Recurrent Processes Emulate a Cascade of Hierarchical Decisions: Evidence from Spatio-Temporal Decoding of Human Brain Activity.

Laura Gwilliams[1, 2] Jean-Remi King[1, 3, 4, 5]
[1]Department of Psychology, New York University, [2]NYU Abu Dhabi Institute, [3]Frankfurt Institute for Advanced Studies, [4]Ecole Normale Supérieure, [5]CNRS

Abstract

Mounting evidence suggests that perception depends on a largely-feedforward brain network. However, the discrepancy between (i) the latency of the corresponding feedforward responses (150-200 ms) and (ii) the time it takes human subjects to recognize brief images (often >500 ms) suggests that recurrent neuronal activity is critical to visual processing. Here, we use magneto-encephalography to localize, track and decode the feedforward and recurrent responses elicited by brief presentations of variably-ambiguous letters and digits. We first confirm that these stimuli trigger, within the first 200 ms, a feedforward response in the ventral and dorsal cortical pathways. The subsequent activity is distributed across temporal, parietal and prefrontal cortices and leads to a slow and incremental cascade of representations culminating in action-specific motor signals. We introduce an analytical framework to show that these brain responses are best accounted for by a hierarchy of recurrent neural assemblies. An accumulation of computational delays across specific processing stages explains subjects' reaction times. Finally, the slow convergence of neural representations towards perceptual categories is quickly followed by all-or-none motor decision signals. Together, these results show how recurrent processes generate, over extended time periods, a cascade of hierarchical decisions that ultimately predicts subjects' perceptual reports.

Keywords: Perceptual decision making, Magneto-encephalography, Recurrent processes

1. Introduction

To process the rich sensory flow emanating from the retina, the brain recruits a hierarchical network originating in the primary visual areas and culminating in the infero-temporal, dorso-parietal and prefrontal cortices. [1, 2, 3, 4].

In theory, the feedforward recruitment of this neural hierarchy could suffice to explain our ability to recognize visual objects. For example, recent studies demonstrate that artificial feedforward neural networks trained to categorize objects generate similar activations patterns to those elicited in the infero-temporal cortices [5, 6].

However, feedforward architectures have a fixed number of processing stages, and are thus unable to explain a number of neural and perceptual phenomena. For example, the time it takes subjects to recognize objects considerably varies from one trial to the next [7]. In addition, the neural responses to visual stimuli generally exceed the 200 ms feedforward recruitment of the visual hierarchy [8, 9].

A large body of research shows that recurrent processing accounts for such behavioral and neural dynamics [9, 10, 11, 12, 13, 14]. In this view, recurrent processing would mainly consist in accumulating sensory evidence until a decision to act is triggered [13].

However, the precise neuronal and computational organization of recurrent processing remains unclear at the system level. In particular, how distinct recurrent assemblies implement series of hierarchical decisions remains a major unknown.

To address this issue, we use magneto-encephalography (MEG) and structural magnetic-resonance imaging (MRI) to localize, track and decode, from whole-brain activity, the feedforward (0-200ms) and recurrent processes (>200 ms) elicited by variably ambiguous characters briefly flashed on a computer screen. We show that the late and sustained neural activity distributed along the visual pathways generates, over extended time periods, a cascade of categorical

decisions that ultimately predicts subjects' perceptual reports.

2. Results

2.1. Subjective reports of stimulus identity are categorical

To investigate the brain and computational bases of perceptual recognition, we used visual characters as described in [15]. These stimuli can be parametrically morphed between specific letters and digits by varying the contrast of their individual edges, hereafter referred to as pixels (Fig.1A-B).

To check that these stimuli create categorical percepts, we asked eight human subjects to provide continuous subjective reports by clicking on a disk after each stimulus presentation (Experiment 1. Fig.1A). The radius and the angle of the response on this disk indicated the subjective visibility and the subjective identity of the stimulus respectively. We then compared (i) the reported angle with (ii) the stimulus evidence (i.e. the expected angle given the pixels) for each morph separately (e.g. 5-6, 6-8, etc). Subjective reports were categorical: cross-validated sigmoidal models better predicted subjects' responses (r=0.49+/-0.05, p=0.002) than linear models (r=0.46+/-0.043, p=0.002, sigmoid>linear: p=0.017 Fig.1B-C).

We adapted this experimental paradigm for an MEG experiment by modifying three main aspects (Experiment 2). First, we used stimuli that could be morphed between letters and digits, to trigger macroscopically distinguishable brain responses in the visual word form area (VWFA) and number form area (NFA) [16, 17]. Second, we added two task-irrelevant flankers next to the target stimulus (Fig.1D) to increase our chances of eliciting recurrent processes via crowding [18]. Third, a new set of seventeen subjects reported subjective identity via a two-alternative forced-choice button press. The identity-response mapping was orthogonal to the letter/digit category and changed on every block of 48 trials. There were 1920 trials total, 320 of which were presented passively and did not require a response.

