Title: Spatial and feature-selective attention have distinct effects on population-level tuning

Authors: Erin Goddard^{i,ii}
 Thomas A. Carlson^{ii,iv}
 Alexandra Woolgar^{ii,iii}

Affiliations: (i) McGill Vision Research Group,
 McGill University,
 Montreal, QC, H3G1A4, Canada

(ii) ARC Centre of Excellence in Cognition and its Disorders (CCD), Macquarie University,
 Sydney, NSW, 2109, Australia

(iii) Perception in Action Research Centre (PARC)
 and Department of Cognitive Science
 Macquarie University,
 Sydney, NSW, 2109, Australia

(iv) School of Psychology,
 University of Sydney,
 Sydney, NSW, 2006, Australia

Corresponding Author: Erin Goddard
 McGill Vision Research Group,
 1650 Cedar Ave., Rm L11 513
 Montreal, QC, H3G1A4, Canada
 erin.goddard@mail.mcgill.ca

Conflict of interest: None to report

Abstract

Selective attention is fundamental to cognitive activity and can be deployed in different ways. Non-human primate data suggests that spatial and feature-based visual attention have qualitatively different effects on neural tuning, but this has been challenging to assess in humans. Using multivariate decoding of MEG data, we tracked the effects of spatial and feature-selective attention on population-level coding of novel objects. We found that spatial and feature-selective attention interacted multiplicatively to enhance object representation. Moreover, the two types of attention induced qualitatively different patterns of enhancement in occipital cortex, and these differences were accounted for by the principles of response-gain and tuning curve sharpening derived from single-unit work. A novel information flow analysis further showed that stimulus representations in occipital cortex were Granger-caused by coding in frontal cortices earlier in time. We find that human spatial and feature-selective attention rely on qualitatively different, interacting, neural mechanisms.

At any moment, there is far more information available from our senses than we can possibly process at once. Accordingly, only a subset of the available information is processed to a high level, making it crucial that brain can dynamically devote greatest processing resources to the most relevant information. Our ability to selectively attend to relevant information is remarkably flexible. For instance, we can adapt our attentional state by directing our attention in space (spatial attention, e.g. attend left), to a specific feature dimension (feature-*selective* attention, e.g. detect changes in color across a scene) or based on a particular feature value along that feature dimension (feature-*based* attention, e.g. find all the red objects), using the definitions of Chen et al. (2012). Each of these types of attention can change behavior, improving performance related to the attended location or feature-dimension, while decreasing performance on the ignored dimension/location (Pestilli and Carrasco, 2005; Rossi and Paradiso, 1995; Saenz et al., 2003; Carrasco, 2011), consistent with neural resources being redistributed.

What is the neural basis for this important ability, and to what extent do the same mechanisms give rise to spatial and feature-based attentional enhancements? Shifts in attention induce changes in the responses of individual neurons (Sprague et al., 2015; Reynolds and Heeger, 2009; Maunsell, 2015), change the overall responsiveness of brain

34 regions (Corbetta et al., 1990; Chawla et al., 1999; Saenz et al., 2002, 2003; Serences
35 and Boynton, 2007; Gouws et al., 2014), and change the information carried by a
36 population response (Guggenmos et al., 2015; Woolgar et al., 2015; Vaziri-Pashkam
37 and Xu, 2017). The most marked difference between spatial and feature-based
38 attention is that the effects of spatial attention vary according to the part of the visual
39 field to which a cell responds, whereas feature-based attention is spatially diffuse,
40 changing the responses of neurons (Treue and Martinez-Trujillo, 1999; McAdams and
41 Maunsell, 2000; Martinez-Trujillo and Treue, 2004) and voxels (Saenz et al., 2002;
42 Serences and Boynton, 2007) across the visual field, rather than being restricted to the
43 attended location or the stimulus location.

44 The reported effects of spatial attention on the tuning of individual neurons are diverse:
45 its effects have been characterized as multiplicative response gain (McAdams and
46 Maunsell, 1999; Treue and Martinez-Trujillo, 1999; Lee and Maunsell, 2010b), contrast
47 gain (Li and Basso, 2008; Martinez-Trujillo and Treue, 2002; Reynolds et al., 2000), or
48 a combination of these effects (Williford and Maunsell, 2006). There have also been
49 mixed results regarding the effect of spatial attention on contrast response functions
50 measured with fMRI (Buracas and Boynton, 2007; Li et al., 2008). Fewer studies have
51 investigated the effects of feature-based attention, and only a subset of these where
52 shifts in feature-based attention were not accompanied by changes in spatial attention
53 (Maunsell and Treue, 2006). Intriguingly, feature-based attention may affect the tuning
54 of individual neurons in a subtly different manner to spatial attention. In an influential
55 electrophysiological study Martinez-Trujillo and Treue (2004) found effects at the
56 single-unit level which would lead to a ‘sharpening’ of the population response around
57 the attended feature value across the visual field. In a recent MEG study Bartsch et al.
58 (2017) reports similar sharpening of the population response with attention to color.
59 However, even this difference in the effects of spatial and feature-based attention does
60 not eliminate the possibility of a unified attentional system, where stimulus location is
61 treated as one of many stimulus features that can potentially be selected with attention
62 (Treue and Martinez-Trujillo, 1999; Maunsell and Treue, 2006; Maunsell, 2015).

63 While there is an increasing body of work investigating the effects of spatial and
64 feature-based or feature-selective attention, there are few studies that directly compare

65 these two attention types. In one of the few previous studies that simultaneously
 66 manipulated both spatial and feature-selective attention Cohen and Maunsell (2011)
 67 implied highly similar processes of spatial and feature-selective attention, affecting the
 68 same subpopulations of neurons. The main difference between their effects was across
 69 hemispheres: for feature-selective but not spatial attention the effects were correlated
 70 across hemispheres. Directly comparing attention types is critical for resolving whether
 71 and how these attention types produce different effects on the population code, and
 72 how their effects interact.

73 The overlapping characteristics of these two attention types make them difficult to
 74 separate, as does the diversity of their reported effects. Another complicating factor is
 75 that much of our current understanding of attention comes from work exploring its
 76 effects on individual neurons, but attention can also induce changes in the information
 77 represented by a population of neurons that will not be revealed in the tuning curves of
 78 individual neurons (Sprague et al., 2015). For instance, attention has been shown to
 79 decrease response variance (e.g. Mitchell et al. 2007), and decrease (Cohen and
 80 Maunsell, 2009) or increase (Ruff and Cohen, 2014) the correlation between pairs of
 81 neurons. It can be difficult to predict how each of changes should affect the
 82 information represented by the population response, for example, predicting how
 83 changes in correlation across neurons will affect population codes is non-trivial
 84 (Moreno-Bote et al., 2014). There is a need, then, to complement measurements of the
 85 effects of attention on single-unit responses with measurements of its effects on
 86 information carried by a population of cells (Sprague et al., 2015), via simultaneous
 87 multi-electrode recordings (Cohen and Maunsell, 2011) or neuroimaging. Multivariate
 88 classification analyses, applied to multi-electrode recordings or neuroimaging measures,
 89 provide a means of measuring the overall stimulus-related information that is carried
 90 by a population response. Unlike the tuning of single neurons, any signal or noise
 91 correlations that could decrease or increase information carried by the population
 92 response (Moreno-Bote et al., 2014) should affect classifier accuracy. This sensitivity to
 93 additional factors make classifier accuracy an ideal intermediate level of description for
 94 linking single-unit responses to the information in the population response which is
 95 available for readout by other brain regions, and to the organism's percept/behaviour
 96 (Carlson et al., 2018).

97 Another key question for understanding attentional modulation of visual information is
 98 to identify the regions that drive these changes in processing, and when and how they
 99 influence visual cortical areas. There is evidence that some prefrontal cortical (PFC)
 100 regions are critically involved in visual attention and task-based modulations in
 101 response, including the frontal eye fields (Moore et al., 2003; Gregoriou et al., 2012;
 102 Zhou and Desimone, 2011), the ventral prearcuate region (Bichot et al., 2015), the
 103 superior precentral sulcus (Jerde et al., 2012) and lateral PFC (Tremblay et al., 2015;
 104 Luo and Maunsell, 2018). Selective prioritisation of task-relevant information in
 105 prefrontal cortex (e.g. Duncan 2001) may provide a source of bias, driving processing
 106 in visual cortices in favour of task relevant information (Desimone and Duncan, 1995;
 107 Dehaene et al., 1998; Miller and Cohen, 2001). But precisely what this influence is, and
 108 when it occurs, remains unknown.

109 Here we measured the effects of spatial and feature-selective attention within the same
 110 datasets of magnetoencephalography (MEG) recordings (n=20), enabling us to directly
 111 compare and contrast their effects. We obtained fine timescale measures of
 112 stimulus-related information in two large regions of interest (ROIs): visual cortex and
 113 frontal/prefrontal cortex. For both ROIs, we found strong, multiplicative effects of
 114 spatial and feature-selective attention, but these only emerged relatively late (>200ms
 115 after stimulus onset). We used an information flow analysis to test for how the two
 116 ROIs were interacting over time: we measured Granger-causal relationships between
 117 their stimulus-related information. This revealed that for visual cortex, the strongest
 118 attentional modulation occurred after the onset of feedback from frontal regions. We
 119 also tested whether spatial and feature-selective attention induced different effects on
 120 the population response. We predicted that both types of attention would enhance
 121 stimulus-related information, but that feature-selective attention would induce
 122 sharpening of the population response around the attended feature value, whereas
 123 spatial attention would induce a more generalized enhancement across feature values.
 124 In line with these predictions, we found that spatial attention produced relatively more
 125 enhancement of discriminability for stimulus pairs that were far apart in feature space,
 126 while the effects of feature-selective attention were relatively stronger for stimulus pairs
 127 that were closer in feature space.

128 Results

129 Performance on behavioral task

130 Participants (n=20) viewed a series of stimuli while we recorded their neural activity
 131 using MEG. On every trial there were two objects on the screen, one on the left and
 132 one on the right of fixation (Figure 1A). Participants were instructed to covertly
 133 attended either to the stimulus on the left or right of fixation (spatial attention
 134 manipulation), and they were required to make a judgment based on the target
 135 object's color or shape (feature-selective attention manipulation). As shown in Figure
 136 1B, there were four stimulus colors ranging from red to green, and four shapes ranging
 137 from strongly X-shaped to strongly non-X-shaped. The four feature values along each
 138 dimension meant that for both tasks the stimuli were either far from the decision
 139 boundary (e.g. strongly red; 'easy' trials) or closer to the decision boundary (e.g.
 140 weakly red; 'hard' trials). As expected, participants were faster and more accurate at
 141 identifying color and shape for objects that were far from the decision boundary
 142 relative to those that were near the decision boundary. For the color task, the average
 143 accuracy was 95.6% (std 3.6%) on the easy trials, and 85.2% (std 7.3%) on the hard
 144 trials, while median reaction time was 0.69s on the easy trials and 0.81s on the hard
 145 trials. Similarly, for the shape task the average accuracy was 94.1% (std 3.5%) on the
 146 easy trials, and 74.1% (std 4.7%) on the hard trials, while median reaction time was
 147 0.74s and 0.82s on the easy and hard trials respectively.

148 Decoding attentional state

149 We trained classifiers to make a series of orthogonal discriminations in order to
 150 quantify neural information about the participant's task and the stimulus. First, we
 151 trained classifiers to discriminate the participant's attentional set: the attended
 152 location (left versus right) and feature (color versus shape). Second, we trained
 153 classifiers to discriminate the stimuli and compared the strength of discrimination
 154 between attentional conditions.

155 Our first question concerned the timecourse with which we could decode information
 156 about the participant's attentional state. For both ROIs we asked whether we could

Figure 1: Visual stimuli, showing attention conditions (**A**) and stimulus dimensions (**B**). **Attention conditions (A):** At the start of each block of trials, participants were told the location to which they should direct their attention (left or right of fixation), and which task they should perform for that block of stimuli: either reporting on the target object's shape ('X-shaped' or 'non-X-shaped') or color (reddish or greenish). Two objects appeared on each trial, and participants covertly attended to one while we used eye tracking to monitor their fixation. The example above illustrates how the same stimulus configuration was used in each of the four attention/task conditions. The dotted circle indicates the location of spatial attention, and was not visible during the experiment. **Stimulus dimensions (B):** Each object varies systematically along 2 dimensions, color and shape. In the color task, participants categorized the attended object as either 'greenish' or 'reddish'. In the shape task, participants categorized the attended object as either 'X-shaped' or 'non-X-shaped', based on the orientations of the object's spikes. To encourage participants to attend to the overall object shape rather than (for example) the orientation of a single spike, on each trial the object was randomly selected from 100 exemplars with the target shape statistics, and there were variations between exemplars in the location, length and orientation of the spikes. This is illustrated above in the shape variation between objects in the the same column.

Figure 2: Regions of interest (A) and classifier performance across participants (n=20) for decoding attention condition using occipital sources (B) or frontal sources (C). A: the ‘Occipital’ (cyan) and ‘Frontal’ (yellow) regions of interest shown on the partially inflated cortical surface of the ICBM152 template brain. B and C: At each timepoint, classifiers were trained to discriminate the location and feature to which participants were attending. The shaded error bars indicate the 95% confidence interval of the between-subject mean. At the top of each plot, boxes indicate the time of the stimulus presentation (shaded area indicates onset until the median duration of 92ms), the reaction time (RT) distribution (shaded area includes RTs within the first and third quartiles, black line indicates median RT), and the time during which participants received feedback on their accuracy on those trials where their RT was <1s (77% of trials). On trials where RT was >1s (23% of trials), the 200 ms feedback started at the time of response. Classification performance can be above chance in the pre-stimulus period since attentional condition was blocked: participants knew which attentional condition to perform before the stimulus appeared. Nonetheless, decoding of attentional condition improved dramatically after the stimulus was presented, and peaked earlier when classifiers were decoding attended location (270 ms and 390 ms after stimulus onset for occipital and frontal ROIs respectively) than when decoding attended object feature (455 ms after stimulus onset for both ROIs). Shaded gray region around x-axis indicates the 95% confidence intervals of the same classifications when performed on permuted data (chance performance level). Colored crosses below the plot indicate that at every time point classifier accuracy was significantly above the average chance performance level (chance $d' = 0.0001$ (A), 0.0000 (B); $p < 0.05$ in a one-tailed t -test of the between-subject mean, FDR corrected at $q < 0.05$ for multiple comparisons across time points).

