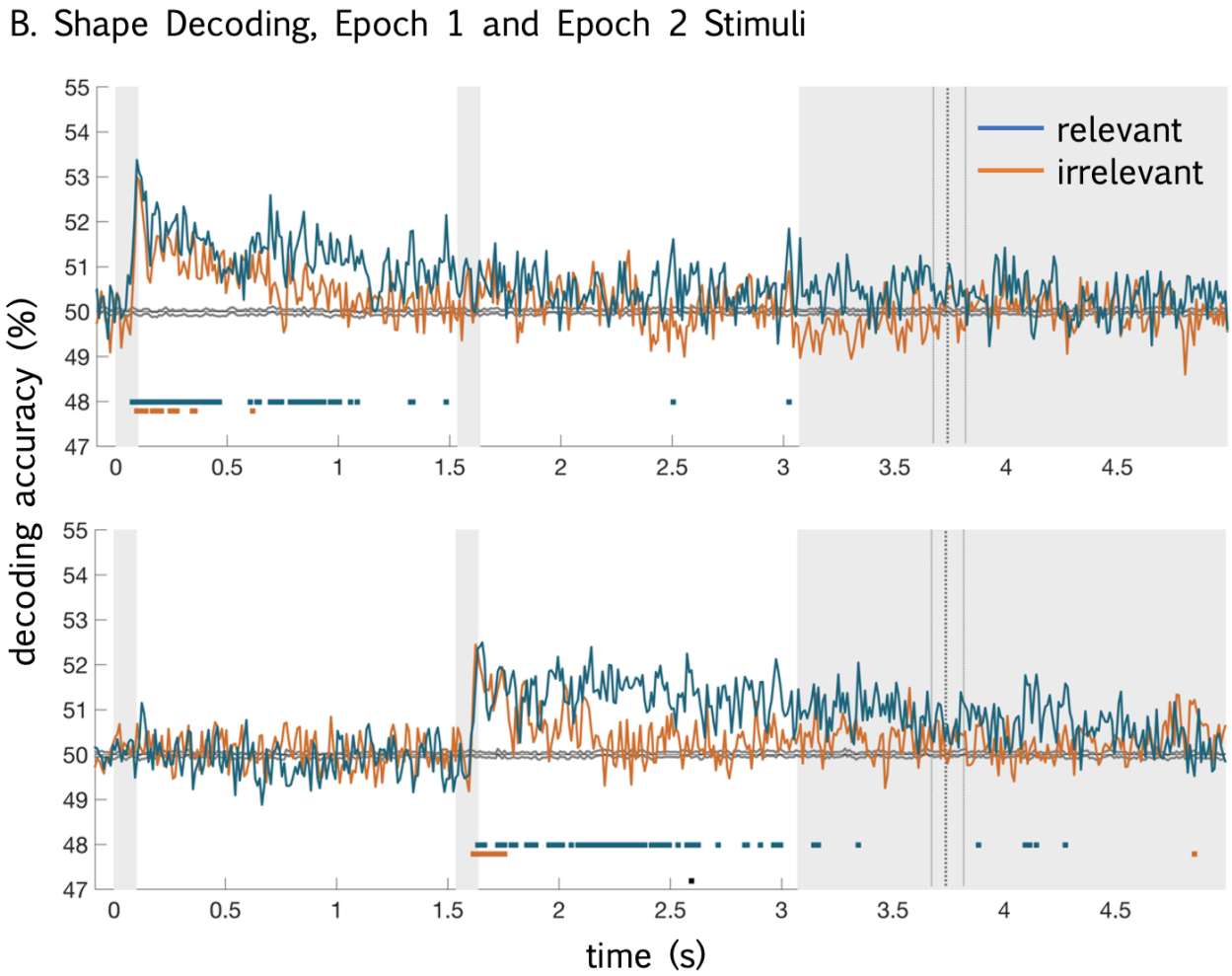
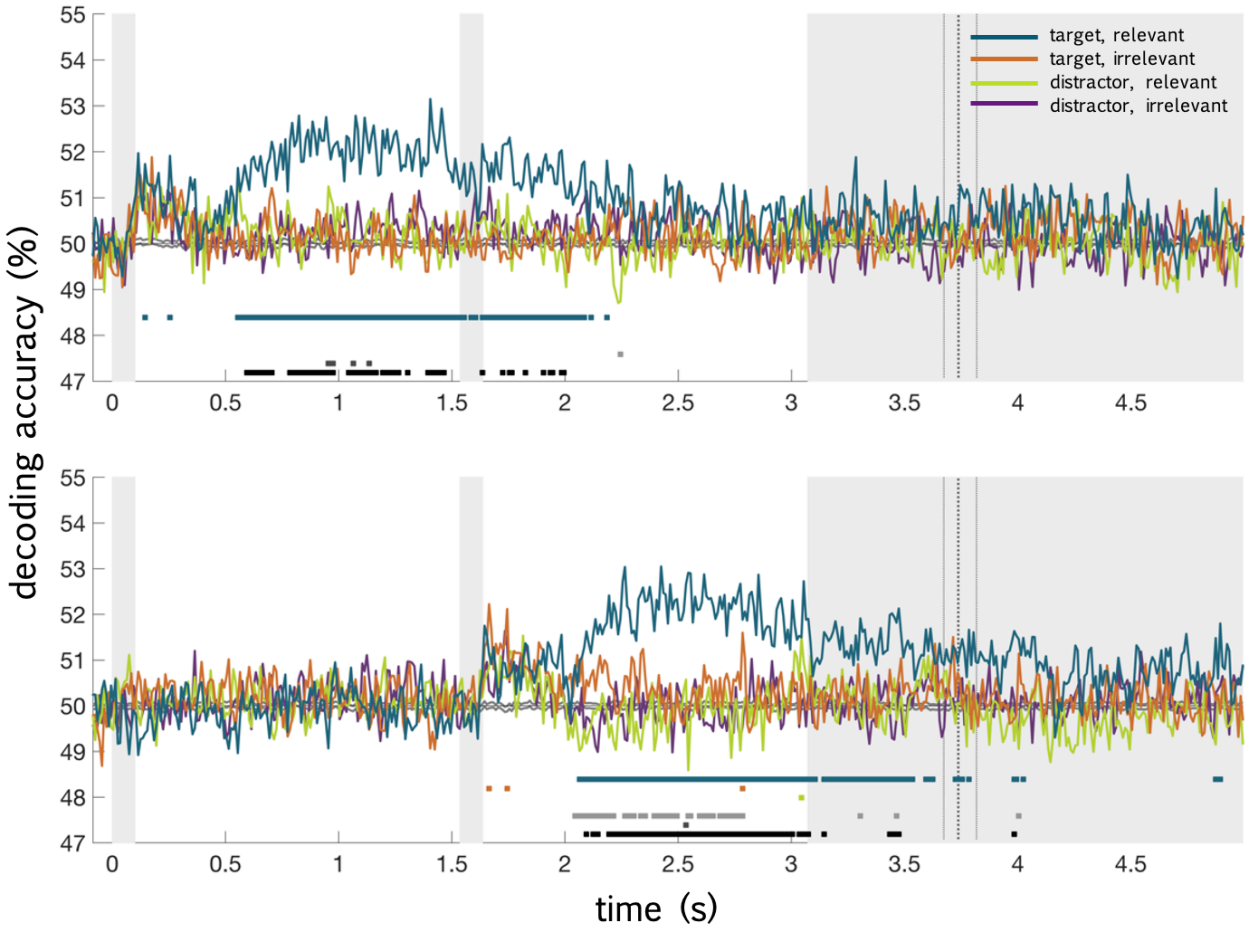


Figure 5. Colour (A) and shape (B) decoding for Experiment 2. A and B show decoding traces for the first and second targets in the upper and lower panels. Decoding accuracies are shown for each feature when it was relevant (blue) or irrelevant (orange) for the task. Grey bars mark the stimulus and response display durations. Vertical lines show the median response time, ± one quartile. Times at which decoding was greater than chance, $p < 0.05$, using a cluster-based correction for multiple comparisons, are marked below each trace in the corresponding colour. Times at which relevant information coding was reliably above coding for the irrelevant target feature (threshold-free cluster correction, $p<.05$) are marked in black.

Colour and Shape Decoding in Experiment 2 With Distractor Object Traces

A. Colour Decoding, Epoch 1 and 2 Stimuli



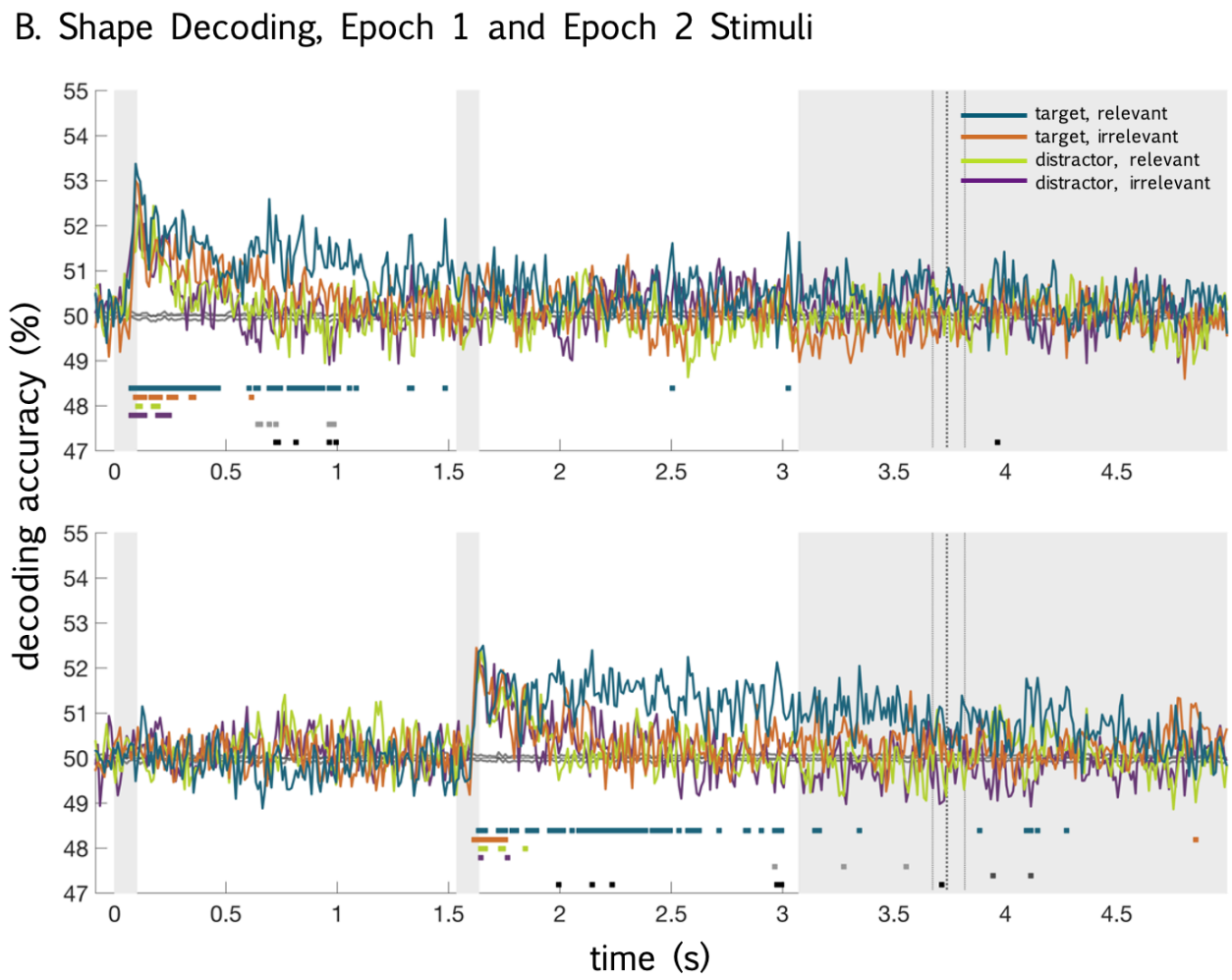
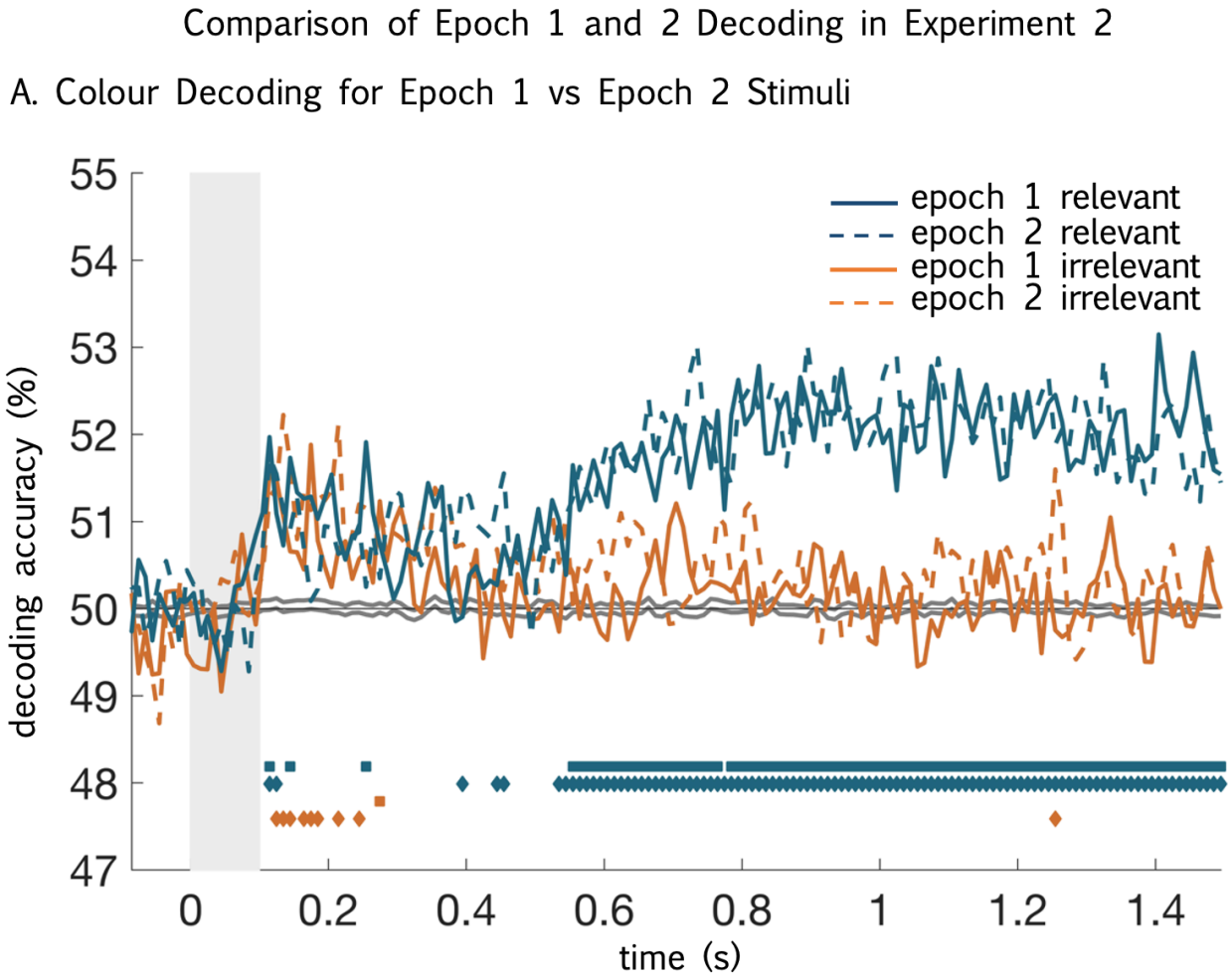


Figure 6. Experiment 2 colour (A) and shape (B) decoding for the target and distractor objects on each display. Traces represent decoding accuracy for colours or shapes at the attended location (blue=relevant feature, orange=irrelevant feature), data repeated from Figure 5, as well as at the unattended location (green=attended feature, purple=unattended feature). Times at which each trace was reliably different to chance, at $p<0.05$ with a threshold-free cluster correction for multiple comparisons, are marked in the corresponding colour. Greyscale markers indicate times with a statistically reliable effect of spatial attention (target vs distractor, light grey), feature attention (attended vs unattended feature, dark grey), or interaction between spatial and feature attention (relevant feature of target vs all other features, black).

2.3.5. Rapid coding of features across epochs

To compare the dynamics of attentional prioritisation across the two epochs, I took the decoding traces for the target in each epoch of Experiment 2 and aligned them in time. I anticipated that the effect of attention (enhancement of relevant information) might develop later in Epoch 2, which reflected a sub-trial shift of

attention when participants had less time to prepare what they would attend to. However, preferential coding for relevant information in Epoch 2 was comparable to Epoch 1 (Figure 7). I did not observe a main effect of epoch, or an interaction between epoch and relevance. This does not rule out the possibility that shifting attention mid-trial incurs some delay in preferential coding in other circumstances, for example with more difficult tasks or a shorter within-trial inter-stimulus interval. However, it demonstrates that humans can rapidly reconfigure their neural codes to prioritise coding of a new stimulus dimension mid-trial, even while holding the previously attended stimulus information in mind. Commensurate with non-human primate work, this highlights our capacity to dynamically code task-relevant information.



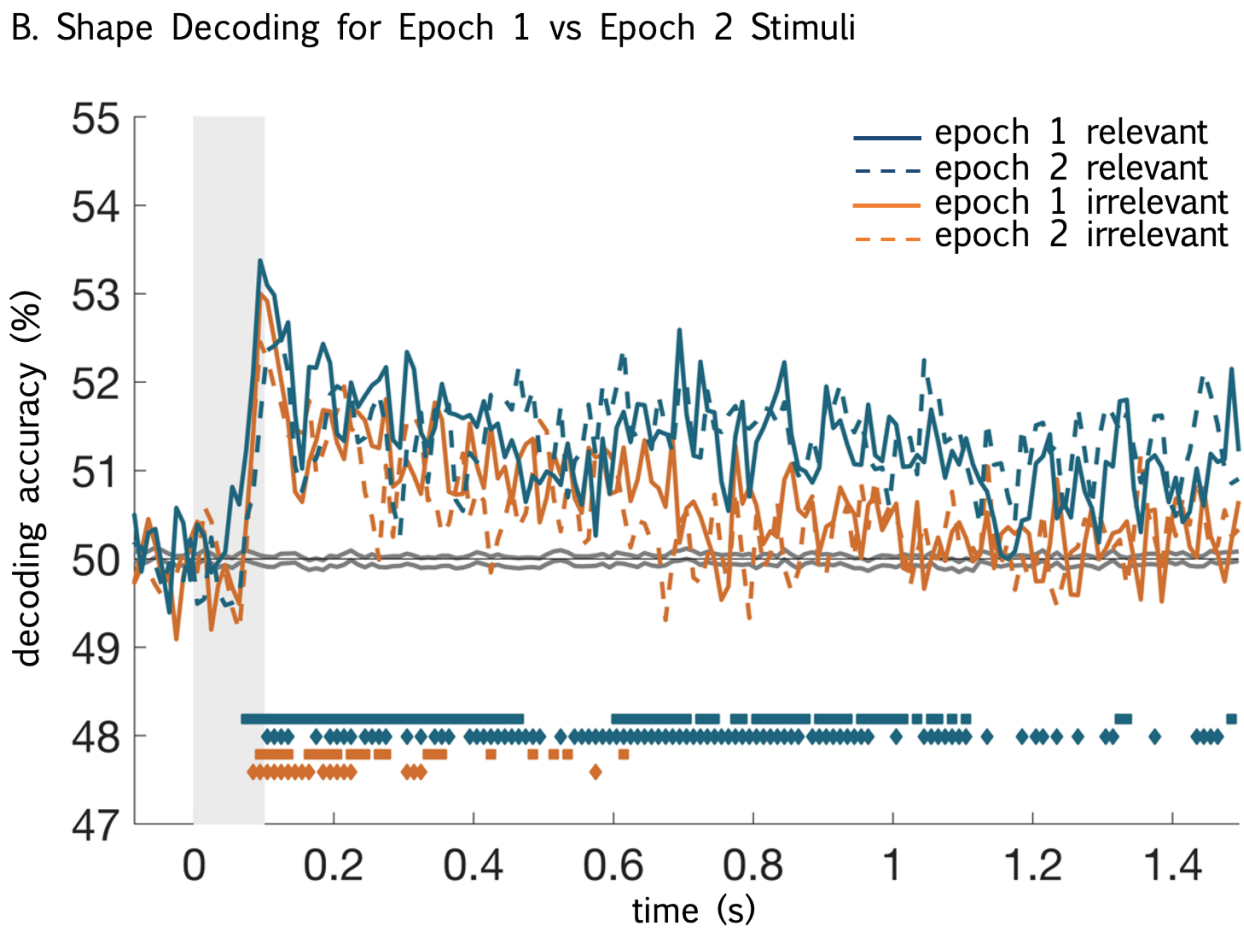


Figure 7. Colour (A) and shape (B) decoding for both epochs superimposed. Blue and orange colour indicate relevant and irrelevant features, and solid and dotted lines indicate epoch 1 and 2, respectively. For each trace, timepoints that reliably differ from chance are marked with coloured squares (epoch 1) or diamonds (epoch 2). There was no reliable difference between epochs, or interaction between epoch and relevance.

2.4. Discussion

Understanding how task-sensitive neural codes reconfigure is a key step in tracing how the brain supports adaptive behaviour. Here, I conducted two experiments to ask whether the brain can rapidly reconfigure neural codes for relevant stimulus features when what is relevant changes. In both experiments, participants judged the shape, then colour, or vice versa, of two targets presented in sequence. When shape and colour judgements were easy (Experiment 1), I observed strong coding of all object information. I found no reliable evidence for preferential

coding of task-relevant features. By contrast, when the shape and colour judgements were difficult and additional distractors were present (Experiment 2) I did see preferential coding for the relevant feature. Crucially, stronger coding for the relevant feature occurred in both phases of the trial, even though participants were shifting attention between features mid-trial.

Tracing this process with MEG allowed us to see the temporal evolution of preferential coding in the human brain, showing with millisecond resolution how attention emerges and redirects. Even with this precise temporal detail, Experiment 2 demonstrated a remarkably similar timecourse for selection of relevant information for the first and second stimulus. We might expect that preferential encoding of the relevant feature in the second epoch would be slower and/or less selective than in the first. For example, a lag or reduction in selectivity could reflect residual attention to the feature that was relevant for the first epoch, or time taken to transition to selective encoding of the second feature. Instead, I did not find any evidence of slower or reduced selectivity in the second epoch, suggesting that, in this paradigm, reconfiguration was fast enough for the relevant feature of the second stimulus to be selected as efficiently as for the first. These findings indicate that, when adaptive coding is engaged, task-relevant information is preferentially coded with remarkable speed even as task demands change within single trials. This provides possible infrastructure for the fast, sub-trial switching of attentional sets necessary for goal-directed behaviour (Duncan, 2013).

Although participants successfully performed both tasks, Experiment 1 did not elicit reliably increased neural coding of the relevant stimulus. Curiously, both tasks showed strong and sustained representation of the rule ("attend colour, then shape"), even though only one task showed an effect of rule on stimulus coding. Current explanations of top-down control emphasise both maintaining task information and enhancing relevant stimulus information. For example, both rule and relevant stimulus information can typically be decoded from MD regions in human fMRI (Jackson et al., 2016; Woolgar, Afshar, et al., 2015; Woolgar,

Thompson, et al., 2011; Woolgar & Zopf, 2017) and from frontal cortex in non-human primate single-unit recordings (Everling et al., 2006; M. G. Stokes et al., 2013). Disrupting prefrontal function causes reduction in task-relevant information coding (Jackson et al., 2021), and incorrect rule or stimulus information coding predicts incorrect behavioural responses (Woolgar et al., 2019). Moreover, the structure of frontal stimulus information predicts subsequent occipital stimulus information as attentional selection of relevant features emerges (Goddard et al., 2021). In view of these findings, it is plausible that selection occurs through rule information that is maintained by domain-general regions, which in turn selectively enhance relevant stimulus information in both domain-general and task-specific regions. In contrast, in Experiment 1 I observed a dissociation: clear rule coding, but no evidence of enhanced coding of the relevant stimulus features, even though the rule defined which stimulus features to attend to. Rule decoding increased after the stimulus displays in both tasks, particularly in Experiment 2. These increases could reflect neural responses diverging as participants applied the feature rule to the stimuli, in a way that did not enhance coding of the relevant stimulus features to an extent that my methods could reliably detect. Conversely, increases in rule decoding could be related to a more general shift, such as the widespread reduction in cortical response variance at the onset of a stimulus (Churchland et al., 2010). This highlights the utility of tracing both attentional rule information and rule-related changes in stimulus information, to characterise the impact of the rule on attentional selection. As Experiment 1 shows, the presence of decodable attentional rules does not necessarily translate to preferential coding of relevant stimulus information.

There were several differences between the two experiments that may have contributed to the different results. Experiment 2 was more difficult: participants responded well above chance level in both tasks, but overall performance was lower in Experiment 2 even after intensive training on the task. In Experiment 1, stimuli were drawn from a set of four objects, with strongly differentiated colours and

shapes, and a single object was shown on each display. Because of this small stimulus set, on 25% of trials the objects on Display 1 and Display 2 were identical, making the task trivial. On the remaining trials, participants had to select differential information from each display to respond accurately. However, there was significantly less information on each display, and less confusability among colours and shapes, than in Experiment 2. Thus, responding to the relevant information could well engage different attentional mechanisms across the two tasks.

Increased selection with increased stimulus complexity is a common theme in many theories of attention. For example, behavioural data demonstrate that although participants can find and respond to targets more quickly in simple displays compared to complex displays, they are also more easily influenced by salient distractors (Lavie, 1995; Lavie & Tsal, 1994). Neuroimaging evidence also suggests that distractors are not processed as deeply when a task becomes more difficult: BOLD activity associated with a distractor stimulus category no longer differentiates repeating and unrepeating distractors when target visibility drops (Yi et al., 2004). Load theory (Lavie, 1995; Lavie et al., 2014), takes these findings to argue that selection is qualitatively different for simple and complex stimuli. In simple environments, perceptual capacity not spent on relevant information spills over to other stimuli. As complexity increases, through the number, similarity, or visibility of the stimuli, we voluntarily direct our fixed capacity toward relevant features and ignore salient distractors.

Load theory does not strictly specify that all features that fall within perceptual capacity limits are equally represented. Based on behavioural responses to distractors under low load, we might predict that relevant and irrelevant features in simple displays are equally encoded, so that preferential coding only occurs when we exceed our perceptual capacity. The differential findings in Experiments 1 and 2 could be consistent with this view, if Experiment 1 displays fell within most participants' perceptual capacity while Experiment 2 displays exceeded it. However,

neuroimaging data so far do not support the idea that we require complex displays to engage preferential coding. Indeed, multivariate analyses of fMRI data show that relevant feature coding in visual cortex (V1 and LOC) can be enhanced in simple displays, with this enhancement extending to frontoparietal cortex when stimulus discrimination is difficult (Jackson et al., 2016; Woolgar, Williams, et al., 2015). Recent sensor-space MEG data also show enhanced coding of the relevant stimulus category (objects or letters) even though the displays contained only two easily distinguishable objects (Grootswagers et al., 2021). Based on these previous results, we might predict that feature-selective attention produces a relative enhancement of relevant perceptual information in simple displays, even though both relevant and irrelevant information can be perceived and recalled. This raises an interesting question: if both simple and complex displays can elicit preferential coding (that we can detect with both fMRI and MEG), why is stimulus coding in Experiment 1 unaffected by relevance?

Theories focusing on the object-based nature of attention (Baldauf & Desimone, 2014; Z. Chen, 2012a) may offer a better explanation for why coding two features of a single object, as in Experiment 1, and coding two objects, as in Grootswagers et al. (2021), would follow different rules. Behavioural studies demonstrate that we can often report irrelevant features of a target object without any apparent performance cost, suggesting that all features of the object are processed in parallel before we chose specific elements to respond to (Chen, 2012; Duncan, 1984). Under this object-based account of attention, it is unsurprising that I did not observe different responses to the same visual feature when it was the relevant or irrelevant dimension of a target object. Rather, we should expect to see preferential coding of the target object over the distractor. We can see this in Goddard et al. (2021), in which a spatial attention effect emerges before coding of the relevant target feature outstrips all other traces. This same pattern is suggested by my secondary analyses, where brief main effects of spatial attention emerge before preferential coding of the relevant target feature (Figure 6, epoch 2 colour

and epoch 1 shape). However, object-based accounts struggle to account for the preferential coding of single dimensions of stimuli (for example, Jackson et al., 2016; Jackson & Woolgar, 2018), that I observed at later timepoints in Experiment 2.

Biased competition (Desimone & Duncan, 1995; Kastner et al., 1998; J. H. Reynolds et al., 1999) provides a possible unifying framework for the load-driven and object-based characteristics of attention. Similar to load theory, this account proposes that complex stimuli trigger attentional selection. Rather than appealing to a threshold for perceptual capacity, biased competition suggests that, as distinct representations of stimulus features in early visual cortex feed forward to shared neural populations in higher visual cortex, competition emerges for what feature will be represented at the higher level, forcing selection to occur (Desimone & Duncan, 1995; J. R. Reynolds et al., 2012; Scalf et al., 2013). Because integration co-occurs with broadening receptive fields, even spatially segregated shapes can project to the same neurons and compete for in-depth processing. In the present study, the two-object displays of difficult-to-discriminate stimuli in Experiment 2 might elicit more competition than the single-object displays in Experiment 1, creating the opportunity for selection, even within the target objects.

Importantly, Duncan (2006) integrates space-, object-, and feature-based attention under the biased competition framework, highlighting that competition drives selection across disparate forms of attention, which can operate independently or in concert. This broader perspective of attention as a family of processes implemented through biased competition has since been embraced by Kravitz and Behrmann (2011), who demonstrate that space-, object-, and feature-based attention can combine to enhance object processing. Combined effects of spatial and feature-based attention have also been observed in non-human primates' lateral intraparietal area (LIP; Ibos and Freedman, 2016). Goddard et al. (2021) similarly show multiplicative effects of spatial and feature-selective attention give rise to selective coding of only the relevant feature at the relevant location.

Using the same stimuli, I replicated this finding, showing that coding of the relevant feature at the relevant location is enhanced relative to the irrelevant feature at that location (Figure 5) and the relevant and irrelevant features of the distractor (secondary analyses, Figure 6).

From a broader perspective, each of these theories incorporates the suggestion that selection processes are not always engaged. This selection-free zone could be a basic perceptual capacity; object binding that lets us process multiple features of the same object in parallel; or differences in location, colour, and orientation allowing for simultaneous processing without competition. Neural network simulations additionally offer some insight into the cost of selection, showing that strong coding of currently relevant task features induces slow reconfiguration to code subsequently relevant information (Musslick et al., 2018). Therefore, there may be a computational benefit to avoiding re-configuration of attentional sets (e.g., within trials) where possible. An adaptive system may be characterised not only by the ability to flexibly prioritise processing of currently-relevant information, but the flexibility to only do so when processing demands require it.

Here I have shown that human adaptive population codes can reconfigure within a single trial. This supports current theory, which emphasises the potential of focusing on each step in a task to produce complex and creative behaviour. Surprisingly, where attention effects were seen, the dynamics were comparable for between trial and within-trial shifts of attentional focus. This provides a potential neural substrate for the rapid creation of attentional episodes in multi-part tasks. However, significant effects of attention were only obtained in a demanding version of the task. Although many factors differed between the experiments, the difference could reflect the inherent cost of reconfiguring attention, meaning that it is not always an optimal strategy to engage. Future work will be important to identify what conditions push us toward preferentially coding the relevant information. Spatio-temporally resolved methods, such as source reconstructed MEG or MEG-

fMRI fusion (Cichy et al., 2016; Moerel et al., 2021; Mohsenzadeh et al., 2019), paired with systematic manipulation of task difficulty, could further elucidate how domain-general and task-specific brain regions interact to select relevant information under varying task demands. Rapid stimulus streams or self-directed attention shifting could further probe how rapidly the brain can reconfigure neural codes for preferential processing. Furthermore, relating the speed of reconfiguration to measures of fluid ability could clarify the functional importance of adaptive coding timescales. Together with my findings, this will offer rich insight into the biological bases of a mind that adapts to connect our goals with the world around us.

Chapter 3

Adaptive Coding for Visual Attention: Reconfiguring Top-Down Bias

Flexibility arises across the brain in many different forms. In Chapter 2, we saw that flexible coding of task-relevant features in a sequence, previously measured in monkey lateral PFC, can be observed in humans with sensor-space MEG. Many theories propose or assume that selective attention to stimulus features relies on top-down control from the PFC. More specifically, the attentional episodes theory proposes that highly dynamic coding across the MDN drives moment-by-moment focus on what is currently relevant. In this chapter, I resolve the data from Chapter 2's Experiment 2 into source space, to understand how ventral visual cortex and the MDN contribute to prioritising relevant stimulus features throughout a multi-step task. Decoding within source-reconstructed ROIs shows that coding of the behaviourally-relevant stimulus feature in the MDN co-occurs with delay period maintenance of that information in ventral visual cortex. Information flow analysis shows that colour and shape information are fed back from the MDN as the working memory delay begins. Models of graded stimulus representations poorly describe the MDN's representational structure, relative to ventral visual cortex, suggesting that its role is to prioritise relevant information rather than veridically reflect perceptual features. Curiously, stimulus information coding for a second task step, seen across Chapter 2's sensor-space decoding, was present in ventral visual cortex but not the MDN. I suggest some ways to more precisely probe what information in the MDN is or is not used to orient perception towards the problem at hand.

3.1. Introduction

Flexible cognition is tightly linked to domain-general cortex. Frontoparietal networks connect with sub-cortical and cingulate systems to code our current goal, task structure, and the value of switching focus (Arulpragasam et al., 2018; Farooqui & Manly, 2019; Sayalı & Badre, 2018; Wang et al., 2021; Waskom et al., 2014; Wen et al., 2020). Among these adaptive systems, the frontoparietal control network, or "multiple-demand" network (MDN), stands out for its extreme flexibility; neurons within the MDN adapt to code what is currently relevant across a wide range of tasks (Duncan, 2010; Erez et al., 2020; Fedorenko et al., 2013; Jackson & Woolgar, 2018; Kadohisa et al., 2013; M. G. Stokes et al., 2013), even as what is relevant changes sub-trial (Rao et al., 1997). At the same time, brain activity in visual regions adapts to the task. Local field potentials in non-human primate extrastriate visual cortex synchronise under working memory demand, driving an increase in alpha-beta band power that predicts task performance (Bahmani et al., 2018). Functional magnetic resonance imaging (fMRI) of human primary visual cortex shows that task-relevant stimuli are coded in highly-discriminable patterns, relative to task-irrelevant stimuli (Jackson et al., 2016). Object-selective fusiform cortex adapts to code current and prospective targets in anti-correlated patterns, showing sensitivity to both task-relevance and task structure (van Loon et al., 2018). Thus, both the MDN and visual cortex can preferentially encode what is relevant for the current task.

Many theories propose or assume that adaptive visual attention emerges through goal-oriented frontoparietal codes biasing sensory representations within the visual system (D'Esposito, 2007; Duncan, 2010; Erez & Duncan, 2015; Humphreys et al., 1998; Li et al., 2007; Lorenc et al., 2018; Miller & D'Esposito, 2005; Scalf et al., 2013). Evidence from neuroimaging supports this idea. Both the MDN and the lateral occipital complex preferentially code relevant visual features within a task (Jackson et al., 2016), with source-reconstructed MEG showing that

this preference for behaviourally relevant stimuli emerges in synchrony in frontal and occipital ROIs (Goddard et al., 2021). Detailed explorations of MD and visual cortex interactions suggest that their association is non-trivial. Task-sensitive connectivity within the ventral visual stream increases with frontoparietal engagement (Hwang et al., 2018), plausibly reflecting top-down communication of task goals. Cooling non-human primate dorsolateral PFC elicits a drop in inferotemporal cortex delay activity (Fuster et al., 1985), further reinforcing the idea that adaptive coding in the MDN supports the visual system to maintain and prioritise relevant stimulus features.

However, claims about what brain region drives another's function are plagued by worries that the connections we interpret as feedforward or feedback are just echoes of what is really driving information flow. The brain is intricately connected, so that when two regions code the same features, it is difficult to say whether they are critically involved in shaping each other's representational structure, or whether we are reading out a pattern that has emerged independently in both regions – perhaps through an intermediate source. While the MDN and visual cortex simultaneously code what is relevant, these data do not speak to whether those regions depend on each other. And, while MDN engagement predicts ventral stream adaptation, we do not know what information these regions share.

To complicate the issue further, our ideas about information flow encompass not just what information from frontoparietal cortex might influence visual representations, but when. Theories of visual perception go back and forth about what selective responses are embedded in visual cortex, or fed back to visual cortex later in perception. For example, load theory assumes that frontoparietal cortex drives early, perceptual capacity limits (Kelley & Lavie, 2011). By contrast, a biased-competition reframing of load theory proposes that early competition between stimuli emerges first within visual cortex, with top-down bias from frontoparietal cortex coming into play in later stages of selective (Scalf et al., 2013). Precise timing is important if we want to put these theories to the test.