Perceptual reports followed a similar sigmoidal pattern to Experiment 1: performance was worse for more ambiguous trials (65%) as compared to unambiguous trials (92%, p<0.001). In addition, reaction time slightly, and consistently, increased with difficulty. For example, highly ambiguous stimuli were identified within 690 ms, whereas nonambiguous stimuli were identified within 624 ms (z=-21.68, p<0.001 (Fig.1E-F).

2.2. Neural representations are functionally organized over time and space

Here, we aimed to decompose the sequence of decisions that allow subjects to transform raw visual input into perceptual reports. To this aim, we localized the MEG signals onto subjects' structural MRI with dynamic statistical parametric mapping (dSPM, [19]), and morphed these source estimates onto a common brain coordinate [20, 21]. The results confirmed that the stimuli elicited, on average, a sharp response in the primary visual areas around 70 ms, followed by a fast feedforward response along the ventral and dorsal visual pathways within the first 150-200 ms. After 200 ms, the activity appeared sustained and widely distributed across the associative cortices up until 500-600 ms after stimulus onset (Fig.1G and Supplementary Video 1).

To separate the processing stages underlying these neural responses, we applied i) mass-univariate; ii) temporal decoding and iii) spatial decoding analyses based on the five orthogonal features varying in our study: (1) the position of the stimulus, (2) its identity, (3) its perceived category, (4) its difficulty and (5) its corresponding button press.

First, we aimed to identify when and where low-level visual features would be represented in brain activity. To do so, we estimated, at each time sample separately, the ability of an l2-regularized logistic regression to predict, from all MEG sensors, the position of the stimulus on the computer screen (left versus right). Stimulus position was decodable between 41 and 1,500 ms and peaked at 120 ms (AUC=0.94; SEM=0.007; p<0.001 as estimated with second-level non-parametric temporal cluster test across subjects, (Fig.2C). These signals peaked in the early visual cortex (mean MNI (x=27.59; y=-74.15; z=-1.07)), and propagated along the ventral and dorsal streams during the first 200 ms (Fig.2A, supplementary video). To summarize where stimulus position was represented in the brain, we implemented 'spatial decoders': l2-regularized logistic regressions fit across all time samples (0 - 1,500 ms) for each estimated brain source separately. Spatial decoding peaked in early visual areas and was significant across a large variety of visual and associative cortices as estimated with a second-level non-parametric spatial cluster test across subjects (Fig.2B), confirming the retinotopic organization of the visual hierarchy [22, 23].

Second, we aimed to isolate more abstract representations related to stimulus identity. Stimulus identity can be analyzed either from an objective referential (what stimulus is objectively presented?) or from a subjective referential (i.e. what stimulus did subjects report