Figure 3: Classifier performance across participants (n=20) for decoding object features. For both occipital (A) and frontal (B) regions of interest, classifiers were trained to discriminate the color (upper plots) and shape (lower plots) of attended and unattended objects. Classifier performance is shown for each attention condition separately: attended location, attended feature (aLaF); attended location, unattended feature (aLuF); unattended location, attended feature (uLaF); and unattended location, unattended feature (uLuF). Shaded error bars indicate the 95% confidence interval of the between-subject mean, and boxes at the top of the plot show relevant trial events. The shaded gray region around the x-axis indicates the 95% confidence intervals of the four classifications when performed on randomly permuted data (the empirical null distribution). Small dots below each plot indicate timepoints at which the classification of matching color was above chance level (FDR corrected, $q < 0.05$). Below these, crosses indicate timepoints at which there was a significant effect (FDR corrected, $q < 0.05$) of spatial attention (blue asterisks), feature-selective attention (red asterisks) or an interaction of the two (black asterisks).

157 decode where participants were attending (left or right) and what task they were
158 performing (color or shape) at each timepoint. Figure 2 shows that attentional state
159 could be decoded from both occipital and frontal sources at most time points (at most
160 time points the between-subjects mean was above zero when tested with a one-tailed
161 t -test, $p < 0.05$, FDR corrected at $q < 0.05$ for multiple comparisons across time points,
162 (Genovese et al., 2002)). The period of above-chance classifier performance for
163 attended location included time points before the onset of the stimulus, when
164 participants knew their task and were waiting for the stimulus to appear: classifier
165 performance at this time was low but significantly above chance for both ROIs.
166 Although we do not have a behavioral measure of the participant's attentional state at
167 this time, these pre-stimulus effects suggest that neural activity differed with the
168 location to which participants were covertly attending, or to which they were preparing
169 to covertly attend. This interpretation is consistent with previous work demonstrating
170 the pre-stimulus effects of spatial attention on neural coding (Kastner et al., 1999; Ress
171 et al., 2000).

172 Decoding of both attended location and attended object feature increased substantially
173 once the stimulus appeared. This presumably reflects changes in neural activity
174 associated with enhancing the neural representation of the attended object and the
175 task-relevant feature and/or suppressing the neural representation of the unattended
176 object and the task-irrelevant feature of the attended object. Classification of attended
177 feature was above chance from 70ms and 135ms after stimulus onset in the occipital and
178 frontal ROIs respectively. Classifier performance peaked earlier when classifiers were
179 decoding attended location (270ms and 390ms after stimulus onset for the occipital and
180 frontal ROIs respectively) than when decoding attended object feature (455ms after
181 stimulus onset for both the occipital and frontal ROIs). These timing differences could
182 reflect differences between the timing of spatial and feature-selective attention
183 processes, but these data are not conclusive (particularly for the onset times) since the
184 lower overall accuracy for decoding of attended feature may have contributed to the
185 delay in onset of above-chance classifier performance (Grootswagers et al., 2017).

186 Decoding object features: color and shape

187 We next asked whether we could use the neural signal to decode the features of the
 188 attended and unattended stimuli, and how this information varied over time and
 189 attentional state. Our design included simultaneous manipulations of both attended
 190 feature and attended location, enabling us to ask how these different types of attention
 191 interact. By balancing the training trials across irrelevant features and creating
 192 averaged ‘pseudo-trials’ (see Methods), we were able to train classifiers to discriminate
 193 the color and shape of both the attended and non-attended object. Figure 3 shows the
 194 decoding of object color and shape for each attention condition, in each case averaged
 195 across 6 pairwise comparisons, and transformed classifier weights, showing the most
 196 informative locations in each ROI, are summarized in Figure S7.

197 For both decoding object color and object shape, 2-way ANOVAs revealed significant
 198 main effects of spatial attention and feature-selective attention, and significant
 199 interactions between these effects, at the times indicated by blue, red and black crosses
 200 respectively in Figure 3 ($p < 0.05$, in each case FDR corrected at $q < 0.05$ across time
 201 points). In the occipital ROI, for both object shape and object color, we found an
 202 initial peak of robust classifier performance which showed a small effect of spatial
 203 attention, followed by a selective increase in the neural information concerning the
 204 relevant feature of the attended object, while all other information was attenuated.
 205 Around the initial peak of stimulus decoding spatial attention produced a small but
 206 significant increase in decoding of both color and shape (blue crosses $< 100ms$ in
 207 Figure 3A, at $75ms$ for decoding color and 90 and $105ms$ for decoding shape). After the
 208 initial peak, the representation of task-relevant stimulus-related information was
 209 sustained, persisting beyond the offset of the stimulus (median: $92ms$) and beyond the
 210 median response time ($770ms$). In the frontal ROI, above-chance decoding accuracy
 211 emerged later than for the occipital ROI, and was only seen for the attended feature at
 212 the attended location. This is consistent with frontal areas prioritizing representation
 213 of task-relevant information.

214 Interestingly, for both occipital and frontal regions, the effects of spatial and
 215 feature-selective attention interacted with each other, consistent with their effects
 216 combining in a multiplicative rather than an additive manner. For both occipital and

frontal ROIs, whenever both spatial and feature-selective attention had significant effects there was generally also an interaction. The interaction reflected the selective boost in the decoding of the attended feature at the attended location, with little enhancement in classifier performance for spatial attention in the absence of feature-selective attention or for feature-selective attention in the absence of spatial attention. We think it is unlikely that the lack of an independent effect of feature-selective attention in our data reflects a true absence of any effect of feature-selective attention at the unattended location, since there are numerous reports of feature-based attention having effects at unattended locations (e.g. Treue and Martinez-Trujillo 1999; McAdams and Maunsell 2000; Martinez-Trujillo and Treue 2004; Saenz et al. 2002; Serences and Boynton 2007; Ipata et al. 2012; Bichot et al. 2015). However, there are two differences between our results and this previous work which may reflect a genuine difference. Firstly, these modulations are typically reported during responses to stimuli at the unattended location (Treue and Martinez-Trujillo, 1999; McAdams and Maunsell, 2000; Martinez-Trujillo and Treue, 2004; Saenz et al., 2002), whereas here the effects are predominantly after stimulus offset (but see Serences and Boynton 2007). Secondly, in our experiment the participants were attending to a feature dimension (feature-selective attention) rather than a particular feature value (feature-based attention), so the absence of an effect of feature-selective attention at the unattended location may reflect a difference between these types of feature attention.

Despite these protocol differences, a more parsimonious explanation is that any effects of feature-selective attention on the representation of the unattended stimulus were too small to detect. For both feature-based and feature-selective attention, a weak effect of feature attention at unattended locations is also predicted where feature attention is spatially diffuse but there is a multiplicative interaction between feature and spatial attention. The normalization model of Reynolds and Heeger (2009), which is considered in greater detail below, includes versions with either additive or multiplicative interactions between spatial and feature-based attention. The multiplicative version of their model, which is most consistent with our data, predicts a strong interaction between the effects of spatial and feature-based attention, and a very small effect of feature-selective attention alone (see Figure 7B and discussion below),

249 which may have been too small to detect here. This interaction between spatial and
250 feature-selective attention demonstrates that the neural information was highly
251 adapted to the participant's task, and that the brain is efficiently selecting only
252 relevant information for sustained processing.

253 The earliest peaks in classifier performance for the occipital ROI showed only a slight
254 modulation with attention. For decoding object color, the initial peak was at
255 105 – 110ms after stimulus onset in all attention conditions, and there was no significant
256 effect of attended location or attended feature at either time point (2-way ANOVAs,
257 with subject as a random factor, at 105ms and 110ms: $F_{(1,19)} = 2.54, 2.20$, $p = .13, .15$
258 for effect of attended location; $F_{(1,19)} = .26, .40$, $p = .62, .54$ for effect of attended
259 feature). For decoding object shape, the initial peak was at 95 – 100ms after stimulus
260 onset in all conditions, and there was a small increase in classifier performance at the
261 attended location which reached significance at 95ms ($F_{(1,19)} = 4.48$, $p = .048$, $q < .05$
262 with FDR correction), and approached significance at 100ms ($F_{(1,19)} = 4.36$, $p = .051$).
263 There was no significant effect of attended feature on decoding of shape at either 95ms
264 or 100ms ($F_{(1,19)} = .18, .41$, $p = .68, .53$). The weak effects of attention on classifier
265 performance in the occipital ROI suggest that at the time of the initial peak the object
266 representation in visual cortex is primarily stimulus-driven. This is consistent with the
267 lack of above-chance decoding in the frontal ROI at this time. Previous work shows
268 that attention tends to have a greater effect on the sustained part of neural responses
269 than on onset transients (Fries et al., 2001; Cohen and Maunsell, 2009; Lee and
270 Maunsell, 2010a) (although the temporal dynamics of attentional modulation vary
271 according to task requirements (Ghose and Maunsell, 2002)). The short duration of
272 our stimulus (median: 92ms) means that we cannot confidently separate the sustained
273 part of the stimulus-driven response from responses reflecting short-term memory and
274 response preparation following stimulus offset, but our finding that the initial transient
275 is largely unaffected by attentional task is consistent with these previous results.

276 There was also a secondary early peak in the occipital ROI for decoding color
277 (~ 165 – 240ms after stimulus onset), but not for decoding shape. During this second
278 early peak for decoding color there was a significant effect of spatial attention, with
279 stronger decoding at the attended location than at the unattended location, but

280 classifier performance in all attention conditions remained relatively high compared to
281 later times, where there was a marked attenuation of classification performance for all
282 conditions except the attended feature, attended location condition.

283 At later time points there were stronger effects of both spatial and feature-selective
284 attention for both stimulus features at both ROIs, and an interaction between the
285 effects of the two types of attention. In the occipital ROI, the effect of spatial attention
286 preceded that of feature-selective attention. For decoding object color there was a
287 sustained effect of spatial attention from 165ms after stimulus onset, while the earliest
288 significant effect of feature-selective attention was 385ms after stimulus onset. For
289 shape there was a sustained effect of spatial attention from 285ms after stimulus onset,
290 and an effect of feature-selective attention from 335ms after stimulus onset. In both
291 cases (color and shape), the sustained effects of spatial and feature-selective attention
292 interacted multiplicatively (seen in the selective enhancement of the aLaF condition,
293 and the black crosses in Figure 3).

294 Information about the attended feature at the attended location (dark red lines in
295 Figure 3) had later, local peaks in the vicinity of 600ms post-stimulus onset for both
296 stimulus features in both ROIs: decoding of both color had local peaks at 540ms and
297 630ms for the occipital ROI, 595ms and 695ms for the frontal ROI; decoding of shape
298 peaked at 590ms in the occipital ROI and 595ms in the frontal ROI. Each of these
299 peaks are well after the offset of the stimulus (92ms) and just prior to the median
300 response time (770ms), suggesting that classifier performance around this later peak
301 may be associated with the participant's decision and/or the remembered feature
302 value. Since we balanced the response mapping (by switching the keys associated with
303 each response pair on half the runs) it is unlikely that the motor preparation associated
304 with the participants' response contributed to this effect.

305 For both occipital and frontal ROIs, classification of the attended feature of the
306 attended object remained above chance well after the median response time. Sustained
307 classification of the task-relevant information could reflect processing of the feedback
308 presented to participants after 1000ms (see Methods). To limit the scope of the present
309 study we include only data from 0 – 1000ms after stimulus onset in our next analyses,
310 excluding any effects due to the feedback.

311 In summary, classification performance of the occipital ROI contain early peaks in the
 312 decoding of both color and shape that showed little or no modulation with attention.
 313 At later times, both spatial and feature-selective attention had robust effects in both
 314 ROIs, and these effects were multiplicative rather than additive. In the following
 315 analyses we consider how these effects vary across classifications of varying feature
 316 difference, and we test for evidence of information exchange between the occipital and
 317 frontal ROIs.

318 Decoding object features: effect of physical difference between stimuli **319 and task difficulty**

320 Next we considered how classifier performance varied with the physical difference in
 321 the stimuli being discriminated (i.e. with task difficulty). Our design included stimuli
 322 that were far apart in feature space (e.g. ‘strongly red’ vs ‘strongly green’) and stimuli
 323 that were close in feature space (e.g. ‘strongly green’ vs ‘weakly green’). Since we
 324 included 4 steps along both color and shape dimensions, the pairs of object stimuli that
 325 classifiers were trained to discriminate could be either 1, 2 or 3 steps apart along either
 326 dimension. These pairs also differ in task difficulty: for those that are 3 steps apart the
 327 stimuli being discriminated were both from ‘easy’ trials, while those of 1 or 2 steps
 328 difference contained at least one stimulus from a ‘hard’ trial. In Figure 4 we separately
 329 consider classifier performance for pairs of different step size separation, where
 330 participants were attending to the stimulus feature and location (pairs of different step
 331 size were averaged in Figure 3).

332 For both decoding of object color and shape, in the occipital ROI performance at the
 333 early peak (at around 100 ms after stimulus onset) clearly increased with increasing
 334 step sizes. This is consistent with the classifier performance at this early time being
 335 driven by predominantly stimulus-driven neural responses in these cortical visual areas.
 336 In the case of decoding object shape this ordering persisted throughout the first 1000
 337 ms after stimulus onset for the occipital ROI, and was also seen in the frontal ROI
 338 when classification performance emerged.

339 For decoding object color in the occipital ROI, the order continued until around 350
 340 ms after stimulus onset, when classifier performance on the ‘strongly red’ vs ‘strongly

Figure 4: Effect of feature step size on the decoding of object color (upper) and shape (lower) for the occipital (A) and frontal (B) ROIs. In both cases, classifiers were trained to discriminate the shape or color of the object at the attended location, when participants were performing the task relevant to the decoded feature (aLaF condition). Shaded gray region around x-axis indicates the 95% confidence intervals of the same classifications when performed on permuted data (chance performance level, averaged across classifications). Colored crosses below the plot show time points at which classifier accuracy was significantly above the chance performance level ($p < 0.05$ in a one-tailed t -test of the between-subject mean, FDR corrected at $q < 0.05$ for multiple comparisons across time points). Shaded error bars indicate the 95% confidence interval of the between-subject mean.