A new method of information-based connectivity, called information flow analysis (IFA; Goddard et al., 2016), offers a possible solution. This approach combines three methods to enable rich insight into communication through the brain. First, it uses time-resolved neuroimaging (magnetoencephalography, MEG; or electroencephalography, EEG) to trace how representations transform on a millisecond timescale. Data from sensors or electrodes can be combined with structural MRIs and current flow models to estimate how cortical regions of interest (ROIs) gave rise to the time-resolved data.

Second, it uses Granger "causality" to test dependence between two brain regions over time. Granger analysis asks whether a source variable (such as the MDN) predicts something about the future state of a target variable (such as visual cortex) that the history of the target variable cannot account for. This will not tell us whether the MDN is necessary for some activity in visual cortex, or whether the structure they share is critical for behaviour. However, it is an accessible way to move past measuring correlations alone, and to test long-standing assumptions about the cascade of processes that make up visual perception.

Third, IFA combines Granger causality with information-based analyses of brain data. It conceptualises what each region is doing through a representational dissimilarity matrix (RDM): a correlation matrix that maps how a brain region differentiates visual or task features. Clusters in the matrix can reflect a region's organisational principles, for example, that it divides stimuli along an animate/inanimate boundary (Kriegeskorte, Mur, Ruff, et al., 2008). In an information flow framework, we can then ask when that representational structure is passed from region to region, mapping how information is transformed from perception to action.

Three key studies have used information flow analysis to extract detailed information about visual processing from non-invasive neuroimaging data. The first used source-reconstructed MEG and a large matrix of distances (1-Pearson's correlation) between visual stimulus conditions. Within the ventral visual stream,

they observed reliable feedforward information flow from around 70 ms post-stimulus onset, from V1-3 to V4/lateral occipital complex, and from V4/lateral occipital complex to inferotemporal and parahippocampal cortex (Kietzmann et al., 2019). This feedforward sweep co-occurs with feedback shortly after, around 110 ms for V4/lateral occipital complex to V1-3, and from 140 ms for inferotemporal/parahippocampal cortex to V4/lateral occipital complex (Kietzmann et al., 2019). A second study, again using source-reconstructed MEG, formed their dissimilarity matrices from pairwise stimulus decoding in each of four attention conditions, so that the matrices retained information about both stimuli and task. They found that frontal to occipital feedback begins to dominate information flow at the same time that the strongest period preferential coding of relevant stimulus information begins in the occipital ROI (Goddard et al., 2021). The third study used peri-frontal and peri-occipital electrodes in sensor-space EEG, with dissimilarity matrices (1-Spearman's correlation) based on visual features as in Kietzmann et al., but with a separate matrix for each level of stimulus coherence to capture how the balance of feedforward and feedback information flow changes with perceptual difficulty. They showed that feedback from peri-frontal to peri-occipital sensors is stronger for low coherence relative to high coherence face stimuli (Karimi-Rouzbahani et al., 2021), suggesting that the extent to which information is fed back could adapt to task demand.

Discussions of feedforward and feedback information flow tend to focus on a single event, but we know that attending to multiple things in sequence brings up challenges and behavioural tendencies that we do not see in isolated events. This could have knock-on effects for information flow. For instance, feedback information flow could be fast with more preparation but slow when we shift attention. Adapting functional connectivity to meet new task demands, and changing information conveyed between brain regions, could take time and effort to make sure the transition is smooth. Conversely, communication between domain-general regions and the ventral visual stream could rapidly reconfigure, with behavioural limits

that we see during task-switching reflecting bottlenecks for planning and action rather than for attention.

The previous chapter showed that coding of task-relevant information can be prioritised quickly within a task. Here, I resolve those data into source space to probe how frontal and visual brain regions contributed to those effects. I build on the findings of Chapter 2 to ask what part top-down information flow plays in preferential coding of relevant information, as what is relevant changes.

3.2. Methods

3.2.1. Participants and task

I used data previously described in Chapter 2 under Experiment 2. I did not include data from Chapter 2's Experiment 1, as those subjects had no structural MR scans. Magnetoencephalography (MEG) data were acquired from 20 right-handed subjects (16 female, 4 male, mean age 31±12 years). All participants had normal or corrected-to-normal visual acuity and normal colour vision, had no history of neurological damage or current psychoactive medication, and were MEG-compatible. One subject was excluded from re-analysis because they had no structural MRI data, leaving n=19 (age=31±11, 15 female, 4 male). Participants were recruited from the volunteer panel at the MRC Cognition and Brain Sciences Unit (Cambridge) and gave written informed consent in each session (training, MEG, MRI). Ethical approval was obtained from the Psychology Research Ethics Committee at the University of Cambridge (PRE.2018.101).

During training, participants were asked to view novel spiky objects (Figure 1) and classify them by colour or shape. Colour varied over four steps from red to green. Shape varied over four steps from upright (more spikes pointing vertically) to flat (more spikes pointing horizontally). Participants completed 100 training trials on each dimension and received feedback on each trial. They then progressed to training on the task.

In the task, participants saw two displays in sequence on each trial. Each display held two of the novel spiky objects, one to the left and one to the right. Participants were instructed to attend to one feature on each display, for example, the colour of the left object on Display One and the shape of the right object on Display Two. They then selected a symbol at a choice display that represented the combination of the two target features. For example, a red X represented the combination of red and upright-oriented spikes. Figure 1 shows an example trial with the choice display options. Participants trained to 80% accuracy on the task before moving to the MEG. Once in the MEG scanner, participants completed four blocks of 256 trials each. Instructions were blocked. They spanned all combinations of left, right, colour, and shape across blocks, always switching location and feature between displays. Block order was balanced across participants.

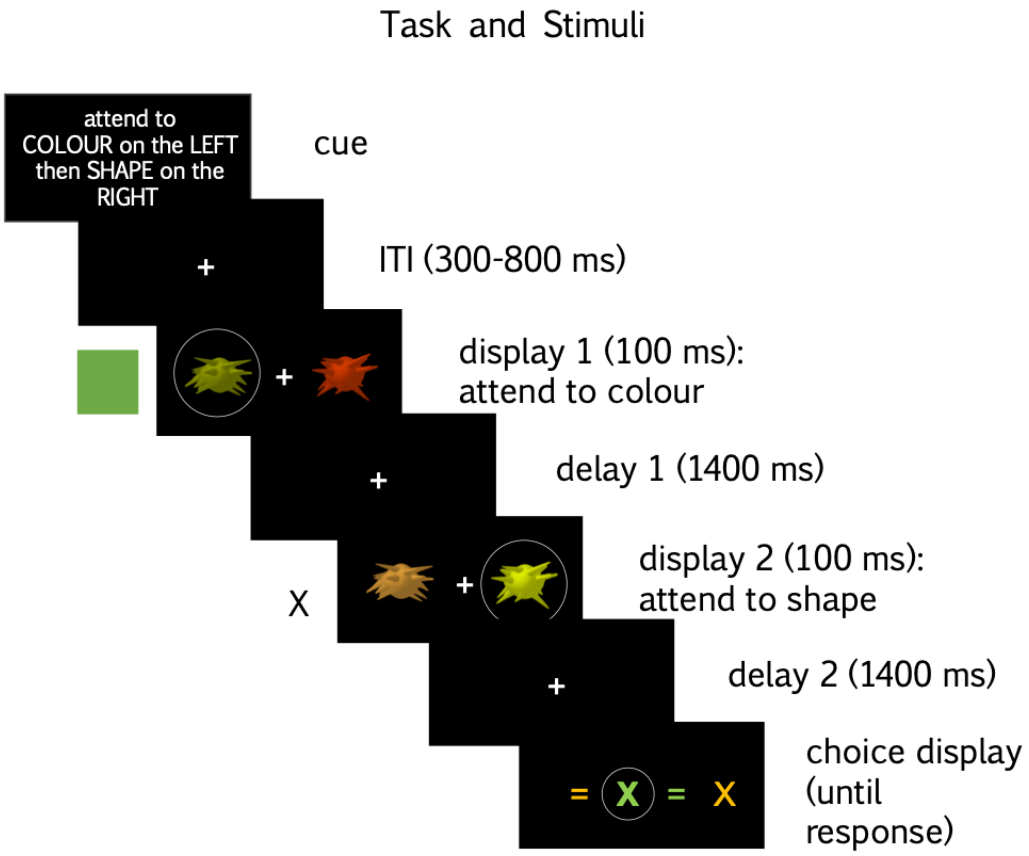


Figure 1. An example trial.

3.2.2 Data acquisition

Data were acquired with a 306-channel (204 planar gradiometer, 102 magnetometer) Elekta Neuromag Vectorview system in upright position. Acquisition details are described in Chapter 2, section 2.2.5.2. Preprocessing remained largely the same as for Chapter 2's sensor-space analysis. I first spatiotemporally filtered the data with temporal signal space separation to remove recording artefacts, using Neuromag's proprietary Maxfilter software (*Maxfilter*, 2010). I used the same software to compensate for motion, based on head position coil values, to re-orient each block the subject's initial head position, and to remove high-frequency artefacts that are caused by the head position coils. I then bandpass filtered from .1-45 Hz, and down-sampled to 100 Hz (Fieldtrip software; Oostenveld et al., 2011). Trials were epoched from zero to five seconds around the onset of the first display. In contrast to Chapter 2, I applied further preprocessing steps to more strictly clean the functional MEG data before source reconstruction. I baseline corrected the trials with the mean of the 100 ms immediately prior to stimulus onset.

I then used the Autoreject toolbox (Jas et al., 2017) to repair or reject bad trials. This automated process finds bad data segments by flagging, for each trial and sensor, where the peak-to-peak amplitude change exceeds a pre-set threshold. Trials with bad sensor data are repaired through interpolation, or rejected if too many sensors are unreliable. Both the peak-to-peak amplitude threshold, and the number of bad sensors that can be interpolated without rejecting the trial, are then iteratively adjusted until the split-half reliability of the repaired dataset is maximised. I applied this process separately to gradiometer and magnetometer data to allow different thresholds for peak-to-peak amplitude, as the two sensor types can differ in scale by an order of magnitude. After the parameters were adaptively set, gradiometer and magnetometer thresholds averaged .4 across the group, with the maximum number of bad sensors interpolated averaging 32 for gradiometers and one for magnetometers (median values). The number of rejected trials remained

below 10% across the group (mean: 58.36, median: 10). Three of the 19 participants lost more than 10% of their trials (178, 261, and 374 trials rejected out of 1024 total). I nevertheless chose to include these datasets in the analysis, as the sample was limited and no other measures suggested that these datasets were outliers.

The impact of including unbalanced datasets is difficult to anticipate, but I have detailed some of the risks below, along with information on the rejected trials. Trials included multiple colours and shapes, so that excluding a given trial should not specifically skew the dataset towards representing, for example, neural responses to red items. Excluding trials within a certain attention condition could disadvantage it relative to the others, which could have consequences for measuring attentional influences on colour and shape. For example, if a large proportion of the rejected trials fell within an "attend colour, then shape" condition, this could undermine a true advantage for attended colour information in Epoch 1. For attended location, 534 "attend left, then right" trials and 575 "attend right, then left" trials were rejected across the group. For attended feature, 484 "attend colour, then shape" trials and 625 "attend shape, then colour" trials were rejected across the group.

T1-weighted MPRAGE structural scans were also acquired (slice thickness 1.0 mm, resolution 1.0 x 1.0 mm). All but two subjects' scans were initially acquired for other projects at a scanning facility on-site, with acquisition dates ranging from 0 to 2 years prior to the MEG scan. These subjects (n=17) gave permission at the time of testing for their MR scan to be used in future studies. One subject did not have an MR scan prior to this study, but returned for an MR scan one year after their MEG session. They gave written informed consent and received £20 for their time. One subject's MRI was acquired for a project hosted elsewhere (though scanned at the same facility). I obtained explicit permission from them and from the external study's lead investigator to access the scan. Structural images were converted from proprietary DICOM format to NIFTI format, visually inspected

(MRIcron, https://github.com/neurolabusc/MRIcron), and reconstructed with Freesurfer's recon-all (*FreeSurfer*, https://surfer.nmr.mgh.harvard.edu/).

3.2.3. Source reconstruction

I reconstructed sources with minimum norm estimation (Hämäläinen & Ilmoniemi, 1994) implemented in MNE-Python (Gramfort et al., 2013). For an overview of the steps involved, together with subsequent analyses, see Figure 2.

3.2.3.1. Forward model

Boundary element model and source space. I estimated volumetric conduction from structural scans, using a single-shell boundary element model based on the inner skull. I replaced the inner skull boundary extracted by FreeSurfer with a Fieldtrip alternative, as collaborators have previously found that the Fieldtrip extracts this information more accurately. The source space consisted of 20,484 points estimated by FreeSurfer as sitting along the boundary of grey and white matter. I defined sources as surface normals with loose orientation.

MEG-MRI co-registration. Structural and functional data were co-registered using MNE-Python's interactive viewer, mne_analyze. For each subject, I loaded a MR surface file along with digitiser data collected during the MEG acquisition session. I manually mapped digitised fiducials to corresponding head points on the MR skull surface. I then iteratively adjusted the co-registration with an automated process that optimised the fit of MEG digitiser points to the MR skull surface.

3.2.3.2. Inverse model

I calculated the sensor noise covariance using baseline data from -100 to 0 ms relative to the first stimulus onset, and regularised the covariance matrix with the Ledoit-Wolf method to moderate extreme values. I estimated source contributions using minimum norm estimation, which selects the source-to-sensor mapping that

minimises the "norm", or sum of squared values, of the sources contributing to a sensor.

Lastly, I took the surface normal of the three source dipoles at each vertex to produce a single activation value at each vertex. This approach is sensitive to meaningful negative values, such that current flow toward the cortical surface on one side of a gyrus registers as current flow away from the surface on the gyrus's opposing side.

3.2.3.3. Parcellation and regions of interest

I used cortical parcellations defined in the Glasser atlas (Glasser et al., 2016). To do this, we took Glasser atlas annotation files available through the Human Connectome Workbench (Van Essen et al., 2013) and converted them to a FreeSurfer-compatible format. I mapped these FreeSurfer-compatible annotation files to each individual's structural data with spherical averaging, allowing us to define regions of interest in individual data.

I defined two regions of interest: one ventral visual, and one frontoparietal. The ventral visual ROI was based on Kietzmann et al. (2019), and spanned early visual cortex, V4-LOC, and ventral temporal cortex (Glasser areas V1, V2, V3, V4t, LO1, LO2, LO3, TE1p, TE2p, FFC, VVC, VMV3, VMV2, PHA3, PHA2, and PHA1). The frontoparietal ROI was based on Assem et al.'s (2020) "core" multiple-demand network, defined by a large-scale, multi-model parcellation (Glasser areas i6-8, p9-46v, 8C, IFJp, a9-46v, AVI, IP1, IP2, PFm, SCEF, and 8BM). For each individual, I removed any vertices that fell within two cortical areas and so could not be unambiguously assigned to one area.

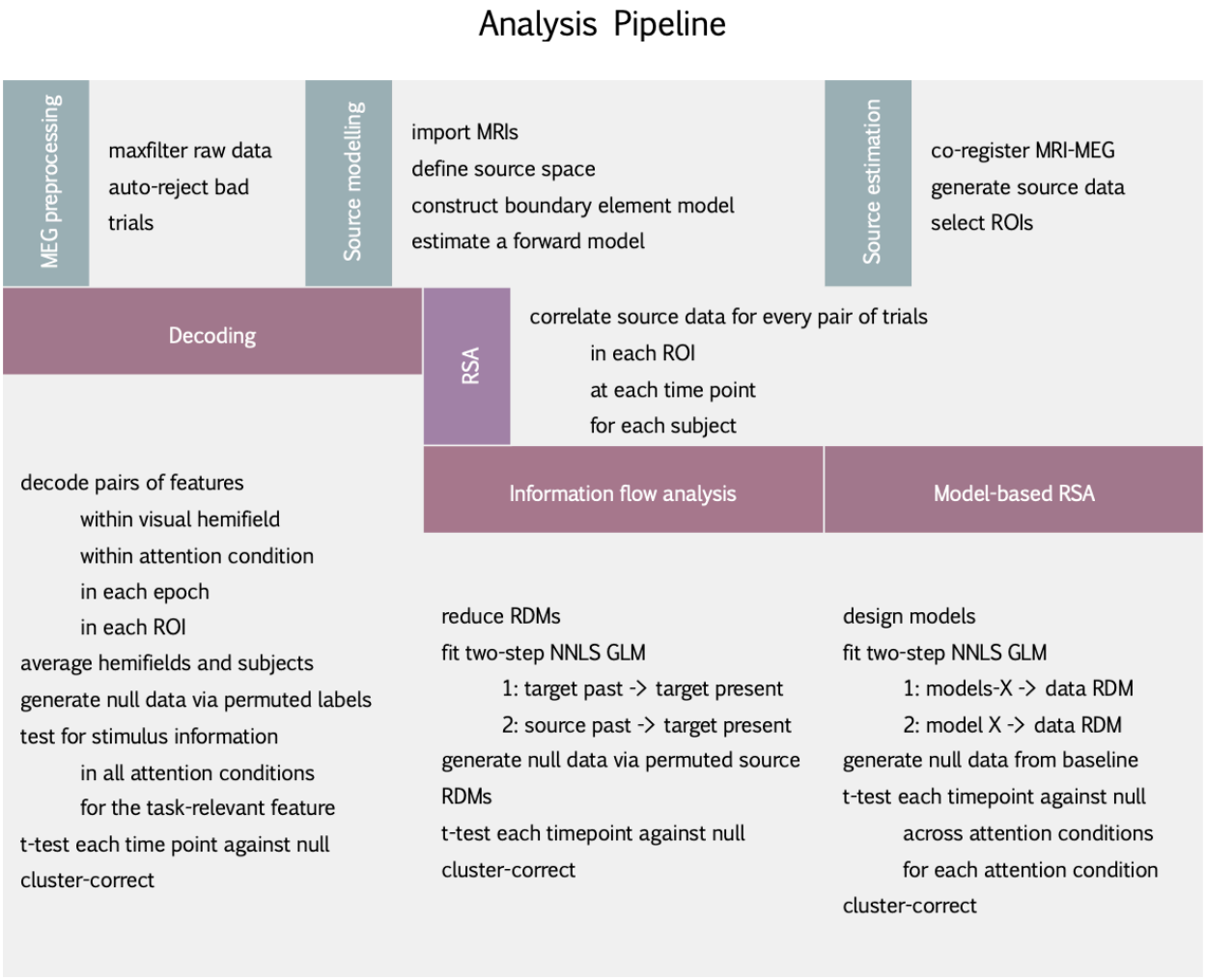


Figure 2. Visual depiction of the analysis pipeline. Blue headings reflect steps that supported the source estimation, lavender the RDM creation, and mauve the analyses that I used to estimate task feature information within (decoding and model-based RSA) or between (IFA) ventral visual and MD regions.

3.2.4. Decoding analysis

Next, I conducted a decoding analysis with the source reconstructed data. This served two purposes. First, I wanted to validate the source reconstruction by confirming that the source reconstructed data held information about the task, and that the two regions of interest captured distinct components of the visual feature information. Second, I wanted to use the opportunity that source reconstruction presents to observe how visual information unfolds over time in the visual system, compared to the domain-general MDN.

I followed the same steps as in Experiment 2, Chapter 2. I decoded the attended feature (colour first or shape first) to extract a measure of rule information across the trial. For Epoch 1, I decoded each pair of colours and shapes within each visual hemifield and run. I then averaged colour decoding traces that represented target colour (left colour when attending to the left) when it was the attended feature and unattended feature, doing the same for distractor colour. I generated the same four traces for shape, and repeated the process for Epoch 2. Classification parameters were matched to Chapter 2.

The first aim of this analysis was to investigate when stimulus information emerged in each ROI. I plotted each stimulus decoding trace, and compared it to zero at every time point. As in Chapter 2, I generated permutation-based null data by taking the class labels that the classifier gave at each trial and timepoint, shuffling them across trials, calculating the decoding accuracy, and repeating this until I had obtained 10,000 null traces. Together, these individual-level null data formed a group-level null distribution by sampling one trace from each participant, averaging across participants to create a group-level null trace, and repeating the process until there were 10,000 group-level null traces. I compared the true decoding trace to the null data with a one-sided t-test at each time point. Again following Chapter 2, I identified clusters with a threshold-free cluster statistic (Stelzer et al., 2013) implemented in the CoSMoMVPA toolbox. Threshold-free cluster enhancement adaptively finds a threshold that best separates clusters (in this case, peaks in beta weights), and is useful for its ability to identify effects that are low and sustained or large and transient within the same timecourse. Then, I corrected for multiple comparisons at the cluster level.

As a second step, I tested whether the task-relevant trace (target shape when reporting shape) differed from the average of all other traces. This would allow me to see to what extent the preferential coding that I observed in Chapter 2 was present in ventral visual cortex and the MDN over the course of the trial. I implemented this with a one-way ANOVA on the difference score (task-relevant -

task-irrelevant decoding accuracy) at each time point, using a normal distribution for the null, and identified above-chance clusters with the procedure described in Chapter 2.

3.2.5. Representational dissimilarity matrices

In order to analyse the information flow between regions of interest, I transformed the data into "representational dissimilarity matrices" (RDMs; Kriegeskorte, Mur, & Bandettini, 2008). RDMs represent trial data in terms of the relationships between conditions. We can measure the "distance" (for example, 1-correlation; or cross-validated distance such as decoding accuracy) between the brain's response for each pair of conditions, and place all these distance values into a conditions x conditions matrix. This allows us to align data across participants and regions in a shared representational space; to test our intuitions of how experimental conditions influence the brain's response; and to trace how distances between conditions develop over time and space.

For each epoch separately, I sorted trials by attended location, attended feature, target colour, distractor colour, target shape, and distractor shape. Each trial uniquely combined these attention conditions and visual features. For each person and ROI, at each time point, I calculated the dissimilarity in pattern of activation over vertices (1-Pearson's correlation) between a pair of trials. I repeated this for every pair of trials, building up a trials x trials RDM. Across timepoints, these RDMs formed a movie that reflected the change in representational structure over time. I created time-resolved RDM movies for each participant and ROI.

These full, trial by trial RDMs retained information about where and what people were attending to, and what stimulus features were present. They could also carry information specific to each participant's experimental session, such as the trial sequence or that participant's focus and fatigue. These features would not be captured when the data RDMs are compared to theoretical models (below) but could potentially contribute to the analysis of information flow where one data RDM is

compared to another. To avoid this, to assess information flow I reduced the full RDMs to smaller matrices that separately represented colour or shape. For each colour pair (for example, red target and orange distractor), I took the dissimilarity between trials with that colour pair and trials for each other colour pair (red target, red distractor; red target, yellow distractor; and so on). The average dissimilarity between two colour pairs (i.e., how dissimilar neural responses were to red-red vs green-green displays) became a single cell in the reduced RDM. In this way, I built a symmetrical 16 (4 target * 4 distractor colours in Epoch 1) by 16 RDM, then repeated this for Epoch 2, and for shape in each epoch.

Representational dissimilarity matrices in other studies are often larger than 16 by 16. For context, Kriegeskorte et al. (2008) used a set of 92 objects to construct a 92 by 92 RDM in fMRI data. The same stimulus set and RDM dimensions have been used since then in MEG and behavioural studies (Carlson et al., 2013; Cichy et al., 2014; Mur et al., 2013). Grootswagers, Robinson, and Carlson (2019) used a larger set of items to generate a 200 by 200 RDM in EEG data. The RDMs in the current study are nearer to those reported in Moerel, Rich, and Woolgar (2021), who also reduced their full trial set to a 16 by 16 matrix to represent task-relevant and task-irrelevant visual features. We necessarily lose some variance by averaging the full RDMs into smaller RDMs that represent a subset of the task features. On the other hand, we gain specificity about what drives information flow.

Condition balancing in the design ensured that there were equal proportions of each target and distractor colour pair, and the same for shape, at each display. However, I should highlight that because of the limit on the total number of trials, I was not able to perfectly balance how often a target and distractor colour pair in Epoch 1 was followed by each pair of colours at Epoch 2. Thus, information flow could reflect some facets of both epochs.

To separate information flow in Epochs 1 and 2 more stringently, I also created matrices that represented the task features in a purely epoch-specific way. Target colour, distractor colour, target shape, and distractor shape were each

carefully balanced between Epoch 1 and Epoch 2 displays for each subject. I averaged across dissimilarities for a pair of target colours, repeating this for each pair to create 6 cells, each of which represented the average dissimilarity between two colours (red-orange, red-yellow, red-green, etc.). I did the same for distractor colours, target shapes, and distractor shapes to create a 6 by 4 matrix at each time point, for each subject, ROI, and epoch. Thus, each column was an independent variable – the colour or shape of the target or distractor objects – with its rows capturing how dissimilar the neural responses were to each pair of its levels. The matrices were therefore not true RDMs, as the columns and rows captured different elements of the task. By creating these unconventional matrices, I could extract separate time courses of task information for each epoch, and estimate how this epoch-specific information was fed forward and back between ventral visual and MD regions.

3.2.6. Information flow analysis

Next, I applied a Granger-causal analysis to the source-reconstructed data to investigate the information flow between regions. For this, I took one of the reduced RDM movies produced in the previous step within the ventral visual and core MD ROIs. I fit a two-step GLM at every time point for each subject, ROI, and feature (e.g., colour in Epoch 1), using non-negative least squares following Kietzmann et al (2019). First, for a given "target" ROI at time *t*, I identified the beta weight for the past representational structure within that same ROI at times *t*-120 to *t*-20ms. Then, I identified the beta weight for the remaining ROI (the "source") by fitting its past structure (time *t*-120 to *t*-20ms) to the residual variance in the target ROI (time *t*) that was not explained by the target ROI's past. If the source ROI's past explained something about the target ROI's present, over and above what could be explained by the target ROI's past, I inferred that some information was passed from the source to the target ROI. In the context of visual and frontoparietal regions of interest, this provided a window into when information was fed forward (visual to

frontoparietal) or fed back (frontoparietal to visual). I defined the past of the target and source ROIs as a window from t-120 to t-20 ms, following Kietzmann et al. (2019).

For each source-target pair (visual-MDN and MDN-visual), I compared the beta weight for the source ROI to zero at each timepoint, using an empirical null. I generated the null data by shuffling the rows and columns of the RDM for the past of the source ROI before fitting it to the residual variance in the target's present. I repeated this 100 times for each target time point and subject, then iteratively sampled from and averaged those individual subject nulls to generate 1000 null traces at the group level. I accounted for multiple comparisons over the time course with the threshold-free cluster statistic and cluster-based correction described in Chapter 2 and under the decoding analysis above.

I applied this process to each reduced RDM in turn: the 16 by 16 colour and shape RDMs oriented towards Epoch 1 and 2, and the 6 by 4 colour and shape matrices specific to each epoch. As mentioned above for representational similarity analysis, the small matrices are important to specifically measure whether colour, shape, Epoch 1, or Epoch 2 information is shared between regions. However, it is worth noting that small matrices could particularly handicap information flow analysis. We expect the variance within a region to be better explained by its own history than by another region. If there is very little variance to explain, as with a small matrix, there could be even less opportunity to see interplay between regions.

3.2.7. Model-based representational similarity analysis

Beyond knowing that information *is* exchanged between brain regions, we would ideally like to know *what* aspects of information are shared. Moreover, knowing that two conditions can be decoded gives us no information about the structure of representation within each condition. With this in mind, I conducted an exploratory supplementary analyses designed to unpack further what task

information is represented within, and therefore potentially communicated between, ventral visual cortex and the MD network.

I was interested in testing an approach to measuring the information represented within RDMs, with the hope that I could extend it to information flow analysis. To do this, I fit explicit models to the representational geometry of each region of interest. The purpose of these models is to formalise theory or intuition about what may influence our conditions, by predicting which conditions will elicit dissimilar brain responses. Because the models explain variance in representational space, they could potentially be included as a predictor in an RDM-based information flow analysis, allowing us to see what information is passed between regions. For example, we could compare how well MDN representational structure predicts later ventral visual representational structure when attended location information is or is not modelled out.

First, I fit stimulus and rule models to the data RDMs for each region of interest. Model-based representational similarity analyses are conceptually similar to decoding, in that they track the distance between some task features in multivariate space. However, decoding pairs of conditions can give us a limited view of the overall relationship between conditions. Fitting models to the pattern across condition pairs allows us to imagine and implement different explanations for how the conditions relate. For example, we might predict that stimuli will be represented according to their physical differences. Two red items could be similar, a red and green item dissimilar, with red-orange and red-yellow falling in between. We can test this by creating a conditions x conditions matrix, in which the dissimilarity is 0 for cells with the same colour, and increases in equal steps as colours become physically more distinct. We can compare this model to the data at each time point, and extract an estimate of how well colour information predicts the data as we move through the trial. We can fit multiple models together to test how task features (such as rule, colour, and shape information) uniquely explain the brain's response over time, scaling pairwise decoding comparisons to predict specific

patterns across conditions. Crucially, we can adapt the models to our hypotheses, whether we expect colour representations to be response-oriented ("red" response stimuli i.e. red and orange, dissimilar to "green" response stimuli i.e. yellow and green), same-different (red vs all other colours), or stepwise (dissimilarity increasing gradually from red to green; see Figure 3).

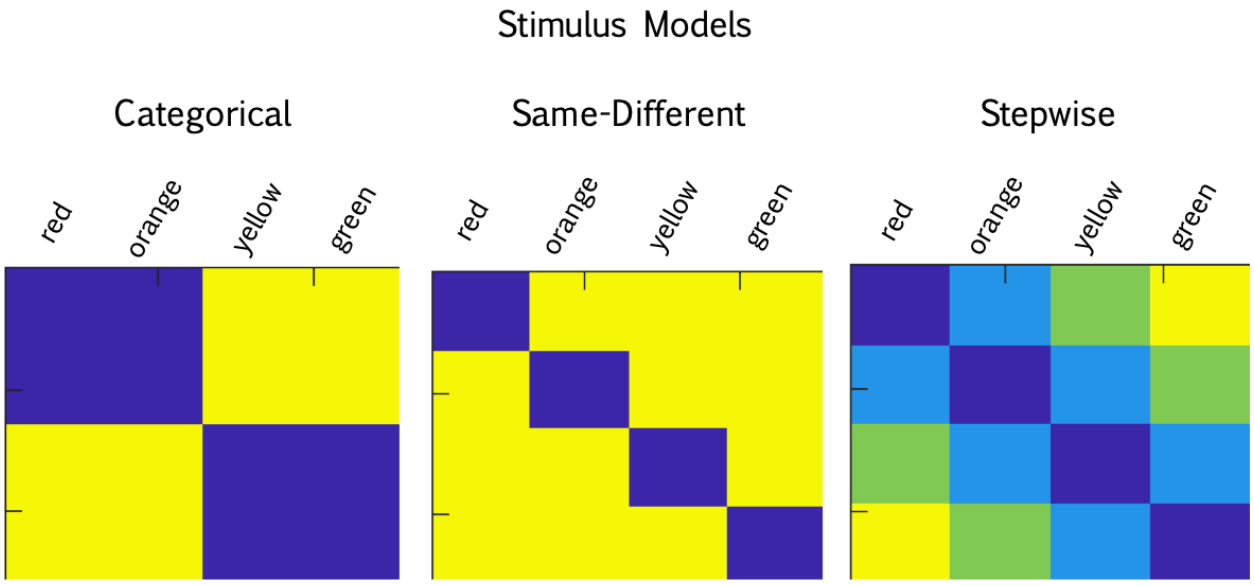


Figure 3. Example predictions about the representational structure of colour, expressed as dissimilarity models. Blue indicates minimum dissimilarity, yellow maximum dissimilarity. Conditions are grouped so that the top-left to bottom-right diagonal contains identical conditions. The categorical model predicts that four colours from red to green are grouped by whether they are more red or more green (for example, in a task where responses are based on colour category). The same-different model assigns minimum dissimilarity for the same colour, but maximum dissimilarity for all other colours, ignoring graded differences between colours. The stepwise model specifies that colours will differ according to their physical dissimilarity, with red nearer to orange than to yellow, and nearer to yellow than to green.

In this case, I created two core models. I predicted that dissimilarity in the neural response would reflect physical similarity in the colour and shape of the stimuli. To implement this, I set the values in the models to four equal steps between 0 and 1, corresponding to the difference in feature space between pairs of four colours (red, orange, yellow, green), or four shapes (upright, slightly upright, slightly flat, flat; as in Figure 3C). Thus, each model made the same prediction: that neural responses would be more dissimilar when stimuli were physically more

dissimilar. I fit the stimulus models to target colour, distractor colour, target shape, and distractor shape, for each epoch and attention condition separately (Figure 4).

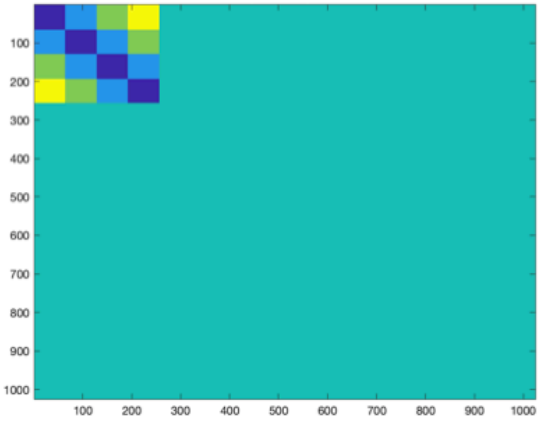
Since I wanted to compare the model fits between attention conditions, I did not predict anything about the representational dissimilarity of visual features between stimuli in different attention conditions (e.g., target shape vs distractor shape). However, I generated two secondary models that spanned attention conditions. These models represented my predictions for the effect of attended location and attended feature. Both models predicted low dissimilarity for the same attention condition (e.g., for all "attend left first" cells in the RDM) and high dissimilarity for opposing attention conditions (e.g., for all "attend left first" x "attend right first" cells in the RDM), represented as zeros and ones (Figure 5). These same-different models of attention map very closely to what I probed by decoding attended features, so I expected to see a similar timecourse.

For all models, cells without a prediction were left empty. Before fitting the models to the data, I converted the model values to z-scores and set empty cell values to zero.

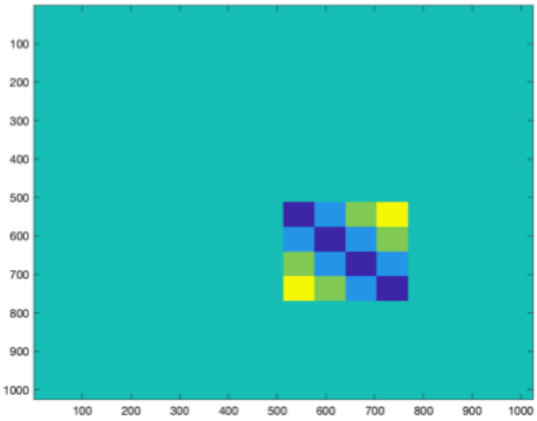
Stepwise Stimulus Models for Object Features

Target Colour, "Attend Colour"

object presented in left hemifield

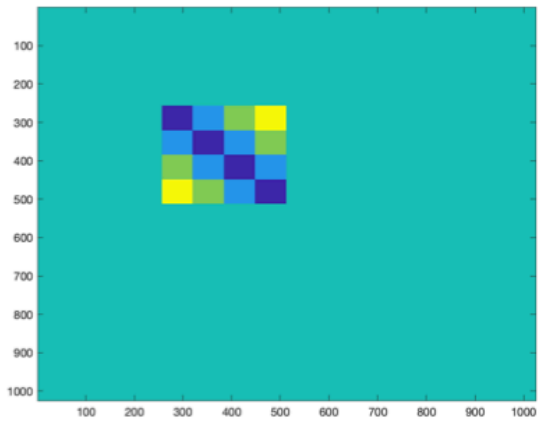


object presented in right hemifield

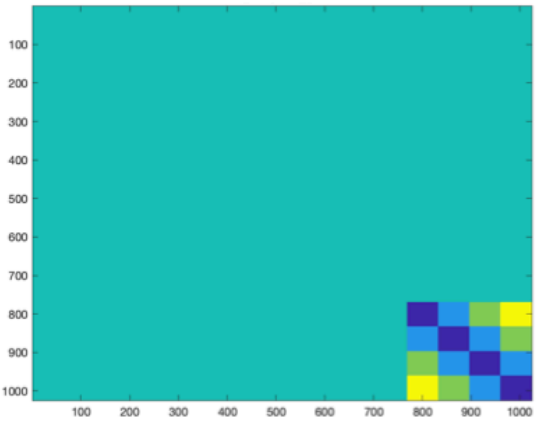


Target Colour, "Attend Shape"

object presented in left hemifield

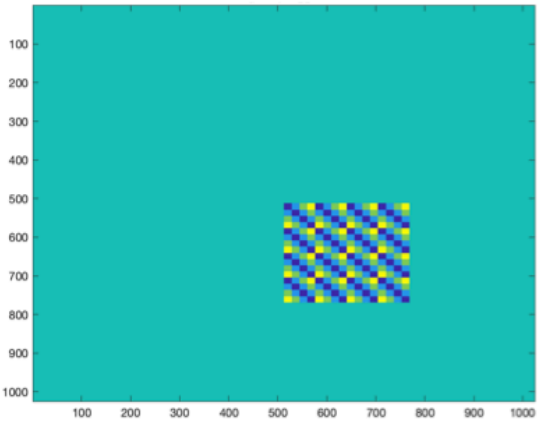


object presented in right hemifield



Distractor Colour, "Attend Colour"

object presented in left hemifield



object presented in right hemifield

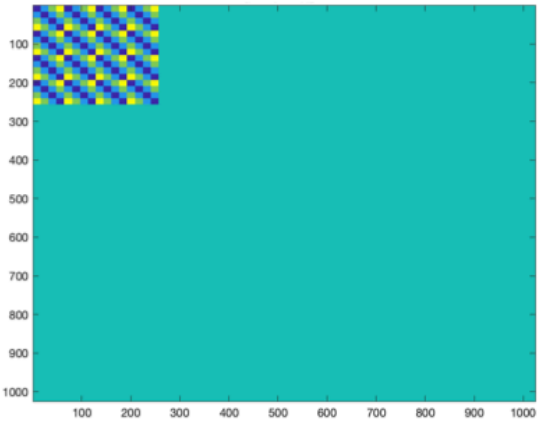


Figure 4. Stepwise models formatted for the data RDMs. RDMs were sorted by attended location (left or right) and attended feature (colour or shape) to form four combinations of attended location and feature along the diagonal. These represented the four experimental attention conditions. Within each attention condition, conditions were sorted by target colour, distractor colour, target shape, and distractor shape, with distractor colour embedded within target colour and so on. Thus, each quadrant along the diagonal could be described simultaneously in terms of its target and distractor colour and shape. I did not predict anything about how stimuli related across attention conditions. The first row shows stepwise models fit to target colour within the two attention conditions for which colour was relevant. The next row depicts stepwise models fit to target colour, in attention conditions for which colour was irrelevant. The third row depicts how stepwise models for distractor colour were tiled to describe the same trials captured by target colour models in the first row, but reversed to describe the opposing (task-irrelevant) object. Target and distractor shape models were further tiled within these four attention conditions.