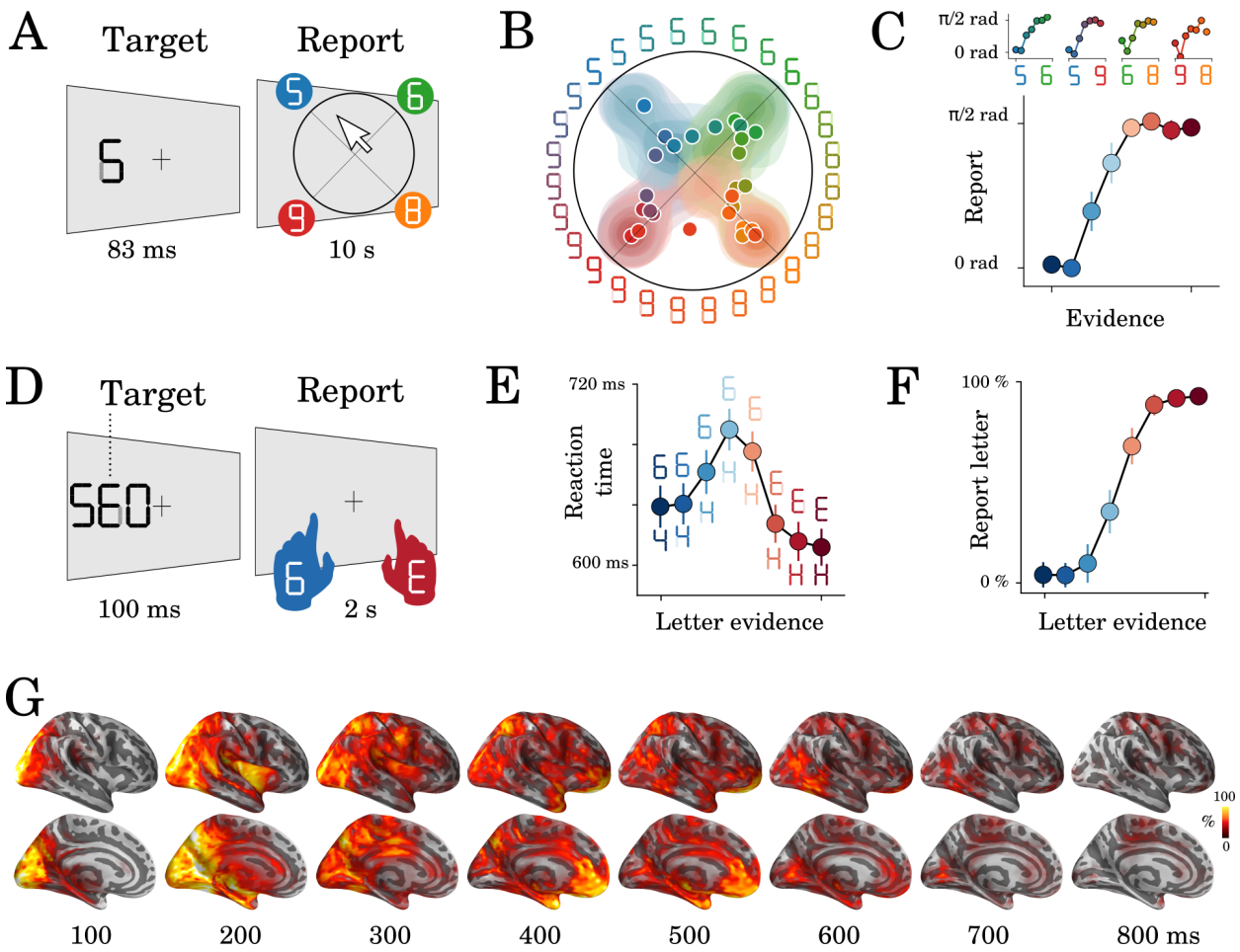

Figure 1: **Experimental protocol and behavioral results.** Experiment 1: 8 human subjects provided perceptual judgments on variably ambiguous digits briefly flashed at the center of a computer screen (A). Reports were made by clicking on a disk, where (i) the radius and (ii) the angle on the disk indicate (i) subjective visibility and (ii) subjective identity respectively. (B) Distribution (areas) and mean response (dots) location for each color-coded stimulus. (C) Top plots show the same data as B, broken down for each morph set. The x-axis indicates the expected angle given the stimulus pixels (color-coded), hereafter referred to as evidence. The y-axis indicates the angle of the mean response relative to stimulus evidence. The bottom plot shows the same data, grouped across morphs. (D) Experiment 2: 17 subjects categorized a briefly flashed and parametrically manipulated-morph using a two-alternative forced-choice. Stimulus-response mapping changed on every block. (E) Mean reaction times as a function of categorical evidence (the extent to which the stimulus objectively corresponds to a letter). (F) Mean probability of reporting a letter as a function of categorical evidence. (G) Evoked activity estimated with dSPM and estimated across all trials and all subjects. These data are also displayed in Supplementary Video 1. Error-bars indicate the standard-error-of-the-mean (SEM) across subjects.

having seen?). We first focus on decoding features of the stimulus that are not ambiguous, such that subjective and objective representations are confounded. To this aim, we grouped stimuli along common continua (e.g. The eight stimuli along the 4-H continuum belong to the same morph and are here considered to share a common identity) and fit logistic regression classifiers across morphs (i.e. E-6 versus 4-H). The corresponding stimulus identity was decodable between 120 and 845 ms and peaked at 225 ms (AUC=0.59; SEM=0.01; p<0.001). These effects peaked more anteriorly than those of stimulus position (mean MNI: x=27.75; y=-62.75; z=-1.55; p<0.001).