341 green' discrimination decreased while classifier performance on 1 or 2 step
 342 discriminations increased. Similarly, when classification of color emerged in the frontal
 343 ROI performance was weakest for stimuli that were 3 steps apart. The weaker classifier
 344 performance at later time points for 'strongly red' vs 'strongly green' could be related
 345 to the participants taking less time to decide their response when judging color on easy
 346 trials compared with hard trials. However, this explanation does not account for why
 347 there was not a similar effect for decoding of object shape, where reaction times and
 348 accuracy on the easy and hard tasks were comparable to that for color. Another
 349 possibility is that for the 'easy' color trials the participants' decision was based on
 350 neural signals related to the categorization of object color, by an area such as VO
 351 (Mullen et al., 2007) or a more anterior area along the ventral temporal processing
 352 stream (Lafer-Sousa et al., 2016), with little involvement of frontal areas. Whereas for
 353 the more difficult color task trials, and for the shape task, which is unlikely to
 354 correspond to a feature dimension of relevance in the occipital cortex, there could have
 355 been more involvement by prefrontal areas, which would be consistent with the higher
 356 classifier performance in the frontal ROI in these cases.

357 The relationship between step size and classifier performance was remarkably consistent
 358 across the occipital and frontal ROIs. Classifier performance in the frontal ROI did not
 359 include the early peak, suggesting that there was good separation of the signals from
 360 these different brain regions. But when classifier performance emerged in the frontal
 361 ROI the occipital and frontal ROIs showed a very similar pattern of variation across
 362 step size, consistent with functional connectivity between these ROIs and the ongoing
 363 transfer of stimulus-related information between these brain regions.

364 **Frontal influence on the occipital representation of object shape and** 365 **color**

366 To characterize the exchange of stimulus-related information between the occipital and
 367 frontal ROIs we used an information flow analysis (Goddard et al., 2016). Since we
 368 have fine temporal resolution measures of each pairwise classification, in each attention
 369 condition, we used the pattern of classification performance across these measures as a
 370 summary of the structure of representational space at each timepoint, and tested for

371 evidence of Granger causal interactions between the ROIs (see Methods for details).
372 Note that by applying this analysis to patterns of classification accuracy (unlike typical
373 Granger causality analyses, which are applied to raw signals), we are not simply testing
374 for evidence of connectivity between brain regions, but are specifically testing for
375 evidence of the exchange of stimulus-related information between areas.

376 The results of this analysis are plotted in Figure 5. For both color and shape, we found
377 that the earliest time points were dominated by feedforward information flow
378 (FF>FB), consistent with the early visual responses in occipital cortex being relayed
379 to frontal regions. These early periods where feedforward information flow dominated
380 were followed by periods of feedback information flow, starting at 285ms and 185ms for
381 color and shape respectively. In both cases, the information flow is biased towards the
382 feedback direction until $\sim 400ms$ after stimulus onset. Interestingly, for both color and
383 shape the timing of the feedback information flows align with the onsets of the largest
384 differences in stimulus decoding across attention condition, despite the later onset of
385 these effect for color than for shape. This is seen in Figure 5B, where the large
386 divergence between the dark red line (aLaF condition) and the other conditions starts
387 around the onset of the first red region (FB>FF), for both color (upper panel) and
388 shape (lower panel). This is compatible with the suggestion that frontal feedback to
389 occipital regions drives the larger attentional effects observed later in the
390 timecourse.

391 The timing differences between color and shape also shed light on the nature of these
392 feedforward and feedback information flows. For color the early period of FF>FB
393 persisted later than for shape (until 240ms and 115ms after stimulus onset
394 respectively). This extra period of feedforward information flow for color appears to
395 correspond to the second early peak in decoding performance ($\sim 165 - 240ms$ after
396 stimulus onset), and could be related to higher-order processing of color information by
397 occipital cortex at this time, such as the ventral temporal occipital areas (Mullen et al.,
398 2007; Lafer-Sousa et al., 2016). Conversely, since the shape dimension we constructed
399 for this study is highly artificial and unlikely to correspond to a feature dimension of
400 relevance in the occipital cortex, it could be that the earlier feedback signal in this case
401 is related to the frontal cortex's involvement in storing information about the shape

task and in modifying the responses of occipital areas in such a way that the object's position along the shape dimension can be read out.

Note that while our results are consistent with a late dominance of feedback from frontal to occipital regions, it is possible that the feedback could originate in another area. As with any correlation, it is possible that our partial correlations reflect correlation with another (untested) area. It is also possible that our source reconstruction did not accurately isolate frontal and occipital regions, and that either of these include signals from nearby regions. However, note that if, for example, any parietal signals were present in both frontal and occipital ROIs, or in the unlikely event that frontal signals were present in the occipital ROI or vice versa, this would tend to reduce the measures of feedforward and feedback information flows, rather than introduce false positives, making this a conservative analysis. Indeed, the presence of significant feedforward and feedback information flows provides evidence that the ROIs were well segregated from one another, as does the absence of early classification performance in the frontal ROI.

Later oscillations between feedforward and feedback information flows ($> 400ms$ after stimulus onset) are more difficult to interpret. Before the median response time ($690 - 820ms$ across conditions) there is a period with a trend towards feedforward information flow for shape ($400 - 500ms$), but not for color. This may reflect the 'read-out' of object shape from occipital cortex, after the occipital responses have been modified by the earlier feedback from frontal cortex: future work may explore this possibility.

Differential effects spatial and feature-selective attention across feature step size

Figure 3 shows the effects of both the attended location and attended feature on the decoding of object features, and Figure 4 shows that decoding accuracy also varied with how far apart the stimuli were along the relevant feature dimension. We next asked whether there was an interaction between these effects. We reasoned that if feature-selective attention sharpens the population response to the attended feature while spatial attention does not, then they would likely produce qualitatively different

432 patterns of enhancement across stimulus pairs of varying feature difference.

433 To predict the direction of such an interaction, we used a normalization model of
434 attention (Reynolds and Heeger, 2009) to model the effects of spatial and
435 feature-selective attention on classifier performance. A number of groups have proposed
436 models including normalization to describe the effects of attention on neuronal
437 response properties (Reynolds and Heeger, 2009; Boynton, 2009; Lee and Maunsell,
438 2009). The normalization model of Reynolds and Heeger 2009 predicts that neuronal
439 responses are given by a stimulus drive that is divided (normalized) by a suppressive
440 drive that varies with the stimulus drive. In the model, the effect of attention on
441 neuronal responses is an ‘attention field’ that varies with spatial position and the
442 stimulus feature dimension, to incorporate the effects of both spatial and feature-based
443 attention. The attention field affects the stimulus drive, and in turn the suppressive
444 drive. Depending on the relative sizes of the stimulus and the attention field, the model
445 can predict changes in both response gain and contrast gain in the response to the
446 attended stimulus. The model also accounts for the sharpening of tuning curves across
447 the visual field with feature-based attention (Martinez-Trujillo and Treue, 2004).

448 Here we tested whether this normalization model could also predict the effects of
449 spatial and feature-selective attention for our population-level measures of
450 stimulus-related information. Normalization models are based on the average effect of
451 attention on the responses of single neurons, ignoring the heterogeneity of effects across
452 neurons, and the effects of factors such as signal and noise correlations (Sprague et al.,
453 2015; Moreno-Bote et al., 2014). We tested whether this model was useful for
454 predicting patterns of classifier performance despite these simplifications. The model
455 predictions for our experimental design are illustrated in Figure 6**A-B**. Figure 6**A**
456 shows the effects of spatial and feature-selective attention on the population response
457 for an example set of parameters, illustrating the predicted sharpening of the
458 population response with feature-selective attention, compared to a more general
459 facilitation across the population response with spatial attention. Details of the model
460 predictions, including further illustrations, are found in the Methods section (see
461 Figure 7). Since the model is descriptive (Reynolds and Heeger, 2009), with a large
462 number of free parameters, we systematically generated model predictions for a wide

range of model parameter sets, 172,800 in total. Across these different parameter sets, there was variation in the predicted magnitude of the effects of spatial attention and feature-selective attention, and there was also variation in which stimulus pair feature distances (step sizes) showed the greatest enhancement. However, when compared with spatial attention, feature-selective attention tended to produce relatively more enhancement of small stimulus feature differences than larger ones, as seen in the average difference across all model parameter sets (Figure 6B). As seen in Figure S2, a majority of model parameter sets (83%) showed this qualitative pattern of relative enhancement across attention types. Furthermore, there were some combinations of spatial and feature attention excitatory and inhibitory widths for which this same qualitative pattern was found for all 400 combinations of the remaining model parameters (bright red cells in Figure S2).

If feature-selective attention especially enhances the discrimination of small differences along that feature dimension, then we should see a larger effect of feature-selective attention (compared with spatial attention) for pairs of stimuli that differ by only one step, rather than 2 or 3, along the relevant feature dimension. That is, we should see the qualitative pattern from Figure 6B in our data. Alternatively, if spatial and feature-selective attention produce qualitatively similar enhancements in the population representation of the stimulus features, we would expect this difference measure ($\text{Diff} = \text{SpatAtt} - \text{FeatAtt}$) to be constant across stimulus step size.

To test this prediction, for stimulus pairs of each step size difference we calculated metrics summarizing the effects of spatial attention (**SpatAtt**, Eqn 1 in Methods) and feature-selective attention (**FeatAtt**, Eqn 2), as shown in Figure 6C, for the decoding of color in the occipital ROI. In Figure 6C-E and in subsequent figures we plotted data as ‘tuning curves’ across step size, mirror-reversing the data from 1 and 2 steps difference to visually highlight differences between spatial and feature-selective attention in their influence on the shape of these curves. For all statistical analyses we used data without the mirror reversals.

While our key prediction concerns the difference between **SpatAtt** and **FeatAtt**, in order to give a more complete depiction of the data we plotted these two metrics separately in Figure 6C, including data from every step size and time point. In these

494 color plots, cyan to lime indicates that there was little or no effect of attention on
495 classifier performance, while yellow through to red indicates a small to large increase in
496 discriminability. While it is possible for the metrics to have a negative (dark blue)
497 value, which would indicate decreased classifier performance with attention, this was
498 not seen in the data.

499 If spatial and feature-selective attention produced qualitatively similar effects on neural
500 responses, then the plots in Figure 6C should look similar, and the regions of
501 yellow-red should have a similar shape. Instead, visual comparison of the plots in
502 Figure 6C reveals differences between the two types of attention in their effects on
503 decoding of color in the occipital ROI. Consistent with the data in Figure 3, the effect
504 of spatial attention emerges earlier than that of feature-selective attention: at ≈ 200
505 ms there is a band of yellow for spatial but not feature-selective attention. Critically,
506 there was also a systematic difference between spatial and feature-selective attention in
507 their relative effects on classifier performance across step size. In 6C this is seen most
508 clearly in the ‘convex’ versus ‘concave’ shape of the yellow-red regions from 300 ms
509 after stimulus onset in the upper and lower plots. Furthermore, while spatial attention
510 tended to produce the greatest increase in classifier performance (the largest red area)
511 for stimuli separated by 2 steps in feature space, feature-selective attention tended to
512 produce greatest enhancement for stimuli separated by only 1 step along the relevant
513 feature dimension (the stimulus pairs that were most similar).

514 To identify times at which spatial and feature-selective attention differed in their
515 effects across step size we performed a 2-way ANOVA, with subject as a random
516 factor, at each time point. Clusters of time points at which there was a significant
517 interaction between attention type and step size ($p < 0.05$, at least 2 consecutive time
518 points) are indicated by the black crosses in Figure 6C. The earliest cluster began after
519 the second peak in classification performance, at 340ms after stimulus onset. To
520 visualize the interaction at these times, and in order to plot the inter-subject
521 variability, for each cluster we plotted the average effect of spatial and feature-based
522 attention (Figure 6D), including 95% confidence intervals of the between-subject mean.
523 For every cluster of timepoints for which there was a significant interaction between
524 attention type and step size the effect went in the same direction: spatial attention had

525 a greater effect than feature-selective attention at the largest step size, while
 526 feature-selective attention had a larger effect than spatial attention at the smallest step
 527 size. This is illustrated most clearly in the difference plots (**SpatAtt-FeatAtt**) of
 528 Figure 6E. As an additional control, we confirmed that the same pattern of results
 529 persists when excluding participants with any bias toward the attended location in
 530 their average fixation location (Figure S4). These data suggest a robust qualitative
 531 difference between spatial and feature-selective attention in the way they enhance the
 532 color information in occipital areas.

533 For the decoding of shape in the occipital ROI, the effects of spatial and
 534 feature-selective attention were more uniform across step sizes (see Figure S5), and
 535 there were no clusters of time points with a significant interaction between attention
 536 type and step size. This was also true for the frontal ROI, for decoding of both color
 537 and shape (data not shown). In order to test if there any interaction between attention
 538 types and step size for object shape when data from the entire brain was included, we
 539 also calculated **SpatAtt** and **FeatAtt** for the decoding of object shape based on sensor
 540 data (before any source localization). In this case, there were 2 clusters of consecutive
 541 time points where there was a significant interaction between attention type and step
 542 size (Figure S6), and the earliest of these began at 365ms after stimulus onset. Notably,
 543 where these interactions occurred, the effects were also in the predicted direction,
 544 despite variation in the effect of step size on decoding color and shape (Figure 4). This
 545 suggests a general qualitative difference between spatial and feature-selective attention
 546 in the way they enhance the information that is carried by neural population codes,
 547 which aligns with that predicted by a normalization model.

Figure 5: Analysis of feedforward and feedback interactions between occipital and frontal cortices. **A** FF (see Eqn 3) minus FB (see Eqn 4) based on classification performance on decoding stimulus color (upper plot) and shape (lower plot). Time points at which the difference is significantly above or below zero (FF>FB, or FF<FB) are shown in blue and red respectively (p -values based on bootstrapped distribution, FDR corrected to $q < 0.05$). Shaded error bars indicate the 95% confidence interval of the between-subject mean. In **B** the occipital classification performance in each attention condition is replotted from Figure 3A. The background of the plot is colored according to the data from **A**, as indicated by the colorbar. Time points where FF-FB was significantly different from zero are also replotted, here with black crosses.