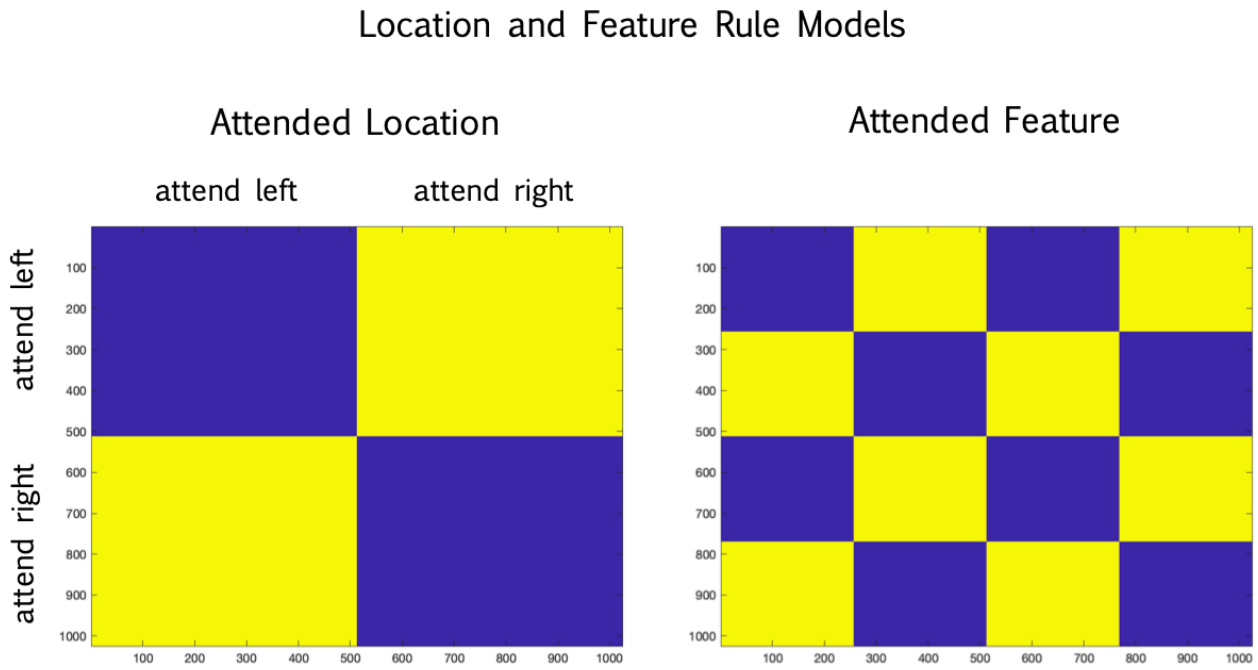


Figure 5. Attention models. The first panel predicts that conditions within the same attended location (top-right and lower-left quadrants, blue) were minimally dissimilar, and that conditions with different attended locations were maximally dissimilar (yellow). The second panel shows the same conceptual prediction, now predicting that conditions with the same attended feature (colour or shape) would be minimally dissimilar.

As with the information flow analysis, I used a non-negative least squares two-step GLM to fit the models to the data at each timepoint. Under this framework, we can fit multiple models, and so extract estimates of how colour and shape uniquely explain the representational structure over the course of a trial. I first fit all the models but one, along with a constant, to the data for a single

timepoint. I then fit the excluded model to the residual variance and estimated the beta weight, that is, how strongly the model predicted the data once we had accounted for the variance explained by other models. I repeated the two-step process until we had estimated beta weights for each model at this point in time. I then repeated this over time to extract traces of the colour and shape information over time.

Next, I applied inferential statistics to test whether, and when, the models reliably predicted the data at the group level. Because adding parameters to a GLM can often increase predictive power, I judged the beta weights during the trials relative to the beta weights during a pre-trial baseline (-100 to 0 ms). I permuted the baseline data to form an empirical null distribution that represented how well the models could predict data if those data had no true stimulus effects. For the first timepoint in the trial, I randomly selected one baseline timepoint for each subject. I did this 1000 times to build a distribution of null values. I then repeated the process over time, building a new null distribution from the baseline data at each timepoint within the trial. Under the null hypothesis, there would be no true stimulus effects throughout the trial, and the models would explain the trial data no better than they explained the baseline data. Thus, if the null hypothesis were true, we would expect the beta values during the trial to fall within the distribution of null values. Following this logic, I compared the model fits to this distribution at each time point. I again corrected for multiple comparisons over time points using the non-parametric threshold-free cluster test (Stelzer et al., 2013) described in Chapter 2, implementing this process in CoSMoMVPA.

I ran this model fitting three times: once with models describing the visual features in the first display, a second time with the same models mapped to the visual features of the second display, and a third time with models describing the attended location and attended feature. In all cases, I fit the models to each time point in the trial.

Across both ROIs and task epochs, I expected that the dissimilarity between visual features would explain dissimilarity in the brain's response. Beyond this, I expected that the stimulus-driven models would fit the data better when they described the task-relevant feature, indicating that the brain responds more dissimilarly to features that we are actively trying to discriminate.

3.3. Results

3.3.1. Decoding

3.3.1.1. Ventral visual ROI

In Epoch 1, colour information coding peaked rapidly after stimulus onset in all attention conditions (Figure 6). Shortly after, around 750 ms, information coding re-emerged for colour when it was task-relevant, while decoding traces for task-irrelevant colour information returned to near chance. This patterned echoed the pattern that I observed in the sensor-space decoding of Chapter 2, with an early stimulus-driven followed by preferential coding of the relevant feature. In Epoch 2, colour information did not visibly or statistically increase for all attention conditions, but only for the target colour when attending to shape. The task-relevant feature visibly exceeded chance during the post-stimulus delay, but this was not statistically reliable.

Shape information coding similarly peaked rapidly after stimulus onset and then dropped, though here I did not see a reliable second peak for the task-relevant information (Figure 7). This was repeated in Epoch 2. Shape information coding for three of the four attention conditions reliably exceeded chance shortly after the onset of the second display. The task-relevant shape information from Display 2 was coded above chance for longer than the other attention conditions, but was not statistically different to them.

Attended feature decoding (attend colour vs attend shape) was strong and sustained within each task epoch (Figure 8).

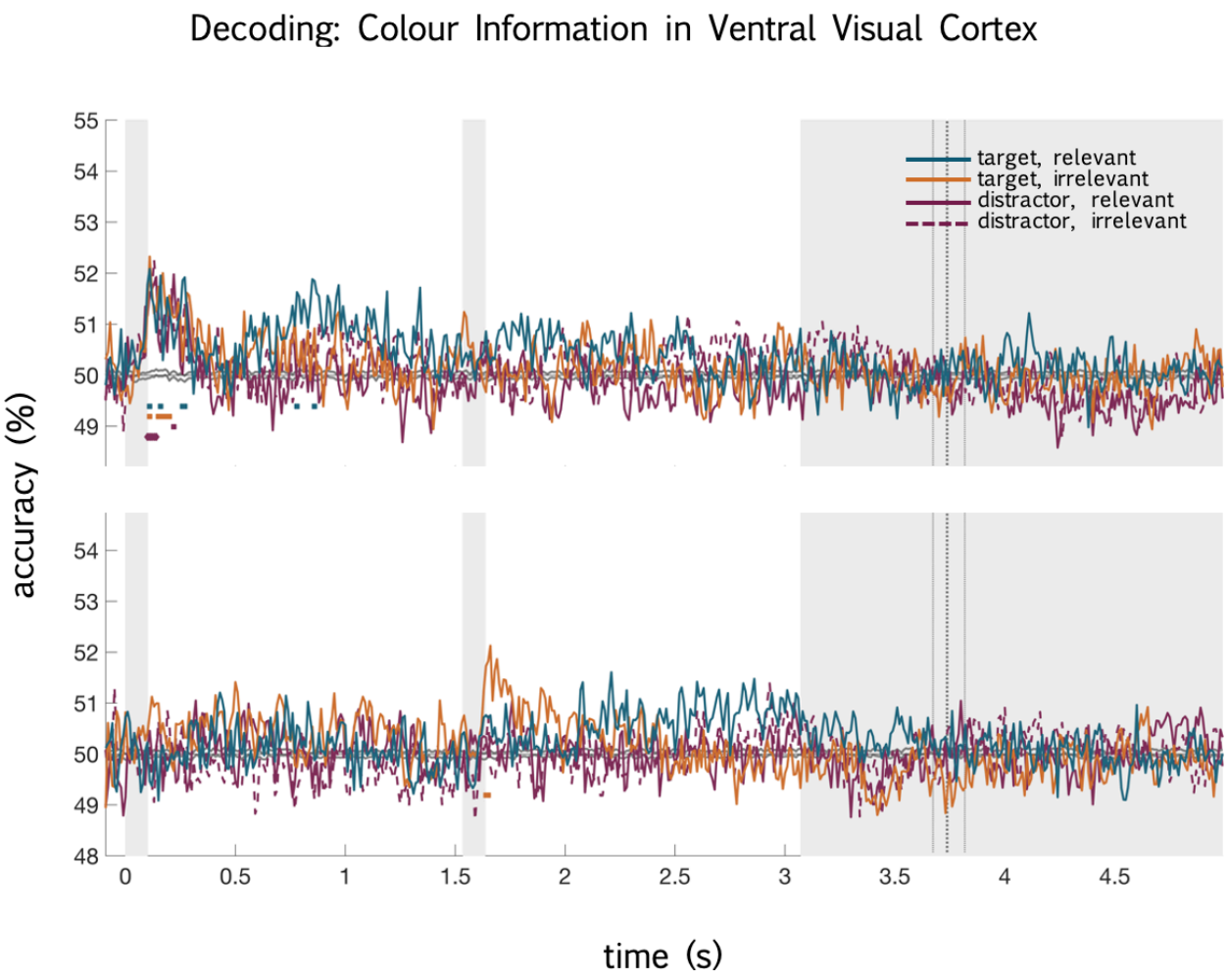


Figure 6. Group-level decoding accuracy (%) for ventral visual ROI colour information in Epoch 1 (upper panel) and Epoch 2 (lower panel), shown across the full trial. Decoding was applied separately to each colour pair, location on screen, and run, then averaged to produce these four traces representing target and distractor colour in attend colour and attend shape conditions. The task-relevant colour information coding (target colour when colour was task-relevant) is in blue. Information coding for target colour when it was irrelevant is in orange, and for distractor traces in red. Coloured squares mark timepoints at which decoding for the corresponding-coloured trace was reliably different to chance, based on a cluster-based permutation test. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

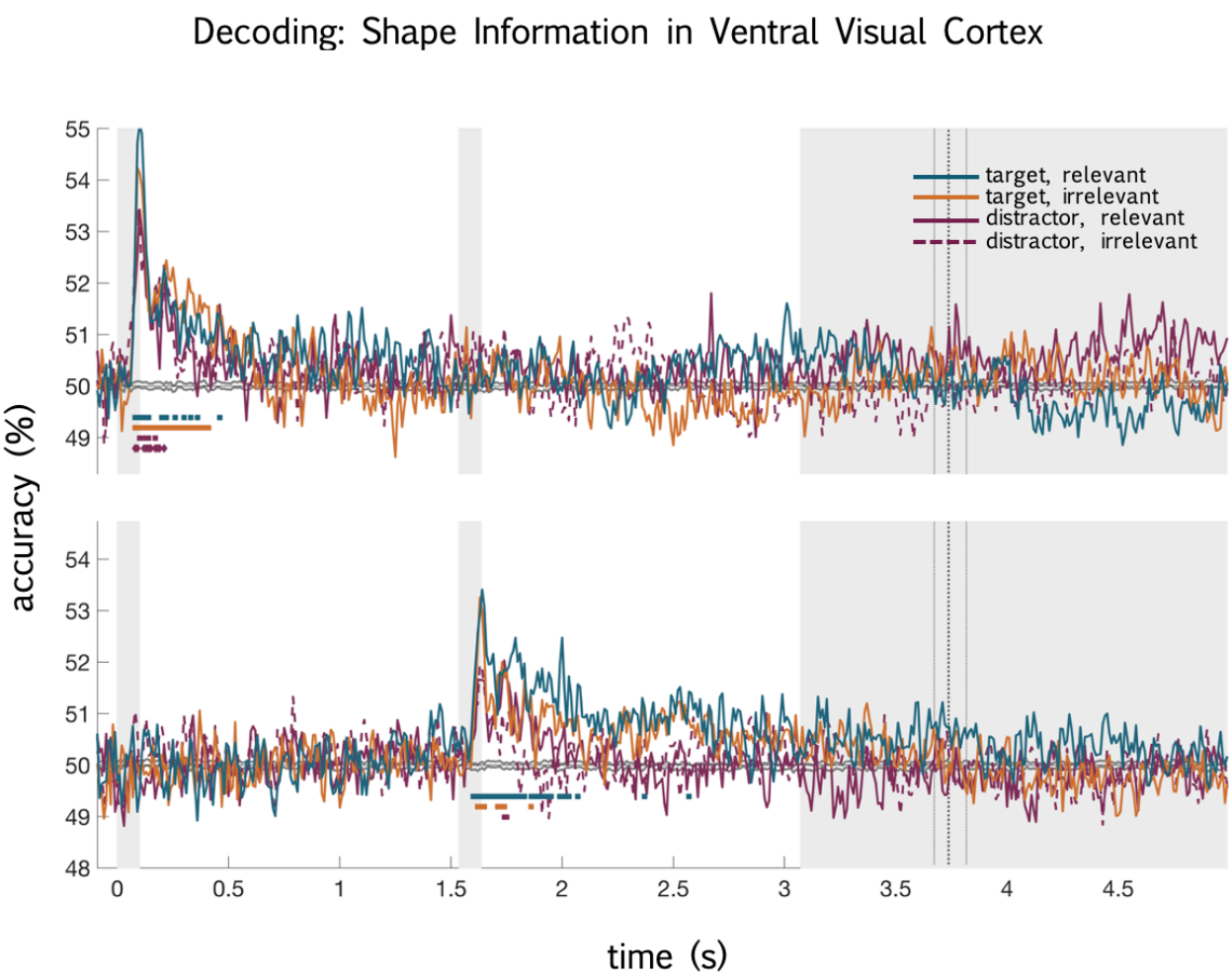


Figure 7. Group-level decoding accuracy (%) for ventral visual shape information in Epoch 1 (upper panel) and Epoch 2 (lower panel). The task-relevant shape information coding is in blue. Information coding for target shape when it was irrelevant is in orange, and for distractor traces in red. Coloured squares mark timepoints at which decoding for the corresponding-coloured trace was reliably different to chance, based on a cluster-based permutation test. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

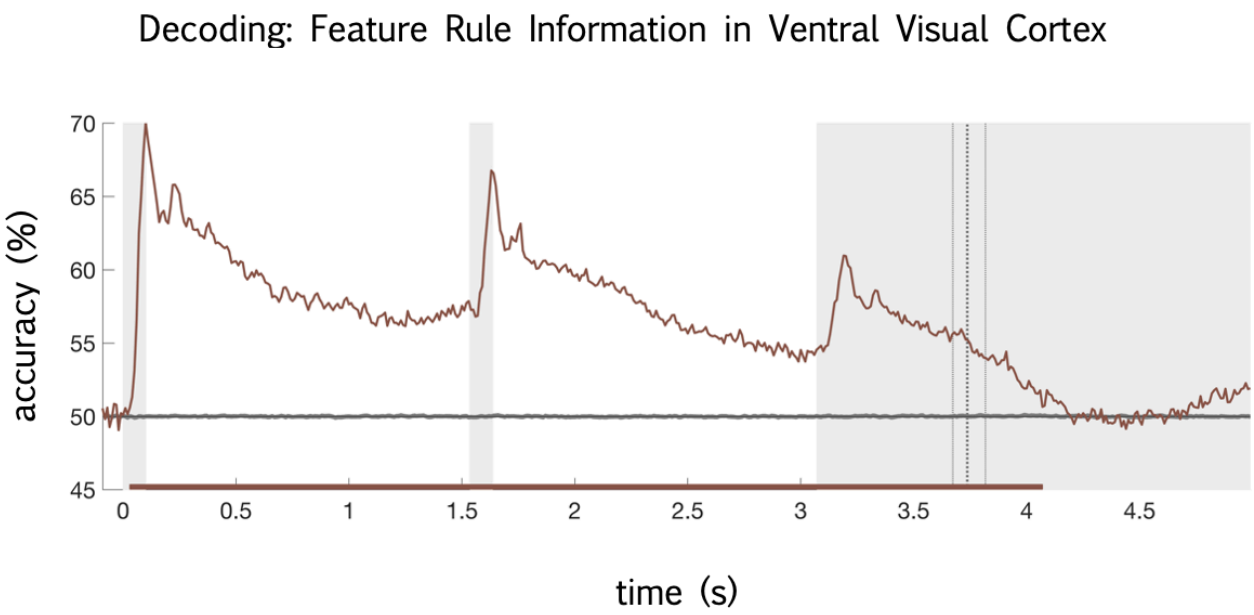


Figure 8. Group-level decoding accuracy (%) for ventral visual attended feature information. Coloured squares mark timepoints at which attended feature decoding was reliably different to chance, based on a cluster-based permutation test. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

3.3.1.2. Multiple-demand ROI

In contrast to the ventral visual ROI, the multiple-demand ROI showed no visible or statistically reliable early stimulus-driven effect for colour, and a reduced early stimulus-driven effect for shape. Instead, selective information coding for the relevant feature emerged after stimulus offset for both features in Epoch 1.

Coding for the relevant colour information in Epoch 1 appeared at approximately 600 ms, shortly before its preferential coding in the ventral visual ROI (Figure 9). In Epoch 2, there was no visible or statistical evidence for colour information coding, for the task-relevant or task-irrelevant features.

Coding for the relevant shape information in Epoch 1 was also reliably different to chance after stimulus offset, and statistically different to the other attention conditions at approximately 650 ms (Figure 10). As with the colour information coding, Epoch 2 shape information coding was not reliably detected,

whether for task-relevant or task-irrelevant items. Attended feature decoding was strong and sustained within each task epoch (Figure 11).

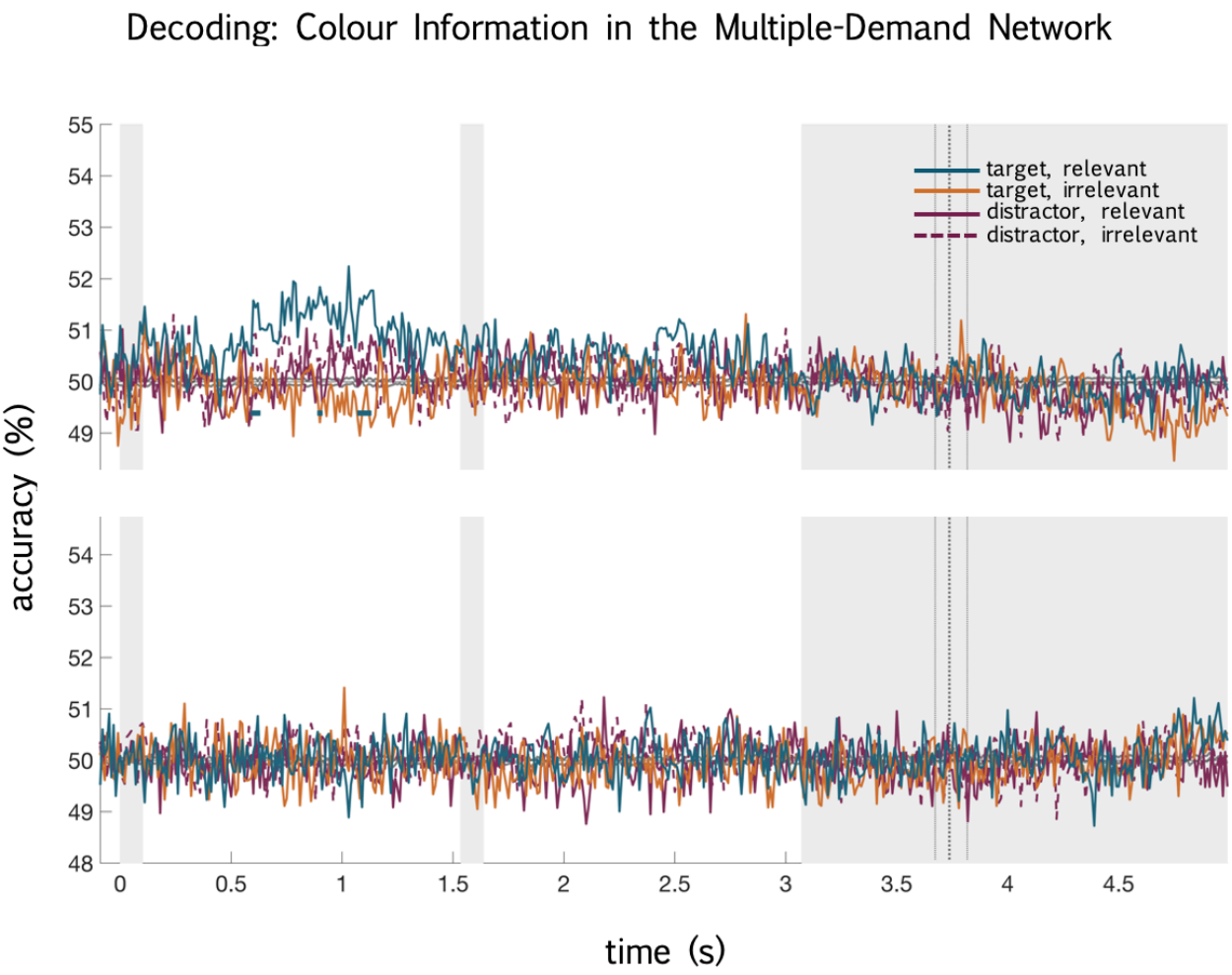


Figure 9. Group-level decoding accuracy (%) for MDN colour information in Epoch 1 (upper panel) and Epoch 2 (lower panel). The task-relevant colour information coding is in blue. Information coding for target colour when it was irrelevant is in orange, and for distractor traces in red. Coloured squares mark timepoints at which decoding for the corresponding-coloured trace was reliably different to chance, based on a cluster-based permutation test. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

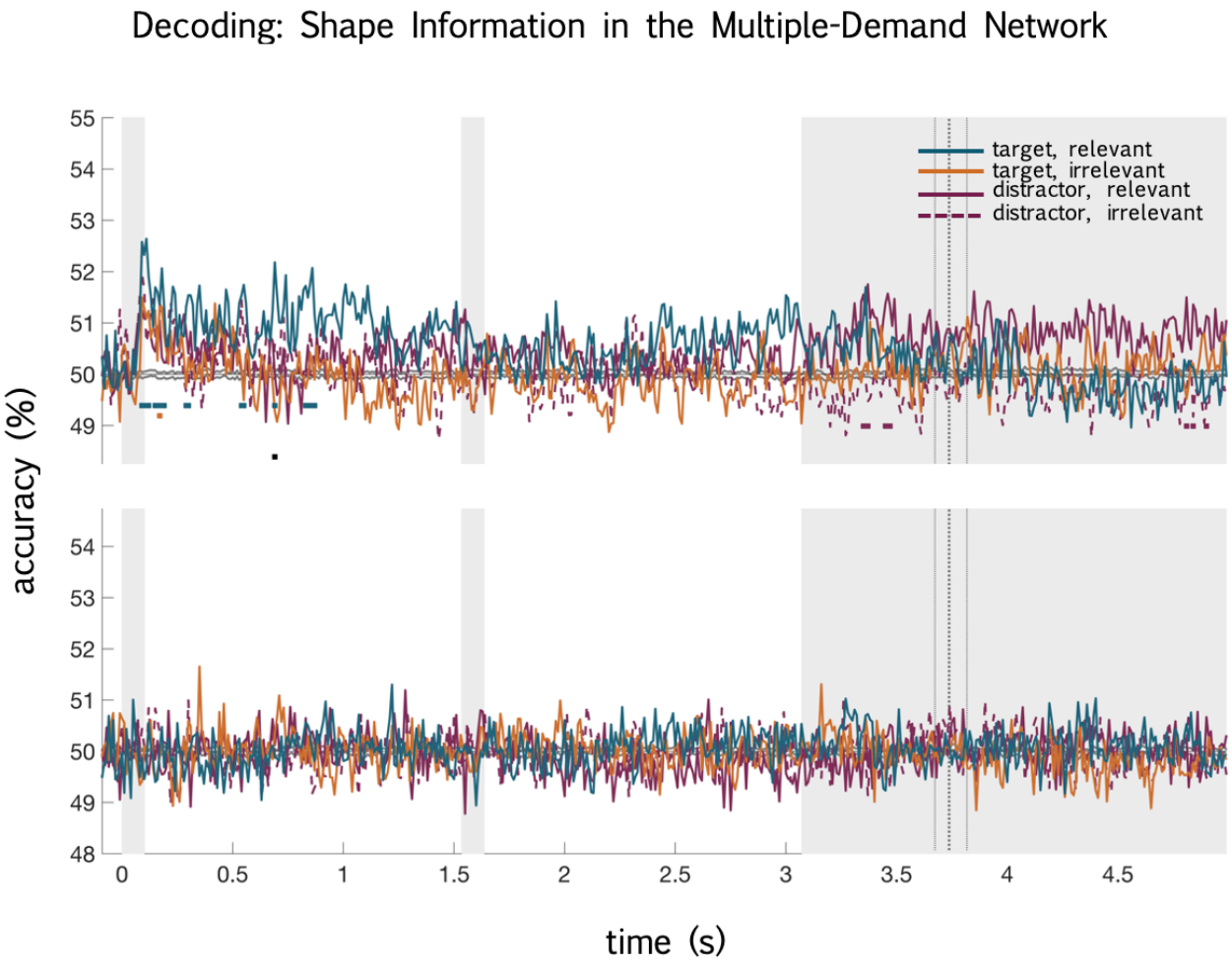


Figure 10. Group-level decoding accuracy (%) for MDN shape information in Epoch 1 (upper panel) and Epoch 2 (lower panel). The task-relevant shape information coding is in blue. Information coding for target shape when it was irrelevant is in orange, and for distractor traces in red. Coloured squares mark timepoints at which decoding for the corresponding-coloured trace was reliably different to chance, based on a cluster-based permutation test. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

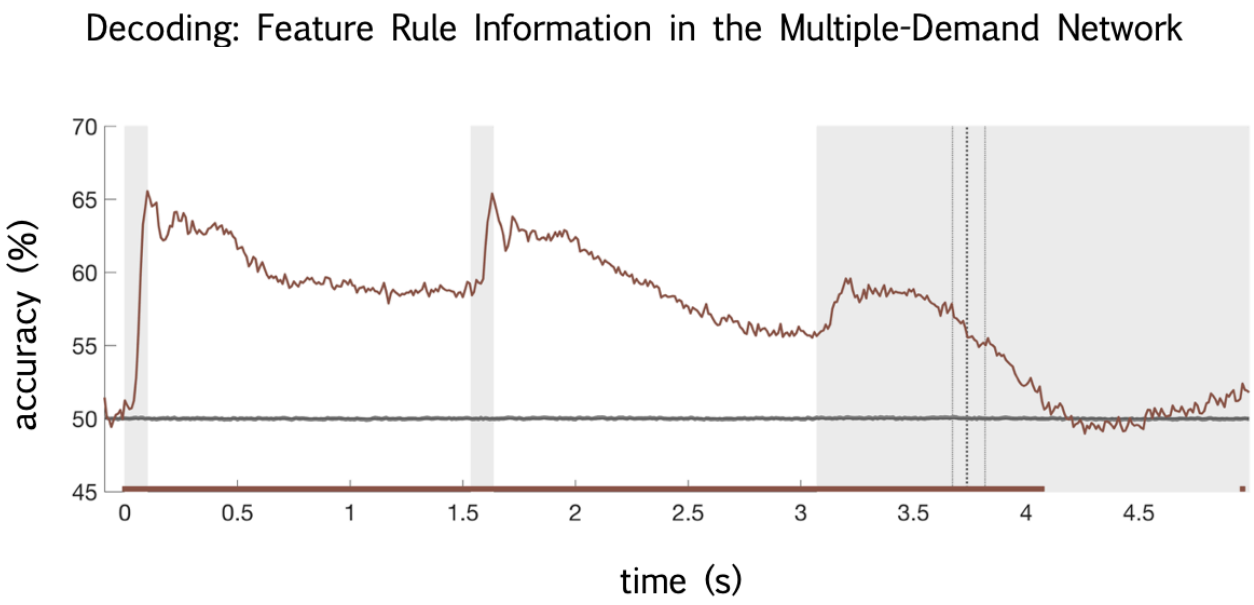


Figure 11. Group-level decoding accuracy (%) for MDN attended feature information. Coloured squares mark timepoints at which attended feature decoding was reliably different to chance, based on a cluster-based permutation test. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

3.3.2. Information flow

Granger-causal analysis revealed subtle but statistically reliable feedforward and feedback information flow (Figure 12). For Epoch 1, representational structure in the multiple demand ROI predicted the structure within the ventral visual ROI 120-20 ms later (i.e., feedback), during the first stimulus display and delay periods. Representational structure in the ventral visual ROI similarly predicted the structure in the multiple demand ROI (feedforward) from the first display and into the first delay period. Results for Epoch 2 colour and shape RDMs closely mirrored those for Epoch 1, with significant information flow during the first display and delay phases, though feedback and feedforward information flow were also visible and sporadically statistically reliable after the second display.

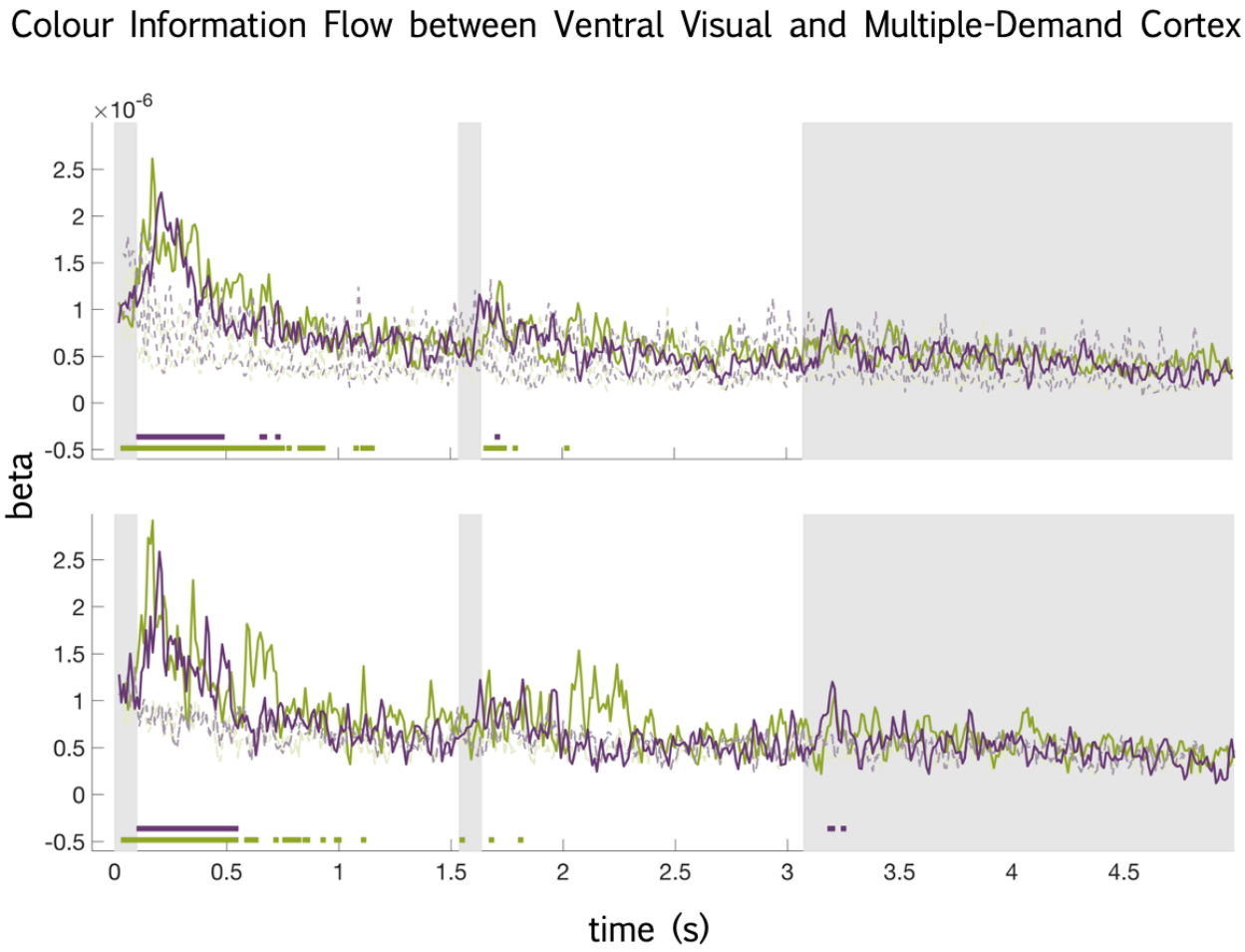


Figure 12. Granger-causal information flow for colour information oriented to Epoch 1 (upper panel) and Epoch 2 (lower panel). Green represents feedforward of information from ventral visual to multiple demand regions. Purple represents feedback from multiple demand to ventral visual regions, with 120-20 ms lag between "source" and "target" ROIs. Source predictions are visualised on the target time point, with the first time point starting 20 ms into the first stimulus display (120 ms from the start of the baseline). Dashed green and purple lines mark the upper and lower bounds of the null distribution's 95% confidence interval, for feedforward and feedback predictions. Coloured squares mark timepoints at which the corresponding-coloured trace was reliably different to zero, based on a cluster-based permutation test. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

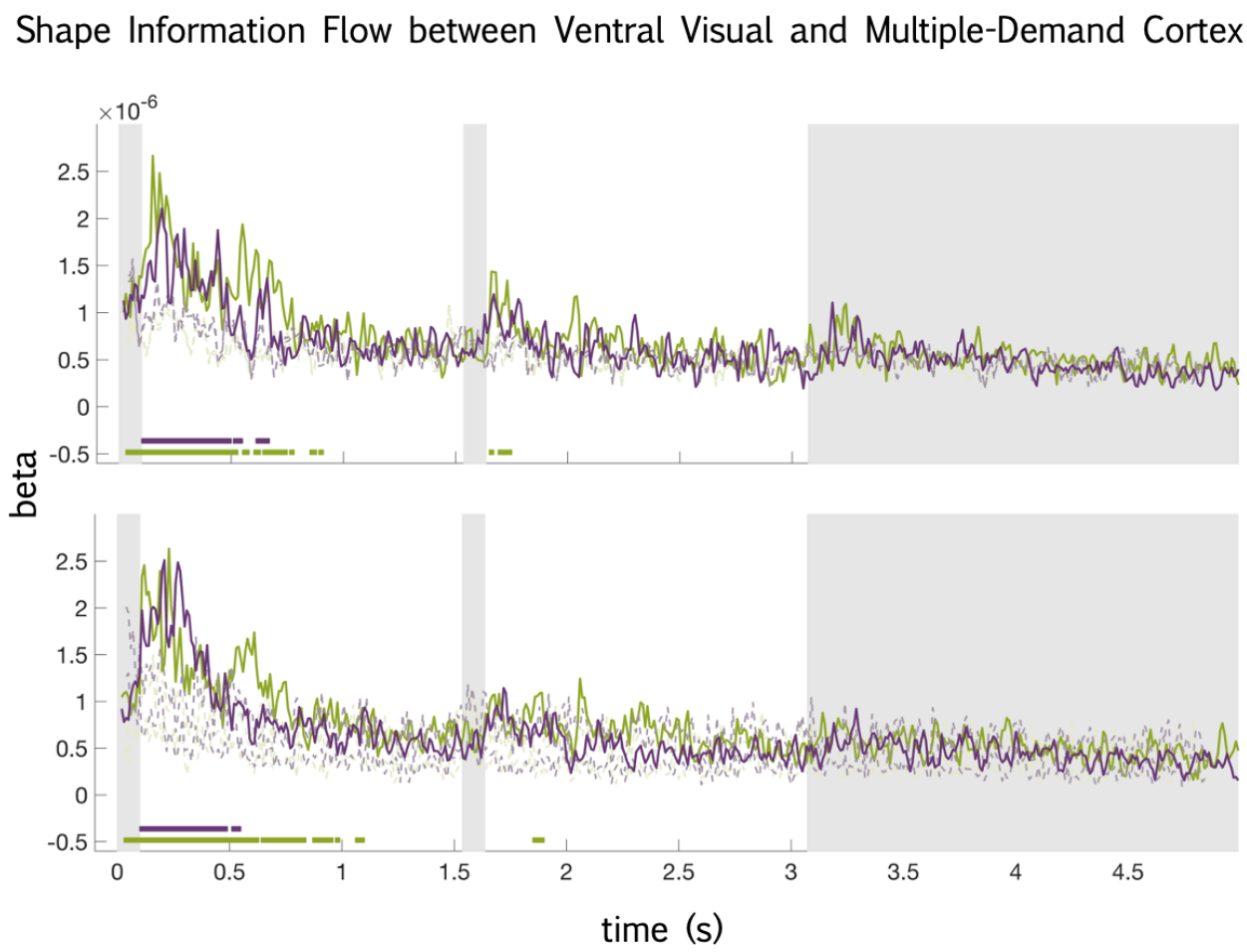


Figure 13. Granger-causal information flow for shape information oriented to Epoch 1 (upper panel) and Epoch 2 (lower panel). Green represents feedforward of information from ventral visual to multiple demand regions. Purple represents feedback from multiple demand to ventral visual regions, with 120-20 ms lag between "source" and "target" ROIs. Source predictions are visualised on the target time point, with no predictions for the target during the baseline period. Coloured squares mark timepoints at which the corresponding-coloured trace was reliably different to zero, based on a cluster-based permutation test. Dashed green and purple lines mark the upper and lower bounds of the null distribution's 95% confidence interval, for feedforward and feedback predictions. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

3.3.3. Epoch-specific information flow

I further reduced the full trial by trial RDM to a 6 by 4 matrix representing the colour and shape features in each epoch that could not predict the other. Information flow for the representational structure derived from Epoch 1 stimulus features was negligible, for feedforward and feedback directions (Figure 14, upper

panel). Beta weights for ventral visual cortex structure predicting later MDN structure were small and statistically unreliable throughout the trial (green trace). The same was true for feedback information flow (purple trace).