Figure 6: Effects of spatial and feature-selective attention on the decoding of object color in the occipital ROI. **A:** The predicted effects of spatial and feature-based attention on a population of neuronal responses, for an example set of model parameters. According to the model, spatial attention tends to boost the response of all neurons as a multiplicative scaling of the original response, while feature-based attention produces both facilitation of neurons which prefer the attended value, and suppression of neurons preferring nearby values, which leads to sharpening of the population response around the attended value. **B:** Predicted difference between the effects of spatial (**SpatAtt**, Eqn 1) and feature-selective attention (**FeatAtt**, Eqn 2) on classifier performance across pairs of stimuli with different feature differences, averaged over all 172,800 sets of model parameters we tested. **C:** The effects of spatial attention (upper plot) and feature-selective attention (lower plot) on decoding of stimulus color were calculated by taking the difference in classifier accuracy (d') between the relevant attended and unattended conditions, normalized by the accuracy in the aLaF condition at each time point, for each step size (see Equations 1 and 2). Data from three epochs of interest were averaged and plotted in the insets below (**D**). In **E** the difference between the two attention effects (from the same time points as in **D**) are plotted, and p-values indicate the result of the significance of the interaction between attention type and step-size in each case. The difference values plotted in **C** correspond to the prediction from the model in **B**. Shaded error bars indicate the 95% confidence interval of the between-subject mean.

548 Discussion

549 Attentional selection is critical for fast and accurate processing of behaviorally relevant
 550 visual information. There are different methods by which we can select a subset of
 551 visual information for further processing, but the extent to which these are
 552 implemented by similar or different neural processes, and how these attentional effects
 553 interact, remains unclear. Spatial and feature-selective attention have rarely been
 554 directly compared within the same experiment, and to our knowledge this is the first
 555 test of two key predictions regarding their interaction: that spatial and feature-selective
 556 attention interact in a multiplicative way in their effects on neural coding, and that
 557 they induce qualitatively different patterns of enhancement across fine and coarse
 558 feature differences. We found that a normalization model of attention, designed
 559 primarily to account for the effects of attention on individual neurons, predicts these
 560 effects of attention on the information carried by a population neural signal.

561 Previous neuroimaging work has revealed some of the effects of spatial (Brefczynski
 562 and DeYoe, 1999; Jehee et al., 2011; Guggenmos et al., 2015; Sprague and Serences,
 563 2013) and feature-selective (Corbetta et al., 1990; Chawla et al., 1999; Saenz et al.,
 564 2002; Serences and Boynton, 2007; Saproo and Serences, 2014; Jackson et al., 2016;
 565 Vaziri-Pashkam and Xu, 2017) attention at a population level. Like some previous
 566 fMRI studies, we used classifier accuracy as an intuitive means of measuring the effects
 567 of attention: using classifier accuracy as a proxy for the amount of information that is
 568 potentially available in the neural response. Here we applied this decoding approach to
 569 MEG data, which allowed us to explore the timecourse of these effects using the
 570 millisecond resolution of MEG. We found evidence that both spatial and
 571 feature-selective attention boost the stimulus-related information in the population
 572 response, and we were able to measure these effects in both frontal and occipital
 573 regions. In both frontal and occipital regions, the effects of spatial attention emerged
 574 earlier than those of feature-selective attention. Through our information flow analysis
 575 of Granger-causal relationships between occipital and frontal regions, we found that
 576 stimulus-related activity in frontal regions influenced occipital representations from as
 577 early as 185ms after stimulus onset, and that the onset of this influence coincided with
 578 the largest magnitude attentional effects in occipital regions. In addition, we found

evidence confirming two predictions relating to how the effects of spatial and feature-based attention interact, and how they differ in their relative enhancement of the discriminability small versus large stimulus feature differences. We consider each of these findings below.

Earliest responses in occipital areas modulated by spatial but not feature-selective attention

For the decoding of both color and shape, we found that spatial attention had only a small effect and feature-selective attention had no significant effect on the initial peak of classifier performance in the occipital ROI ($\sim 100ms$), but much larger effects at later times. The effect of feature-selective attention on occipital stimulus representation was only significant from 335 – 385ms: at least 200ms after the effect of spatial attention. Furthermore, while there was a small effect of spatial attention around the initial peak in classifier performance ($\sim 100ms$ after stimulus onset) there was no significant effect of feature-selective attention, consistent with another report that the earliest occipital responses are not affected by feature-based attention (Bartsch et al., 2017). This finding that spatial attention effects preceded those of feature attention is consistent with previous results from electrophysiological recordings in V4 and FEF (Zhou and Desimone, 2011; Bichot et al., 2015), although the delay observed here is longer than in this previous work. For both features, feature-selective attention had an impact on classifier performance in the occipital only after feedback from the frontal ROI began to dominate the information flow (FB>FF). Since information flow analysis specifically measures the exchange of stimulus-related information, this result suggests that the effects of feature-selective attention in occipital cortex may rely on feedback of stimulus-related information from frontal areas.

The degree to which subjects are engaging attention prior to stimulus onset could also have contributed to the pre-stimulus decoding of attentional task for spatial but not feature-selective attention (Figure 2) and to the earlier effects of spatial attention, relative to feature-selective attention on the stimulus representation (Figure 3). For example, it may be easier to prepare to attend to a location than to prepare to attend to a feature dimension. A previous study reported that feature-based attention can

modulate event-related potentials (ERPs) much earlier than in our data, within 100ms of the stimulus onset (Zhang and Luck, 2009) (in contrast to 335 – 380ms onset in the our results). This discrepancy may reflect a difference between feature-based attention (attending to a feature value, e.g. ‘red’) and feature-selective attention (attending to a feature dimension, such as ‘color’). Another critical difference between these studies is in stimulus design: Zhang and Luck (2009) recorded responses to a flashed probe stimulus of red or green dots while subjects attended to dots of one color in another covertly attended stimulus, where dots of both colors were always present. In our experiment, the stimuli were always preceded by a blank screen, so that subjects were planning to attend to a particular stimulus feature rather than already attending to it. In our data, decoding of attention condition became much stronger once stimuli appeared and the participants were actively performing the task. We hypothesize that these stimulus differences account for the later onset of feature-selective attention’s effect on stimulus representation here, and that the early effects of feature-based attention reported by Zhang and Luck (2009) are only present when the subject is already engaged in attending to one feature value (or suppressing the irrelevant feature value, see Moher et al. 2014; Andersen and Müller 2010).

Information flow analysis: the role of frontal feedback in attentional modulation

The earliest responses of the occipital cortex showed little modulation with attentional condition, consistent with a stimulus-driven response. Shortly after these initial responses there were large effects of both attention types: attention changes the stimulus information represented by the population response in occipital cortex. What regions drive the effects of attention on the occipital population response? Within occipital cortex, previous work suggests that attentional effects are present first in higher-order visual areas that induce a top-down modulation of earlier areas (Buffalo et al., 2010), but this leaves open the possibility that effects in higher-order visual areas are driven by another region. Our information flow analysis suggests a contribution from frontal areas, with stimulus-related information coding in occipital cortex appearing to follow from the information coding in the frontal lobe shortly

639 beforehand. A class of models of prefrontal function converge on the proposal that
640 prefrontal cortex implements cognitive control by affecting processing in more
641 specialised cortices (Duncan, 2001; Desimone and Duncan, 1995; Dehaene et al., 1998;
642 Miller and Cohen, 2001). By tracking the dynamics of information exchange between
643 frontal and occipital cortex we were able to test this suggestion and resolve the
644 timecourse of the proposed top-down effects.

645 We found that information flow was initially dominated by feedforward propagation of
646 information from occipital to frontal lobe, then later dominated by information flowing
647 in the opposite direction, with information coding in the frontal ROI predicting
648 subsequent information coding in occipital cortex (see also, Goddard et al. 2016;
649 Karimi-Rouzbahani 2018). Moreover, the onset of feedback dominating the flow of
650 information between frontal and occipital cortex corresponded to the time at which the
651 occipital lobes showed a divergence between task-relevant and task-irrelevant
652 information. For decoding color, where there was a second early peak in classifier
653 performance, this period was later (285ms) than for decoding shape (185ms), but in
654 both cases it aligned with the time at which information processing in the occipital
655 lobes became dominated by the task-relevant information (classifier performance in the
656 attended location, attended feature condition remained steady or increased, while
657 performance in other conditions was strongly attenuated).

658 Our finding that prefrontal cortex appears to shape responses in occipital areas is
659 consistent with work demonstrating that the responses of frontoparietal regions contain
660 stimulus-related information (for example, Freedman et al. 2001), that increases with
661 spatial (Woolgar et al., 2015) and feature-selective (Jackson et al., 2016) attention, and
662 that attentional effects in frontal cortices precede those in sensory cortex (e.g. Lennert
663 and Martinez-Trujillo 2013). One prominent model of prefrontal cortex function
664 (biased competition model Desimone and Duncan 1995; Duncan 2006) proposes that
665 the prefrontal cortex biases processing in more specialized (visual) cortices in favor of
666 task-relevant information. In line with such a proposal, our data suggest that after an
667 initial feedforward sweep of information, feedback from frontal to occipital cortices
668 drives the selective representation of information in the occipital cortex.

669 Future work could build on these findings in two ways. First, we chose not to resolve

670 into more fine-grained parcellations of the frontal lobe here because of the limitations
 671 of not having individual MRI scans and concerns about the inverse problem. This
 672 presents an interesting avenue for future work using the methods described here,
 673 perhaps using concurrent EEG and individualized MRI scans to constrain the inverse
 674 problem. Second, with better source estimation it would be interesting to examine the
 675 role of other brain regions, particularly the parietal lobe (which is known to have
 676 important roles in attention, e.g. Duncan 2010; Woolgar et al. 2011; Hebart et al. 2018;
 677 Jerde et al. 2012). In the context of information flow analyses such finer parcellations
 678 could identify cases in which correlations between two brain regions are likely mediated
 679 by both areas correlating with a third.

680 **Differential effects of spatial and feature-selective attention as** 681 **predicted by a normalization model of attention**

682 Much of our knowledge of spatial and feature-selective attention comes from studies
 683 that have investigated their effects in separate experiments. As such, the results
 684 presented here provide valuable new insight into how these two types of attention
 685 interact. We found that where there were effects of both types of attention there also
 686 tended to be an interaction between them, which is consistent with a multiplicative
 687 rather than an additive combination of attentional effects. In the normalization model
 688 of attention presented by Reynolds and Heeger (2009), they modeled all but one of the
 689 results with a multiplicative rather than additive interaction¹

690 We used results from single-unit work to predict how differences in the effects of spatial
 691 and feature-selective attention might manifest in population-level codes for stimulus
 692 features. Specifically, we predicted that feature-selective attention would produce
 693 relatively more enhancement of classifier performance for small feature differences than
 694 for large feature differences, as compared with the effects of spatial attention. We
 695 confirmed this intuition by using a normalization model (Reynolds and Heeger, 2009)
 696 to generate predictions for our data. Normalization models of attention are primarily
 697 based upon the electrophysiological study of the effects of spatial and feature-based
 698 attention on tuning of individual cells, yet here we demonstrate that the same model

¹In Reynolds and Heeger (2009) the parameter ‘Ashape’ was set to ‘oval’ rather than ‘cross’ for all but one of their figures, but to our knowledge our result is the first test of this prediction.

699 can account for population level data, and can be extended to predict the effects of
 700 feature-selective attention. It is particularly important that we understand how the
 701 effects of attention manifest at a population level since there are significant effects at a
 702 population level that cannot be captured by measuring the tuning curves of individual
 703 cells (Sprague et al., 2015; Cohen and Maunsell, 2009). The results of our classification
 704 analyses based on the MEG data revealed that spatial and feature-selective attention
 705 have distinct effects on stimulus-related information coding at a population level, and
 706 these differences were consistent with the predictions of the normalization model.

707 The fact that classifier performance was consistent with the predictions of the
 708 normalization model does not definitively identify what information the classifier
 709 analysis is using to decode stimulus color and shape, which is difficult to pin down in
 710 any case where classifiers are used to measure stimulus-related information from
 711 neuroimaging data (Carlson et al., 2018). However, this result suggests that the
 712 information that is accessible to the classifier varies in signal strength in a manner that
 713 is consistent with what we expect based on the effects on single-unit tuning predicted
 714 by the normalization model. Additionally, differences between decoding of color and
 715 shape are broadly consistent with color (but not ‘X-shaped-ness’) being a feature
 716 dimension that is explicitly encoded by visual cortex. We found the most marked
 717 difference between the attention types in the decodability of stimulus color in the
 718 occipital ROI. Of the two feature dimensions we manipulated (shape and color) it is
 719 more plausible for color that there are single-units with response functions that
 720 approximate those included in the normalization model. Neurons in a range of visual
 721 cortical areas are tuned for color (for example, Komatsu et al. 1992; Hanazawa et al.
 722 2000), and attention to color is a form of feature-based and feature-selective attention
 723 that has been investigated in single-unit work (for example, Motter 1994; Bichot et al.
 724 2005; Chen et al. 2012). In contrast, the shape dimension (from ‘X-shaped’ to
 725 ‘non-X-shaped’) is an artificial, more complex dimension than color. It is possible that
 726 this dimension could align with the feature selectivity of some neurons in an area with
 727 intermediate to high level shape selectivity, such as the in area V4 (see review by
 728 Pasupathy 2006), but it is unlikely that there is population tuning for this shape
 729 dimension in the same way that we expect a population code for the color dimension.
 730 Although the tuning differences between spatial and feature-selective attention were

731 weaker for shape than for color, where these differences were significant (in the
732 sensor-level decoding) the effect was in the same direction as for color. This suggests
733 that a population tuning curve framework may be helpful for understanding the effects
734 of attention on arbitrary, higher level feature dimensions as well as for lower-level
735 ones.

736 Normalization models of attention can account for a range of the effects of attention
737 observed at the level of a single neuron (Boynton, 2005; Reynolds and Heeger, 2009;
738 Boynton, 2009; Lee and Maunsell, 2009). Although designed to model single-neuron
739 effects, these models can be used to predict attention-based changes in the information
740 carried by the population response, such as in the implementation used in the present
741 study. For both single-unit and population responses these models are primarily
742 descriptive rather than quantitative, but in selecting ranges of model parameters we
743 considered parameters that are feasible for single-unit responses and found that these
744 same parameters could account for population-level effects. Our results demonstrate
745 that the same principles that describe phenomena at the single-unit level, such as
746 multiplicative scaling in spatial attention, and sharpening of the population response in
747 feature-selective attention, can account for population level changes, particular in the
748 encoding of color by occipital areas. Notably, the normalization model successfully
749 predicted these population-level effects despite the fact that the model does not
750 incorporate any heterogeneity of effects across neurons, nor any effects of signal or
751 noise correlations, which could have caused differences between single-unit and
752 population-level effects (Sprague et al., 2015; Moreno-Bote et al., 2014). This opens the
753 possibility of using such models as an explanatory bridge between levels of description:
754 if future work constrains model parameters for the normalization model at either the
755 single-unit or the population level this may generate predictions that can be tested at
756 other, to further characterize the similarities and differences between these levels of
757 description. When model parameters are further constrained by data, another direction
758 for future work is to test quantitative as well as qualitative predictions of these
759 models.

760 Conclusions

761 We used multivariate pattern analysis of MEG recordings to measure the effects of
 762 spatial and feature-selective attention on the amount of stimulus-related information
 763 decodable from large populations of neurons. We manipulated both spatial and
 764 feature-selective attention simultaneously in order to compare these attention types
 765 within the same dataset, and to test how these attention types interact in their effects
 766 on population-level representation of visual stimuli. We found that both spatial and
 767 feature-selective attention enhanced the representation of visual information and that
 768 the two types of attention interacted in a multiplicative way to yield an adaptive
 769 neural representation which prioritised the task relevant feature of the attended object.
 770 An information flow analysis suggested that the largest attentional effects in occipital
 771 areas may be driven by feedback from frontal areas.

772 We further found that modelling the distinct effects of spatial and feature attention at
 773 the level of single cells predicted the qualitative differences between spatial and
 774 feature-selective attention in our population level recordings. The success of the
 775 modelling was remarkable given that the model only included the effects of attention
 776 on tuning properties, without modelling, for example, any influence of attention on the
 777 correlation structure of the population. Specifically, consistent with a normalization
 778 model of attention in which feature-selective attention results in tuning curve
 779 sharpening and spatial attention predominantly yields response gain, we found that for
 780 decoding of color in occipital cortex, feature-selective attention produced more
 781 enhancement of the neural representation of small stimulus feature differences than
 782 spatial attention did, while spatial attention resulted in greater discrimination of large
 783 stimulus feature differences.

784 Our ability to direct our attention to different locations and to different features of the
 785 environment appears to rely on interacting attentional mechanisms that induce
 786 qualitatively distinct changes in population-level neural responses in sensory
 787 cortices.

788 **Materials and Methods**

789 **Participants**

790 20 volunteers (14 female, 6 male) participated in this study, and were paid \$50 as
 791 compensation for their time. Participants ages ranged from 18-32 years (mean 22.4
 792 years). All were right-handed, had normal or corrected to normal vision, had no
 793 history of neurological or psychiatric disorder, and were naïve to the purposes of the
 794 study. All participant recruitment and experiments were conducted with the approval
 795 of the Macquarie University Human Research Ethics Committee.

796 **Visual stimuli**

797 Visual stimuli were generated and presented using Matlab (version R2014b) and
 798 routines from Psychtoolbox Brainard (1997); Pelli (1997). We created novel object
 799 stimuli that varied in color and in their shape statistics (see Figure 1B), using custom
 800 code. The shapes were variants of ‘spikie’ stimuli used in previous work (Op de Beeck
 801 et al., 2006; Woolgar et al., 2015; Jackson et al., 2016). All our ‘spikie’ shapes had a
 802 common almost spherical body and 16 spikes varying in location, length and
 803 orientation. All shapes were rendered with diffuse illumination and a direct (upper left)
 804 illuminant source, and presented on a black background. We varied the spike
 805 orientation statistics to create four classes of ‘spikie’ objects: strongly ‘X-shaped’,
 806 weakly ‘X-shaped’, weakly ‘non-X-shaped’, and strongly ‘non-X-shaped’ (Figure 1B).
 807 When performing the shape-based task participants categorized the target object as
 808 either ‘X-shaped’ or ‘non-X-shaped’. We created 100 unique versions of each shape
 809 class by adding random variation in the spike locations, lengths and orientations, to
 810 ensure that participants could not perform the task by attending to a single feature,
 811 and to encourage them to attend to the object’s overall shape.

812 In color, there were also four classes: ‘strongly red’, ‘weakly red’, ‘weakly green’ and
 813 ‘strongly green’ (Figure 1B). When performing the color-based task participants
 814 categorized the target object as either ‘reddish’ or ‘greenish’. Each object had a
 815 maximum luminance of 108.1 cd/m^2 , and constant u’v’ and xy chromaticity coordinates
 816 (Wyszecki and Stiles, 1982). The chromaticity coordinates were as follows; strongly red

817 $u'v'$: 0.35, 0.53 (xy: 0.56, 0.38); weakly red $u'v'$: 0.27, 0.54 (xy: 0.50, 0.44); weakly
818 green $u'v'$: 0.23, 0.55 (xy: 0.45, 0.48) and strongly green $u'v'$: 0.16, 0.56 (xy: 0.36
819 0.57). The weak red and weak green colors were defined as lying on a line joining the
820 strong red and strong green coordinates in $u'v'$ space, and their distance from the line's
821 midpoint was 30% of the distance between the midpoint and the relevant
822 endpoint.

823 During MEG sessions, stimuli were projected through a customized window by an
824 InFocus IN5108 LCD back-projection system (InFocus, Portland, Oregon, USA)
825 located outside the Faraday shield, onto a screen located above the participant.
826 Participants, lying supine, viewed the screen from 113cm. Individual 'spikie' objects
827 each had a central body of 195 pixels (5.8 degrees visual angle [dva]) wide x 175 pixels
828 (5.2 dva) high. Their total size varied with their spikes, but the spikes never reached
829 the border of the object image (403x403 pixels). On each trial, the stimulus included 2
830 'spikie' object images side-by-side (total size 806 pixels wide x 403 pixels high: 24 x 12
831 dva). A white fixation cross, with height and width of 1 dva, was drawn in the center
832 of the screen (Figure 1A). The display system was characterized in situ using a Konica
833 Minolta CS-100A spectrophotometer and calibrated as described previously (Goddard
834 et al., 2010).

835 **Experimental design: MEG and eye tracking**

836 MEG data were collected with a whole-head MEG system (Model PQ1160R-N2, KIT,
837 Kanazawa, Japan) consisting of 160 coaxial first-order gradiometers with a 50 mm
838 baseline (Kado et al., 1999; Uehara et al., 2003). Prior to MEG measurements, five
839 marker coils were placed on the participant's head. Marker positions, nasion, left and
840 right pre-auricular points, and the participant's head shape were recorded with a pen
841 digitizer (Polhemus Fastrack, Colchester, VT), using a minimum of 2000 points.
842 Each participant's MEG data were collected in a single session of approximately 90
843 minutes, at a sampling frequency of 1000Hz. On each trial participants responded using
844 a Fiber Optic Response Pad (fORP, Current Designs, Philadelphia, PA, USA).
845 We tracked participant's eye movements using an EyeLink 1000 MEG-compatible

846 remote eye-tracking system (SR Research, 500Hz monocular sampling rate). Before
847 scanning we tested participants for their dominant eye (usually right), and focused the
848 eye-tracker on this eye.

849 **Experimental design: participant's task**

850 Each participant's MEG session was divided into 8 blocks, where the location of the
851 attended object (left or right of fixation) and the task (reporting the attended object's
852 shape or color category) was constant within each block. Figure 1A illustrates the four
853 different attention conditions. Before the experiment, each participant was familiarized
854 with the 'X-shaped' and 'non-X-shaped' object categories and completed a training
855 session on a laptop outside the MEG scanner where they practiced the color and shape
856 tasks.

857 On every trial we presented two objects, one each on the left and right of fixation. We
858 presented the objects simultaneously since both spatial attention (Reynolds and
859 Desimone, 1999; Sundberg et al., 2009) and feature-selective attention (Saenz et al.,
860 2003) effects are stronger when attended and unattended stimuli simultaneously
861 compete for access to perceptual processing. Within each block every pairing of the 16
862 objects in Figure 1B was included once, giving 256 (16x16) trials. These 256 trials
863 were presented in a counterbalanced order within each block, so that objects of each
864 shape and color were equally likely to precede objects of all shapes and colors. A
865 different counterbalanced order was used for each block, and to this sequence of 256
866 trials the last trial was added to the beginning, and the first trial was added to the
867 end, giving a total of 258 trials in each block. Data from these first and last trials were
868 discarded.

869 The participant's task alternated between shape and color on every block, and the
870 location of the attended object alternated after the 2nd, 4th and 6th blocks. Starting
871 location and task were counterbalanced across participants. Within each pair of blocks
872 where the attention condition was the same (e.g. blocks 1 and 5), the buttons
873 corresponding to the two response options were switched, so that response mappings
874 were counterbalanced across blocks.

875 Every block commenced with an instruction regarding the attended object, the task,
876 and the response mapping for that block. Before the first trial participants were
877 required to identify the response buttons correctly with a keypress. Participants also
878 repeated the eye-tracker 5-point calibration, before the block commenced.

879 Every trial began with the fixation marker alone while the participant's fixation was
880 verified using the eye tracker. Participants had to fixate within 1 dva of the fixation
881 marker for at least 300 ms before the stimulus would appear. During the stimulus
882 (maximum 150ms) a 50x50 pixel white square appeared in the bottom right corner of
883 the projected image (outside the stimulus region), which was aligned with a
884 photodetector, attached to the mirror, whose signal was recorded with the MEG signal
885 from the gradiometers. We used the photodiode signal to accurately align MEG
886 recordings with stimulus timing during data analysis. When eye-tracking showed
887 participants were no longer fixating during the 150ms stimulus presentation, the
888 stimulus was removed from the screen. Due to eye tracker variability (e.g. eye tracker
889 missing frames), this resulted in an unexpectedly high number of shorter trials: the
890 median stimulus duration was 92ms, and the first and third quartiles were 64 and
891 126ms. Since this affected a majority of trials, we included all trials in our analysis,
892 but ran an extra analysis to check that variability in stimulus duration did not account
893 for our results (see below). After stimulus offset, the fixation marker remained white
894 until participants responded to the appropriate task via a button press. After the
895 participant's response, but no sooner than 1000 ms from the onset of the stimulus, the
896 fixation marker changed for 200 ms to provide feedback: dimming to gray for 'correct',
897 or turning blue for 'incorrect'. After feedback, there was a variable inter-trial interval
898 (300-800ms), which comprised the fixation check for the subsequent trial. We used a
899 variable inter-trial interval to avoid expectancy effects. Across participants, the median
900 reaction time was 0.77s (shape task: 0.78s; color task: 0.75s); on 77% of trials the
901 reaction time was shorter than 1 s and the feedback onset was 1 s. The first and third
902 quartiles of the distributions of reaction times are shown in Figures 2 and 3.

903 **MEG data analysis: Source reconstruction**

904 Forward modeling and source reconstruction were performed using Brainstorm (Tadel
905 et al., 2011), which is documented and freely available for download online
906 (<http://neuroimage.usc.edu/brainstorm>). First, we created a model of each
907 participant's brain by manually aligning the ICBM152 template brain (Fonov et al.,
908 2011) to their head shape using nasion, pre-auricular points, and head shape data.
909 Once aligned, we applied nonlinear warping to deform the template brain to the
910 participant's head shape, which provides a superior model to an unwarped canonical
911 template (Henson et al., 2009). We generated a forward model for each model by
912 applying a multiple spheres model (Huang et al., 1999) to the individually warped
913 template brain and their measured head location.

914 Functional data were preprocessed in Brainstorm with notch filtering (50, 100 and
915 150Hz), followed by bandpass filtering (0.2-200Hz). Cardiac and eye blink artifacts
916 were removed using signal space projection (SSP): cardiac and eye blinks events were
917 identified using default filters in Brainstorm, manually verified, then used to estimate a
918 small number of basis functions corresponding to these noise components, which were
919 removed from the recordings (Uusitalo and Ilmoniemi, 1997). From these functional
920 data we extracted two epochs for each trial: first, a measure of baseline activity (-100
921 to -1ms relative to stimulus onset), and secondly the evoked response (0 to 1000ms).
922 We used the baseline measures to estimate the noise covariance for each run, then
923 applied a minimum norm source reconstruction to the evoked data. For each source
924 reconstruction, we used a 15,000 vertex cortical surface (standard for the ICBM152
925 template, with atlas information). Dipole orientations in the source model were
926 constrained to be normal to the cortical surface, the noise covariance was regularized
927 using the median eigenvalue and all other options were set to their default values. We
928 visually inspected the quality of the source reconstruction: the average trial data
929 included an initial ERP at the occipital pole and subsequent ERPs at sources within
930 the occipital cortex but lateral and anterior to the occipital pole, consistent with
931 extrastriate areas along the ventral visual pathway (see Supplementary Figure
932 S1).

933 MEG data analysis: Preprocessing and dataset definitions

934 For classification analyses we generated three datasets: the first included preprocessed
935 data from all sensors, without source reconstruction. The second included sources in
936 occipital, occipito-temporal, and inferior-temporal cortices ('Occipital' ROI, 3302
937 vertices) in the atlas for the ICBM152 template, and the third included frontal and
938 prefrontal cortices ('Frontal' ROI, 3733 vertices), as shown in Figure 2A.