Information flow specific to the Epoch 2 stimulus features was also small and unreliable (Figure 14, lower panel). Here, I saw small peaks in the information explained by the source ROI in both task epochs, despite stimulus structure being independent between epochs and participants being unable to predict what stimuli would appear in Epoch 2. This suggests that visible, statistically unreliable hints of information flow were either artefactual, or reflected imperfect counterbalancing between Epoch 1 and Epoch 2 information after some experimental trials were removed for noise. In either case, we cannot resolve epoch-specific estimates of information flow from these results.

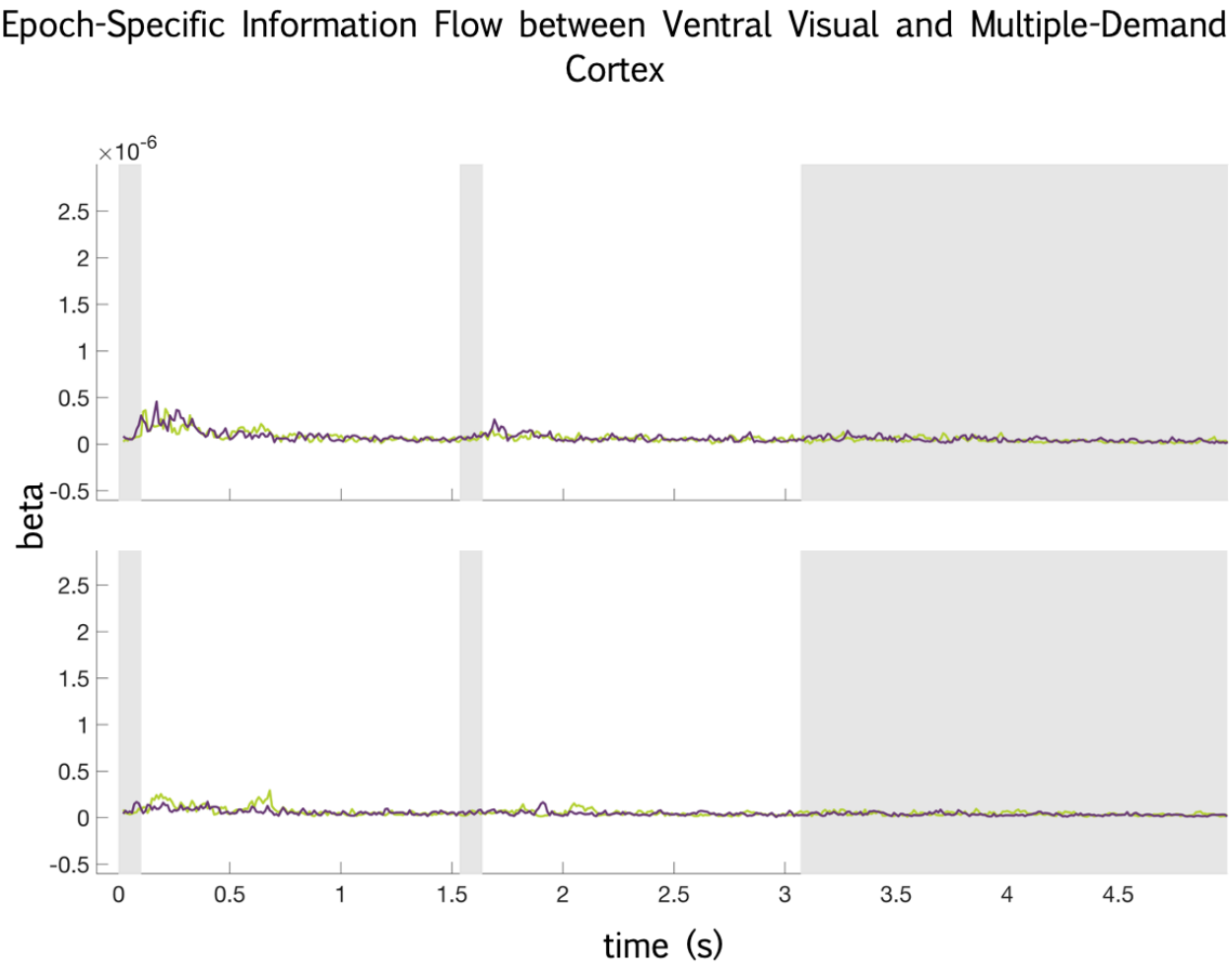


Figure 14. Granger-causal information flow for information specific to Epoch 1 (upper panel) and Epoch 2 (lower panel). Green represents feedforward of information from ventral visual to multiple demand regions. Purple represents feedback from multiple demand to ventral visual regions, with 120-20 ms lag between "source" and "target" ROIs. Source predictions are visualised on the target time point. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display. There were no time points at which the traces reliably differed from zero.

3.3.4. Model-based representational similarity analysis

Next, I explored how the task information was represented within regions, and therefore what task features could have contributed to information flow between them. The model-based representational similarity analysis probes not just how colour and shape information emerge over the course of the trial, but whether

that information takes a form specified by a theoretical model. In turn, this could give us insight into what representations – whether response-oriented or reflecting stimulus properties – could be shared between visual and domain-general brain regions.

3.3.4.1. Ventral visual ROI

Within the ventral visual ROI, I saw evidence that representations of colour and shape reflected graded differences in the stimuli. For shape, this was true across all attention conditions and in both epochs (Figure 15). Graded differentiation of shape information peaked rapidly after stimulus onset (100 ms) before returning to near baseline.

In Epoch 1, this peak in the graded differentiation of shape information was enhanced for the behaviourally-relevant aspect: target shape when shape was relevant, relative to distractor shape or task-irrelevant target shape information. The advantage for the behaviourally-relevant shape information was also sustained relative to when it was task-irrelevant, from approximately 1 s following stimulus onset.

For colour, I observed the same pattern. An initial peak followed the stimulus display in both epochs, reflecting that graded difference in the colours that people saw on each display could reliably predict a graded difference in the response within the visual ventral stream (Figure 16).

Stepwise stimulus models continued to predict the representational structure, selectively for task-relevant colour, in both epochs. Task-relevant colour information was sustained relative to when it was task-irrelevant from approximately 500 ms following stimulus onset in Epoch 1, and from approximately 300 ms following stimulus onset in Epoch 2.

Attended location and attended feature information was reliable within the first 100 ms of the trial (Figure 17). It peaked rapidly after each stimulus display and remained above zero until the response period.

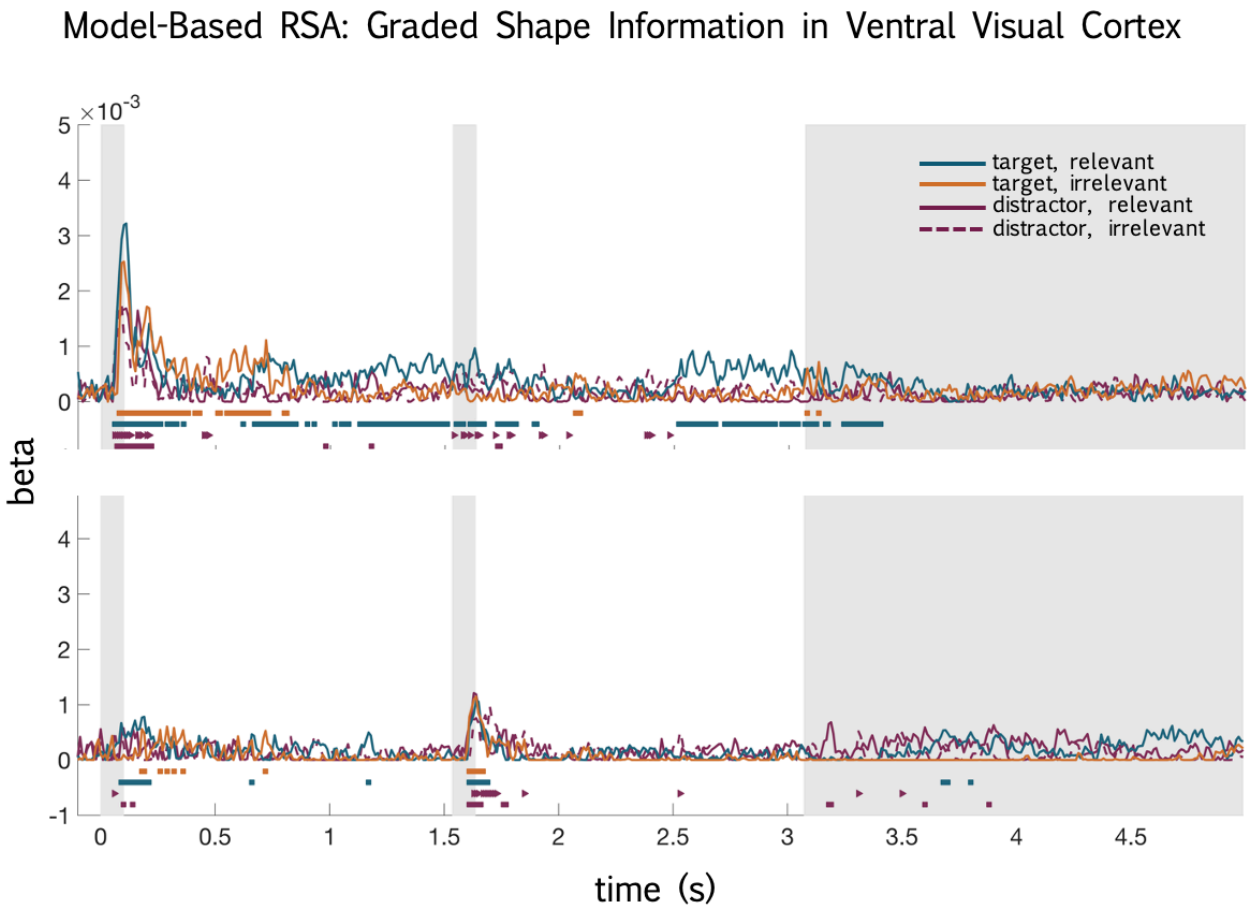


Figure 15. Group-level model fits (betas) for models predicting that the visual cortex response was increasingly dissimilar for shapes that were more discriminable, with stepwise changes in visual cortex response across four levels of shape similarity, for Epochs 1 (top panel) and 2 (bottom panel). Models were fit separately for target and distractor shapes, when shape was the task-relevant or task-irrelevant feature dimension, producing these four traces. The model fit for task-relevant shape information (target shape when shape was task-relevant) is in blue. Fits for target shape when it was irrelevant are in orange, and for distractor traces in red. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

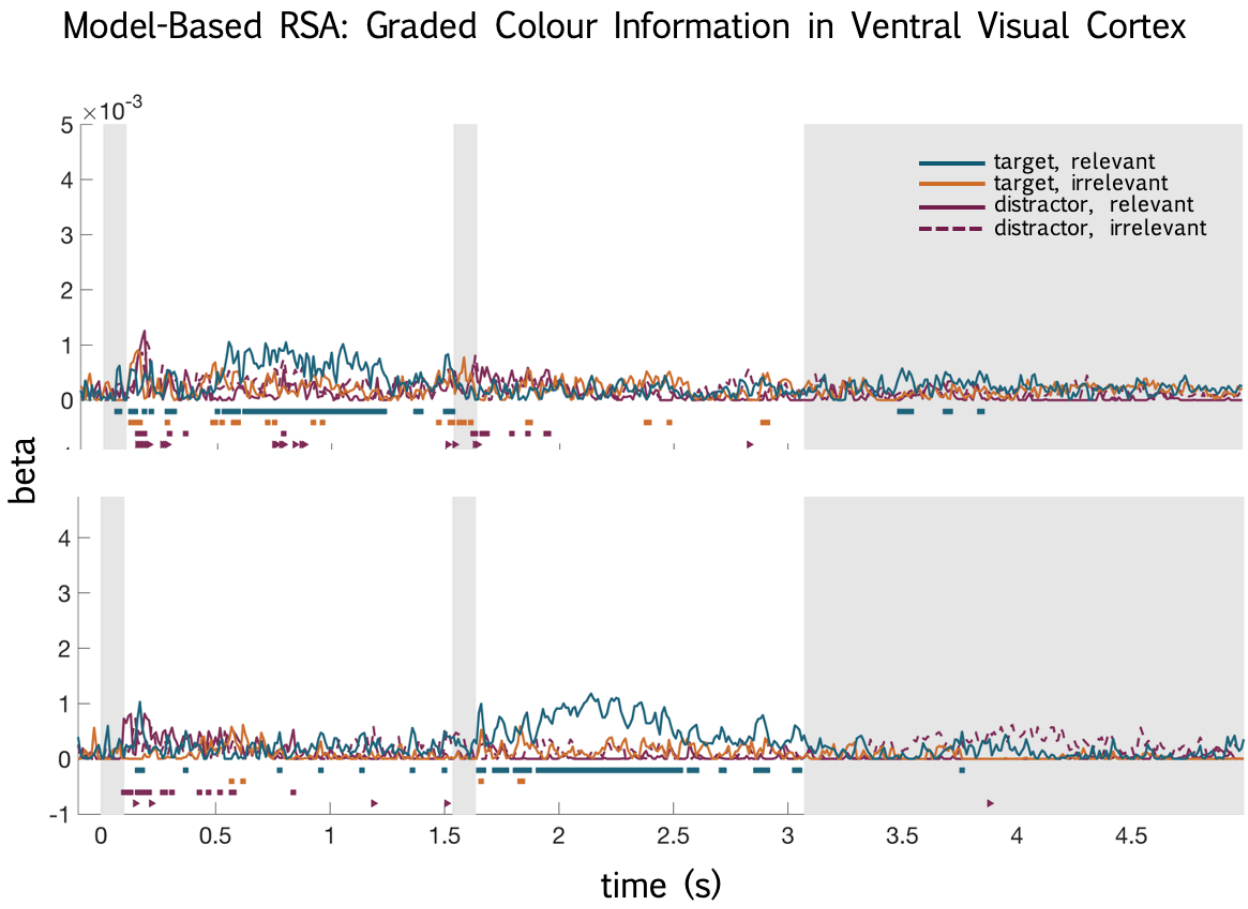


Figure 16. Group-level model fits (betas) for models predicting that the visual cortex response was increasingly dissimilar for colours that were more discriminable, with stepwise changes in visual cortex response across four levels of colour similarity, in Epochs 1 (top panel) and 2 (bottom panel). Models were fit separately for target and distractor colour, when colour was the task-relevant or task-irrelevant feature dimension, producing these four traces. The model fit for task-relevant colour information (target colour when colour was task-relevant) is in blue. Fits for target colour when it was irrelevant are in orange, and for distractor traces in red. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

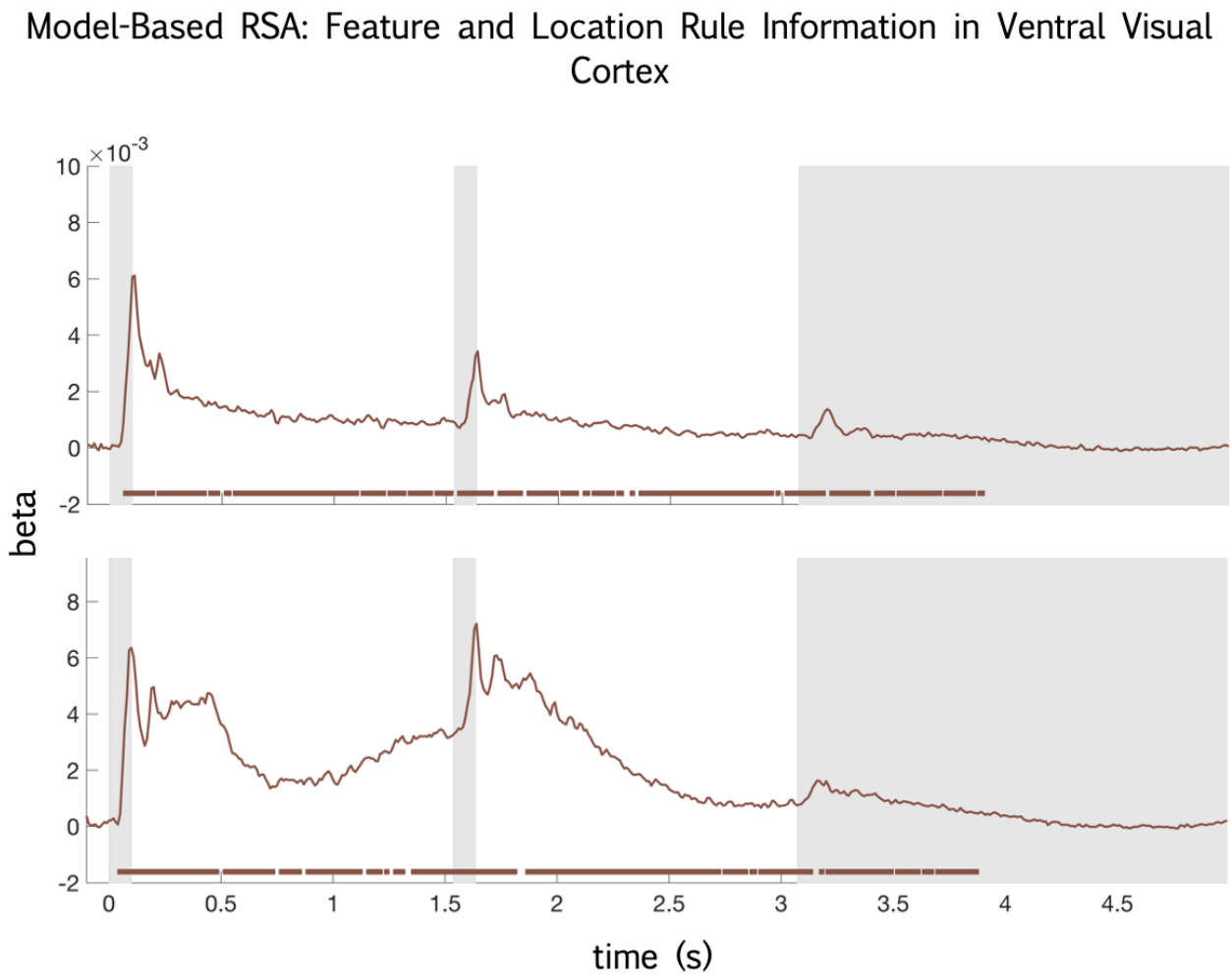


Figure 17. Group-level model fits (betas) for attended feature (A) and attended location (B) models within the ventral visual ROI.

3.3.4.2. Multiple-demand ROI

Next, I turned to the multiple-demand ROI to ask how far the representations within the MDN could be described by graded differences in colour and shape stimuli. In contrast to the ventral visual ROI, I saw only weak evidence that responses within the MDN differed proportionally to the stimuli's physical similarity, for shape (Figure 18) or colour (Figure 19). Stimulus information, as measured by the stepwise models, was sporadically statistically different to zero for colour and shape across the timecourse of the trial. Whereas the decoding analysis

showed that task-relevant colour and shape information was present within the MDN in Epoch 1, models predicting graded discrimination of stimuli according to their physical similarity did not visibly better fit the MDN's representational structure for the task-relevant feature in either epoch. However, attended location and feature information emerged early in the trial and was sustained throughout (Figure 20).

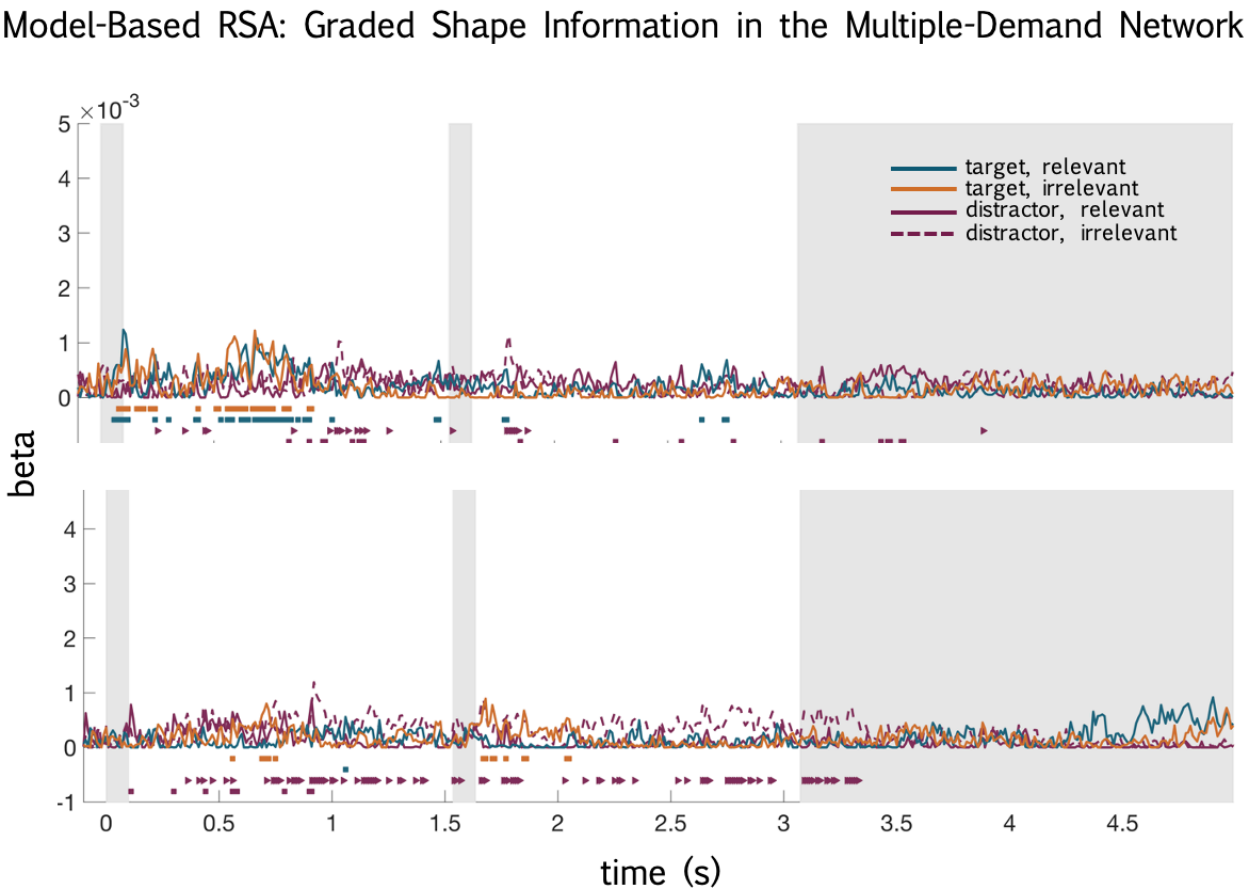


Figure 18. Group-level model fits (betas) for models predicting that the MDN response was increasingly dissimilar for shapes in Epoch 1 (upper panel) and Epoch 2 (lower panel) that were more discriminable, with stepwise changes in MDN response across four levels of stimulus similarity. Models were fit separately for targets and distractors, when shape was task-relevant or task-irrelevant, producing these four traces for each epoch. The model fit for task-relevant information (target shape when shape was task-relevant) is in blue. Fits for the target shape when it was irrelevant are in orange, and for distractor traces in red. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

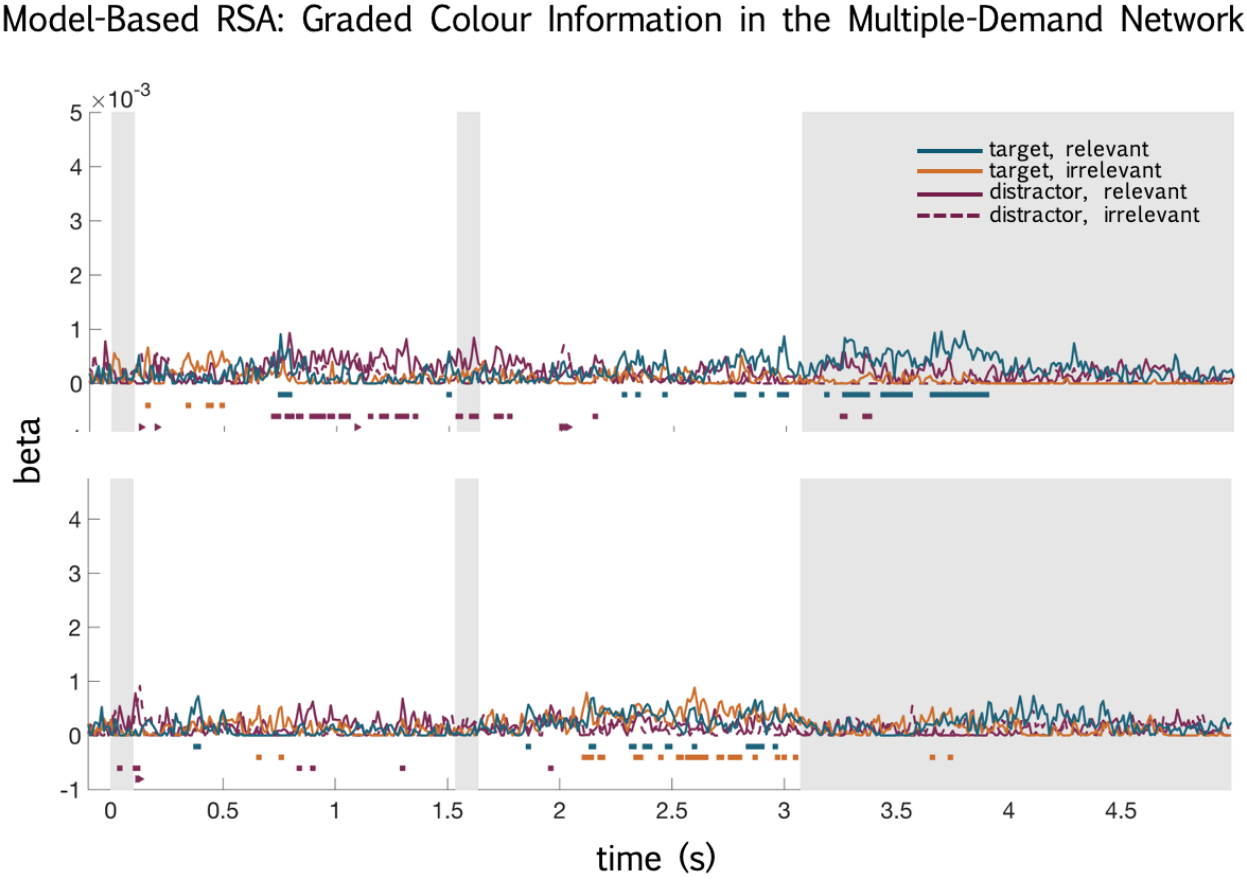


Figure 19. Group-level model fits (betas) for models predicting that the MDN response was increasingly dissimilar for colours in Epoch 1 (upper panel) and Epoch 2 (lower panel) that were more discriminable, with stepwise changes in MDN response across four levels of stimulus similarity. Models were fit separately for targets and distractors, when colour was task-relevant or task-irrelevant, producing these four traces for each epoch. The model fit for task-relevant information (target colour when colour was task-relevant) is in blue. Fits for the target colour when it was irrelevant are in orange, and for distractor traces in red. Vertical grey patches mark the two stimulus display durations and the maximum duration of the response display.

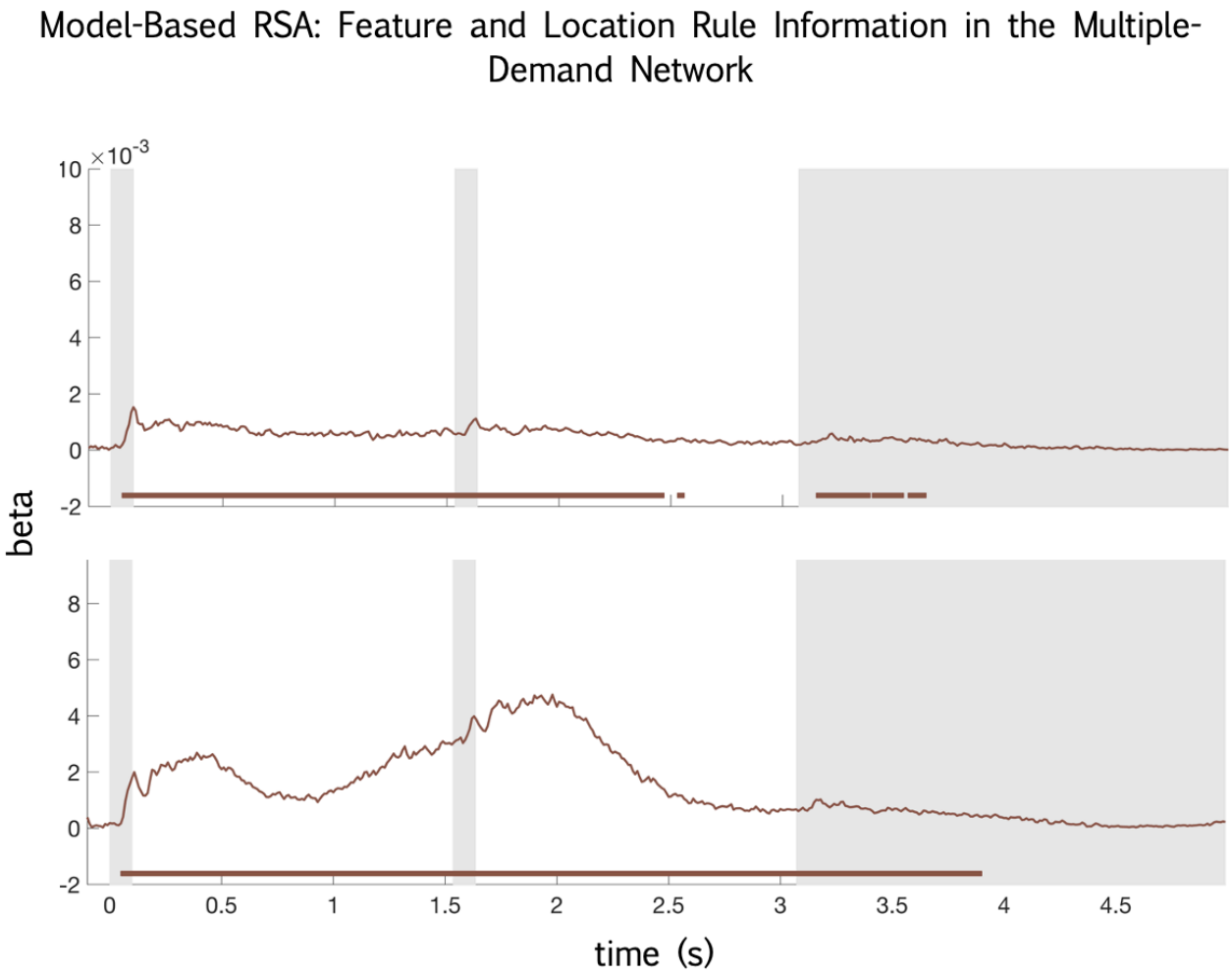


Figure 20. Group-level model fit (betas) for attended feature (A) and attended location (B) models in the multiple-demand ROI.

3.3.4.3. Model overlap

These models predict graded differences in neural responses with graded differences in stimuli. That is, for the four colours red, orange, yellow, and green, the models predict that red will be more different to yellow and green than to orange. This means that the graded difference models overlap with categorical models. More specifically, graded difference models could fit data that are better explained by categorical models. We can test this by probing how well within-category differences (as in the graded difference models, but not categorical models)

describe the data. I have tested this, and included the within-category model fits in Appendix A.

3.3.4.4. Statistical significance

It is worth noting that visibly small peaks in model fits are flagged as reliably different to chance. For example, Figure 18A shows statistically significant fits for Epoch 1 distractor shape models off and on through both epochs, though the actual magnitude of the model fits is very small. This is especially clear in the Epoch 2 model fits (lower panel of Figures 15 to 19), where fits that easily could be treated as noise are marked as statistically significant. A possible reason for this is that the permutation-based null distribution was generated by averaging individual-level permutation samples from every participant. Because of this, the null distribution represents a narrower selection of results that are plausible under the null hypothesis than would be covered by the individual-level data. A better strategy going forward could be to use the individual-level permutations, rather than group-averaged null data, to more conservatively represent the null hypothesis. Further, the model-based RSA differed from the other analyses presented here by relying on the pre-trial baseline period to provide null data. This restricted the range of values that could form the null distribution to 10 baseline beta values per subject and model fit. Using baseline model fits as null data could also underestimate the range of values that are plausible under the null hypothesis if events during the trial elicited higher variation in dissimilarities, so that model fits to permuted RDMs could be strong or weak while model fits to baseline data were homogeneous.

3.4. Discussion

Tracking information flow is important for testing how disparate brain regions work together to produce visual attention. Information about the timescale of feedback can be important for other cognitive questions, such as whether conscious awareness is possible with only feed-forward information flow (see for

example Maguire & Howe, 2016; Potter et al., 2014). It can also lay the groundwork for brain stimulation investigations of causality by allowing researchers to predict when feedforward, feedback, and recurrent processing dominate information flow, and so target specific processes to perturb.

Here, I traced the information coded within ventral visual cortex and the MDN, and the timecourse of information flow between them. I found that ventral visual and MDN showed different patterns of information coding, with a stronger stimulus-driven response in visual regions, followed by preferential coding of task-relevant information and strong representation of task rules in both networks. Granger-causal information flow analysis showed that the representational structure within the MDN shortly after stimulus offset explained subsequent ventral visual representational structure, and vice versa, in both task epochs; though this was most apparent early in the trial.

3.4.1. Information flow

Past studies have shown an initial feedforward sweep, followed by feedback shortly after stimulus offset (Goddard et al., 2021; Karimi-Rouzbahani et al., 2021). In line with these findings, here I show a rapid onset of feedforward flow for colour and shape information in the first 20 to 200 ms of the trial. Feedback information flow emerged shortly after, with frontal representations during stimulus presentations predicting visual cortex representations early in the post-stimulus delay. This suggests that stimulus information can be quickly organised and communicated top-down, even as people are shifting their attention.