939 For each dataset, we extracted data from -100 ms to +2000 ms relative to the stimulus
940 onset of each trial. We then reduced each data set, comprising 2100 ms of data for each
941 of 2048 trials and up to 160 sensors or up to 3733 sources, using PCA. We retained
942 data from the first n components which accounted for 99.99% of variance (mean, std n :
943 85.3, 6.9 for frontal ROI; 76.6, 5.8 for occipital ROI; and 157.2, 1.1 for whole brain
944 sensor data) and down-sampled to 200Hz using the Matlab 'decimate' function.

945 MEG data analysis: Classifier analyses

946 We used classification analyses to measure the extent to which brain activity could
947 predict attention condition and the color and shape of the stimuli on each trial. For
948 every classification we repeated the analysis at each time point (each 5ms bin) to
949 capture how the information carried by the neural response changed over time: we
950 trained classifiers to discriminate between two categories of trial and tested on held-out
951 data. We report results obtained with a linear support vector machine (SVM)
952 classifier, using the Matlab function *fitcsvm* with 'KernelFunction' set to 'linear'. We
953 also repeated our analyses with a linear discriminant analysis (LDA), using the Matlab
954 function *classify* with 'type' of 'diagLinear' and obtained very similar results (not
955 shown).

956 For each classification we created 'pseudo-trials' by averaging across trials with the
957 same value on the dimension-of-interest, but with differing values along other
958 dimensions. We used pseudo-trials in order to increase signal-to-noise along the
959 dimension-of-interest (e.g. see Guggenmos et al. 2018; Grootswagers et al. 2017). For
960 example, when classifying the attended location, we took the 4 blocks of 256 trials
961 where the participant attended to the object on the left, and generated 256

962 pseudo-trials, each the average of 4 trials with one randomly-selected trial from each
 963 block. This meant that each pseudo-trial included data from an equal number of trials
 964 from the attended feature conditions (attend to color and attend to shape). For each
 965 classification we generated 100 sets of pseudo-trials, updating the random assignment
 966 of trials for each set, and averaged classification performance across these.

967 Features that were balanced across pseudo-trials varied with the feature-of-interest
 968 being classified. As mentioned above, when classifying attended location pseudo-trials
 969 were balanced across attended feature. Similarly, for classifying attended feature
 970 pseudo-trials were balanced across attended location. When training classifiers to
 971 discriminate object color and shape, we trained and tested within a single attention
 972 condition (e.g. attend left, report color), comprising two blocks (512 trials). We
 973 trained classifiers separately on each pair of the 4 levels along each feature dimension,
 974 at each object location, using pseudo-trials to balance across irrelevant dimensions. For
 975 example, when classifying ‘strongly green’ versus ‘weakly green’ objects on the left of
 976 fixation, we balanced pseudo-trials across left object shape, and right object color and
 977 shape. Since balancing across all 3 of these irrelevant dimensions would not provide
 978 sufficient data for classifier training (only 2 pseudo-trials per category), we instead
 979 created pseudo-trials that were balanced across 2 of these 3 irrelevant dimensions, and
 980 randomized across the third (allowing 8 pseudo-trials per category). As before, we
 981 generated 100 sets of the pseudo-trials, each with a different randomization.
 982 Additionally, we repeated this entire process 3 times, balancing across different pairs of
 983 irrelevant features. For each of set of pseudo-trials, we trained a classifier using 7 of the
 984 8 pseudo-trials in each condition and tested using the remaining pair of trials,
 985 repeating 8 times. We averaged classifier performance across these 8 classification
 986 boundaries, and across the 300 sets of pseudo-trials.

987 For color and shape we performed the classification analysis pairwise for each pair of
 988 feature values, then averaged classifier performance across feature differences of the
 989 same ‘step size’. Since both dimensions had 4 values, pairs were either 1, 2 or 3 steps
 990 apart along the given feature dimension. Pairs 2 or 3 steps apart belonged to opposite
 991 categories in the participant’s task (‘greenish’ vs ‘reddish’ and ‘X-shaped’ vs
 992 ‘non-X-shaped’). Pairs 1 step apart could be within or across these categories; we did

993 not find any differences between these data (data not shown) so averaged across these
994 when reporting our results.

995 For all analyzes we expressed average classifier accuracy in d' (a unit-free measure of
996 sensitivity) which provides an intuitive measure of effect size: a d' value of 0
997 corresponds to no stimulus-related information, which was useful when calculating the
998 effects of spatial and feature-selective attention (below). To test whether classifier
999 performance was above chance performance, we repeated each classification analysis for
1000 data where trial labels were randomly permuted. We repeated this 10 times for data
1001 from every 4th time bin (one every 20ms). In statistical tests we tested whether the
1002 observed classification performance exceeded the average chance performance across
1003 time bins. Across classifications, average chance performance varied from $d'=0.000$ to a
1004 maximum of $d'=0.015$.

1005 Additionally, to predict the effect of variable trial duration, we repeated each
1006 classification of stimulus feature using the stimulus state (on or off) at each time point.
1007 Across time points, the maximum average classifier accuracy was $d'=0.4$ for this data,
1008 indicating that stimulus variability could have made a small contribution to overall
1009 accuracy. However, there was very little difference between this decoding for different
1010 attention conditions or across step sizes. When we performed the statistical tests
1011 reported in Figures 3 on the trial duration data, the only significant result (effect of
1012 attended location for decoding stimulus color) was in the opposite direction (decoding
1013 was higher for unattended than attended locations).

1014 For each stimulus classification boundary, we averaged the classifier weights across each
1015 set of pseudo-trials to generate an estimate of the classifier weights for each
1016 participant's data, at each time point. The magnitudes of raw classifier weights can
1017 vary with both signal strength and noise magnitude, making maps of raw weights
1018 difficult to interpret (Haufe et al., 2014). To obtain more informative maps we followed
1019 a method used previously (Haufe et al., 2014; Wardle et al., 2016) to transform the
1020 classifier weights: For each vector ($\mathbf{W}$) of average classifier weights across occipital or
1021 frontal vertices, we obtained the transformed weights ($\mathbf{W}'$) using the covariance matrix
1022 of the n pseudo-trials that constituted the classifier training/test data
1023 ($\text{cov}(\text{pseudotrials})$), using $\mathbf{W}' = \text{cov}(\text{pseudotrials}) * \mathbf{W}$. We averaged these

transformed weights ($\mathbf{W}'$) across all pairwise comparisons before multiplying the weights by the subject-specific PCA coefficients, and finally averaging across participants.

To summarize the effects of spatial attention (**SpatAtt**) and feature-selective attention (**FeatAtt**), we used the following metrics, based on classifier performance (d') in the attended location, attended feature (*aLaF*) condition, the attended location, unattended feature (*aLuF*) condition, the unattended location, attended feature (*uLaF*) condition, and the unattended location, unattended feature (*uLuF*) condition. For both attention effects, we normalized the effects by the classifier accuracy in the *aLaF* condition to minimize the influence of overall classifier accuracy on the estimates of attention effects.

$$\mathbf{SpatAtt} = (aLaF + aLuF - uLaF - uLuF)/aLaF; \quad (1)$$

$$\mathbf{FeatAtt} = (aLaF + uLaF - aLuF - uLuF)/aLaF; \quad (2)$$

Modeling the effects of spatial and feature-selective attention on population representations of shape and color

We used a normalization model of the effects of attention at the cellular level to make predictions of how attention would affect stimulus-related information in the population response. Intuitively, we expected that if feature-based attention sharpens the population response to the attended feature, then feature-selective attention should particularly increase classifier performance for stimulus pairs with small feature differences. Conversely, spatial attention, which is not thought to sharpen population responses, should produce relatively more enhancement of classifier performance for larger feature differences. To formalize this intuition we implemented the Reynolds and Heeger (2009) normalization model of attention to generate predictions, as illustrated in Figure 7 and detailed in the Supplementary Methods.

Figure 7: Summary of a normalization model of attention (Reynolds and Heeger, 2009), as implemented here to predict the effects of spatial and feature-selective attention on classifier performance. A: an illustration of each of the model elements from Reynolds and Heeger (2009), Figure 1, for a set of example model parameters, where each grayscale image depicts a matrix of values varying along a spatial dimension (horizontally) and a feature dimension (vertically). For each set of model parameters we generated a single ‘stimulus drive’, and two versions of the ‘attention field’, which lead to subtly different ‘suppressive drives’ and ‘population responses’. From these two population responses we derived curves predicting the population response as a function of each neuron’s preferred feature value for each of the four attention conditions (the columns of the matrix indicated with different colored vertical lines in **A**). These population responses are plotted again as lineplots in **B**. In **C** (redrawn from Figure 6A) the predicted effects of spatial and feature-based attention on the population response are summarized as the difference between relevant population curves from **B**. **D:** We predicted classifier performance in each attention condition by centering the population response from **B** on 4 different stimulus feature values and predicting classifier performance when discriminating between population responses to stimuli of that were either 60, 40 or 20 (arbitrary) units apart along the feature dimension, to simulate the population response to stimuli that were 3, 2 or 1 steps apart in either color or shape. We predicted classifier performance (d') using the separation of the two population responses, in a manner analogous to that used in signal detection theory (see Supplementary Methods for details)

MEG data analysis: Granger analysis of feedforward and feedback information flows

We tested for temporal dependence between the patterns of classifier performance in occipital and frontal datasets, seeking evidence of information flows from occipital to frontal cortices (feedforward) and from frontal to occipital cortices (feedback), following the rationale we developed in earlier work (Goddard et al., 2016). Specifically, we tested for Granger causal relationships between the patterns of classifier performance based on the occipital and frontal datasets. We summarized the color and shape information for each region (occipital and frontal), at each timepoint, as a 6x4 dissimilarity matrix (DSM) of classifier performances. For both color and shape, the 6x4 DSM was defined as each pairwise comparison (6 classifications across the 4 levels of the feature), by 4 attention conditions (aLaF, aLuF, uLaF, uLuF).

The logic of Granger causality is that time series X ‘Granger causes’ time series Y if X contains information that helps predict the future of Y better than information in the past of Y alone (for a recent review of its application in neuroscience, see Friston et al. (2013)). We performed a sliding-window analysis of a simplified (special case) of Granger causality, using the partial correlations in Equations 3 and 4 to define ‘Feedforward’ (FF) and ‘Feedback’ (FB) information flows at each time point (t).

$$FF(t, d, w) = \rho DSM_{(frontal,t)} DSM_{(occipital,t,d,w)} \cdot DSM_{(frontal,t,d,w)} \quad (3)$$

$$FB(t, d, w) = \rho DSM_{(occipital,t)} DSM_{(frontal,t,d,w)} \cdot DSM_{(occipital,t,d,w)} \quad (4)$$

where $DSM_{(loc,t)}$ is the DSM based on the sources at location loc at time tms post stimulus onset, and $DSM_{(loc,t,d,w)}$ is the DSM based on the sensors at location loc , averaged across all time points from $t - dms$ to $t - (d + w)ms$ post stimulus onset. We calculated FF and FB for 30 overlapping windows: for 5 window widths ($w = 10, 20, 30, 40$ or 50 ms) for each of 6 delays ($d = 50, 60, 70, 80, 90$ or 100). We tried a range of values for w and d in order to capture interactions between occipital and frontal cortices that may occur at different timescales. Since the results were broadly similar

1072 across values of w and d (see Figure S8) we report FF and FB values averaged across
1073 all values of w and d .

1074 We report the results of this analysis in terms of the difference between the feedforward
1075 and feedback information flows (FF-FB). To assess whether this difference was
1076 significantly above or below chance, we generated a null distribution of this difference
1077 at every timepoint by performing the same analysis on 1000 bootstraps of data from
1078 each subject where the exemplar labels were randomly permuted for each of the DSMs
1079 used in Equations 3 and 4.

1080 **Data availability**

1081 All the raw data and the results of our classification analyses are available on an Open
1082 Science Framework project (after publication we will make this project publically
1083 accessible and include the DOI for the project in our manuscript).

1084 **Acknowledgments**

1085 This project was funded under an Australian Research Council Future Fellowship
1086 (FT120100816) awarded to TC, ARC Discovery Projects (DP160101300) awarded to
1087 TC and (DP170101840) awarded to AW, an ARC Future Fellowship (FT170100105)
1088 awarded to AW, MRC (U.K) intramural funding SUAG/035/RG91365 awarded to AW,
1089 and the ARC Centre of Excellence in Cognition and its Disorders (CE110001021). We
1090 thank Erika Contini and Elizabeth Magdas for their assistance with MEG data
1091 collection.

References

- Andersen SK, Müller MM (2010) Behavioral performance follows the time course of neural facilitation and suppression during cued shifts of feature-selective attention. *Proc. Natl. Acad. Sci. USA* 107:13878–13882.
- Bartsch MV, Loewe K, Merkel C, Heinze HJ, Schoenfeld MA, Tsotsos JK, Hopf JM (2017) Attention to Color Sharpens Neural Population Tuning via Feedback Processing in the Human Visual Cortex Hierarchy. *J. Neurosci.* 37:10346–10357.
- Bichot NP, Rossi AF, Desimone R (2005) Parallel and serial neural mechanisms for visual search in macaque area V4. *Science* 308:529–534.
- Bichot NP, Heard MT, DeGennaro EM, Desimone R (2015) A Source for Feature-Based Attention in the Prefrontal Cortex. *Neuron* 88:832–844.
- Boynton GM (2009) A framework for describing the effects of attention on visual responses. *Vision Res.* 49:1129–1143.
- Boynton GM (2005) Attention and visual perception. *Curr. Opin. Neurobiol.* 15:465–469.
- Brainard DH (1997) The Psychophysics Toolbox. *Spat. Vis.* 10:433–436.
- Brefczynski JA, DeYoe EA (1999) A physiological correlate of the ‘spotlight’ of visual attention. *Nat. Neurosci.* 2:370–374.
- Buffalo EA, Fries P, Landman R, Liang H, Desimone R (2010) A backward progression of attentional effects in the ventral stream. *Proc. Natl. Acad. Sci. USA* 107:361–365.
- Buracas GT, Boynton GM (2007) The effect of spatial attention on contrast response functions in human visual cortex. *J. Neurosci.* 27:93–97.
- Carlson T, Goddard E, Kaplan DM, Klein C, Ritchie JB (2018) Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *Neuroimage* 180:88–100.
- Carrasco M (2011) Visual attention: The past 25 years. *Vision Res.* 51:1484–1525.
- Chawla D, Rees G, Friston KJ (1999) The physiological basis of attentional modulation in extrastriate visual areas. *Nat. Neurosci.* 2:671–676.

- 1120 Chen X, Hoffmann KP, Albright T, Thiele A (2012) Effect of feature-selective attention
1121 on neuronal responses in macaque area MT. *J. Neurophysiol.* 107:1530–1543.
- 1122 Cohen MR, Maunsell JHR (2009) Attention improves performance primarily by
1123 reducing interneuronal correlations. *Nat. Neurosci.* 12:1594–1600.
- 1124 Cohen MR, Maunsell JHR (2011) Using neuronal populations to study the mechanisms
1125 underlying spatial and feature attention. *Neuron* 70:1192–1204.
- 1126 Corbetta M, Miezin FM, Dobmeyer S, Shulman GL, Petersen SE (1990) Attentional
1127 modulation of neural processing of shape, color, and velocity in humans.
1128 *Science* 248:1556–1559.
- 1129 Dehaene S, Kerszberg M, Changeux JP (1998) A neuronal model of a global workspace
1130 in effortful cognitive tasks. *Proc. Natl. Acad. Sci. U.S.A.* 95:14529–14534.
- 1131 Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. *Annu.*
1132 *Rev. Neurosci.* 18:193–222.
- 1133 Duncan J (2001) An adaptive coding model of neural function in prefrontal cortex.
1134 *Nat. Rev. Neurosci.* 2:820–829.
- 1135 Duncan J (2006) EPS Mid-Career Award 2004: Brain mechanisms of attention. *Q J*
1136 *Exp Psychol (Hove)* 59:2–27.
- 1137 Duncan J (2010) The multiple-demand (MD) system of the primate brain: Mental
1138 programs for intelligent behaviour. *Trends Cogn. Sci.* 14:172–179.
- 1139 Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL, Group BDC
1140 (2011) Unbiased average age-appropriate atlases for pediatric studies.
1141 *Neuroimage* 54:313–327.
- 1142 Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation
1143 of visual stimuli in the primate prefrontal cortex. *Science* 291:312–316.
- 1144 Fries P, Reynolds JH, Rorie AE, Desimone R (2001) Modulation of oscillatory
1145 neuronal synchronization by selective visual attention. *Science* 291:1560–1563.
- 1146 Friston K, Moran R, Seth AK (2013) Analysing connectivity with Granger causality
1147 and dynamic causal modelling. *Curr. Opin. Neurobiol.* 23:172–178.

- 1148 Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in
1149 functional neuroimaging using the false discovery rate. *Neuroimage* 15:870–878.
- 1150 Ghose GM, Maunsell JHR (2002) Attentional modulation in visual cortex depends on
1151 task timing. *Nature* 419:616–620.
- 1152 Goddard E, Mannion D, McDonald J, Solomon S, Clifford C (2010) Combination of
1153 Subcortical Color Channels in Human Visual Cortex. *J. Vis.* 10:25, 1–17.
- 1154 Goddard E, Carlson TA, Dermody N, Woolgar A (2016) Representational dynamics of
1155 object recognition: Feedforward and feedback information flows.
1156 *Neuroimage* 128:385–397.
- 1157 Gouws AD, Alvarez I, Watson DM, Uesaki M, Rodgers J, Rogers J, Morland AB
1158 (2014) On the role of suppression in spatial attention: Evidence from negative
1159 BOLD in human subcortical and cortical structures. *J. Neurosci.* 34:10347–10360.
- 1160 Gregoriou GG, Gotts SJ, Desimone R (2012) Cell-type-specific synchronization of
1161 neural activity in FEF with V4 during attention. *Neuron* 73:581–594.
- 1162 Grootswagers T, Wardle SG, Carlson TA (2017) Decoding Dynamic Brain Patterns
1163 from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to
1164 Time Series Neuroimaging Data. *J Cogn Neurosci* 29:677–697.
- 1165 Guggenmos M, Sterzer P, Cichy RM (2018) Multivariate pattern analysis for MEG: A
1166 comparison of dissimilarity measures. *Neuroimage* 173:434–447.
- 1167 Guggenmos M, Thoma V, Haynes JD, Richardson-Klavehn A, Cichy RM, Sterzer P
1168 (2015) Spatial attention enhances object coding in local and distributed
1169 representations of the lateral occipital complex. *Neuroimage* .
- 1170 Hanazawa A, Komatsu H, Murakami I (2000) Neural selectivity for hue and saturation
1171 of colour in the primary visual cortex of the monkey. *Eur. J. Neurosci.* 12:1753–1763.
- 1172 Haufe S, Meinecke F, Görden K, Dähne S, Haynes JD, Blankertz B, Bießmann F
1173 (2014) On the interpretation of weight vectors of linear models in multivariate
1174 neuroimaging. *Neuroimage* 87:96–110.
- 1175 Hebart MN, Bankson BB, Harel A, Baker CI, Cichy RM (2018) The representational
1176 dynamics of task and object processing in humans. *Elife* 7.

- 1177 Henson RN, Mattout J, Phillips C, Friston KJ (2009) Selecting forward models for
1178 MEG source-reconstruction using model-evidence. *Neuroimage* 46:168–176.
- 1179 Huang MX, Mosher JC, Leahy RM (1999) A sensor-weighted overlapping-sphere head
1180 model and exhaustive head model comparison for MEG. *Phys. Med.*
1181 *Biol.* 44:423–440.
- 1182 Ipata AE, Gee AL, Goldberg ME (2012) Feature attention evokes task-specific pattern
1183 selectivity in V4 neurons. *Proc. Natl. Acad. Sci. U.S.A.* 109:16778–16785.
- 1184 Jackson J, Rich AN, Williams MA, Woolgar A (2016) Feature-selective Attention in
1185 Fronto-parietal Cortex: Multivoxel Codes Adjust to Prioritize Task-relevant
1186 Information. *J. Cogn. Neurosci.* pp. 1–12.
- 1187 Jehee JFM, Brady DK, Tong F (2011) Attention improves encoding of task-relevant
1188 features in the human visual cortex. *J. Neurosci.* 31:8210–8219.
- 1189 Jerde TA, Merriam EP, Riggall AC, Hedges JH, Curtis CE (2012) Prioritized maps of
1190 space in human frontoparietal cortex. *J. Neurosci.* 32:17382–17390.
- 1191 Kado H, Higuchi M, Shimogawara M, Haruta Y, Adachi Y, Kawai J, Ogata H, Uehara
1192 G (1999) Magnetoencephalogram systems developed at KIT. *IEEE Trans. Appl.*
1193 *Supercond.* 9:4057–4062.
- 1194 Karimi-Rouzbahani H (2018) Three-stage processing of category and variation
1195 information by entangled interactive mechanisms of peri-occipital and peri-frontal
1196 cortices. *Sci Rep* 8:12213.
- 1197 Kastner S, Pinsk MA, Weerd PD, Desimone R, Ungerleider LG (1999) Increased
1198 activity in human visual cortex during directed attention in the absence of visual
1199 stimulation. *Neuron* 22:751–761.
- 1200 Komatsu H, Ideura Y, Kaji S, Yamane S (1992) Color selectivity of neurons in the
1201 inferior temporal cortex of the awake macaque monkey. *J. Neurosci.* 12:408–424.
- 1202 Lafer-Sousa R, Conway BR, Kanwisher NG (2016) Color-Biased Regions of the Ventral
1203 Visual Pathway Lie between Face- and Place-Selective Regions in Humans, as in
1204 Macaques. *J. Neurosci.* 36:1682–1697.

1205 Lee J, Maunsell JHR (2009) A normalization model of attentional modulation of single
1206 unit responses. *PLOS ONE* 4:e4651.

1207 Lee J, Maunsell JHR (2010a) Attentional modulation of MT neurons with single or
1208 multiple stimuli in their receptive fields. *J. Neurosci.* 30:3058–3066.

1209 Lee J, Maunsell JHR (2010b) The effect of attention on neuronal responses to high and
1210 low contrast stimuli. *J. Neurophysiol.* 104:960–971.

1211 Lennert T, Martinez-Trujillo JC (2013) Prefrontal neurons of opposite spatial
1212 preference display distinct target selection dynamics. *J. Neurosci.* 33:9520–9529.

1213 Li X, Lu ZL, Tjan BS, Doshier BA, Chu W (2008) Blood oxygenation level-dependent
1214 contrast response functions identify mechanisms of covert attention in early visual
1215 areas. *Proc. Natl. Acad. Sci. USA* 105:6202–6207.

1216 Li X, Basso MA (2008) Preparing to move increases the sensitivity of superior
1217 colliculus neurons. *J. Neurosci.* 28:4561–4577.

1218 Luo TZ, Maunsell JHR (2018) Attentional Changes in Either Criterion or Sensitivity
1219 Are Associated with Robust Modulations in Lateral Prefrontal Cortex.
1220 *Neuron* 97:1382–1393.e7.

1221 Martinez-Trujillo J, Treue S (2002) Attentional modulation strength in cortical area
1222 MT depends on stimulus contrast. *Neuron* 35:365–370.

1223 Martinez-Trujillo JC, Treue S (2004) Feature-based attention increases the selectivity
1224 of population responses in primate visual cortex. *Curr. Biol.* 14:744–751.

1225 Maunsell J (2015) Neuronal Mechanisms of Visual Attention. *Annu Rev Vis*
1226 *Sci* 1:373–91.

1227 Maunsell JHR, Treue S (2006) Feature-based attention in visual cortex. *Trends*
1228 *Neurosci.* 29:317–322.

1229 McAdams CJ, Maunsell JH (1999) Effects of attention on orientation-tuning functions
1230 of single neurons in macaque cortical area V4. *J. Neurosci.* 19:431–41.

1231 McAdams CJ, Maunsell JH (2000) Attention to both space and feature modulates
1232 neuronal responses in macaque area V4. *J. Neurophysiol.* 83:1751–1755.

- 1233 Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu.*
1234 *Rev. Neurosci.* 24:167–202.
- 1235 Mitchell JF, Sundberg KA, Reynolds JH (2007) Differential attention-dependent
1236 response modulation across cell classes in macaque visual area V4.
1237 *Neuron* 55:131–141.
- 1238 Moher J, Lakshmanan BM, Egeth HE, Ewen JB (2014) Inhibition drives early
1239 feature-based attention. *Psychol. Sci.* 25:315–324.
- 1240 Moore T, Armstrong KM, Fallah M (2003) Visuomotor origins of covert spatial
1241 attention. *Neuron* 40:671–683.
- 1242 Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, Pouget A (2014)
1243 Information-limiting correlations. *Nat. Neurosci.* 17:1410–1417.
- 1244 Motter BC (1994) Neural correlates of attentive selection for color or luminance in
1245 extrastriate area V4. *J. Neurosci.* 14:2178–2189.
- 1246 Mullen KT, Dumoulin SO, McMahon KL, de Zubicaray GI, Hess RF (2007) Selectivity
1247 of human retinotopic visual cortex to S-cone-opponent, L/M-cone-opponent and
1248 achromatic stimulation. *Eur. J. Neurosci.* 25:491–502.
- 1249 Op de Beeck HP, Baker CI, DiCarlo JJ, Kanwisher NG (2006) Discrimination training
1250 alters object representations in human extrastriate cortex. *J.*
1251 *Neurosci.* 26:13025–13036.
- 1252 Pasupathy A (2006) Neural basis of shape representation in the primate brain. *Prog.*
1253 *Brain Res.* 154:293–313.
- 1254 Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming
1255 numbers into movies. *Spat. Vis.* 10:437–442.
- 1256 Pestilli F, Carrasco M (2005) Attention enhances contrast sensitivity at cued and
1257 impairs it at uncued locations. *Vision Res.* 45:1867–1875.
- 1258 Ress D, Backus B, Heeger D (2000) Activity in primary visual cortex predicts
1259 performance in a visual detection task. *Nat. Neurosci.* 3:940–945.

- 1260 Reynolds JH, Desimone R (1999) The role of neural mechanisms of attention in solving
1261 the binding problem. *Neuron* 24:19–29, 111–25.
- 1262 Reynolds JH, Pasternak T, Desimone R (2000) Attention increases sensitivity of V4
1263 neurons. *Neuron* 26:703–714.
- 1264 Reynolds JH, Heeger DJ (2009) The normalization model of attention.
1265 *Neuron* 61:168–185.
- 1266 Rossi A, Paradiso M (1995) Feature-specific effects of selective visual attention. *Vision*
1267 *Res.* 35:621–634.
- 1268 Ruff DA, Cohen MR (2014) Attention can either increase or decrease spike count
1269 correlations in visual cortex. *Nat. Neurosci.* 17:1591–1597.
- 1270 Saenz M, Buracas GT, Boynton GM (2002) Global effects of feature-based attention in
1271 human visual cortex. *Nat. Neurosci.* 5:631–632.
- 1272 Saenz M, Buracas GT, Boynton GM (2003) Global feature-based attention for motion
1273 and color. *Vision Res.* 43:629–637.
- 1274 Saproo S, Serences JT (2014) Attention improves transfer of motion information
1275 between V1 and MT. *J. Neurosci.* 34:3586–3596.
- 1276 Serences JT, Boynton GM (2007) Feature-based attentional modulations in the
1277 absence of direct visual stimulation. *Neuron* 55:301–312.
- 1278 Sprague TC, Saproo S, Serences JT (2015) Visual attention mitigates information loss
1279 in small- and large-scale neural codes. *Trends Cogn. Sci.* 19:215–226.
- 1280 Sprague TC, Serences JT (2013) Attention modulates spatial priority maps in the
1281 human occipital, parietal and frontal cortices. *Nat. Neurosci.* 16:1879–1887.
- 1282 Sundberg KA, Mitchell JF, Reynolds JH (2009) Spatial attention modulates
1283 center-surround interactions in macaque visual area V4. *Neuron* 61:952–963.
- 1284 Tadel F, Baillet S, Mosher JC, Pantazis D, Leahy RM (2011) Brainstorm: A
1285 user-friendly application for MEG/EEG analysis. *Comput. Intell.*
1286 *Neurosci.* 2011:879716.
- 1287 Tremblay S, Pieper F, Sachs A, Martinez-Trujillo J (2015) Attentional filtering of

1288 visual information by neuronal ensembles in the primate lateral prefrontal cortex.
1289 *Neuron* 85:202–215.