Beyond existing research, I was curious to know how feedforward and feedback information flow was affected by multi-step task demands, to understand how these visual and domain-general regions might co-operate in complex tasks. I saw that feedforward information flow also emerged quickly in the second task epoch, as probed by RDMs oriented toward Epoch 2 colour and shape information. However, information flow estimates were very similar for RDMs oriented toward

the two epochs. The reduced RDMs could not definitively separate information from the first and second displays, meaning that part of the variance predicted by the source ROI in one epoch could reflect the other epoch. Thus, while information flow followed the stimulus displays, we cannot be confident that the information being transferred was about a given task epoch.

One important question is what impact MDN feedback could have on task information coding in the visual system. Feedback here increased sharply from the start of the post-stimulus delay, perhaps commensurate with the MDN supporting the maintenance of information in working memory. Past studies that trace information flow for stimulus features point to this feedback playing a specific role in the enhancement of relevant stimulus features. Goddard et al. (2021) showed that delay period coding of a behaviourally-relevant feature was strongest in frontal and visual ROIs when feedback of stimulus information dominated information flow. Thus, a possible function of MDN feedback could be selectively enhancing the discriminability of task-relevant visual features. Similar to their findings, preferential coding of relevant stimulus information in the decoding analyses here emerged during the post-stimulus delay period, at the same time or earlier in the multiple demand relative to the ventral visual ROI. Observing the same cascade in this study reinforces the idea that the MDN primarily engages after the first feedforward sweep of stimulus information, and that it may have a role in supporting working memory and selective attention.

However, the time at which the MDN predicts visual cortex representations is only one indicator of the role it plays. Ultimately, we want to understand what kind of information is fed back, and so grasp the mechanics of generating diverse, organised behaviour from our busy brains. One path towards this kind of understanding is to simply control what information we retain in our data, before trying to observe how it is passed between brain regions. I presented two approaches to this, with larger RDMs and smaller matrices that were designed to be either more sensitive or more specific. In this case, the larger RDMs were able to

uncover times at which visual cortex or MDN representations predicted each other, over and above what the history within that ROI could capture, but they also gave limited insight into what information was being shared. The smaller matrices were deliberately designed to capture information that people could access at different times in the task, so that I could associate any feedback with a specific stimulus display. Due to the constraints of the task, however, these epoch-specific matrices were very small, leaving very little variance to explain with either the target or the source ROI's past. The length of the dual-epoch trials meant that I could only match the number of trials in the original study using this task (Goddard et al. 2021) even with intensive testing sessions. A flexible dataset, with many repetitions for each condition, and enough trials that imbalance from trial rejection does not harm the overall picture, could better support this kind of multi-faceted approach.

Another path towards understanding how MDN representations impact ventral visual representations is through explicit models. In theory, we could explain away facets of the source or target regions representations with model-based representational similarity analysis. For instance, we could compare how the MDN predicts ventral visual representations while attended feature information is either retained within the MDN, or modelled out. This offers the exciting possibility that we could isolate conceptually different facets of a task in two brain regions (such as rule information in the MDN, and preferential coding within the ventral visual stream), and begin to understand how distinct properties of two regions interact to enable adaptive behaviour. However, this also raises a non-trivial hurdle: we need to first derive models that capture the information that is represented within each region. Unfortunately, I found that ventral visual and MDN representations were difficult to capture with explicit theoretical models derived from first principals, so could not perform this analysis. Again, this limitation could partly stem from the limited number of stimulus repetitions that I could obtain with this task. It is made more complex by the fact that representational similarity analyses in time-resolved data are still somewhat new,

so we cannot rely on an existing body of research to tell us what models will be best. The strength of this approach is that forcing researchers to convert their verbal predictions into numbers highlights how ambiguous our predictions can be, but it does not give us the solution. For colour alone, we could expect it to be represented according to actual or perceived gradations in tone, by a strict separation between identical and non-identical colours, or by what colours are relevant for the task. Each of these predictions might change as we think about different brain regions and different moments in visual processing.

At their heart, the methods I have presented here are simple to think about. Yet facing the practical challenges of designing a task that can provide rich enough information, and translating our idea of how information is transferred within the brain to measurable representations and numerical models, is an essential part of using these methods well. As we have seen above, working towards precise measures of information flow also brings up many analysis decisions that can quickly inflate the parameter space. This uncertainty and flexibility creates opportunities to mistake parameters that work well with one dataset for parameters that will reliably help us capture meaning in noisy brain data. But it is also inherent to developing a new analysis, where there is no clear best approach. Thus, an important practice for future studies will be to test a wide range of models and parameters but cross-validating findings within a dataset, so that we can push further towards precise measures of communication within the brain while scrutinising our conclusions.

3.4.2. Stimulus information in MD cortex

While there was evidence for information flow, the two regions appeared to be dominated by different representational structures. Unlike in ventral visual regions, representational structure within the MDN only weakly reflected gradations in visual features. Decoding results in the MDN ROI also showed reduced early responses to stimuli across attention conditions, relative to the

ventral visual ROI. By contrast, non-human primate studies show that both relevant and irrelevant stimuli are coded early in lateral PFC, with the relevant feature gradually gaining preference (Erez et al., 2020; Erez & Duncan, 2015; Kadohisa et al., 2013). Human fMRI data reinforce this, with many studies showing some coding of relevant and irrelevant stimuli as well as the characteristic preference for features that are behaviourally relevant (Jackson et al., 2016, 2021; Woolgar, Hampshire, et al., 2011, 2011; Woolgar, Williams, et al., 2015). Multivariate pattern analysis of source-space MEG has also shown that coding of target identity and location emerges rapidly in the lateral PFC, just tens of milliseconds after it can be detected in visual cortex (Wen et al., 2019). With this in mind, we could expect to see both early stimulus-driven information coding, and preferential coding of the relevant feature, in time-resolved MDN data.

One problem may be a narrow ROI definition. Coupled with individual differences in MDN localisation (Assem et al., 2020; Fedorenko et al., 2012, 2013), the uncertainty inherent in source estimation, and the dispersed nature of the MDN, the narrow ROI could overlook regions of the MDN that contain useful information. Recent evidence suggests that strict ROI definitions may not benefit multivariate analyses (Shashidhara et al., 2020), meaning that broad ROI definitions could be a safer choice.

Another possibility is that the MDN may not be oriented towards graded differences between stimuli. Visual information can be detected in the MDN through multiple methods. Invasive recordings in NHPs reliably show that a substantial proportion of sampled neurons encode behaviourally relevant features (see for example Freedman et al, 2003: 18-19% of sample PFC neurons category-selective during stimulus and delay periods; Rao et al., 1997: 64% of sampled lateral PFC neurons sensitive to target identity, location, or both during working memory delay). Human MEG-fMRI fusion also shows that stimulus coding is strong, relative to decision coding, in portions of the MEG signal that share representational structure with the MDN in fMRI, further suggesting that coding relevant stimuli is

a substantial part of the MDN's role (Moerel et al., 2021). However, visual feature coding in lateral PFC (the recording site in many NHP studies) may overrepresent the stimulus-driven nature of the wider MD network, as anterior insula/frontal operculum and pre-supplementary motor/anterior cingulate regions may be more strongly associated with stable goal maintenance than trial-to-trial adaptation (Dosenbach et al., 2007). More specifically, the early responses to stimuli in both decoding and representational similarity analyses of the ventral visual stream, along with graded responses to colours and shape, may not be a priority within task-sensitive multiple-demand cortex. In the decoding analysis, I saw a suggestion that the MDN briefly differentiated all shape information shortly after stimulus onset in Epoch 1, but more consistently, I saw coding of rule, and of the task-relevant feature. Decoding data from a human MEG task near-identical to this study (Goddard et al., 2021), resolved to frontal cortex, show little to no response to stimulus onset, suggesting that what I observed could be a consistent feature of decoding from source-reconstructed frontal cortex.

A more nuanced answer may be that MDN representations, while sensitive to stimulus features, encode them based on behavioural relevance rather than real-world similarities. Multivariate analysis of target detection data from non-human primate lateral PFC shows that responses selectively discriminate target and distractor categories, but not various target categories, highlighting that even higher-level coding of stimulus category is oriented toward the distinctions that are most important for decision and action (Erez et al., 2020). Human fMRI similarly shows that frontoparietal brain regions code the conjunction of features (in this case, form and motion) that are critical for behaviour (Li et al., 2007). If the MDN represents stimuli along behaviourally relevant divisions, the stimuli in this categorical judgement task could have elicited categorical (red vs green) rather than stepwise differences (red, orange, yellow, green) in the network's response. Goddard et al. (2021) showed that feature differences dominate the early visual cortex response, but not the later frontal responses. Functional MRI suggests that MDN

activity could even emphasise small physical differences over large ones (Woolgar, Hampshire, et al., 2011). Thus, real-world stimulus similarities may not be a central driver of MDN representations.

3.4.3. MDN engagement in multiple task steps

Experiment 2 of Chapter 2 showed that relevant stimulus information was quickly preferentially encoded in both task steps. The source-reconstructed data localised preferential coding of colour and shape information to the MDN. However, decoding analysis of the second task epoch showed no reliable coding of relevant or irrelevant stimulus features in the MDN, and representational dissimilarity analyses showed weak and highly variable fits to stimulus models in the MDN across features and epochs. Ventral visual cortex showed decoding across attention conditions in both task epochs, though this was unreliable for colour information in Epoch 2. Many studies show that the MDN codes whatever information is relevant (for example, Jackson et al., 2016; Woolgar et al., 2011, 2015; Woolgar & Zopf, 2017), and it is easy to infer from this that the MDN conveys information about the relevant stimulus to sensory cortex. In fact, concurrent brain stimulation and fMRI has showed that disrupting an MDN node both reduces coding of relevant information across the MDN, and reduces the difference in coding between relevant and irrelevant information in early visual cortex (Jackson et al., 2021). From this, we could expect that preferential coding of relevant stimuli in sensor-space MEG would correspond to coding of that information within the MDN. The absence of any stimulus information in the decoding analysis of Epoch 2 raises the possibility that the MDN's relevant feature coding was not critical for the sensor-space findings. Instead, dependence between the MDN and visual cortex shown through brain stimulation could reflect that rule information coding in the MDN, and not coding of the relevant stimulus feature, is important for top-down attention. Relevant stimulus coding in the MDN could be non-essential, perhaps what is selected and feedforward from ventral visual cortex, and could drop away without harming

preferential coding elsewhere. On the other hand, failing to detect relevant stimulus information within the MDN during Epoch 2 certainly does not rule out that the information was present – particularly considering that colour and shape information appeared to be exchanged between the MDN and ventral visual cortex in both task phases. Like any null result, this should be interpreted with caution.

3.4.4. Conclusion

This study demonstrated one way that we can probe the timecourse of information flow between two functionally distinct brain systems. It extended the findings of Chapter 2 into source space, showing distinct time-courses of information coding within ventral visual cortex and the MDN. This laid the groundwork for a novel analysis, which probed how information was passed between ventral visual cortex and the MDN throughout a multi-step task. It showed that colour and shape information are quickly fed forward, with feedback following shortly after, commensurate with a role for the MDN in maintaining information through a working memory delay. The dynamic flow of information throughout multiple task epochs once again highlights how flexibly we can direct our brain's resources as we move through parts of a task. In turn, precisely capturing what information we as observers can read out from the brain goes hand in hand with understanding how that information is communicated and transformed by the brain. The approaches that I have showcased here, along with the challenges I faced, are part of a small but exciting step towards understanding how the brain co-ordinates adaptive behaviour.

Chapter 4

Characterising Everyday Behaviour: Is Temporal Modularity Normal?

Many of our everyday tasks require us to integrate information from multiple steps to make a decision. Dominant accounts of flexible cognition suggest that we are able to navigate complex tasks by attending to each step in turn, yet few studies measure how we direct our attention to immediate and future task steps. Here, I used a two-step task to test whether participants are sensitive to information that is relevant to a future task step. Participants viewed two displays in sequence. Each display contained two superimposed moving dot clouds. On each trial, cues indicated the colours of the relevant dot clouds (for example, "orange, then blue"). Participants reported the average direction of the two target dot clouds at a response screen. In a subset of trials, I presented a "decoy" distractor: the second cued target colour appeared as the distractor in the first display. I tracked how this future-relevant distractor influenced responses, compared to never-relevant, recently relevant, and globally relevant distractor baselines. Across four experiments, I found that responses reflected what was immediately relevant, as well as the broader history of the distractors. Relevance for a future task step did not reliably influence attention, suggesting that attention in multi-step tasks can be precisely directed to each step in turn.

4.1. Introduction

In our everyday tasks, we often need to perform multiple, related, processing steps. For instance, to decide what fruit you would like to buy at the supermarket, you might find the fruit section, focus on the oranges, then hold the price of oranges in your mind while you look at the apples. Making an informed decision relies on having focused on each fruit in turn, maintained the relevant pricing information, and integrated it to select the best option.

This challenge of selecting and integrating multiple features is amplified by the fact that multiple goal-relevant features can be present at once. In the example of choosing fruit to buy, you may know from the start that you intend to choose either apples or oranges, so you need to extract some information about each. You then need to decide where to direct your focus at each moment so that information about the apples is not mixed up with information about the oranges.

Brain imaging, computational modelling, and behavioural research have highlighted that our capacity to self-impose periods of focus, or "attentional episodes", around parts of a task, could be essential for fluid intelligence (Duncan, 2010, 2013; Duncan et al., 2017, 2020; Yang et al., 2019). This ability in turn powerfully predicts performance across a wide range of tasks (Pagani et al., 2017; Primi et al., 2010; Wray et al., 2020; Wrulich et al., 2014). Focusing exclusively on simple task parts could be especially important in novel or difficult tasks, where we cannot keep the whole task in mind (Duncan, 2013; Laird et al., 1986).

Conversely, approaching a sequential task by exclusively focusing on what is relevant at each moment could be burdensome. Neural network simulations show that engaging deeply with the current task can make it difficult to reconfigure the network and engage in a new task (Musslick et al., 2018). This is reinforced by a wealth of behavioural research on task switching, in which changing tasks typically produces slow and error-prone responses (Kray, 2006; Longman et al., 2017; Mayr & Keele, 2000; Meiran et al., 2000; Monsell, 2003; Rogers & Monsell, 1995). However, these findings only indirectly address the question of whether we proactively direct

our attention to each task part in turn, or to all relevant task features. As yet we do not have a reliable measure of how much we attend to non-immediate task parts.

In this study, I investigated people's ability to form periods of focus around subsets of task-relevant information. I presented an integrated decision-making task (Rangelov & Mattingley, 2020), modified to test how far responses are biased by a target that appears outside the period in which it is relevant. Across four experiments, I found that attention was strongly biased by how often an item was relevant over the course of a few minutes. Once I accounted for this effect, however, attention was not reliably biased toward targets that appeared outside their period of relevance, that is, information that would be relevant in the subsequent epoch. I propose that performance in simple sequential tasks is characterised by awareness of features that are relevant in the broader experimental context, together with sharp focus on each task part in turn.

4.2. Experiment 1

4.2.1. Methods

4.2.1.1. Participants

I recruited 25 participants (age=26.48±5.25, 17 female, 8 male) from the volunteer panel at the MRC Cognition and Brain Sciences Unit (MRC CBU). Participants could only access the study if they had previously reported that they were fluent in English, had normal or corrected to normal vision, had normal colour vision, and were between the ages of 18 and 65. Two participants failed to meet the criteria for behavioural performance (see below) and were excluded, leaving 23 participants in the final sample (age = 26.39±4.97 years, 15 female, 8 male). Participants gave written informed consent before participating. All participants were given £6 per hour or pro rata, in minimum increments of £1.50. The project

was approved by the Psychology Research Ethics Committee at the University of Cambridge (PRE.2018.101).

4.2.1.2. Task

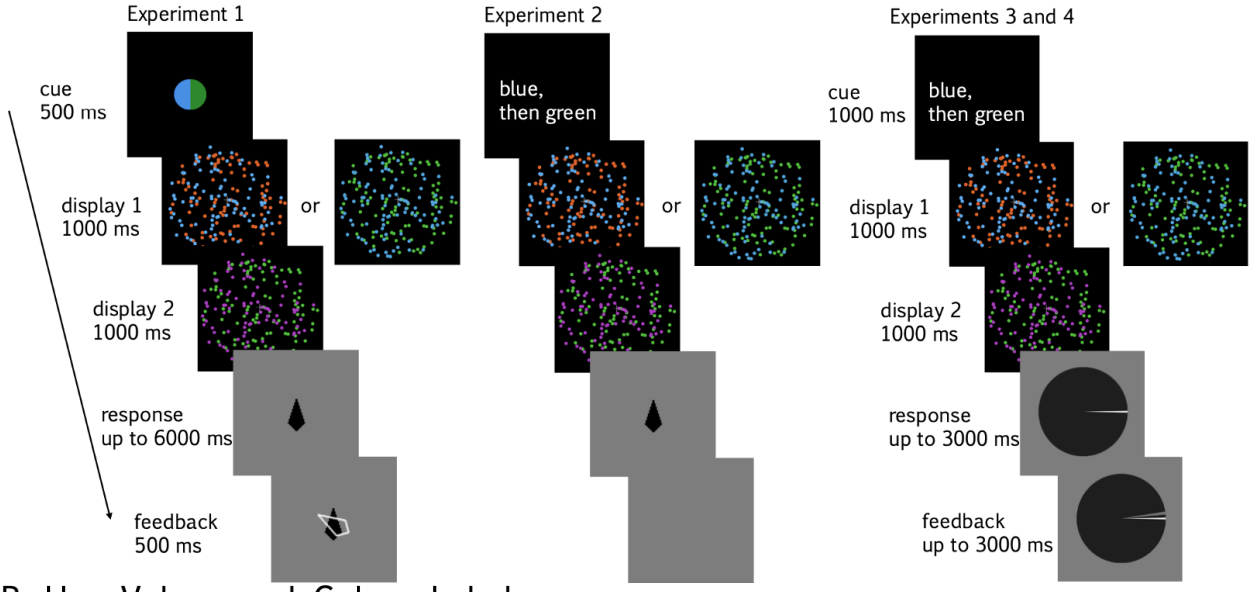
Participants completed a computer-based behavioural task. Each trial began with a fixation cross for an intertrial interval between 500 and 1500 ms. A circular cue, presented for 500 ms at fixation, showed the target colours for the trial. The first target colour was shown in the left semicircle and the second target colour in the right.

The cue was followed by two 1000 ms displays. Each display contained two moving dot clouds, one in a cued colour (the target). Participants remembered the motion direction of the target dot cloud on each display. After the dot displays disappeared, participants reported the average motion direction of the two target dot clouds by moving a black, obelisk-shaped response dial with a joystick. If participants did not respond within a 6000 ms time limit, the task continued automatically. After the response, or the maximum response time, a white feedback dial appeared imposed over the response dial, illustrating the correct response in a contrasting colour for 500 ms.

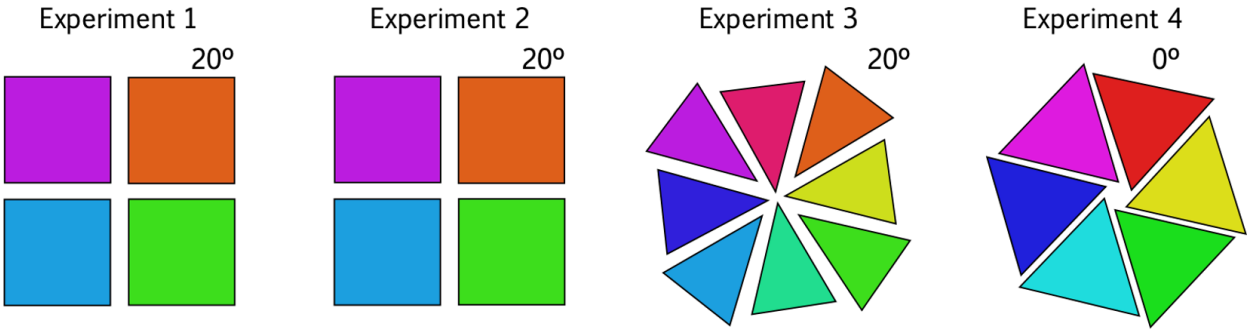
I embedded two core conditions by manipulating the task-relevance of the un-cued dot clouds (the distractors). In the baseline condition, Distractor 1 or Distractor 2 were presented in a non-target colour that was irrelevant both to the current trial and to the task as a whole, that is, baseline distractors were "consistent non-targets". In the decoy condition, Distractor 1 was presented in the anticipated Target 2 colour (Figure 1, left panel). Cues were blocked, while baseline and decoy trials were randomly mixed within each block with equal proportions of each condition. As I was primarily interested in cognitive processing of stimuli that would be relevant in the future, the decoy distractors, when present, always appeared in the first epoch.

Task and Stimuli for Experiments 1 to 4

A. Example Trials



B. Hue Values and Colour Labels



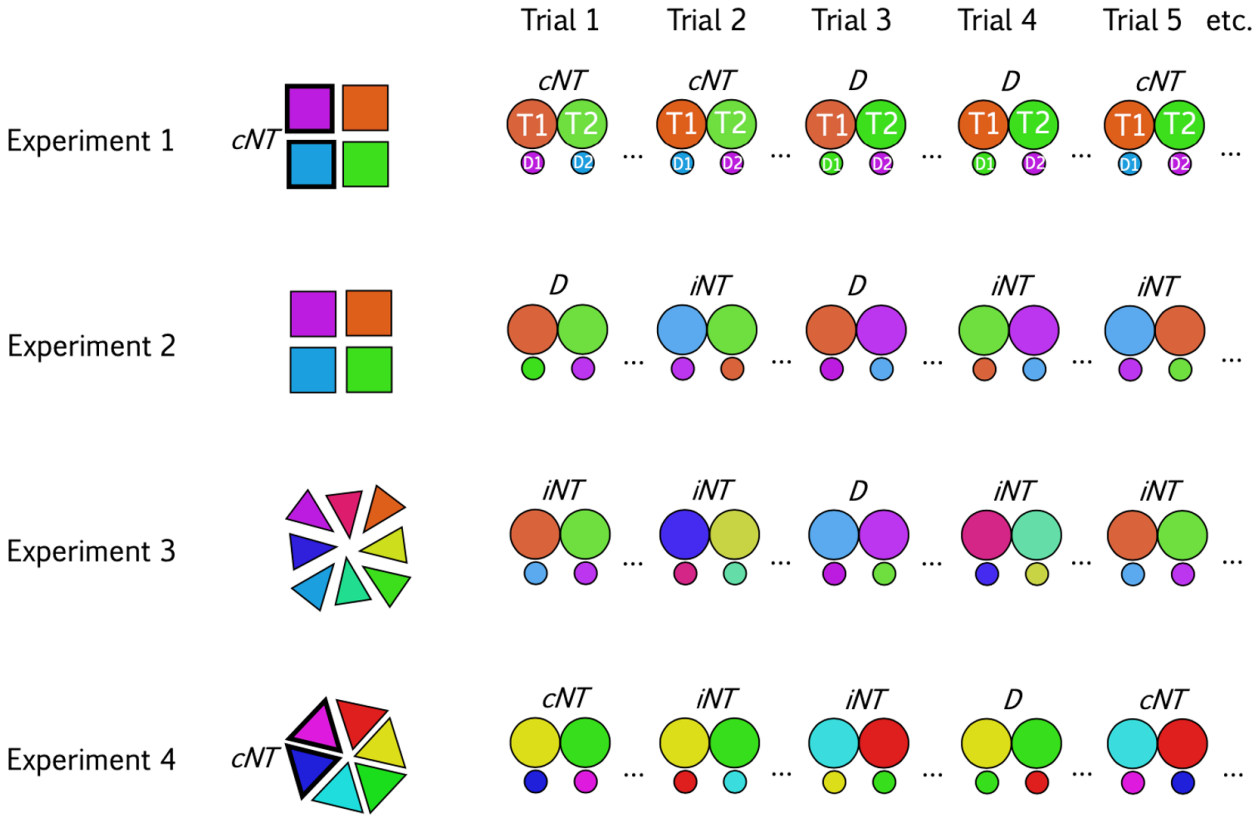


Figure 1. (A) Example trials for Experiment 1 (left), Experiment 2 (middle), and Experiments 3 and 4 (right). The panel for Experiment 1 also shows example stimuli for baseline and decoy conditions. Experiments 1 and 2 had the same timings. (B) Colours for each experiment. The hue value is marked for the first colour in each set moving clockwise from 0º in HSL colour space. Subsequent hues in the set were equally spaced around the circle (i.e., 90º for Experiments 1 and 2, 45º for Experiment 3, and 60º for Experiment 4). Colour names below the colour sets were the labels used for written cues, where applicable. For Experiment 3, outer colours formed Set 1, and inner colours Set 2. (C) An example trial sequence for each experiment. Large circles represent Target 1 (left) and Target 2 (right) colours, with small circles representing distractors. Trials are labelled D for decoy, and cNT or iNT to indicate a baseline trial with never-relevant colours ("consistent non-targets") or target colours ("inconsistent non-targets") as distractors. Colours that are never relevant in the example sequence are labelled on the colour wheel where applicable (Experiments 1 and 4).

4.2.1.3. Stimuli

I generated random dot stimuli in PsychoPy (Peirce, 2007; Peirce et al., 2019). Dot clouds consisted of 80 dots, of which 40% moved in a coherent direction

(i.e., contained signal) and the remaining 60% followed a random walk. This ensured that participants could not rely on individual dots to indicate the motion direction, as any dot had a .4 probability of containing signal in one epoch and .16 probability of containing signal across two epochs. Dots moved at 19 degrees of visual angle per second. Each dot was randomly assigned a duration between 0 and 100 ms. When a dot reached its assigned duration, it was replaced by a new dot at a random location.

I selected a range of colours to distinguish superimposed dot clouds. All colours were set at 75% saturation and 50% luminance in HSL colour space (hue, saturation, luminance). I selected four colours with hues 90° apart, starting from 20º (Figure 1, Panel B). I adjusted these colour values for colour distortion to ensure that saturation and brightness were matched on the lab computer monitors. Motion directions were selected from 0 to 359° in steps of 1°. Target and distractor motion within an epoch (e.g., Target 1 and Distractor 1) as well as the two target motions between epochs (e.g., Target 1 and Target 2) could differ by between 30 and 150º. This ensured that the average of the target dot clouds' motion direction gave an unambiguous answer (that is, target clouds were never separated by 180º). Stimuli were presented on a 19-inch Dell 1908FP LCD monitor at 1280x1024 resolution and refresh rate of 60Hz. Stimuli subtended 15.94 degrees of visual angle at a viewing distance of 50 cm.

4.2.1.4. Procedure

Participants were seated comfortably at a computer monitor. All participants gave basic demographic information in a pre-experiment questionnaire (age, sex, vision). They then began training on the task. This consisted of three blocks of 24 trials each. In the first block, participants saw only a single epoch of coloured dots and reported the target motion direction presented in that epoch while ignoring concurrently presented distractor motion. In the subsequent two blocks, they practiced the dual-epoch task. Cue durations began at 1000 ms in the first two

blocks, reducing to 500 ms in the third training block, to ease participants into the task. The experimenter discussed errors with the participant between training blocks and explained any aspects that were unclear.

Following training, participants began the main session. They completed 12 blocks of 384 trials. Each participant was presented all possible cues. That is, each of the four colours was paired once with each other colour (orange with green, orange with blue, etc.) to form six cues. These cues were reversed (green with orange, blue with orange, etc.) to form six more cues and create 12 blocks. Block sequence was randomised within and between participants.

4.2.1.5. Analyses

Exclusions. Participants with incomplete data, or whose mean absolute error exceeded 45°, were excluded from the analysis. Two participants were excluded for high errors.

Ordinary least-squares regression. For each participant, I regressed their responses on the true motion directions of the target and distractor dot clouds. Since motion directions were angles, and real-valued angles are circular (-π is the same as π), I represented angles for the predictors (design matrix, X) and responses (R) as complex values (Eq. 1). I then fit a linear model (Eq. 2), using ordinary least squares regression (Eq. 3) to estimate the weight vector ($\hat{B}$) that optimally fit the four predictors to the responses.

$$\tilde{X} = \cos X + \sin X \times 1i \quad (1)$$

$$\tilde{R}_{N\times 1} = \tilde{X}_{N\times 4} B_{4\times 1} + E_{N\times 1} \quad (2)$$

$$\hat{B} = (\tilde{X}' \times \tilde{X})^{-1} \tilde{X}' \times \tilde{R} \quad (3)$$

The weight matrix, like the predictors and responses, was complex valued. I took the absolute values of the weight vector (Eq. 4). These values represent an expansion factor for each predictor, so that a high weight for a given predictor

represents a strong influence on the response. I refer to these absolute weights as "decision weights" because they reflect an estimate of how the target and distractor dot motions influenced participants' choices at the response screen.

$$|\hat{B}| = \sqrt{real(\hat{B})^2 + imag(\hat{B})^2} \quad (4)$$

Permutation testing. Next, I obtained single-subject measures of attention to Target 1 and Target 2. Decision weights could not be negative, meaning that chance values could be greater than zero but not below. Rather than test the decision weights against a normal distribution, which assumes that chance values are distributed around the null value, I tested the decision weights against individual-specific, permutation-based null distributions. I took each participant's predictor values and randomly permuted them, so that target and distractor dot directions for a given trial could be assigned to any trial. I calculated the decision weights as above for these permuted data, and repeated the process until I had decision weights for 10,000 permutations. This formed the empirical null distribution of decision weights for that participant. I then compared the decision weights for the correctly labelled (unpermuted) data to the null distribution. Decision weights that exceeded the 95th percentile of the null distribution were considered unlikely under the null hypothesis, and reliably different to zero.

The main purpose of this analysis was to confirm that participants were doing the task as instructed, using the target decision weights as a secondary measure of accuracy. In particular, participants could respond somewhat accurately by attending to one of the two targets on each trial. A strategy of attending to either Target 1 or Target 2 would make it difficult for us to judge how a decoy distractor, presented in the Target 2 colour, influenced focus on Target 1. Thus, I used these individual-level tests to exclude participants whose baseline target weights were not both reliably different to zero.

Group baseline analysis. Next, I compared baseline target and distractor weights to zero at the group level. As individual participants were excluded if their target weights did not both exceed zero, target weights should always exceed zero across the group. However, just as target decision weights provide information about the source of accuracy, distractor decision weights can provide information about the source of inaccuracy: in this case, whether errors were random, or reflected attentional capture by Distractor 1 or Distractor 2. Understanding this could provide context for comparing conditions, for which I expected distractor weights to play a role. Thus, I tested whether each decision weight was greater than zero in the baseline. For each condition and dot cloud, I collated the individual-level null data and estimated a cumulative distribution for the null across the group. If the group's average decision weight exceeded the null value whose probability was 5% in the null distribution, the decision weight could be considered reliably different to chance at the alpha = .05 level. To account for multiple comparisons, I applied a family-wise correction. I set alpha at .013 (0.05/4) for one-sided comparisons of each decision weight to zero.

Condition comparisons. Finally, I compared target and distractor weights between the conditions with a paired t-test on baseline and decoy conditions. If the future-relevance of an upcoming target affected the current attentional set, I expected that decision weights would be higher for Distractor 1 in the decoy condition relative to baseline. On the same assumption, I expected lower weights for Target 1 in the decoy condition relative to baseline. Consequently, I planned two one-sided t-tests, setting alpha at .025. For completeness, I analysed Target 2 and Distractor 2 weights in the same way, but do not interpret these data as I did not have targeted hypotheses for these comparisons.

4.2.2. Results

4.2.2.1. Baseline decision weights

All individual subjects' baseline target weights exceeded zero, compared to their individual permutation-based null distributions. Baseline target and distractor decision weights across the group were also reliably different from zero (all $p<.001$, alpha=.013).

4.2.2.2. Decoy effect on decision weights

I predicted that if momentary attentional focus was affected by the upcoming attentional set, Target 1 weight would decrease, and Distractor 1 weight would increase, in the presence of a 'decoy' distractor.

Indeed, Target 1 weights decreased, and Distractor 1 weights increased, when the distractor in Epoch 1 was the anticipated Target 2 colour, relative to a non-target baseline (Figure 2). That is, responses were less influenced by relevant information, and more influenced by irrelevant information, when the distracting information was relevant to the subsequent task epoch. I also observed a similar pattern for Target 2, whereby having just ignored a particular colour in the preceding epoch (decoy condition) reduced weighting of that colour in the current epoch.

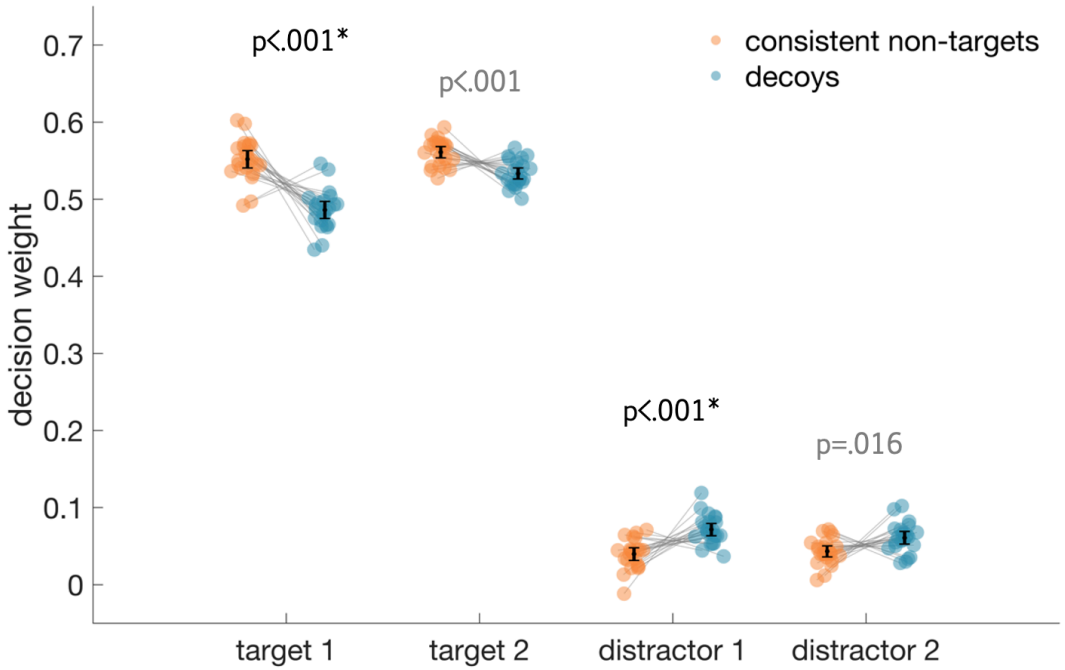


Figure 2. Decision weights for target and distractor dot clouds for Experiment 1, shown separately for consistent non-target baseline (orange) and decoy (blue) conditions. Decision weights are normalised for visualisation, and show within-subject effects centred on the group mean (Cousineau, 2005). Grey lines between conditions connect individual participants. Black bars indicate the 95% confidence interval. P-values are derived from paired t-tests and are uncorrected for multiple comparisons. Between-condition comparisons for Target 2 and Distractor 2 are in grey. Statistically significant condition differences are marked with an asterisk.

These findings are consistent with the idea that current attentional focus is influenced by the upcoming relevance of related information; that is, that attention is not allocated in strictly discrete temporal chunks. It appeared that participants would down-weight relevant information, and up-weight irrelevant information, when the irrelevant information was presented in the anticipated target colour of the subsequent epoch. This pattern also extended to a second task epoch, suggesting that suppressing a future-relevant distractor has negative consequences for attending to that stimulus soon after.

However, distraction by the decoy in this experiment could also be driven by factors besides its imminent relevance in the subsequent epoch. First, cues on each

trial displayed the target colours on screen immediately before the trial, meaning that the decoy distractor was both future-relevant and recently seen. Distraction by a decoy because it was cued is difficult to disentangle from distraction due to its future relevance for the task, since its future relevance is by definition determined by the cue. However, the distinction is important if we want to argue that decoy effects here specifically reflect how we manage future-relevant information.

Second, the task used a single cue (for example, "attend to orange, then blue") for a full block. This meant that the decoy was both the anticipated Target 2 colour, and the Target 2 colour that participants had responded to on the previous trial. It was possible, therefore, that higher Distractor 1 weights for decoys relative to baseline reflected effects of recent history rather than preparation for the immediate future. Finally, the baseline distractor colours were constant over each block, making them globally less relevant to the task than the colours used for targets and decoy distractors. Therefore there were three possible explanations for the decoy effect in Experiment 1: carry over of attentional set from the previous trial (recent relevance), up-weighting colours that are globally relevant across the task (global relevance), or preparing to attend to the upcoming target (future relevance). I carried out a series of experiments to isolate the role of future relevance from these alternate explanations.

4.3. Experiment 2

Although I saw the anticipated effect of decoys in Experiment 1, I could not disentangle the contribution of recent and global relevance from the decoy's future relevance for the trial. I designed Experiment 2 to isolate the influence of future-relevance on attentional processing, by equating recent and global relevance between decoy and baseline conditions.

4.2.1. Methods

4.2.1.1. Participants

I recruited an independent sample of 48 participants (age=33.13±14.95, 30 female, 18 male) for this study, through the MRC Cognition and Brain Sciences Unit Volunteer Panel.

4.2.1.2. Task

As in Experiment 1, I presented distractors in a colour that was irrelevant for the trial (baseline) or in the anticipated target colour (decoy). All task and analysis features remained the same as in Experiment 1, except for three changes, which I introduced to equalise the broader task-relevance of baseline and decoy distractor colours.

First, I presented colour cues in text (for example, "purple, then blue"; see Figure 1, Panel A, centre). I did this to ensure that participants actively constructed their attentional biases, and did not simply respond to targets and decoys because they had recently seen those colours in the cue.

Second, I reduced the opportunity for positive trial-to-trial carry-over of attention (recent relevance), and equalised it between decoy and baseline conditions. In Experiment 1, trials were blocked, so that a decoy distractor in the first epoch – the second target colour on the current trial – was also the second target colour from the previous trial. To minimise this, in Experiment 2 I randomly selected a cue on each trial, so that target colours varied within a block. Randomly assigning each colour to be a target or distractor on each trial meant that carry-over of attentional biases from trial-to-trial could be positive or negative. That is, carry-over could equally help or hinder attention to targets and decoys. More importantly, the opportunity for carry-over was now matched between baseline and decoy conditions. I reasoned that, if participants still upweighted the decoy, this could no longer be easily explained by the decoy's recent history of relevance.

Third, I manipulated the global relevance of the baseline distractors across the task. In Experiment 1, baseline distractor colours were consistent across a block, and were never used as target colours ("consistent non-targets"). In this experiment, the random cue sequence meant that target colours on one trial could become baseline distractor colours on the subsequent trial ("inconsistent non-targets"). Thus, participants could no longer benefit from a tendency to up-weight globally relevant colours, as the same colour appeared equally as often as a distractor or target in the baseline condition.

4.2.2. Results

4.2.2.1. Baseline decision weights

Eight participants were excluded for high errors, leaving 40 participants in the final sample (age = 32.38±15.37, 23 female, 17 male). All remaining subjects' baseline target weights exceeded zero, compared to their individual permutation-based null distribution. Baseline target decision weights across the group were also reliably different from zero ($p<.001$). In contrast to Experiment 1, baseline distractor decision weights were no longer statistically greater than zero.

4.2.2.2. Decoy effect on decision weights

In the blocked designed of Experiment 1, decoy distractors were more distracting than consistent non-target distractors. Here, I used a random trial sequence to ensure that baseline distractors were matched to the decoys' globally relevance, and that decoys would not uniquely benefit from positive carry-over of attention from trial to trial, so that baseline and decoy distractors only differed in whether they were an anticipated target (future relevance). In contrast to Experiment 1, I now saw no reliable evidence of a difference between decoy and baseline conditions in either Target 1 or Distractor 1 weights (Figure 3). This

suggests that the decoy results in Experiment 1 may have reflected the global or past relevance of the decoy colour, rather than its future relevance.

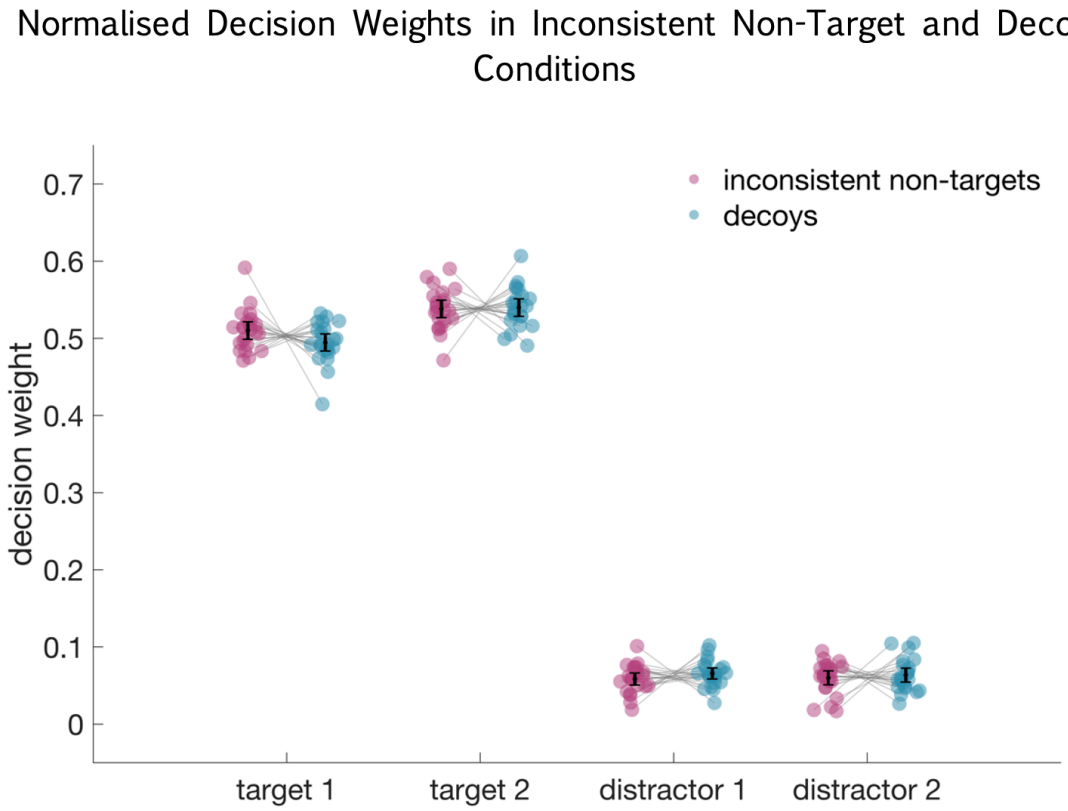


Figure 3. Normalised decision weights for target and distractor dot clouds in Experiment 2, shown separately for inconsistent non-target baseline (pink) and decoy (blue) conditions. Black bars indicate the 95% confidence interval. P-values are derived from one-sided paired t-tests and are uncorrected for multiple comparisons. Between-condition comparisons for Target 2 and Distractor 2 are in grey. There were no statistically significant differences between conditions.

On the other hand, failing to detect a decoy effect in Experiment 2 does not completely rule out the possibility that future-relevant information captures attention. Another possible explanation is that I failed to detect awareness of future-relevant task information in Experiment 2 simply because participants could no longer direct their attention to each display as it appeared. Although target weights were reliably different to zero, many participants spontaneously reported that they struggled to redirect attention to a new pair of target colours. Target weights were indeed lower in Experiment 2 compared to Experiment 1 ($p=.038$;

Figure 4), based on an age-matched sub-sample (n=23, age=25.609±4.846, 13 female, 10 male). Participants could have compensated for the pressure to quickly switch focus between trials by taking a more retroactive approach to the task, relying on the stimuli to engage cognitive control rather than proactively maintaining the targets in mind (Braver, 2012). In a proactive mode, participants would adjust their attentional sets in advance so that only the task-relevant stimuli were encoded and maintained. In a retroactive mode, by contrast, participants would encode both relevant and irrelevant stimuli, and later, when presented with a response display, try to retrieve just the relevant stimuli. The latter might be a low-cost strategy (Braver, 2012), which could make it a natural choice when the participants were overwhelmed by trial-to-trial switching demands. Under this framework, a more proactive approach in Experiment 1 could have enabled participants to quickly enhance anticipated targets and suppress consistent non-target distractors (in line with high performance in the baseline condition), but could fail when conflicting information was unexpected (in line with lower performance in the decoy condition). Thus, a proactive cognitive control strategy in Experiment 1 could produce low awareness of baseline distractors and high awareness of decoys, while a reactive control strategy in Experiment 2 could produce a stable, higher level of distraction across both conditions. If proactive and reactive control strategies produced the different results of Experiments 1 and 2, finding no decoy effect in Experiment 2 could still be consistent with a true decoy effect in Experiment 1.

Figure 4. Average baseline decision weights for Targets 1 and 2 in Experiment 1 (grey) and an age-matched sub-sample of Experiment 2 (green). Black bars indicate the 95% confidence interval. The p-value for a two-sided, independent-samples t-test is shown below the average target weights (far right). I show baseline decision weights for target and distractor dot clouds in Experiment 1 (as in Figure 2) and the Experiment 2 subsample for completeness.

In addition, although there was no statistical decoy effect in Experiment 2 (despite a sample nearly twice that of Experiment 1) numerical trends were in the predicted direction. Within the sample, target weights tended to decrease and Distractor 1 weights tended to increase on decoy trials relative to baseline. Small and inconsistent changes do not support the conclusion that decoys are uniquely distracting, but neither do they rule it out. Given the ambiguity inherent in null effects, it is especially important that we question how robust they are. Considering these two limitations, I designed a third experiment to test the robustness of these findings, using a simplified task to promote proactive cognitive control.

4.4. Experiment 3

So far, we have seen that future-relevant information may draw attention away from currently relevant information (Experiment 1), but that this bias towards future-relevant information is unreliable when we account for the distractor's recent and global relevance (Experiment 2). However, Experiment 2 was difficult, with many participants spontaneously reporting that they struggled to switch focus between cues from trial to trial. I was also wary of over-interpreting the null finding in Experiment 2 without ensuring that I could replicate it. Therefore, in Experiment 3, I aimed to test the robustness of my finding in Experiment 2, using a simplified task and controlled trial sequence.

4.4.1. Methods

4.4.1.1. Participants

I recruited an independent sample of 78 participants (age=28.62±9.69, 27 female, 51 male) for this study. Participants were recruited through the Prolific research recruitment portal (https://www.prolific.co/) and completed the task online. As in the previous experiments, the study was only advertised to participants who had previously reported that they were fluent in English, had normal or corrected to normal vision, had normal colour vision, and were between the ages of 18 and 65. Additionally, participants were asked to only join the experiment if they had access to a desktop computer or laptop (not a tablet or phone). All participants gave informed consent by clicking an online form. Ethical approval and payment remained the same as for Experiments 1 and 2.

4.4.1.2. Task

As in Experiment 2, I used the same colours equally as targets and distractors (inconsistent non-targets), so that participants could not benefit from a

global tendency to up-weight target colours. All methods were the same as in Experiment 2, with the following changes.

In Experiment 2, cues appeared in a random sequence, so that a decoy on the current trial was not consistently the second target on the previous trial. However, all four colours appeared on each trial, meaning that attentional biases toward each colour could carry over from trial to trial, albeit in positive (n and n-1 have the same targets), ambiguous (n and n-1 share only one target), and negative (n and n-1 have different targets) ways. For this experiment, I selected eight colours with hues 45° apart. I divided the eight colours into two overlapping sets, with hues 90° apart within each set (Figure 1B). I chose maximally distinct hues to ensure that, even if they were slightly altered by online participants' computer monitors, they would be easily discriminated. I interleaved the two colour sets to create a train of four trials. Each train comprised (1) a trial from Set 1, for example, "orange, then blue"; (2) a trial from Set 2, for example, "purple, then pink"; (3) the inverse of Trial 1 (orange and blue as distractors); and (4) the inverse of Trial 2. I also increased the cue duration from 500 to 1000 ms to further support participants to prepare their attentional set on each trial.

I did not point out the trial sequence to participants, but they could learn it, explicitly or implicitly. Each block consisted of a single train, so that the exact combination of target and distractor colours repeated every five trials. Thus, I reduced the difficulty inherent in random cueing while making sure that the same colours were targets and distractors in both conditions.

Participants completed six blocks. Set 1 and Set 2 colours each formed six colour pairs (four colours combined without repetition; orange and blue, orange and pink, and so on), and each block contained a target colour pair from each set, so that every individual completed one block with each colour pair as the target colours. For a given block, I selected target colour pairs from each set to maximally vary within subjects what pairs were combined across sets. For example, if in Block 1 the Set 1 target colours were orange-green, and Set 2 target colours were dark blue-yellow,

orange and dark blue would not be selected together as the Target 1 colours in a subsequent block, and the same for green and yellow. Combinations across colour sets were further randomised across participants.

These changes were designed to have two effects. First, I intended the extra preparation time and predictable trial sequence to simplify the task (Altmann, 2004; Longman et al., 2017; Meiran et al., 2000; Vandierendonck et al., 2010). Second, I intended the interleaved trials from distinct colour sets to further reduce the possibility of attentional carry-over, and the difficulty associated with conflicts in stimulus relevance from trial-to-trial (Gilbert & Shallice, 2002; Waszak et al., 2003). Relative to Experiment 1 these features again acted to isolate the effect of future relevance from possible effects of recent or global relevance. As before, I predicted that if momentary attention is affected by keeping track or of planning for a future attentional set, information relevant to a future task part (the decoy) would attract attention and disrupt performance, relative to the baseline distractors.

4.4.1.3. Stimuli

I generated random dot stimuli using bespoke JavaScript code and the jsPsych library (de Leeuw, 2015). I matched the dot cloud parameters (number, coherence, etc.) to Experiments 1 and 2.

For the response screen, I adapted the response dial to be compatible with jsPsych, and to be easy to see and manipulate with a mouse or touchpad. I used a grey circular dial with a white pointer (Figure 1, right). Feedback was presented as a light grey pointer overlaid on the white response pointer. Participants moved the pointer with their mouse or touchpad, either by dragging the pointer or by clicking directly to where on the circle it should go. I reduced the maximum response time from 6000 ms to 3000 ms, as response times greater than this in Experiments 1 and 2 were rare.

Stimuli were presented on participants' personal computers. I estimated each participant's refresh rate by engaging a browser interface method for updating

content on each screen "repaint" (typically the same as a screen refresh; see https://developer.mozilla.org/en-US/docs/Web/API/window/requestAnimationFrame), taking a timestamp on each repaint, and extracting the average time between repaints over five seconds. I also asked participants to match the size of an on-screen box to the size of a credit card. This gave us the pixel dimensions of an object with known true size, allowing us to estimate their screen resolution and match stimulus sizes across participants and devices. I adjusted the distance that each dot travelled on each frame to match speed across refresh rates and screen dimensions. As in the previous experiments, dot clouds covered a 14 cm circular area, so that the stimuli subtended 15.94 degrees of visual angle at a viewing distance of 50 cm. I asked participants to position themselves 50cm from the screen, but as I was unable to see participants during the task, the effective visual angle of the stimuli could have differed based on participants' head position. Due to participants accessing the study remotely, I also could no longer test how their screen distorted the colours. This meant that, although I set all colour values to 75% saturation and 50% luminance, true saturation and luminance may have varied across colours. I chose hues that were maximally distant to each other, to reduce the likelihood that two colours would become indistinguishable because of a screen's colour distortion. However, I presented all colour pairs to each participant and conducted all primary analyses within-subjects, so that between-subject variability in displays would minimally influence the results.

I tested all stimulus presentation scripts locally and through JATOS (Lange et al., 2015), a web application that allows researchers to interact with a web server (for example, generate a study link and download data) through a graphical interface.

4.4.1.4. Procedure

I asked participants to sit 50 cm away from the screen, and to ensure that their screen faced them directly before beginning the task. I also presented a labelled image of the two colour sets (Figure 1B) prior to training, to remove any ambiguity about which cue word indicated which colour.

In Experiments 1 and 2, participants completed three training blocks of 24 trials each, with the cue duration beginning at 1000 ms in blocks one (single-epoch) and two (dual epoch) and reducing to the core session's 500 ms in training block three. In the current experiment, the cue duration remained at 1000 ms throughout training and core sessions. Consequently, the third training block was not necessary to introduce a new cue duration. I therefore reduced the number of training blocks from three to two, but maintained the total number of trials by increasing the number of trials per block from 24 to 32.

Data quality. Comparisons of web- and lab-based data quality have demonstrated that online platforms can be appropriate for cognitive psychology experiments (Germine et al., 2012), and that stimulus presentation times and participant compliance can compare positively to lab studies (using jsPsych and Prolific: Anwyl-Irvine et al., 2021; Peer et al., 2017). However, since I was unable to discuss the task with participants or adaptively respond to any issues they experienced, I implemented some additional measures to encourage and track data quality. Participants were encouraged to re-read the instructions for each training block. I could not offer individualised explanations or answer questions. Instead, I introduced additional tests to filter out participants who misunderstood the task or did not seriously attempt it. First, I included a catch question in the pre-experiment questionnaire ("What year is it?"), which blocked participants from continuing if they answered incorrectly. Next, I used errors on the single-epoch training block to identify underperforming participants. Whereas an error on dual-epoch trials could reflect bias toward one target over the other, errors on single-epoch trials showed that the participant had not accurately identified or perceived the target, making

these errors a fair indicator of whether they understood and seriously attempted the task. Participants whose median absolute error in the first training block exceeded 45° (that is, most responses fell outside the quadrant for the correct response) were prevented from continuing and paid for their time. Lastly, participants who did not respond, or who responded with the default response dial position, for three sequential trials were shown a warning screen prompting them to respond within the time limit. Participants who received this warning three times, either during training or during the core session, were prevented from continuing and paid for their time. These measures were important to reject data from participants who responded with minimal effort, and also to ensure that participants who were struggling did not continue to struggle through the intensive experiment session.

4.4.2. Results

4.4.2.1. Baseline decision weights

Eleven participants were excluded for high errors, leaving 67 participants in the final sample (age = 27.67±8.93, 24 female, 43 male). All remaining subjects' baseline target weights exceeded zero, compared to their individual permutation-based null distribution. Baseline target decision weights across the group were also reliably different from zero ($p<.001$). Baseline distractor decision weights were also reliably different to zero in Epoch 1 ($p = .002$), but not in Epoch 2 ($p = .146$).

4.4.2.2. Decoy effect on decision weights

In contrast to my prediction, and in line with Experiment 2, I observed no reliable change in Target 1 or Distractor 1 weights in the decoy condition (when Distractor 1 was the anticipated Target 2), compared to the baseline distractors (Figure 5). This was true despite low mean absolute errors (30.37º±6.56 , compared to 32.67º±6.32 in Experiment 1). This replicates the null findings of Experiment 2 with an independent sample and an easier task with predictable cues. Together, the

two experiments provide no evidence for the hypothesis that attention is specially directed towards task features that are not immediately relevant.

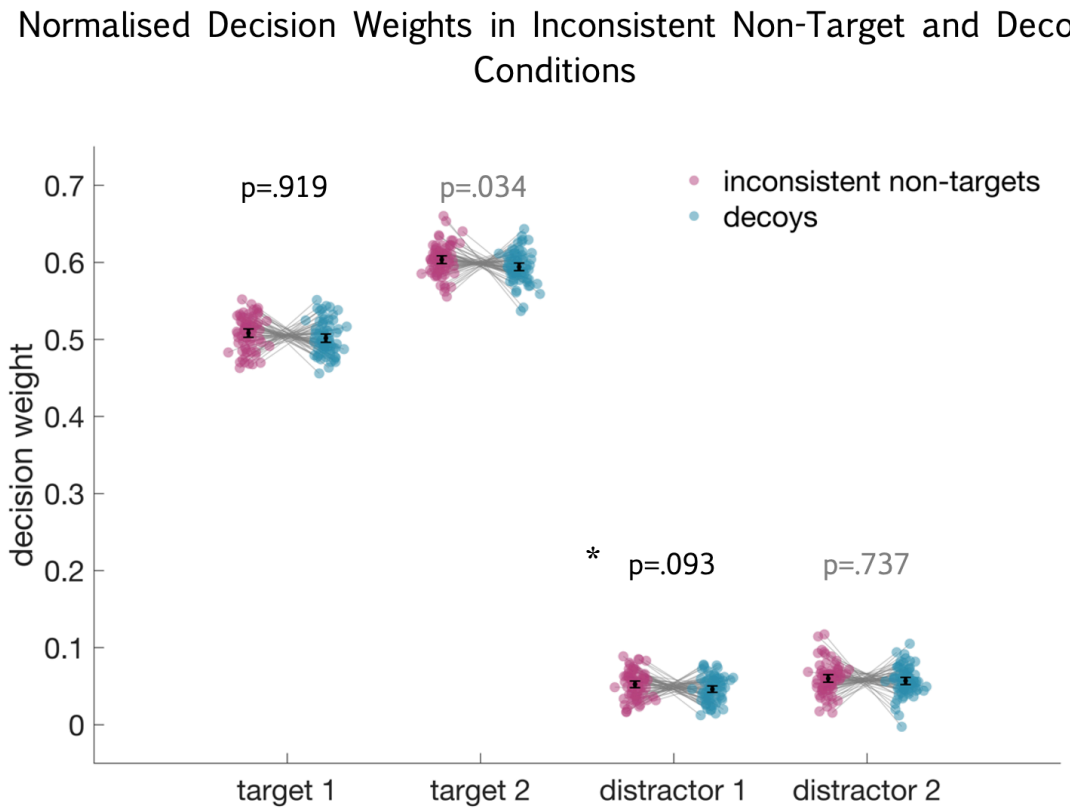


Figure 5. Normalised decision weights for target and distractor dot clouds in Experiment 3, shown separately for inconsistent non-target baseline (pink) and decoy (blue) conditions. Black bars indicate the 95% confidence interval. P-values are derived from one-sided paired t-tests and are reported without correction for multiple comparisons. The critical alpha for inconsistent non-target vs decoy conditions was .025, to account for planned comparisons of Target 1 and Distractor 1. I show outcomes for Target 2 and Distractor 2 (in grey) for completeness. There were no statistically significant differences between conditions.

In light of these findings, my new hypothesis was that future-relevant information does not capture attention beyond what we might expect from its relevance in the broader context of the task. I designed a final experiment to (a) replicate the attentional capture by a decoy, compared to a consistent non-target, that I observed in Experiment 1; (b) directly test whether people indeed upweight globally relevant information (inconsistent non-targets and decoys) over consistent non-targets (as opposed to controlling for global and recent relevance in

Experiments 2 and 3); and (c) directly compare the impact of global or recent relevance and future relevance within a sample.

4.5. Experiment 4

Experiment 4 consisted of three conditions: a baseline condition with consistent non-target distractors (cNT), a baseline condition with inconsistent non-target distractors (colours drawn from the same set as target colours; iNT), and a decoy condition. I simplified the task further relative to Experiment 3 by restricting the target/inconsistent non-target colour set to four colours, and using two (rather than four) cues in each block. As in all previous experiments, I presented every combination of colours to each participant, and randomised block sequence across participants.

In line with my new hypothesis that the broad experimental relevance of each colour determines attentional weighting, with little or no additional contribution of a colour's imminent relevance later on that trial, I now predicted that the iNT condition would elicit lower average target weights, and higher average distractor weights, than the cNT condition. Based on Experiment 1, I predicted that decision weights in the decoy condition would also reflect poorer selectivity (lower target weights and higher distractor weights) compared to the cNT condition. Following Experiments 2 and 3, I now expected that Epoch 1 weights in the decoy condition would not differ from the iNT condition. This set of results would demonstrate that momentary attentional sets are sensitive to the global relevance of distractors, but do not suffer additional interference from information that will become relevant later in the same trial.

4.5.1. Methods

4.5.1.1. Participants

I recruited an independent sample of 76 participants (age=28.92±9.08, 25 female, 50 male, 1 unreported). Participants were recruited through Prolific and completed the task online. All participants gave informed consent by clicking an online form. Ethical approval and payment rate remained the same as for Experiment 1-3. For this experiment, training was offered in a separate 15-minute session, for which participants received an additional £1.50.

4.5.1.2. Task

The task remained the same as in Experiment 3, with the following changes. I selected six colours, with hues 60º apart. For each participant, I assigned a pair of colours to be consistent non-target colours. The remaining four colours formed the target colour set. For each of six blocks, I assigned these four colours to two colour pairs. I balanced colour pairings across blocks so that each colour was paired at least once with the remaining three colours. On each trial, one colour pair was the cued target pair, and the other pair were the distractors. Colours within a pair could appear in any order when they served as distractors, but were always cued in a fixed order when they served as targets. This minimised the number of cue variations per block (set it equal to two), to reduce overall task difficulty. Trials in which the distractor pair were drawn from the held out colours formed a cNT baseline condition (similar to the baseline in Experiment 1). Trials in which the distractor pair were drawn from the target colour set formed an iNT baseline condition (similar to the baseline in Experiments 2 and 3). Decoy trials, in which the first distractor was also the anticipated target colour, made up a third condition. I believed that the iNT condition was a fairer baseline by which to isolate whether people were influenced by future relevant information, as both future relevant and

inconsistent non-target colours were globally relevant for the block. Because of this, I always used inconsistent non-target colours as the second distractor in decoy trials.

Conditions and cues were presented randomly within each block. As the task now included three conditions, I reduced the length of each of the six core task blocks from 64 trials to 60 trials, so that I could include equal numbers of trials in each condition.

4.5.1.3. Stimuli

Screen resolution checks and stimuli were as described in Experiment 3, with the exception of the reduced colour set.

4.5.1.4. Procedure

The procedure was the same as in Experiment 3, with one added data quality assessment, described below.

Data quality. In Experiment 3, participants who passed quality check criteria were free to continue immediately to the core session. In Experiment 4, participants were additionally screened for technical issues (such as lags in the presentation) or personal issues (such as misunderstanding the task, finding it difficult, or being distracted), which they reported at the end of the training session. Participants who reported issues with the task were contacted to clarify the issue before being given the link to the core session. This was done to prevent people from continuing to the core session before they fully understood the task, and so expending substantial time and energy to give unusable data. The time between training and core sessions ranged from one hour to one week. Because of the delay between training and core sessions, the core session began with a full repetition of the instructions and three dual-epoch practice trials.

4.5.1.5. Analyses

As for all previous experiments, I controlled error rate family-wise, separately for baseline against zero, and for between-condition comparisons. I predicted that Target 1 and Distractor 1 weights would not change in the presence of a decoy distractor relative to the iNT baseline. The critical alpha for the decoy effect (iNT vs decoy conditions) was .025 and t-tests were two-sided. To test the impact of global relevance, I also contrasted decision weights between iNT and cNT conditions. Here, I expected that target decision weights would decrease and distractor decision weights would increase in both epochs, in the iNT condition relative to the cNT baseline. Thus, I set alpha at .013 for four planned, one-sided comparisons. I did the same for decoy and cNT conditions, again predicting that target weights would decrease and distractor weights increase in both epochs, in the presence of globally relevant distractors.

4.5.2. Results

4.5.2.1. Baseline decision weights

Five participants were excluded for high errors. One further participant's cNT baseline target weights did not both exceed the 95th percentile in their permutation-based null distribution, meaning that their accuracy for at least one target could not be distinguished from chance. This participant was also excluded from subsequent analyses, leaving 70 participants in the final sample (age = 28.77±9.20, 22 female, 47 male, 1 unreported). Across the group, target and distractor weights in the cNT baseline condition were reliably above zero (all $p<.001$).

4.5.2.2. Decoy effect on decision weights

In line with Experiment 1, the decoy condition showed evidence of increased distraction (lower Target 2 weight, higher distractor weights) relative to the cNT baseline. However, as now predicted, and consistent with Experiments 2 and 3, decoy weights were again not reliably different to the iNT baseline (Figure 6). This experiment also directly confirmed that target weights in the iNT condition were lower, and distractor weights higher, relative to the cNT baseline. These differences were statistically reliable for three out of four comparisons across the two epochs. This supports the suggestion that attention is captured both by what is immediately relevant and by what has been relevant (recently and/or globally). I did not find evidence to suggest that attention is additionally captured by information that will become relevant imminently.

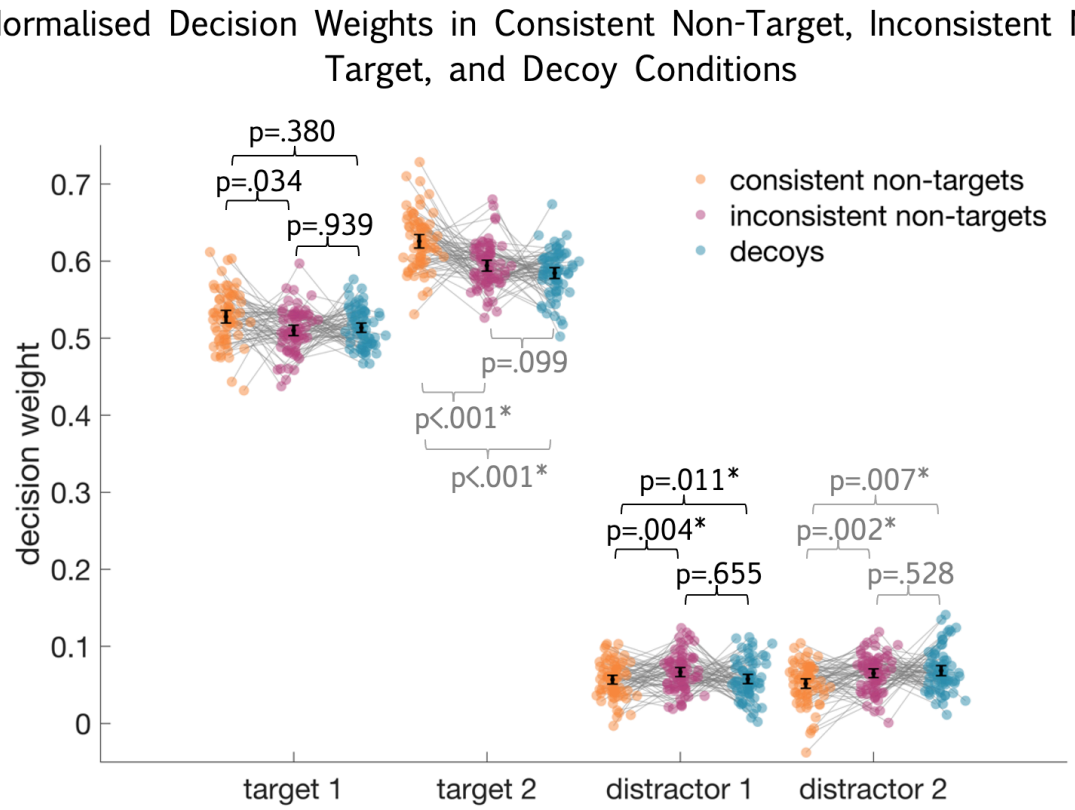


Figure 6. Normalised decision weights for target and distractor dot clouds for Experiment 4, shown separately for consistent non-target baseline (orange), inconsistent non-target baseline (pink), and decoy (blue) conditions. Black bars indicate the 95% confidence interval. P-values are derived from paired t-tests and are uncorrected. Between-condition comparisons for Target 2 and Distractor 2 are in grey. Statistically significant condition differences are marked with an asterisk.

4.5.2.3. Contribution of trial-to-trial attentional set vs global relevance

Through the previous experiments, I found that distractors substantially captured attention when they were future-relevant, recently relevant, and globally relevant (Exp 1). This effect disappeared when I controlled for recent and global relevance (Exps 2 and 3). The current experiment allows us to further separate these influences.

First, reliable differences between iNT and cNT conditions indicate that globally relevant distractors reliably capture attention, relative to distractors that

are never relevant. The global relevance effect persists even when what is relevant changes pseudo-randomly from trial to trial (with equal opportunities for recent relevance to help or hinder current focus), suggesting that the global relevance effect is unlikely to rely on trial-to-trial carry-over of attention.

However, this does not rule out a role for trial-to-trial carry-over of attentional in the current experiment, or in Experiment 1. We can quantify this effect by extracting the decision weights for cNT and iNT conditions, separately for whether targets repeat ("stay") or change ("switch"). For both cNT and iNT conditions, target weights on the current trial could benefit from a repetition ("stay"), or be disadvantaged by a change ("switch"), as the same colours could be enhanced across trials. Distractor weights could also benefit from a repetition, as the same colours could be suppressed across trials. However, the potential for distractor carry-over in switch trials would be different for cNT and iNT conditions. For the cNT condition, distractors could change or repeat in switch trials, but current trial distractors were never previous-trial targets; they were always consistent non-targets. Thus, carry-over from the previous trial could aid in suppressing the distractors (where two cNT trials occurred in sequence), but should not enhance the distractors. For the iNT condition, however, switching targets always co-occurred with the previous trial's target colours becoming the current trial's distractors. If attention to targets on the previous trial drove awareness of them on the current trial, I expected that target weights would decrease in both cNT and iNT conditions in switch trials, relative to stay trials, but that distractor weights would increase only in iNT switch trials relative to iNT stay trials. I excluded the decoy condition from this comparison, because decoy "stay" trials conflated potentially beneficial carry-over (repeating targets) with potentially harmful carry-over (trial n-1 Target 2 becomes trial n Distractor 1).

To understand how recent and global relevance drive attentional capture, I ran a repeated-measures analysis of variance on the decision weights, with global relevance (cNT vs iNT), target carry-over (stay vs switch), and dot cloud (T1, T2,

D1, D2) as within-subject factors. The interaction between all three predictors was statistically significant ($F_{1,67}$=7.90; p=.006), suggesting that global relevance and switch/stay have some effect on the decision weights, but that this effect is inconsistent over the dot clouds (Figure 7).

Recall that while targets could switch or stay in cNT and iNT conditions, distractors on iNT switch trials were also always the previous target colour. This was never the case for distractors on cNT switch trials. If trial-to-trial carry-over of attentional set primarily influenced attention to targets, we could expect to see different target decision weights on stay and switch trials, with no reliable effect among distractors. If trial-to-trial carry-over of attentional set also increased attention to distractors that had recently been targets, we would expect to see high distractor weights on iNT switch trials, relative to all other trials.

So, I conducted follow-up analyses, now considering target and distractor weights independently. First, I re-ran the previous ANOVA, but with dot cloud divided into two factors: target status (targets vs distractors) and epoch (epoch 1 vs epoch 2). I saw a main effect of epoch ($F_{1,67}$=42.45; p<.001) and an interaction between epoch and target status ($F_{1,67}$=52.83; p<.001). However, epoch was not reliably associated with global or recent relevance, in direct, three-way, or four-way interactions. Consequently, I averaged decision weights across epochs to simplify the analysis of relevance effects. I then conducted two ANOVAs, one for target decision weights, and another for distractor decision weights, with global and recent relevance as predictors. I found that target decision weights were additively influenced by whether the distractors were globally relevant ($F_{1,67}$=25.39; p<.001), and by whether targets repeated or changed relative to the previous trial ($F_{1,67}$=10.88; p=.002), with no reliable interaction (F=2.43; p=.124). For distractors, however, the influence of global relevance and target switch was best described by an interaction ($F_{1,67}$=9.70; p=.003). A planned comparison of iNT switch trials to all other conditions revealed that distractor weights increased when they were targets on the previous trial (t_{67}=3.89; p<.001). This highlights two different effects of

recent target relevance. Target weights drop following a switch, regardless of what distractors appear; while distractor weights selectively increase when the previous targets become the current distractors. This highlights the complex role of global and recent history in driving attention towards stimuli that are repeatedly relevant.

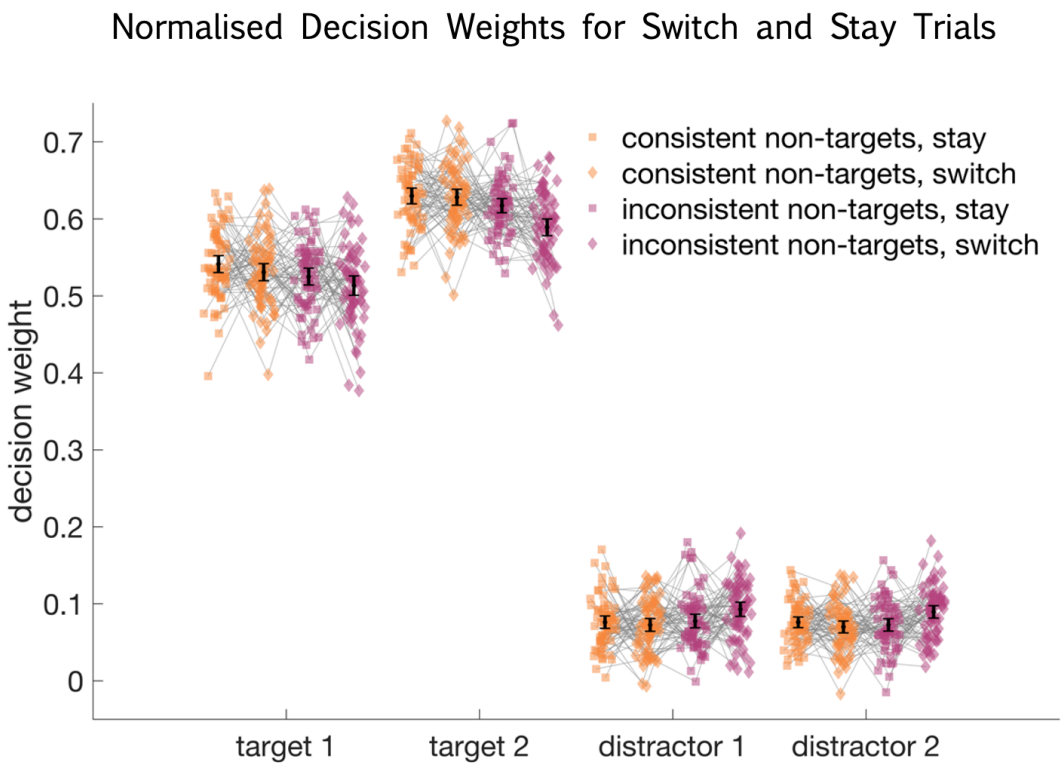


Figure 7. Normalised decision weights for target and distractor dot clouds for Experiment 4, shown separately for consistent non-target (orange) and inconsistent non-target (pink) stay and switch conditions (square and diamond markers respectively). Black bars indicate the 95% confidence interval.