1290 Treue S, Martinez-Trujillo JC (1999) Feature-based attention influences motion
1291 processing gain in macaque visual cortex. *Nature* 399:575–579.

1292 Uehara G, Adachi Y, Kawai J, Shimogawara M, Higuchi M, Haruta Y, Ogata H, Kado
1293 H (2003) Multi-channel SQUID systems for biomagnetic measurement. *IEICE*
1294 *Trans. Electron.* E86-C:43–54.

1295 Uusitalo MA, Ilmoniemi RJ (1997) Signal-space projection method for separating
1296 MEG or EEG into components. *Med Biol Eng Comput* 35:135–140.

1297 Vaziri-Pashkam M, Xu Y (2017) Goal-Directed Visual Processing Differentially Impacts
1298 Human Ventral and Dorsal Visual Representations. *J. Neurosci.* 37:8767–8782.

1299 Wardle SG, Kriegeskorte N, Grootswagers T, Khaligh-Razavi SM, Carlson TA (2016)
1300 Perceptual similarity of visual patterns predicts dynamic neural activation patterns
1301 measured with MEG. *Neuroimage* 132:59–70.

1302 Williford T, Maunsell JHR (2006) Effects of spatial attention on contrast response
1303 functions in macaque area V4. *J. Neurophysiol.* 96:40–54.

1304 Woolgar A, Hampshire A, Thompson R, Duncan J (2011) Adaptive coding of
1305 task-relevant information in human frontoparietal cortex. *J.*
1306 *Neurosci.* 31:14592–14599.

1307 Woolgar A, Williams MA, Rich AN (2015) Attention enhances multi-voxel
1308 representation of novel objects in frontal, parietal and visual cortices.
1309 *Neuroimage* 109:429–437.

1310 Wyszecki G, Stiles WS (1982) *Color Science: Concepts and Methods, Quantitative*
1311 *Data and Formulas.* John Wiley & Sons, New York.

1312 Zhang W, Luck SJ (2009) Feature-based attention modulates feedforward visual
1313 processing. *Nat. Neurosci.* 12:24–25.

1314 Zhou H, Desimone R (2011) Feature-based attention in the frontal eye field and area
1315 V4 during visual search. *Neuron* 70:1205–1217.

1316 **Supplementary Material**

1317 **Supplementary 1: Event related potentials**

Figure S1: Event related potentials. Here the average event related potentials (ERPs), across all 2048 trials, are shown (averaged across 20 subjects), for 8 time points evenly spaced between 0ms and 350ms after stimulus onset. Each column shows the same 5 views of the brain at a single time point, and the ERPs are thresholded separately for each column so that only those values that exceed 10% of the maximum ERP at that time point are shown. The peak ERP values across these time points were found from 150 – 250ms after stimulus onset, and the potentials were of greatest amplitude at this time were around the occipital pole and surrounding cortex, consistent with a visually evoked response.

1318 **Supplementary 2: Modelling: extended methods and results**

1319 We started with the Matlab routines from Reynolds and Heeger (2009) that are freely
 1320 available from <http://www.cns.nyu.edu/heegerlab/>. Since we did not have strong a
 1321 priori predictions for many of the model parameters, we tested a broad range of
 1322 plausible model parameters (see Table 1). For each set of model parameters (172,800
 1323 sets in total) we used the Reynolds and Heeger (2009) model to predict the response of
 1324 the neural population as a function of stimulus feature preference (along the shape or
 1325 color dimension), for each of four cases, illustrated by lines of different colors in Figure
 1326 7A-B. In every case the stimulus was a single feature value (a specific color or shape)
 1327 at 2 fixed locations (left and right of fixation). In two cases, we simulated attention to
 1328 one location in the absence of any feature-based attention (simulating attention to the
 1329 orthogonal feature dimension). In the other two cases we simulated attention to one
 1330 location and attention to the feature value of the stimuli. From these we predicted the
 1331 population response at attended and unattended locations, in the presence and absence
 1332 of feature-based attention. As illustrated in Figure 7C, according to the model spatial
 1333 attention tends to boost the population response as a multiplicative scaling of the
 1334 original response, while feature-based attention produces both facilitation and
 1335 suppression of the response which leads to sharpening of the population response
 1336 around the attended value.

1337 One difference between the Reynolds and Heeger (2009) model and our experiment is
 1338 that the model is designed to capture feature-based attention (attending to a specific
 1339 feature value, e.g. red), whereas we manipulated feature-selective attention (attending
 1340 to a feature dimension, e.g. color). While feature-based attention has received greater
 1341 attention in the electrophysiology literature, feature-selective attention has been
 1342 demonstrated to have similar effects at the level of single neurons (Cohen and
 1343 Maunsell, 2011). We therefore implemented the feature-selective attention
 1344 manipulation in the model by generating population responses to two stimuli of the
 1345 same feature value, and modeling the presence of feature-selective attention as
 1346 feature-based attention to that feature value.

1347 For every predicted population response we predicted classifier performance when
 1348 discriminating responses to stimuli of different feature values. To do this we compared

Model parameter	Parameter description	Values tested
<i>stimWidth</i>	Spatial extent of stimulus	25 (Fixed value)
<i>stimFeatureWidth</i>	Extent of stimulus along feature dimension	25 (Fixed value)
<i>ExWidth</i>	Spread of stimulation field along spatial dimension	30, 40, 50, 60, 70, 80, 90 or 100
<i>EthetaWidth</i>	Spread of stimulation field along feature dimension	30, 40, 50, 60, 70 or 80
<i>IxWidth</i>	Spread of suppressive field along spatial dimension	$=C*ExWidth$, where $C=1.5, 2$ or 2.5
<i>IthetaWidth</i>	Spread of suppressive field along feature dimension	$=C*EthetaWidth$, where $C=1.5, 2$ or 2.5
<i>AxWidth</i>	Extent/width of the spatial attention field	$=ExWidth$
<i>AthetaWidth</i>	Extent/width of the featural attention field	$=EthetaWidth$
<i>ApeakX</i>	Peak amplitude of spatial attention field	2, 4, 6 or 8
<i>ApeakTheta</i>	Peak amplitude of the feature-based attention field	2, 4, 6 or 8
<i>Abase</i>	Baseline of attention field for unattended locations/features	1 (Fixed value)
<i>baselineMod</i>	Amount of baseline added to stimulus drive	0, .1, .3, .5 or 1
<i>baselineUnmod</i>	Amount of baseline added after normalization	0, .1, .3, .5 or 1
<i>sigma</i>	Constant that determines the semi-saturation contrast	$1e-6$ (Fixed value)
<i>Ashape</i>	either 'oval' or 'cross'	'oval' (Fixed value)

Table 1: Model parameters from the normalization model of attention (Reynolds and Heeger, 2009) that we used in model simulations. We defined the stimulus and response matrices as varying from -200 to 200 along both spatial and feature dimensions (arbitrary units). We generated the model predictions for every combination of the above model parameters, resulting in 172,800 sets of model predictions. The process of estimating classifier accuracy from the model predictions is summarized in Figure 7.

1349 two population responses that were identical except that they were centered on
 1350 different feature values, as shown in Figure 7D. To simulate the three steps of stimulus
 1351 difference, we considered cases where the centers of the population responses were
 1352 separated by either 20, 40 or 60 in the arbitrary units of the feature dimension. In the
 1353 case of stimuli varying in color, the chromaticity coordinates of the stimuli varied from
 1354 strongly red $u'v'$: 0.35, 0.53, to strongly green $u'v'$: 0.16, 0.56, which means that for
 1355 the model we were treating a difference of 60 arbitrary units as a distance of
 1356 approximately 0.19 in the $u'v'$ chromaticity plane. For shape the feature dimension is
 1357 defined by the transition from 'X-shaped' to 'non-X-shaped'. We are not asserting that
 1358 there exist neurons tuned to this novel complex shape dimension in the same way as
 1359 there are neurons tuned to color, but for the purposes of the model we treated these
 1360 dimensions as equivalent. Since subject performance was similar for the color and
 1361 shape task, we used the same distances (20, 40 and 60 in the arbitrary units) to avoid
 1362 adding another parameter to the modeling results.

1363 Using the pairs of population responses (such as those in Figure 7D) we predicted
 1364 classifier performance (d') using the separation of the two population responses, in a
 1365 manner analogous to that used in signal detection theory. To determine d' for these
 1366 population responses we calculated a 'hit rate' for an optimal observer detecting a
 1367 signal (stimulus two) amongst noise (stimulus one), where their criterion (c) is at the
 1368 midpoint between the peaks of the two curves. We defined the 'hit rate' ($hits$) as the
 1369 area under the blue curve to the right c , and the 'false alarm rate' (FA) as the area
 1370 under the red curve to the right of c . Then the predicted classifier performance $d' =$
 1371 $\text{norminv}(hits) - \text{norminv}(FA)$. In this way, for each set of model parameters we
 1372 predicted classifier performance in each attention condition, for each of the three step
 1373 sizes in feature difference.

1374 From the predicted classification performance, we summarized the predicted effects of
 1375 spatial attention and feature-selective attention using the **SpatAtt** and **FeatAtt**
 1376 values from equations 1 and 2.

Figure S2: Comparing the model predictions across 4 model parameters. The model predictions across 4 model parameters: the excitation and inhibition width of the spatial and feature-based attention fields ($ExWidth$, $IxWidth$, $EthetaWidth$ and $IthetaWidth$ in Table 1). In each cell, there were 400 sets of model parameters (where other model parameters were varied). For each set of model parameters, we calculated the difference between attention effects ($Diff = SpatAtt - FeatAtt$) across feature difference (as in Figure 6). Here we show number of model parameter sets for which the pattern of results was qualitatively similar to the average model prediction (Figure 6B) and to the data (e.g. Figure 6E). That is, model sets where $Diff$ at 3 steps ($Diff(3)$) minus $Diff$ at 1 step difference ($Diff(1)$) was positive (red cells, 95% of cases). There were also some combinations of excitation and inhibition widths for which all 400 cases followed this pattern (bright red cells, 16% of cases).

1377 **Supplementary 3: Control analysis on the effects of spatial bias in** 1378 **fixation location**

1379 To encourage participants to suppress eye movements we provided explicit instructions
1380 to maintain fixation on the constantly-present fixation cross, and we informed
1381 participants that we were using an eye tracker to measure their eye movements. We
1382 also informed participants that the onset of each trial was contingent on the eye tracker
1383 detecting their fixation. We chose a short stimulus duration (maximum 150ms) to
1384 discourage eye movements after the onset of the stimulus, and if the eye tracker
1385 indicated the participant was no longer fixating the stimulus was removed
1386 immediately.

1387 Due to eye tracker variability we treated fixation locations within 1 dva of the center of
1388 the screen as ‘fixating’ for the purposes of the fixation-contingent onset, in order to
1389 avoid extensive delays in the experiment. Because of this, we could not exclude the
1390 possibility that participants had a small bias to fixate slightly towards the attended
1391 location. From the eye tracking data, we found that most participants (16 of 20)
1392 showed a small bias to fixate slightly towards the attended location (see Figure S3). To
1393 check that this bias was not driving the differences we observed between spatial and
1394 feature-selective attention, we repeated our group analyses including only the 4
1395 participants who had a small bias to fixate towards the unattended location, and found
1396 the same pattern of results as in the main analyses (Figure S4).

Figure S3: Distributions of fixation locations, for individual participants. In each distribution, red lines show the median, and the shaded gray box indicates the first and third quartiles of the distribution of 1024 fixation locations. Participants are ordered by their overall bias, from biased towards unattended to biased towards attended location.

Figure S4: Effects of spatial and feature-selective attention across decoding of object color in occipital ROIs for participants with a slight bias to fixate toward the unattended location. Results for a subset of participants ($n=4$, participants 1-4 in Figure S3). Plotting conventions for **A-C** are as in Figure 6C-E.

Supplementary 4: Effects of spatial and feature-based attention on decoding of shape

Figure S5: Effect of spatial and feature-based attention on the decoding of object shape in the occipital ROI. Plotting conventions as in Figure 6C. In this case, there were no consecutive time points at which there was a significant interaction between attention type and step size.

Figure S6: Effects of spatial and feature-selective attention across decoding of object shape for all MEG sensors. Plotting conventions for **A-C** are as in Figure 6C-E.

Supplementary 5: Information flow analysis, varying averaging window

Figure S7: Average event-related potential (ERP) and transformed classification weights (W'). For both the occipital (A-B) and frontal (C-D) ROIs we spatially binned the ROI into 50 equally spaced bins across two dimensions: posterior to anterior (A,C) and left to right (B,D). In each subplot, the bins span the distance indicated by the red line over the ROI in the leftmost column. In the remaining columns, we plot, as a function of time, the average ERP (2nd column) and average transformed classification weights (W' , see methods) for decoding in the attended location, attended feature condition (columns 3-6). That is, ‘Left color’ (column 3) is the decoding of the color of the left object color when performing the color task on the left object, ‘Right shape’ (column 6) is the decoding of the shape of the right object shape when performing the shape task on the right object, etc. The occipital ROI showed a lateralization consistent with classifier performance being driven by retinotopically organized visual cortex: when decoding of features of the stimulus in the left visual field the classifier tended to give higher weight to right hemisphere locations, and vice versa. The frontal ROI did not show clear evidence of lateralization, consistent with frontal regions containing information about both contra- and ipsilateral visual fields (e.g. Lennert and Martinez-Trujillo (2013)).

Supplementary 6: Information flow analysis, varying averaging window

Figure S8: Information flow analysis across varying averaging windows. Upper and lower plots show, for color and shape respectively, the direction of information flow (FF-FB) for each averaging window, given by 5 window widths ($w = 10, 20, 30, 40$ or 50 ms) for each of 6 delays ($d = 50, 60, 70, 80, 90$ or 100). Lines are colored according to the midpoint of the window, and translucent shaded error bars of the same colour indicate the 95% confidence intervals of each between-subject mean. The thick black line shows the average of these lines, replotted from Figure 5A (see Figure 5A for confidence intervals of this average).