4.6. Discussion

Attending to what is immediately relevant is a useful tool to drive our mental resources toward simple parts of a task. Conversely, attending to features that may shortly become relevant for subsequent task parts could help us shift our focus when we need to. Here, I conducted four experiments to probe whether our attention is captured by irrelevant information if it will shortly become relevant. In Experiment 1, I observed evidence of future-task awareness, when future-task relevant stimuli were also globally and recently relevant. However, in three further experiments, I showed that the effect in Experiment 1 can be accounted for by enhanced attention to distracting information that is broadly relevant in the overall context of the task and/or recently attended, with little or no additional effect of something becoming relevant imminently. That is, we are sensitive to the past and/or global relevance of currently-irrelevant information, but do not appear to give preference to information that is imminently relevant for a future task part. These results emphasise our capacity for both sensitivity to broad attentional relevance, and temporally precise focus, as we move through parts of a task.

Many lines of inquiry have shown that we pay attention to what has been relevant in the past, both globally, over an extended period of time, as well as recently, in the immediately preceding trial. Repeating stimuli prime our perception of their colour, motion, and identity (Ellis et al., 1987; Kristjánsson & Campana, 2010; Maljkovic & Nakayama, 1994), and inter-trial sequence effects can even lead to us misperceive objects in a series as having similar features (Fischer & Whitney, 2014; Maloney et al., 2005). Priming is a perceptual effect, meaning that we perceive current targets more quickly and accurately when they match previous targets. However, rapid and precise perception could in turn drive rapid and precise responses to those stimuli. The current study is consistent with this idea, showing that we select relevant target colours more accurately when cues repeat over trials (Exp 4).

However, a purely perceptual bias is unlikely to drive the consistent preference for recurring targets that I report here. Each trial in this multi-part design included four distinct colours and pseudo-randomly varying motion directions. This means that adjacent perceptual features (for example, trial n-1 Target 2 and trial n Target 1 colour) only reliably repeated in the decoy condition of Experiment 1 and in decoy "stay" trials of Experiment 4. Any serial perceptual effects within the decoy condition of Experiment 4 were not sufficient to reliably increase attentional capture by distracting information, relative to baseline (decoy vs iNT, Figure 6).

Beyond perceptual bias, this study highlights the impact of previous task set on current target and distractor processing. Directing our attention towards new information requires that we disengage from our previous attentional set and reconfigure our attention to prioritise the new targets (Imburgio & Orr, 2021). Task-switching studies have repeatedly demonstrated that this process is non-trivial, with task-set inertia biasing our focus towards what was recently relevant and imposed time limits making it difficult for us to assemble our new task set (Evans et al., 2015; Imburgio & Orr, 2021; Longman et al., 2017; Weiler et al., 2015). The present study offers an added level of specificity to previous studies of errors in task-switching, by separately tracking responses to targets and distractors. Here, I found that responses reflect less target information when targets switch. Curiously, responses did not correspondingly reflect more distractor information on target switch trials. Instead, distractors were only upweighted on switch trials, relative to stay trials, when they had been relevant on the previous trial.

I also observed a tendency to upweight colours that were sometimes targets (inconsistent non-targets), relative to colours that were never targets (consistent non-targets). This cannot be explained by colour salience, as I randomly assigned colours to non-target and inconsistent non-target colour sets for each participant. Instead, it could reflect reinforcement learning, as participants learned that attending to targets and suppressing distractors led to positive feedback at the end

of the trial. This in turn could drive a policy of increasing attention to globally relevant colours (Botvinick, 2012; Sutton & Barto, 2018). In Experiment 4, the reinforced target colours could also be distractors (inconsistent non-targets), but the policy of preferring to respond to those colours would on average be beneficial. These four colours were more likely to be targets than distractors at a rate of 3:2, as they appeared as targets in all three conditions and distractors only in two. While reinforcement learning is often discussed in the context of associating stimuli with explicit actions, this study emphasises our capacity to associate stimuli with mental actions, such as "attend" or "ignore".

Neural data from non-human primates offer a time-resolved interpretation for preferential responses to globally relevant distractors. Neurons in monkey prefrontal cortex commonly code information that is behaviourally relevant (Duncan, 2001; Erez et al., 2020; Erez & Duncan, 2015; Kadohisa et al., 2015; M. Stokes, 2011). However, these neurons initially code relevant and irrelevant information in multi-item displays, before giving priority to the relevant feature (Kadohisa et al., 2013). Interestingly, these data showed that neural coding of globally relevant distracting information (i.e., inconsistent non-targets) disappears less efficiently and completely, compared to coding of features that are never relevant (consistent non-targets).

I should acknowledge that my finding in Experiment 4, that any increased distraction by future-relevant colours can be explained by that colour's previous use as a target, may not fully capture the effect I saw in Experiment 1. In Experiment 1, decoys reliably skewed the response away from the targets and towards the distractors, relative to trials on which the distractor was a consistently irrelevant colour. In Experiment 4, I created conditions to mirror Experiment 1, but retained design changes from Experiments 2 and 3 (written cues, target colours changing throughout the block). While it is difficult to completely rule out the possibility that decoy effects in Experiment 1 partly arose from the decoy's future relevance, based

on the near identical responses to trials containing decoys or inconsistent non-targets in Experiment 4, I believe that this is unlikely.

One limitation of this study comes from the choice of decoy distractor. I have argued that people do not appear to direct their attention toward future-relevant information as they move through parts of a task, based on a task in which I deliberately gave people access to task-relevant information outside the epoch in which it was relevant. But while the colour of the decoy was the future target colour, those dots contained no useful information about the direction of the future target dots. We might better represent the real world by making the information for all task parts available at the same time, and asking people to self-select the sequence of task parts. This self-directed aspect is present in matrix reasoning and block design tasks that are often used to probe fluid ability, and could be a primary reason that those tasks are so difficult (Duncan et al., 2017). However, introducing participant-driven task epochs presents some challenges for controlled research. We often rely on carefully constructed displays, and specific presentation times, to infer what people are thinking of at any moment. In the current task design, showing both target dot directions together would make it difficult to extract the temporal information that we care about, that is, whether attending to each target in strict sequence leads to better performance. Possible ways forward would be to use eye-tracking, interactive stimuli, or time-resolved neural data with encoding models to retain precise information about what is being attended to at each point in time. Along with the current study, this could give us insight into how we are able to direct our mental resources towards a bewildering range of goals as we interact with the dynamic world around us.

Chapter 5

Understanding the Neuroscience of Flexible Behaviour in Multi-Step Tasks

5.1. Overview

Adaptive brain processes orient our behaviour towards what is important to us. Despite the inherent challenges, we are able to engage these processes to construct coherent sequences of thought and action across many different tasks. The attentional episodes account of flexible behaviour argues that we do this by focusing on each task step in turn. Here, I explored this account in three studies, examining how we preferentially encode and act on relevant task features when what is relevant changes sub-trial.

5.1.1. Chapter 2

I began by asking how the brain supports attention in multi-step tasks. Multi-step tasks could elicit different approaches to stimuli to what we observe in single stimulus-response trials, as we need to not only identify some information that is important for our response, but also hold that information securely in mind while we direct our attention to another part of our environment. I tested whether dynamic task requirements impact our ability to code what's relevant, extending work with non-human primates to non-invasive imaging in humans. I showed that we can identify preferential coding of behaviourally relevant information, even as what's relevant changes sub-trial. This came with a caveat: preferential coding of

relevant information was clear and reliable in multiple task steps when the task was difficult, but was not detected in a similar but easy task.

5.1.2. Chapter 3

In the next chapter, I considered the neural generators of the effect observed in Chapter 2. I used source estimation to ask whether preferential coding of task-relevant information pertained to both ventral visual and frontoparietal multiple-demand cortices, and examined communication between the two systems. I found that multiple-demand and ventral visual regions both preferentially encode features that are important for behaviour but do so with different time courses. Information was passed between regions in both parts of the task, emphasising again how fluently the brain co-ordinates activity to support momentary focus.

5.1.3. Chapter 4

One common aspect of everyday behaviour is that we know in advance what we plan to do. Unlike many lab-based tasks, we can anticipate what we are about to do, and choose for ourselves whether we will focus on what is happening now or in the future. Cued multi-step tasks, where we can plan what information we will seek at each step, are an important test-case to develop a full picture of how closely or broadly we direct our focus. In Chapter 4, I developed a new task that let us infer how people attended to currently and imminently relevant information, based on a single response. I used a multi-step task with information from an anticipated task step sometimes presented early, to probe people's sensitivity to imminently relevant information. I found that what is imminently relevant had little to no impact on people's responses, once I accounted for the global and recent relevance of those stimulus features.

5.2. Implications, limitations, and future directions

5.2.1. Multi-step tasks

One of the aims of this series of projects was to understand how the demands of a multi-step task change the way the brain organises itself around a goal. In the attentional episodes framework, we can think of each task part as its own unit, and even propose that organising cognition and neural resources into these units makes a complex task easier to solve. On the other hand, there is a range of possible approaches to multi-step tasks. We can delineate steps at different levels of granularity, and we can choose whether to focus only on the information needed in the current step, or also consider the steps beyond our immediate actions.

Here, I used different multi-step tasks to understand how we encode and use task-relevant information as task demands change. In Chapters 2 and 4, I found that stimulus information was effectively prioritised when it was relevant. People selected what was relevant for an immediate task step without clear distraction by imminently-relevant features (Chapter 4), or lag induced by releasing the previous task step (Chapter 2, Experiment 2). This is consistent with the attentional episodes view, as solving a task bit by bit is only practical if we can focus our attention on subgoals, and shift our attention within a task, with minimal behavioural cost.

However, it also suggests a possible difference between task-switching and shifting attention within a task. Task-switching studies typically show that we respond slowly to new stimulus-response mappings, even when we know what they will be. In the context of the dual-step tasks presented here, we might expect that people would be slower to preferentially encode relevant features of a task as they quickly shift their focus between objects, locations, and stimulus dimensions mid-task. Task-switching studies do raise the possibility that shifting attention alone could be less difficult than shifting stimulus-response mappings, as task switch costs are higher when the stimulus-response mappings conflict between two tasks (Allport & Wylie, 1999). This could explain why our neural data did not indicate a

delay in coding task-relevant information after shifting attention quickly within a trial, relative to after moving between trials.

Thinking more broadly, it is also possible that the way we conceive of a task's boundaries impacts how we direct our attention throughout it. Other work distinguishes the inertia of a stimulus-response mapping from another aspect of attention shifting, typically called reconfiguration (Imburgio & Orr, 2021; Meiran et al., 2000). Reconfiguration describes the process of preparing to maintain and act on a new goal. A task that has a predictable sequence of sub-goals could allow people to group sub-goals within a task, and so prepare in advance. A study of hierarchical task-switching findings has raised similar possibilities (Schneider & Logan, 2006). People completed two tasks in a regular AABB sequence, with some starting on the first A (AABB-AABB) and others on the second (ABBA-ABBA). Both groups were slower to transition between repetitions of the four-step sequence than between A and B subtasks within a sequence. This was true even for the second group, whose sequence ended and began with the same sub-task. Thus, it is plausible that mentally grouping steps within a task will change how we direct our mental resources towards them. In this thesis at least, shifting attention quickly within a task did not appear to incur a cost beyond any that people may have experienced between trials. An open question is whether conceiving of some steps as part of one task reduces reconfiguration costs and helps us to switch between the task parts, or whether we would see reconfiguration costs between task steps in more challenges situations (for example, with a wider range of task steps or less time between them). For example, Experiment 1 of Chapter 2 suggested that we may not always focus on each task step in turn, suggesting that we may avoid having to reconfigure our attention when we have another alternative.

Understanding how we orient our neural resources around a multi-step task moves us closer to understanding complex behaviour more broadly, but there is still a lot to explore. Experiment 1 of Chapter 2 seemed to suggest that people can complete a multi-step task without selecting relevant task features online, as they

become available, or at least as far as it is possible to tell from decoding the MEG signals. On the other end of the spectrum, the experiments of Chapter 4 suggested that we can respond to the information that is relevant for each task step even when there is conflicting, imminently relevant information right in front of us. It is possible that participants in Chapter 4 selected the relevant information online during each epoch, as in Experiment 2 of Chapter 2, prioritising processing of this information in working memory. However, it is also possible that they held both relevant and irrelevant features in memory, and retroactively selected what was relevant for the response. If the first option were true, we might expect that suppressing an imminently relevant colour in one task step could make it more difficult to engage with the goals of the next step. In either situation, knowing that attentional control is limited, we might expect that information presented outside its relevant task step would disrupt focus briefly, though the disruption could be resolved before people act on the relevant information. Although the task in Chapter 4 was developed to probe multiple task steps with a single behavioural response, it is difficult to judge from this the timecourse with which people selected each piece of information, or whether selecting information for one task step was ever disrupted by information pertaining to another. Time-resolved neural data, and behavioural tasks with time pressure, will be important approaches to understanding how we manage multiple task goals in complex behaviour.

Another important avenue is pushing further towards the kinds of situations we face in real life to see how people reflect on and anticipate aspects of a task that they are not currently performing. This might include establishing tasks that are self-paced or not strictly sequential, where people can decide for themselves what they need to focus on at each moment. Interpreting tasks that are self-paced or not sequential can be tricky, especially in neuroimaging, when the experimenter usually needs to know what the participant was doing so that they can group together related pieces of data (i.e., time-lock data to the event boundary). Some solutions may be using participant actions, such as eye or cursor movements, to

define event boundaries, or using a time-unconstrained approach to label periods of data that form a coherent brain "state" (see for example Vidaurre et al., 2016, 2018).

In turn, having a full toolkit of ways to identify event boundaries could help us understand what attentional strategies individuals use to solve complex tasks. The attentional episodes account proposes that we can characterise flexible cognition as a sequence of simple steps (Duncan, 2013). The findings in Chapters 2 and 4 are consistent with the idea that we can and do approach tasks step by step, but they also raise other possibilities. Chapter 2 suggests that attention can be engaged online as things appear, or not, depending on task demands. Chapter 4 suggests that we are able to select what is relevant at a specific moment, but are also influenced by what has been relevant recently or globally. If task demands elicit different ways of attending at the group level, and if our attention can be simultaneously influenced by our immediate goal and by our recent experiences, then it is plausible that people may have different patterns of attentional focus as they piece together a complex behaviour. Exploring how individuals divide up multi-step tasks, with methods that allow people to set their own pace, could show us a much wider range of approaches to complex tasks than we might see at the group level. More specifically, the attentional episodes account emphasises the *value* of moment-by-moment focus, arguing that people who are able to solve complex problems do so by imposing boundaries between steps of a task (Duncan et al., 2017). Testing this means asking whether segregating a task into parts is useful, and not just typical. For instance, we could look at individual differences, predicting that brain "states" will be more internally consistent, and more dissimilar from each other, in people who perform well on complex problem-solving. We could extend this to specifically test the MDN's role in segregating tasks, by asking whether brain "states" are affected (e.g., become more or less distinct) by stimulation of a MDN node (e.g., with transcranial magnetic stimulation). Combining flexible ways to segment data with a diverse pool of participants could

be important if we want to understand how moment-by-moment focus helps or hinders performance.

Multi-step tasks also present some particular methodological challenges. In the MEG studies described in Chapter 2, participants could spend as long as four hours in the lab to practice the task, prepare for the session, and complete enough of the long, multi-step trials to cover a moderate number of experimental conditions. Long experimental sessions can be draining for the participants, as well as decreasing the data quality over the course of the session if participants become fatigued. The long trials also affected the analysis time, as individually analysing multivariate brain responses at every 10 millisecond time point over a five second trial takes substantial computing power. Similarly, long trials can affect statistical power, with many time points to correct multiple comparisons over, and relatively few trial repetitions. For behavioural tasks, a simple solution is to move testing online, so that the burden of generating a strong dataset from a slow task can be easily shared among many participants. Another promising option for multi-step tasks in M/EEG is rapid stimulus presentations, for example with information-based analysis such as decoding or RSA to track the brain's response to each item even as other items are appearing (see for example Grootswagers et al., 2019; Robinson et al., 2019).

5.2.2. Spatio-temporal resolution

Obtaining high spatio-temporal resolution from non-invasive neuroimaging data is an important area of development for cognitive neuroscience. In Chapter 3, I resolved MEG data from a multi-step task to source space, to understand how visual and multiple-demand regions engage as we move through parts of a task. The attentional episodes view proposes that adaptive coding in the MDN drives episode-like focus on simple sub-goals within a task. Non-human primate data are consistent with this view, showing that coding of task-relevant visual features in lateral frontal cortex can quickly reconfigure (Rao et al., 1997). In humans,

obtaining such spatio-temporally resolved information is more difficult. A wealth of fMRI data, measured over the course of minutes, shows that the MDN can flexibly encode what is currently relevant (e.g., Woolgar et al., 2016). Some fMRI findings speak directly to whether the MDN deals with tasks in an episodic way, showing that BOLD activity peaks as we complete goals, proportionally to the breadth of the goal (Farooqui et al., 2012). But if we wish to test the MDN's role in flexibly forming focused episodes around task steps, we need to be able to observe the network with fine temporal detail. Here, decoding with source-reconstructed MEG showed that coding of relevant task features was quickly prioritised within the MDN, as coding of those features was maintained within ventral visual cortex. However, stimulus information coding did not reappear within the MDN in a second task epoch, raising the possibility that reconfiguring focus quickly incurred some costs; though it could equally reflect limits of this particular dataset. In a follow-up representational similarity analysis, both task epochs showed sporadic, noisy coding of stimulus information in the MDN, reinforcing this possibility. Colour and shape information were dynamically exchanged between the MDN and ventral visual cortex following each stimulus display, suggesting that some stimulus information was indeed present within the MDN. Thus, for the first time in humans, we can see this highly adaptive system quickly engaging with a complex task during each step. However, our analysis did not specify whether the MDN specifically, or frontoparietal cortex more generally, prioritises relevant task features during a working memory delay. Future studies could check the specificity of the effect to the MDN by contrasting the responses of different networks (perhaps with individual subject localisers) to further differentiate representations that we can identify across the brain from representations that are uniquely the responsibility of the MD system.

It is important to note that MEG source reconstruction is an imperfect way to obtain spatio-temporal resolution. There are many ways to map data from 100 or so sensor locations to 200 times as many sources, and there is always some uncertainty around how well the sources have been reconstructed from signals

recorded at the scalp. However, these limitations must be balanced against the considerations inherent in alternative approaches such as non-human primate data or intracranial recordings, which can be slow and expensive to collect, as well as asking a lot from the participant, and do not offer insight directly about the healthy human brain. Fortunately, there are ways to test and improve the quality of source reconstruction. One option is to use point-spread and cross-talk values, which are easily generated during minimum-norm source reconstruction, to test how much sources bleed into each other (Hauk et al., 2019). This "resolution analysis" has been developed to estimate how many parcels the data could reasonably support, and so set an upper limit on how far researchers try to localise effects (Farahibozorg et al., 2018). In addition, data from EEG can constrain source reconstruction by adding another dimension to MEG data (Hauk et al., 2019). Another possible way to validate source reconstruction is to compare RDMs of MEG time-courses in a source-reconstructed ROI to the representational structure that is shared fMRI of the corresponding region and sensor-space MEG (i.e. MEG-fMRI "fusion"; see for example Cichy et al., 2014, 2016; Moerel et al., 2021). This approach may be less accessible, as it requires multi-modal imaging data, and it is limited to tracking signals present in both modalities,, but it could be a useful within-study conceptual replication when both MEG and fMRI data are available.

One of the big challenges in resolving neural responses in time and space comes from the fact that we not only need to extract rich information from the data, but also piece it together in a way that we can interpret. In Chapter 3, I combined MEG source reconstruction with RSA and Granger causality to trace information flow in time and space. Information flow analysis takes the opportunity that spatio-temporally resolved data provide to ask questions that have been central to our cognitive theories – such as how top-down control alters emerging sensory representations. Multivariate approaches to connectivity can also allow us to test complex dependencies between regions, rather than assuming that we can represent the information within two connected regions with a single averaged time-course

(Basti et al., 2020). However, their complexity can make them challenging to implement. In the analysis described in Chapter 3, there were many minor decisions to make at each step, making it difficult to choose the best path or to understand how the decisions impacted the outcome without biasing the results. These problems could be solved by a strict pre-planned analysis, reinforced by pre-registration. A more practical solution for a novel analysis could involve exploring the effects of analysis decisions in one half of a dataset, and testing the final pipeline on the remaining half.

Working in an abstracted representational space can be a useful technique for multivariate connectivity analysis. Using representational similarity analysis, I was able to compare how two theoretically distinct brain systems communicated with each other, and infer what information was being shared. I also showed what inferences could be possible when models are introduced to the equation, by testing specific predictions about stimulus representations within each region. Combined representational similarity analysis and connectivity could be further developed to answer a myriad of questions. In particular, model-based RSA could be applied to scrutinise what information is shared between regions and time points, by separating time-courses of information exchange for different task features. Future work could harness the conceptual precision of RSA models to identify what rule, stimulus, and response information the brain relays to shape perception and action.

5.2.3. A broader view of flexible behaviour

Lastly, I should acknowledge that the broad goal of understanding how the brain enables us to construct organised behaviour requires that we think about more than just what people attend to moment by moment. Chapter 3 showed that the multiple-demand network can code what is relevant right now. Patient studies causally link this network with fluid ability (Duncan et al., 1995; Woolgar et al., 2010, 2018). So, we could make a lot of progress in understanding flexible behaviour by studying the dynamics of this network, how it codes and communicates relevant

information. More broadly, the attentional episodes account of flexible behaviour emphasises how important our attention, moment-to-moment, could be for responding adaptively to diverse mental challenges. But flexible behaviour – adapting our focus and actions to reach an overarching goal – also requires that we track our environment, consider what we already know or can guess about the situation, weigh up the value of different choices, and allow our values (what we enjoy, how confident we are in our strategy, or whether we are bored) to direct our attention. While it can be simplest to understand each of these elements separately in the lab, our everyday decisions and achievements come from all of them operating together. Giving participants a reason to engage with a task, then, is important if we wish to see how they approach real-world problems that they are motivated to solve. Some simple steps could include building more breaks into experimental sessions, offering financial incentives where that can be done fairly, and giving encouraging feedback. Self-directed tasks could also help motivation by providing a game-like challenge that people solve in their own way. This will not always be practical, as we are limited to using stimuli and tasks that we as experimenters can interpret. But the methods suggested above for interpreting self-directed tasks, such as tracking gaze or using data-driven brain state analysis to identify mental event boundaries, will be doubly valuable if they mean that we can make participants' lab experiences more engaging and ecologically valid.

5.2.4. Summary

This series of projects explored the neural and behavioural basis of our ability to flexibly attend as we move through parts of a task. They used advanced analysis techniques to unpack time-resolved neural and behavioural data. Together, the projects showcase how focal our attention can be. In Chapter 2, neural data showed that we can rapidly prioritise what is relevant as task demands change. Source reconstruction in Chapter 3 linked this prioritisation to ventral visual cortex and the multiple-demand network, with information dynamically exchanging between

these regions throughout the multi-part task. In behaviour, Chapter 4 showed that we are able to prioritise immediately relevant information over imminently relevant information, even when those features are intertwined. Beyond this, both the neural and behavioural data raised questions about how far our ability to focus on what is immediately relevant extends. The extent to which we prioritise information in an episodic way could depend on many factors, which have yet to be explored. Neural prioritisation of relevant information appeared to depend on task difficulty, suggesting that easy tasks may not elicit the same focus on each task part that a difficult task demands. Behaviourally, imminently relevant information did not bias responses away from what was immediately relevant, but responses were impacted by the broad relevance of distracting information. Creative task designs will be important to better understand the conditions under which we do or do not engage with tasks in an episodic way.

5.3. Conclusions

These studies characterise how we direct our attention in multiple task parts. They show that we are able to our ability to dynamically attend to what is relevant, through prioritising coding of diverse task-relevant features in quick succession, rapidly transmitting information from domain-general to sensory brain regions, and controlling our attention moment by moment to seek task-relevant information in an organised sequence of steps. Together, they demonstrate our extraordinary flexibility in directing our minds and actions toward what will help us reach our goals.

6. Bibliography

Allport, A., & Wylie, G. (1999). Task-switching: Positive and negative priming of task-set. In *Attention, space, and action: Studies in cognitive neuroscience* (pp. 273–296). Oxford University Press.

Altmann, E. M. (2004). Advance Preparation in Task Switching: What Work Is Being Done? *Psychological Science*, *15*(9), 616–622. https://doi.org/10.1111/j.0956-7976.2004.00729.x

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407–1425. https://doi.org/10.3758/s13428-020-01501-5

Arulpragasam, A. R., Cooper, J. A., Nuutinen, M. R., & Treadway, M. T. (2018). Corticoinsular circuits encode subjective value expectation and violation for effortful goal-directed behavior. *Proceedings of the National Academy of Sciences*, *115*(22), E5233–E5242. https://doi.org/10.1073/pnas.1800444115

Assem, M., Glasser, M. F., Van Essen, D. C., & Duncan, J. (2020). A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhaa023

Assem, M., Shashidhara, S., Glasser, M. F., & Duncan, J. (2021). Precise Topology of Adjacent Domain-General and Sensory-Biased Regions in the Human Brain. *Cerebral Cortex*, bhab362. https://doi.org/10.1093/cercor/bhab362

Baddeley, A. (1996). Exploring the Central Executive. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 5–28. https://doi.org/10.1080/713755608

Bae, G.-Y., & Luck, S. J. (2019). Decoding motion direction using the topography of sustained ERPs and alpha oscillations. *NeuroImage*, *184*, 242–255. https://doi.org/10.1016/j.neuroimage.2018.09.029

Bahmani, Z., Daliri, M. R., Merrikhi, Y., Clark, K., & Noudoost, B. (2018). Working Memory Enhances Cortical Representations via Spatially Specific Coordination of Spike Times. *Neuron*, *97*(4), 967-979.e6. https://doi.org/10.1016/j.neuron.2018.01.012

Baldauf, D., & Desimone, R. (2014). Neural Mechanisms of Object-Based Attention. *Science*, *344*(6182), 424–427. https://doi.org/10.1126/science.1247003

Basti, A., Nili, H., Hauk, O., Marzetti, L., & Henson, R. N. (2020). Multi-dimensional connectivity: A conceptual and mathematical review. *NeuroImage*, *221*, 117179. https://doi.org/10.1016/j.neuroimage.2020.117179

Battistoni, E., Kaiser, D., Hickey, C., & Peelen, M. V. (2020). The time course of spatial attention during naturalistic visual search. *Cortex*, *122*, 225–234. https://doi.org/10.1016/j.cortex.2018.11.018

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*(6), 956–962. https://doi.org/10.1016/j.conb.2012.05.008

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Braver, T. S. (2012). The variable nature of cognitive control: A dual-mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198. https://doi.org/10.1038/nrn2575

Camilleri, J. A., Müller, V. I., Fox, P., Laird, A. R., Hoffstaedter, F., Kalenscher, T., & Eickhoff, S. B. (2018). Definition and characterization of an extended multiple-demand network. *NeuroImage*, *165*(Supplement C), 138–147. https://doi.org/10.1016/j.neuroimage.2017.10.020

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1–1. https://doi.org/10.1167/13.10.1

Chafee, M. V., & Goldman-Rakic, P. S. (2000). Inactivation of Parietal and Prefrontal Cortex Reveals Interdependence of Neural Activity During Memory-Guided Saccades. *Journal of Neurophysiology*, *83*(3), 1550–1566. https://doi.org/10.1152/jn.2000.83.3.1550

Chao, L. L., & Knight, R. T. (1998). Contribution of Human Prefrontal Cortex to Delay Performance. *Journal of Cognitive Neuroscience*, *10*(2), 167–177. https://doi.org/10.1162/089892998562636

Chen, T., Liu, H., Ma, Z., Shen, Q., Cao, X., & Wang, Y. (2021). End-to-End Learnt Image Compression via Non-Local Attention Optimization and Improved

Context Modeling. *IEEE Transactions on Image Processing*, *30*, 3179–3191. https://doi.org/10.1109/TIP.2021.3058615

Chen, Z. (2012a). Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics*, *74*(5), 784–802. https://doi.org/10.3758/s13414-012-0322-z

Chen, Z. (2012b). Object-based attention: A tutorial review. *Attention, Perception, & Psychophysics*, *74*(5), 784–802. https://doi.org/10.3758/s13414-012-0322-z

Chevignard, M. P., Taillefer, C., Picq, C., Poncet, F., Noulhiane, M., & Pradat-Diehl, P. (2008). Ecological assessment of the dysexecutive syndrome using execution of a cooking task. *Neuropsychological Rehabilitation*, *18*(4), 461–485. https://doi.org/10.1080/09602010701643472

Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., Bradley, D. C., Smith, M. A., Kohn, A., Movshon, J. A., Armstrong, K. M., Moore, T., Chang, S. W., Snyder, L. H., Lisberger, S. G., … Shenoy, K. V. (2010). Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience*, *13*(3), 369–378. https://doi.org/10.1038/nn.2501

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual

Object Recognition. *Cerebral Cortex*, *26*(8), 3563–3579. https://doi.org/10.1093/cercor/bhw135

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience*, *16*(9), 1348–1355. https://doi.org/10.1038/nn.3470

Cole, M. W., Yarkoni, T., Repovš, G., Anticevic, A., & Braver, T. S. (2012). Global Connectivity of Prefrontal Cortex Predicts Cognitive Control and Intelligence. *Journal of Neuroscience*, *32*(26), 8988–8999. https://doi.org/10.1523/JNEUROSCI.0536-12.2012

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Crittenden, B. M., & Duncan, J. (2014). Task Difficulty Manipulation Reveals Multiple Demand Activity but no Frontal Lobe Hierarchy. *Cerebral Cortex*, *24*(2), 532–540. https://doi.org/10.1093/cercor/bhs333

Crossley, N. A., Mechelli, A., Scott, J., Carletti, F., Fox, P. T., McGuire, P., & Bullmore, E. T. (2014). The hubs of the human connectome are generally implicated in the anatomy of brain disorders. *Brain: A Journal of Neurology*, *137*(Pt 8), 2382–2395. https://doi.org/10.1093/brain/awu132

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

DeSerisy, M., Ramphal, B., Pagliaccio, D., Raffanello, E., Tau, G., Marsh, R., Posner, J., & Margolis, A. E. (2021). Frontoparietal and default mode network connectivity varies with age and intelligence. *Developmental Cognitive Neuroscience*, *48*, 100928. https://doi.org/10.1016/j.dcn.2021.100928

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 761–772. https://doi.org/10.1098/rstb.2007.2086

D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*, *378*(6554), 279–281. https://doi.org/10.1038/378279a0

Dosenbach, N. U. F., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A. T., Fox, M. D., Snyder, A. Z., Vincent, J. L., Raichle, M. E., Schlaggar, B. L., & Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences*, *104*(26), 11073–11078. https://doi.org/10.1073/pnas.0704320104

Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., Burgund, E. D., Grimes, A. L., Schlaggar, B. L., & Petersen, S.

E. (2006). A Core System for the Implementation of Task Sets. *Neuron*, *50*(5), 799–812. https://doi.org/10.1016/j.neuron.2006.04.031

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*(4), 501–517. https://doi.org/10.1037/0096-3445.113.4.501

Duncan, J. (2001). An Adaptive Coding Model of Neural Function in Prefrontal Cortex. *Nature Reviews. Neuroscience; London*, *2*(11), 820–829. http://dx.doi.org/10.1038/35097575

Duncan, J. (2006). EPS Mid-Career Award 2004: Brain mechanisms of attention. *Quarterly Journal of Experimental Psychology*, *59*(1), 2–27. https://doi.org/10.1080/17470210500260674

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, *14*(4), 172–179. https://doi.org/10.1016/j.tics.2010.01.004

Duncan, J. (2013). The Structure of Cognition: Attentional Episodes in Mind and Brain. *Neuron*, *80*(1), 35–50. https://doi.org/10.1016/j.neuron.2013.09.015

Duncan, J., Assem, M., & Shashidhara, S. (2020). Integrated Intelligence from Distributed Brain Activity. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2020.06.012

Duncan, J., Burgess, P., & Emslie, H. (1995). Fluid intelligence after frontal lobe lesions. *Neuropsychologia*, *33*(3), 261–268. https://doi.org/10.1016/0028-3932(94)00124-8

Duncan, J., Chylinski, D., Mitchell, D. J., & Bhandari, A. (2017). Complexity and compositionality in fluid intelligence. *Proceedings of the National Academy of Sciences*, *114*(20), 5295–5299. https://doi.org/10.1073/pnas.1621147114

Duncan, J., Schramm, M., Thompson, R., & Dumontheil, I. (2012). Task rules, working memory, and fluid intelligence. *Psychonomic Bulletin & Review*, *19*(5), 864–870. https://doi.org/10.3758/s13423-012-0225-y

Ellis, A. W., Young, A. W., Flude, B. M., & Hay, D. C. (1987). Repetition priming of face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *39*(2), 193–210. https://doi.org/10.1080/14640748708401784

Erez, Y., & Duncan, J. (2015). Discrimination of Visual Categories Based on Behavioral Relevance in Widespread Regions of Frontoparietal Cortex. *Journal of Neuroscience*, *35*(36), 12383–12393. https://doi.org/10.1523/JNEUROSCI.1134-15.2015

Erez, Y., Kadohisa, M., Petrov, P., Sigala, N., Buckley, M. J., Kusunoki, M., & Duncan, J. (2020). Prefrontal neural dynamics for behavioral decisions and attentional control. *BioRxiv*, 2020.05.06.080325. https://doi.org/10.1101/2020.05.06.080325

Evans, L. H., Herron, J. E., & Wilding, E. L. (2015). Direct Real-Time Neural Evidence for Task-Set Inertia. *Psychological Science*, *26*(3), 284–290. https://doi.org/10.1177/0956797614561799

Everling, S., Tinsley, C. J., Gaffan, D., & Duncan, J. (2006). Selective representation of task-relevant objects and locations in the monkey prefrontal cortex.

European Journal of Neuroscience, *23*(8), 2197–2214. https://doi.org/10.1111/j.1460-9568.2006.04736.x

Farahibozorg, S.-R., Henson, R. N., & Hauk, O. (2018). Adaptive cortical parcellations for source reconstructed EEG/MEG connectomes. *NeuroImage*, *169*, 23–45. https://doi.org/10.1016/j.neuroimage.2017.09.009

Farooqui, A. A., & Manly, T. (2019). Hierarchical Cognition Causes Task-Related Deactivations but Not Just in Default Mode Regions. *ENeuro*, *5*(6), ENEURO.0008-18.2018. https://doi.org/10.1523/ENEURO.0008-18.2018

Farooqui, A. A., Mitchell, D., Thompson, R., & Duncan, J. (2012). Hierarchical Organization of Cognition Reflected in Distributed Frontoparietal Activity. *Journal of Neuroscience*, *32*(48), 17373–17381. https://doi.org/10.1523/JNEUROSCI.0598-12.2012

Fedorenko, E., Duncan, J., & Kanwisher, N. (2012). Language-Selective and Domain-General Regions Lie Side by Side within Broca's Area. *Current Biology*, *22*(21), 2059–2062. https://doi.org/10.1016/j.cub.2012.09.011

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *110*(41), 16616–16621. https://doi.org/10.1073/pnas.1315235110

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*(5), 738–743. https://doi.org/10.1038/nn.3689

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1800755115

Fox, M. D., Corbetta, M., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2006). Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proceedings of the National Academy of Sciences*, *103*(26), 10046–10051. https://doi.org/10.1073/pnas.0604187103

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Essen, D. C. V., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, *102*(27), 9673–9678. https://doi.org/10.1073/pnas.0504136102

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A Comparison of Primate Prefrontal and Inferior Temporal Cortices during Visual Categorization. *Journal of Neuroscience*, *23*(12), 5235–5246. https://doi.org/10.1523/JNEUROSCI.23-12-05235.2003

FreeSurfer. (n.d.). Retrieved August 2, 2021, from https://surfer.nmr.mgh.harvard.edu/

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*(Supplement C), 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Fuster, J. M., Bauer, R. H., & Jervey, J. P. (1985). Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Research*, *330*(2), 299–307. https://doi.org/10.1016/0006-8993(85)90689-4

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Gilbert, S. J., & Shallice, T. (2002). Task Switching: A PDP Model. *Cognitive Psychology*, *44*(3), 297–337. https://doi.org/10.1006/cogp.2001.0770

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Goddard, E., Carlson, T. A., Dermody, N., & Woolgar, A. (2016). Representational dynamics of object recognition: Feedforward and feedback information flows. *NeuroImage*, *128*, 385–397. https://doi.org/10.1016/j.neuroimage.2016.01.006

Goddard, E., Carlson, T. A., & Woolgar, A. (2021). Spatial and feature-selective attention have distinct effects on population-level tuning. *Journal of Cognitive Neuroscience*.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013).

MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *0*. https://doi.org/10.3389/fnins.2013.00267

Grootswagers, T., Robinson, A. K., & Carlson, T. A. (2019). The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, *188*, 668–679. https://doi.org/10.1016/j.neuroimage.2018.12.046

Grootswagers, T., Robinson, A. K., Shatek, S. M., & Carlson, T. A. (2021). The neural dynamics underlying prioritisation of task-relevant information. *Neurons, Behavior, Data Analysis, and Theory*, *5*(1), 1–17. https://doi.org/10.51628/001c.21174

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2016). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Gutnisky, D. A., & Dragoi, V. (2008). Adaptive coding of visual information in neural populations. *Nature*, *452*(7184), 220–224. https://doi.org/10.1038/nature06563

Hall, N. J., Colby, C. L., & Olson, C. R. (2020). *Novel Interaction between Prefrontal and Parietal Cortex during Memory Guided Saccades*. 2020.03.11.985259. https://doi.org/10.1101/2020.03.11.985259

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical & Biological Engineering & Computing*, *32*(1), 35–42. https://doi.org/10.1007/BF02512476

Hauk, O., Stenroos, M., & Treder, M. (2019). Towards an Objective Evaluation of EEG/MEG Source Estimation Methods: The Linear Tool Kit. *BioRxiv*, 672956. https://doi.org/10.1101/672956

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *ELife*, *7*, e32816. https://doi.org/10.7554/eLife.32816

Humphreys, G. W., Duncan, J., Treisman, A., & Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1373), 1245–1255. https://doi.org/10.1098/rstb.1998.0280

Hwang, K., Shine, J. M., & D'Esposito, M. (2018). Frontoparietal Activity Interacts With Task-Evoked Changes in Functional Connectivity. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhy011

Ibos, G., & Freedman, D. J. (2016). Interaction between Spatial and Feature Attention in Posterior Parietal Cortex. *Neuron*, *91*(4), 931–943. https://doi.org/10.1016/j.neuron.2016.07.025

Imburgio, M. J., & Orr, J. M. (2021). Component processes underlying voluntary task selection: Separable contributions of task-set inertia and reconfiguration. *Cognition*, *212*, 104685. https://doi.org/10.1016/j.cognition.2021.104685

Jackson, J., Feredoes, E., Rich, A. N., Lindner, M., & Woolgar, A. (2021). Concurrent neuroimaging and neurostimulation reveals a causal role for dlPFC in coding of task-relevant information. *Communications Biology*, *4*(1), 1–16. https://doi.org/10.1038/s42003-021-02109-x

Jackson, J., Rich, A. N., Williams, M. A., & Woolgar, A. (2016). Feature-selective Attention in Frontoparietal Cortex: Multivoxel Codes Adjust to Prioritize Task-relevant Information. *Journal of Cognitive Neuroscience*, *29*(2), 310–321. https://doi.org/10.1162/jocn_a_01039

Jackson, J., & Woolgar, A. (2018). Adaptive coding in the human brain: Distinct object features are encoded by overlapping voxels in frontoparietal cortex. *Cortex*, *108*, 25–34. https://doi.org/10.1016/j.cortex.2018.07.006

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, *159*, 417–429. https://doi.org/10.1016/j.neuroimage.2017.06.030

Ji, J. L., Spronk, M., Kulkarni, K., Repovš, G., Anticevic, A., & Cole, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, *185*, 35–57. https://doi.org/10.1016/j.neuroimage.2018.10.006

Jocham, G., Neumann, J., Klein, T. A., Danielmeier, C., & Ullsperger, M. (2009). Adaptive Coding of Action Values in the Human Rostral Cingulate Zone. *Journal of Neuroscience*, *29*(23), 7489–7496. https://doi.org/10.1523/JNEUROSCI.0349-09.2009

Jones, J. S., the CALM Team, & Astle, D. E. (n.d.). Segregation and integration of the functional connectome in neurodevelopmentally 'at risk' children. *Developmental Science*, *n/a*(n/a), e13209. https://doi.org/10.1111/desc.13209

Josman, N., Schenirderman, A. E., Klinger, E., & Shevil, E. (2009). Using virtual reality to evaluate executive functioning among persons with schizophrenia: A validity study. *Schizophrenia Research*, *115*(2), 270–277. https://doi.org/10.1016/j.schres.2009.09.015

Jung, K., Min, Y., & Han, S. W. (2021). Response of multiple demand network to visual search demands. *NeuroImage*, *229*, 117755. https://doi.org/10.1016/j.neuroimage.2021.117755

Kado, H., Higuchi, M., Shimogawara, M., Haruta, Y., Adachi, Y., Kawai, J., Ogata, H., & Uehara, G. (1999). Magnetoencephalogram systems developed at KIT. *IEEE Transactions on Applied Superconductivity*, *9*(2), 4057–4062. https://doi.org/10.1109/77.783918

Kadohisa, M., Kusunoki, M., Petrov, P., Sigala, N., Buckley, M. J., Gaffan, D., & Duncan, J. (2015). Spatial and temporal distribution of visual information coding in lateral prefrontal cortex. *European Journal of Neuroscience*, *41*(1), 89–96. https://doi.org/10.1111/ejn.12754

Kadohisa, M., Petrov, P., Stokes, M., Sigala, N., Buckley, M., Gaffan, D., Kusunoki, M., & Duncan, J. (2013). Dynamic Construction of a Coherent Attentional State in a Prefrontal Cell Population. *Neuron*, *80*(1), 235–246. https://doi.org/10.1016/j.neuron.2013.07.041

Karimi-Rouzbahani, H., Ramezani, F., Woolgar, A., Rich, A., & Ghodrati, M. (2021). Perceptual difficulty modulates the direction of information flow in familiar face recognition. *NeuroImage*, *233*, 117896. https://doi.org/10.1016/j.neuroimage.2021.117896

Kastner, S., Weerd, P. D., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of Directed Attention in the Human Extrastriate Cortex as Revealed by Functional MRI. *Science*, *282*(5386), 108–111. https://doi.org/10.1126/science.282.5386.108

Kelley, T. A., & Lavie, N. (2011). Working Memory Load Modulates Distractor Competition in Primary Visual Cortex. *Cerebral Cortex*, *21*(3), 659–665. https://doi.org/10.1093/cercor/bhq139

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863. https://doi.org/10.1073/pnas.1905544116

Kleiner, M., Brainard, D. H., Pelli, D. G., Ingling, A., & Murray, R. (2007). What's new in psychtoolbox-3. *Perception*, *36*.

Kravitz, D. J., & Behrmann, M. (2011). Space-, object-, and feature-based attention interact to organize visual scenes. *Attention, Perception, & Psychophysics*, *73*(8), 2434–2447. https://doi.org/10.3758/s13414-011-0201-z

Kray, J. (2006). Task-set switching under cue-based versus memory-based switching conditions in younger and older adults. *Brain Research*, *1105*(1), 83–92. https://doi.org/10.1016/j.brainres.2005.11.016

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, *60*(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Kristjánsson, Á., & Campana, G. (2010). Where perception meets memory: A review of repetition priming in visual search tasks. *Attention, Perception, & Psychophysics*, *72*(1), 5–18. https://doi.org/10.3758/APP.72.1.5

Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in Soar: The Anatomy of a General Learning Mechanism. *Machine Learning*, *1*(1), 11–46. https://doi.org/10.1023/A:1022639103969

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lavie, N. (1995). Perceptual Load as a Necessary Condition for Selective Attention. *Journal of Experimental Psychology*, *21*(3), 451–468. https://doi.org/10.1037/0096-1523.21.3.451

Lavie, N., Beck, D. M., & Konstantinou, N. (2014). Blinded by the load: Attention, awareness and the role of perceptual load. *Phil. Trans. R. Soc. B*, *369*(1641), 20130205. https://doi.org/10.1098/rstb.2013.0205

Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, *56*(2), 183–197. https://doi.org/10.3758/BF03213897

Li, S., Ostwald, D., Giese, M., & Kourtzi, Z. (2007). Flexible Coding for Categorical Decisions in the Human Brain. *Journal of Neuroscience*, *27*(45), 12321–12330. https://doi.org/10.1523/JNEUROSCI.3795-07.2007

Long, N. M., & Kuhl, B. A. (2018). Bottom-Up and Top-Down Factors Differentially Influence Stimulus Representations Across Large-Scale Attentional Networks. *Journal of Neuroscience*, *38*(10), 2495–2504. https://doi.org/10.1523/JNEUROSCI.2724-17.2018

Longman, C. S., Lavric, A., & Monsell, S. (2017). Self-paced preparation for a task switch eliminates attentional inertia but not the performance switch cost. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(6), 862–873. https://doi.org/10.1037/xlm0000347

Lorenc, E. S., Sreenivasan, K. K., Nee, D. E., Vandenbroucke, A. R. E., & D'Esposito, M. (2018). Flexible Coding of Visual Working Memory

Representations during Distraction. *Journal of Neuroscience*, *38*(23), 5267–5276. https://doi.org/10.1523/JNEUROSCI.3061-17.2018

Maguire, J. F., & Howe, P. D. L. (2016). Failure to detect meaning in RSVP at 27 ms per picture. *Attention, Perception, & Psychophysics*, *78*(5), 1405–1413. https://doi.org/10.3758/s13414-016-1096-5

Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory & Cognition*, *22*(6), 657–672. https://doi.org/10.3758/BF03209251

Maloney, L. T., Martello, M. F. D., Sahm, C., & Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. *Proceedings of the National Academy of Sciences*, *102*(8), 3164–3169.

MATLAB R2012b [Computer Software]. (2012). The MathWorks, Inc.

Maxfilter (2.2). (2010). [Computer software]. Elekta Neuromag.

Mayr, U., & Keele, S. W. (2000). Changing internal constraints on action: The role of backward inhibition. *Journal of Experimental Psychology: General*, *129*(1), 4–26. https://doi.org/10.1037/0096-3445.129.1.4

Meiran, N., Chorev, Z., & Sapir, A. (2000). Component Processes in Task Switching. *Cognitive Psychology*, *41*(3), 211–253. https://doi.org/10.1006/cogp.2000.0736

Miller, B. T., & D'Esposito, M. (2005). Searching for "the Top" in Top-Down Control. *Neuron*, *48*(4), 535–538. https://doi.org/10.1016/j.neuron.2005.11.002

Moerel, D., Rich, A. N., & Woolgar, A. (2021). Selective attention and decision-making have separable neural bases in space and time. *BioRxiv*, 2021.02.28.433294. https://doi.org/10.1101/2021.02.28.433294

Mohsenzadeh, Y., Mullin, C., Lahner, B., Cichy, R. M., & Oliva, A. (2019). Reliability and Generalizability of Similarity-Based Fusion of MEG and fMRI Data in Human Ventral and Dorsal Visual Streams. *Vision*, *3*(1), 8. https://doi.org/10.3390/vision3010008

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P., & Kriegeskorte, N. (2013). Human Object-Similarity Judgments Reflect and Transcend the Primate-IT Object Representation. *Frontiers in Psychology*, *4*, 128. https://doi.org/10.3389/fpsyg.2013.00128

Musslick, S., Jang, S. J., Shvartsman, M., Shenhav, A., & Cohen, J. D. (2018). Constraints associated with cognitive control and the stability-flexibility dilemma. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 806–811.

Nogueira, R., Rodgers, C. C., Bruno, R. M., & Fusi, S. (2021). The geometry of cortical representations of touch in rodents. *BioRxiv*, 2021.02.11.430704. https://doi.org/10.1101/2021.02.11.430704

Norman, D. A., & Shallice, T. (1986). Attention to Action. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and Self-Regulation: Advances*

in Research and Theory Volume 4 (pp. 1–18). Springer US. https://doi.org/10.1007/978-1-4757-0629-1_1

O'Brien, S., Mitchell, D. J., Duncan, J., & Holmes, J. (2020). *Cognitive segmentation and fluid reasoning in childhood*. PsyArXiv. https://doi.org/10.31234/osf.io/dt84m

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2011/156869

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, *10*. https://doi.org/10.3389/fninf.2016.00027

Op de Beeck, H. P., Baker, C. I., DiCarlo, J. J., & Kanwisher, N. G. (2006). Discrimination Training Alters Object Representations in Human Extrastriate Cortex. *Journal of Neuroscience*, *26*(50), 13025–13036. https://doi.org/10.1523/JNEUROSCI.2481-06.2006

Pagani, L. S., Brière, F. N., & Janosz, M. (2017). Fluid reasoning skills at the high school transition predict subsequent dropout. *Intelligence*, *62*, 48–53. https://doi.org/10.1016/j.intell.2017.02.006

Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), 97–107. https://doi.org/10.1038/nrn1603

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peirce, J. W., 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. https://doi.org/10.3758/s13428-018-01193-y

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23–38. https://doi.org/10.1016/j.neuroscience.2005.06.005

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, *76*(2), 270–279. https://doi.org/10.3758/s13414-013-0605-z

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., & Petersen, S. E. (2011). Functional Network Organization of the Human Brain. *Neuron*, *72*(4), 665–678. https://doi.org/10.1016/j.neuron.2011.09.006

Primi, R., Ferrão, M. E., & Almeida, L. S. (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. *Learning and Individual Differences*, *20*(5), 446–451. https://doi.org/10.1016/j.lindif.2010.05.001

Quintana, J., & Fuster, J. M. (1992). Mnemonic and predictive functions of cortical neurons in a memory task. *Neuroreport*, *3*(8), 721–724. https://doi.org/10.1097/00001756-199208000-00018

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(2), 676–682. https://doi.org/10.1073/pnas.98.2.676

Rangelov, D., & Mattingley, J. B. (2020). Evidence accumulation during perceptual decision-making is sensitive to the dynamics of attentional selection. *NeuroImage*, *220*, 117093. https://doi.org/10.1016/j.neuroimage.2020.117093

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of What and Where in the Primate Prefrontal Cortex. *Science*, *276*(5313), 821–824. https://doi.org/10.1126/science.276.5313.821

Reddy, L., Kanwisher, N. G., & VanRullen, R. (2009). Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences*, *106*(50), 21447–21452. https://doi.org/10.1073/pnas.0907330106

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4. *Journal of Neuroscience*, *19*(5), 1736–1753. https://doi.org/10.1523/JNEUROSCI.19-05-01736.1999

Reynolds, J. R., O'Reilly, R. C., Cohen, J. D., & Braver, T. S. (2012). The Function and Organization of Lateral Prefrontal Cortex: A Test of Competing Hypotheses. *PLOS ONE*, *7*(2), e30284. https://doi.org/10.1371/journal.pone.0030284

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Riley, M. R., Qi, X.-L., & Constantinidis, C. (2017). Functional specialization of areas along the anterior–posterior axis of the primate prefrontal cortex. *Cerebral Cortex*, *27*(7), 3683–3697. https://doi.org/10.1093/cercor/bhw190

Robinson, A. K., Grootswagers, T., & Carlson, T. A. (2019). The influence of image masking on object representations during rapid serial visual presentation. *NeuroImage*, *197*, 224–231. https://doi.org/10.1016/j.neuroimage.2019.04.050

Rogers, R., & Monsell, S. (1995). Costs of a Predictable Switch Between Simple Cognitive Tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231. https://doi.org/10.1037/0096-3445.124.2.207

Rorden, C. (2021). *MRIcron* [Pascal]. https://github.com/neurolabusc/MRIcron (Original work published 2015)

Sayalı, C., & Badre, D. (2018). Neural systems of cognitive demand avoidance. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2018.06.016

Scalf, P. E., Torralbo, A., Tapia, E., & Beck, D. M. (2013). Competition explains limited attention and perceptual resources: Implications for perceptual load and dilution theories. *Cognition*, *4*, 243. https://doi.org/10.3389/fpsyg.2013.00243

Schira, M. M., Tyler, C. W., Breakspear, M., & Spehar, B. (2009). The Foveal Confluence in Human Visual Cortex. *Journal of Neuroscience*, *29*(28), 9050–9058. https://doi.org/10.1523/JNEUROSCI.1760-09.2009

Schneider, D. W., & Logan, G. D. (2006). Hierarchical Control of Cognitive Processes: Switching Tasks in Sequences. *Journal of Experimental Psychology*, *135*(4), 623–640.

Shallice, T., Burgess, P., & Robertson, I. (1996). The Domain of Supervisory Processes and Temporal Organization of Behaviour [and Discussion]. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *351*(1346), 1405–1412. https://doi.org/10.1098/rstb.1996.0124

Shashidhara, S., Spronkers, F. S., & Erez, Y. (2020). Individual-subject Functional Localization Increases Univariate Activation but Not Multivariate Pattern Discriminability in the "Multiple-demand" Frontoparietal Network. *Journal of Cognitive Neuroscience*, 1–21. https://doi.org/10.1162/jocn_a_01554

Sigala, N., Kusunoki, M., Nimmo-Smith, I., Gaffan, D., & Duncan, J. (2008). Hierarchical coding for sequential task events in the monkey prefrontal

cortex. *Proceedings of the National Academy of Sciences*, *105*(33), 11969–11974. https://doi.org/10.1073/pnas.0802569105

Siugzdaite, R., Bathelt, J., Holmes, J., & Astle, D. E. (2020). Transdiagnostic Brain Mapping in Developmental Disorders. *Current Biology*, *30*(7), 1245-1257.e4. https://doi.org/10.1016/j.cub.2020.01.078

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., & Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, *106*(31), 13040–13045. https://doi.org/10.1073/pnas.0905267106

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Smith, V., Duncan, J., & Mitchell, D. J. (2021). Roles of the Default Mode and Multiple-Demand Networks in Naturalistic versus Symbolic Decisions. *Journal of Neuroscience*, *41*(10), 2214–2228. https://doi.org/10.1523/JNEUROSCI.1888-20.2020

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. G. (2017). Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *Journal of Neuroscience*, *37*(27), 6503–6516. https://doi.org/10.1523/JNEUROSCI.3364-16.2017

Sporns, O., Chialvo, D. R., Kaiser, M., & Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, *8*(9), 418–425. https://doi.org/10.1016/j.tics.2004.07.008

Spreng, R. N., Sepulcre, J., Turner, G. R., Stevens, W. D., & Schacter, D. L. (2012). Intrinsic Architecture Underlying the Relations among the Default, Dorsal Attention, and Frontoparietal Control Networks of the Human Brain. *Journal of Cognitive Neuroscience*, *25*(1), 74–86. https://doi.org/10.1162/jocn_a_00281

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, *65*, 69–82. https://doi.org/10.1016/j.neuroimage.2012.09.063

Stocco, A., Prat, C. S., & Graham, L. K. (2021). Individual Differences in Reward-Based Learning Predict Fluid Reasoning Abilities. *Cognitive Science*, *45*(2), e12941. https://doi.org/10.1111/cogs.12941

Stokes, M. (2011). The Spatiotemporal Structure of Population Coding in Monkey Parietal Cortex. *Journal of Neuroscience*, *31*(4), 1167–1169. https://doi.org/10.1523/JNEUROSCI.5144-10.2011

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron*, *78*(2), 364–375. https://doi.org/10.1016/j.neuron.2013.01.039

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Tang, C., Herikstad, R., Parthasarathy, A., Libedinsky, C., & Yen, S.-C. (2020). Minimally dependent activity subspaces for working memory and motor preparation in the lateral prefrontal cortex. *ELife*, *9*, e58154. https://doi.org/10.7554/eLife.58154

Tobler, P. N., Fiorillo, C. D., & Schultz, W. (2005). Adaptive Coding of Reward Value by Dopamine Neurons. *Science*, *307*(5715), 1642–1645. https://doi.org/10.1126/science.1105370

Uddin, L. Q., Yeo, B. T. T., & Spreng, R. N. (2019). Towards a universal taxonomy of macro-scale functional human brain networks. *Brain Topography*, *32*(6), 926–942. https://doi.org/10.1007/s10548-019-00744-6

Uehara, G., Adachi, Y., Kawai, J., Shimogawara, M., Higuchi, M., Haruta, Y., Ogata, H., & Kado, H. (2003). Multi-channel SQUID systems for biomagnetic measurement. *IEICE Trans. Electron.*, *E86-C*(1), 43–54. Scopus.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

van Loon, A. M., Olmos-Solis, K., Fahrenfort, J. J., & Olivers, C. N. L. (2018). Current and future goals are represented in opposite patterns in object-selective cortex. *ELife*, *7*, e38677. https://doi.org/10.7554/eLife.38677

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin*, *136*(4), 601–626. https://doi.org/10.1037/a0019791

Vidaurre, D., Myers, N., Stokes, M., Nobre, A. C., & Woolrich, M. W. (2019). Temporally unconstrained decoding reveals consistent but time-varying stages of stimulus processing. *Cerebral Cortex*, *29*(2), 863–874. https://doi.org/10.1093/cercor/bhy290

Vidaurre, D., Quinn, A. J., Baker, A. P., Dupret, D., Tejero-Cantero, A., & Woolrich, M. W. (2016). Spectrally resolved fast transient brain states in electrophysiological data. *NeuroImage*, *126*, 81–95. https://doi.org/10.1016/j.neuroimage.2015.11.047

Wang, X., Gao, Z., Smallwood, J., & Jefferies, E. (2021). Both default and multiple-demand regions represent semantic goal information. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.1782-20.2021

Waskom, M. L., Kumaran, D., Gordon, A. M., Rissman, J., & Wagner, A. D. (2014). Frontoparietal Representations of Task Context Support the Flexible Control of Goal-Directed Cognition. *Journal of Neuroscience*, *34*(32), 10743–10755. https://doi.org/10.1523/JNEUROSCI.5282-13.2014

Waszak, F., Hommel, B., & Allport, A. (2003). Task-switching and long-term priming: Role of episodic stimulus–task bindings in task-shift costs. *Cognitive Psychology*, *46*(4), 361–413. https://doi.org/10.1016/S0010-0285(02)00520-0

Weiler, J., Hassall, C. D., Krigolson, O. E., & Heath, M. (2015). The unidirectional prosaccade switch-cost: Electroencephalographic evidence of task-set inertia in oculomotor control. *Behavioural Brain Research*, *278*, 323–329. https://doi.org/10.1016/j.bbr.2014.10.012

Wen, T., Duncan, J., & Mitchell, D. J. (2019). The time-course of component processes of selective attention. *NeuroImage*, *199*, 396–407. https://doi.org/10.1016/j.neuroimage.2019.05.067

Wen, T., Duncan, J., & Mitchell, D. J. (2020). Hierarchical Representation of Multistep Tasks in Multiple-Demand and Default Mode Networks. *Journal of Neuroscience*, *40*(40), 7724–7738. https://doi.org/10.1523/JNEUROSCI.0594-20.2020

Wilson, C. R. E., Gaffan, D., Browning, P. G. F., & Baxter, M. G. (2010). Functional localization within the prefrontal cortex: Missing the forest for the trees? *Trends in Neurosciences*, *33*(12), 533–540. https://doi.org/10.1016/j.tins.2010.08.001

Woolgar, A., Afshar, S., Williams, M. A., & Rich, A. N. (2015). Flexible Coding of Task Rules in Frontoparietal Cortex: An Adaptive System for Flexible Cognitive Control. *Journal of Cognitive Neuroscience*, *27*(10), 1895–1911. https://doi.org/10.1162/jocn_a_00827

Woolgar, A., Bor, D., & Duncan, J. (2013). Global Increase in Task-related Fronto-parietal Activity after Focal Frontal Lobe Lesion. *Journal of Cognitive Neuroscience*, *25*(9), 1542–1552. https://doi.org/10.1162/jocn_a_00432

Woolgar, A., Dermody, N., Afshar, S., Williams, M. A., & Rich, A. N. (2019). Meaningful patterns of information in the brain revealed through analysis of errors. *BioRxiv*, 673681. https://doi.org/10.1101/673681

Woolgar, A., Duncan, J., Manes, F., & Fedorenko, E. (2018). Fluid intelligence is supported by the multiple-demand system not the language system. *Nature Human Behaviour*, *2*(3), 200–204. https://doi.org/10.1038/s41562-017-0282-3

Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011). Adaptive Coding of Task-Relevant Information in Human Frontoparietal Cortex. *Journal of Neuroscience*, *31*(41), 14592–14599. https://doi.org/10.1523/JNEUROSCI.2616-11.2011

Woolgar, A., Jackson, J., & Duncan, J. (2016). Coding of Visual, Auditory, Rule, and Response Information in the Brain: 10 Years of Multivoxel Pattern Analysis. *Journal of Cognitive Neuroscience*, *28*(10), 1433–1454. https://doi.org/10.1162/jocn_a_00981

Woolgar, A., Parr, A., Cusack, R., Thompson, R., Nimmo-Smith, I., Torralva, T., Roca, M., Antoun, N., Manes, F., & Duncan, J. (2010). Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *107*(33), 14899–14902. https://doi.org/10.1073/pnas.1007928107

Woolgar, A., Thompson, R., Bor, D., & Duncan, J. (2011). Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. *NeuroImage*, *56*(2), 744–752. https://doi.org/10.1016/j.neuroimage.2010.04.035

Woolgar, A., Williams, M. A., & Rich, A. N. (2015). Attention enhances multi-voxel representation of novel objects in frontal, parietal and visual cortices. *NeuroImage*, *109*(Supplement C), 429–437. https://doi.org/10.1016/j.neuroimage.2014.12.083

Woolgar, A., & Zopf, R. (2017). Multisensory coding in the multiple-demand regions: Vibrotactile task information is coded in frontoparietal cortex. *Journal of Neurophysiology*, *118*(2), 703–716. https://doi.org/10.1152/jn.00559.2016

Wray, C., Kowalski, A., Mpondo, F., Ochaeta, L., Belleza, D., DiGirolamo, A., Waford, R., Richter, L., Lee, N., Scerif, G., Stein, A. D., Stein, A., & COHORTS. (2020). Executive functions form a single construct and are associated with schooling: Evidence from three low- and middle- income countries. *PLOS ONE*, *15*(11), e0242936. https://doi.org/10.1371/journal.pone.0242936

Wrulich, M., Brunner, M., Stadler, G., Schalke, D., Keller, U., & Martin, R. (2014). Forty years on: Childhood intelligence predicts health in middle adulthood. *Health Psychology*, *33*(3), 292–296. https://doi.org/10.1037/a0030727

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. https://doi.org/10.1038/s41593-018-0310-2

Yeari, M., & Goldsmith, M. (2010). Is object-based attention mandatory? Strategic control over mode of attention. *Journal of Experimental Psychology. Human*

Perception and Performance, *36*(3), 565–579. https://doi.org/10.1037/a0016897

Yeo, T. B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

Yi, D.-J., Woodman, G. F., Widders, D., Marois, R., & Chun, M. M. (2004). Neural fate of ignored stimuli: Dissociable effects of perceptual and working memory load. *Nature Neuroscience*, *7*(9), 992–996. https://doi.org/10.1038/nn1294

Yip, H. M. K., Cheung, L. Y. T., Ngan, V. S. H., Wong, Y. K., & Wong, A. C.-N. (2021). The Effect of Task on Object Processing revealed by EEG decoding. *BioRxiv*, 2020.08.18.255018. https://doi.org/10.1101/2020.08.18.255018

Zhang, H., Gou, R., Shang, J., Shen, F., Wu, Y., & Dai, G. (2021). Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition. *Frontiers in Physiology*, *12*. https://doi.org/10.3389/fphys.2021.643202

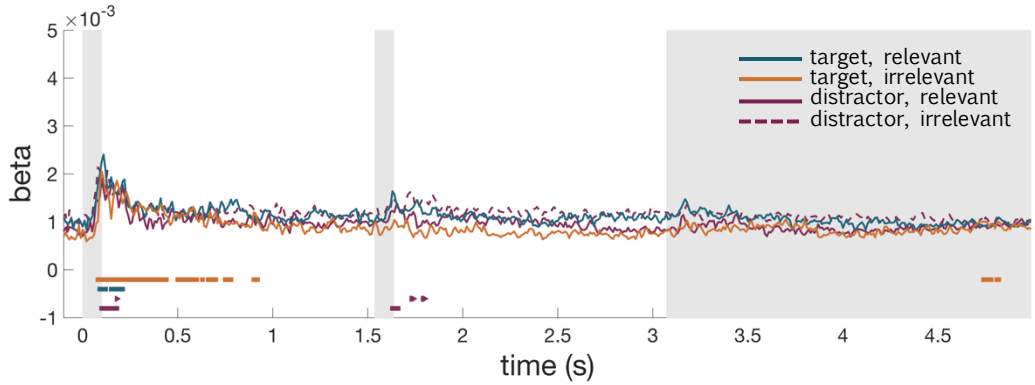
Zink, N., Lenartowicz, A., & Markett, S. (2021). A new era for executive function research: On the transition from centralized to distributed executive functioning. *Neuroscience & Biobehavioral Reviews*, *124*, 235–244. https://doi.org/10.1016/j.neubiorev.2021.02.011

7. Appendices

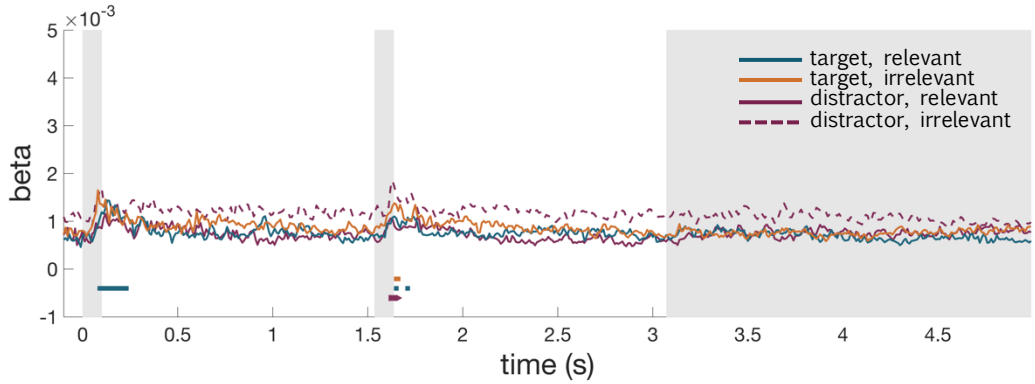
7.1. Appendix A

Information Captured by Models Predicting Different Brain Responses to Different Within-Category Stimuli

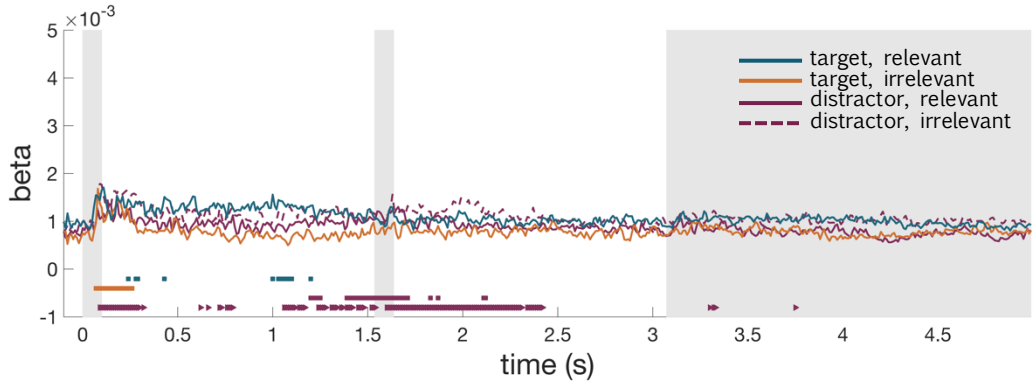
A. Ventral Visual Cortex, Shape Information, Epoch 1



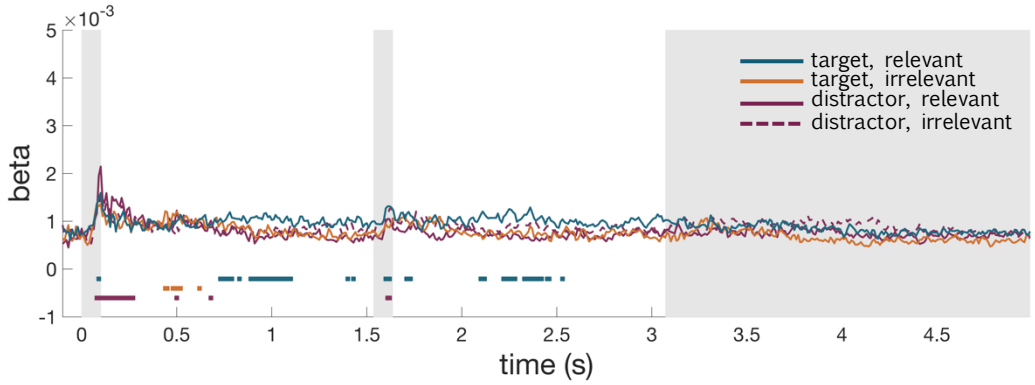
B. Ventral Visual Cortex, Shape Information, Epoch 2



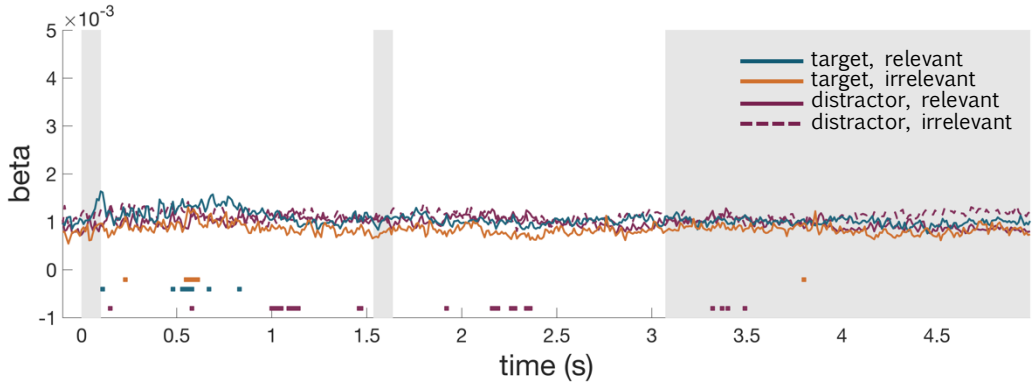
C. Ventral Visual Cortex, Colour Information, Epoch 1



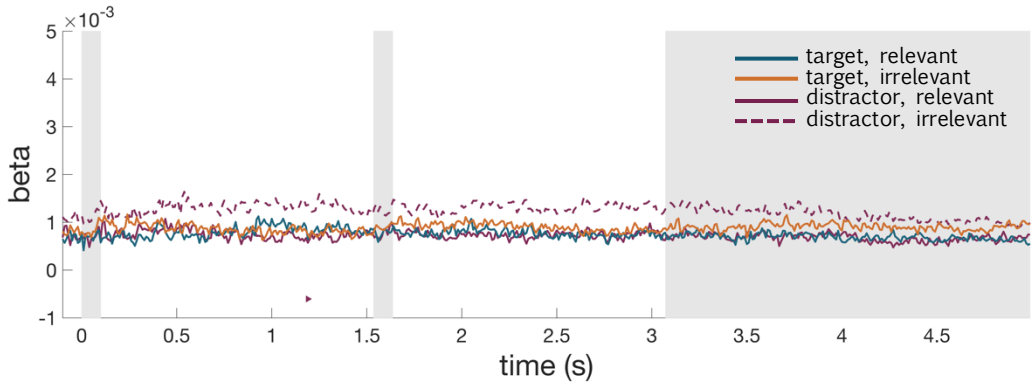
D. Ventral Visual Cortex, Colour Information, Epoch 2



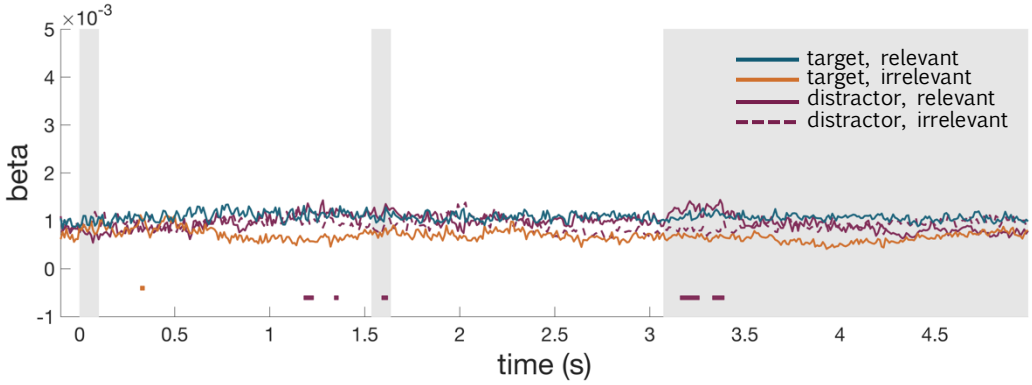
E. Multiple-Demand Network, Shape Information, Epoch 1



F. Multiple-Demand Network, Shape Information, Epoch 2



G. Multiple-Demand Network, Colour Information, Epoch 1



H. Multiple-Demand Network, Colour Information, Epoch 2