Third, we aimed to isolate the neural signatures of subjective perceptual categorization and thus focus on decoding ambiguous pixels. To this aim, we grouped stimuli based on whether the subject reported a digit or to a letter category. Temporal decoders weakly but significantly classified perceptual category from 150 to 940 ms after stimulus onset and peaked at 370 ms (AUC=55%; SEM=0.01; p<0.001, Fig.2C). The corresponding sources also peaked in the inferotemporal cor-

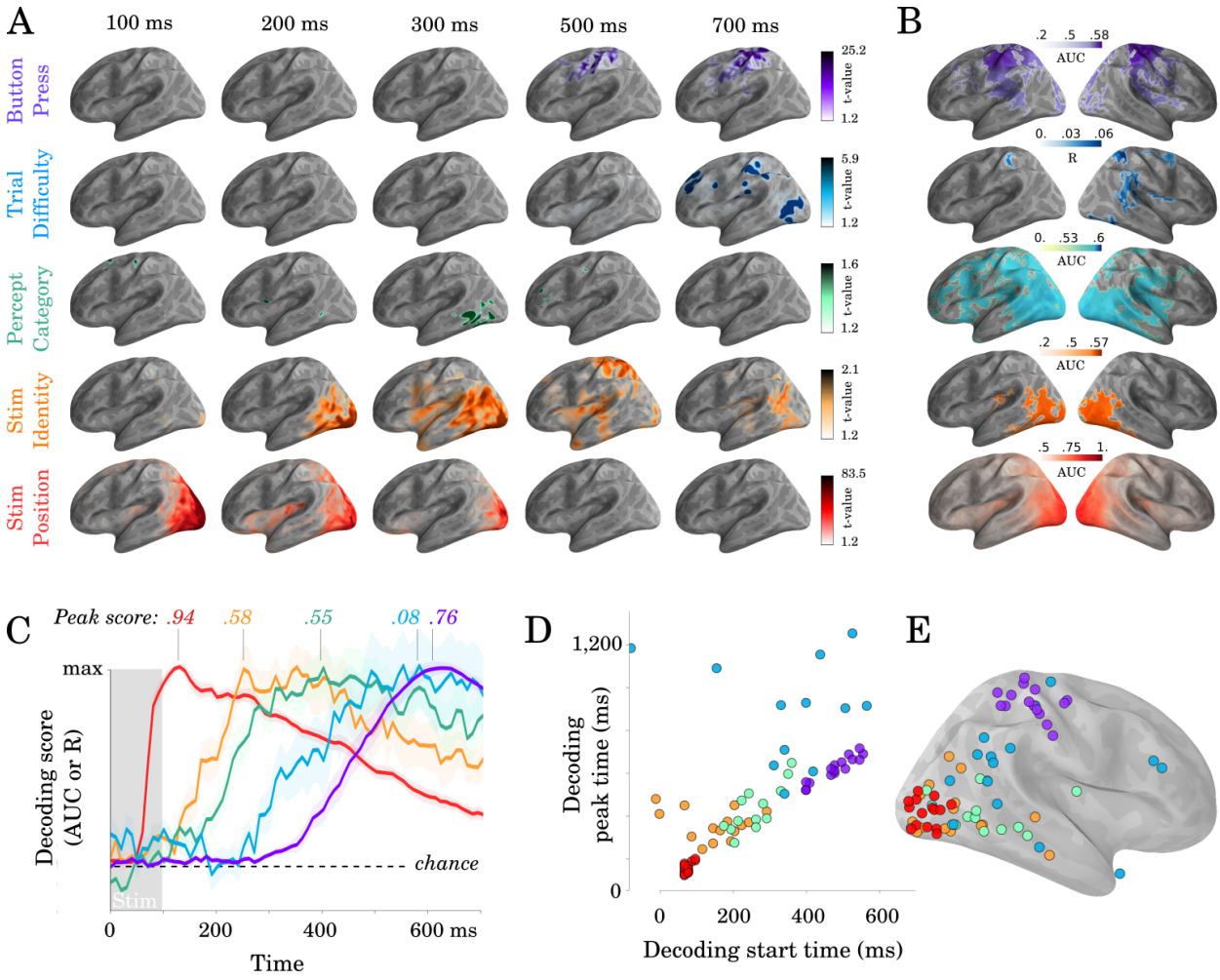

Figure 2: **Spatio-temporal hierarchy.** (A) Mass-univariate statistics. Each row plots the average-across-subjects beta coefficients obtained from regression between single-trial evoked activity and each of the five features orthogonally varying in this study. These results are displayed in Supplementary Video 2. Colors are thresholded based on t-values that exceed an uncorrected p <.1. (B) Spatial-decoders, consisting of linear models fit across all time sample for each source separately, summarize where each feature can be decoded. Lines indicate significant clusters of decoding scores across subjects. (C) Temporal-decoders, consisting of linear models fit across all MEG channels, for each time sample separately, summarize when each feature can be decoded. To highlight the sequential generation of each representation, decoding scores are normalized by their respective peaks. Additional statistics are available in Supplementary Figure 1. (D) The peak and the start of temporal decoding plotted for each subject (dot) and for each feature (color). (E) The peak spatial decoding plotted for each subject (dot) and for each feature (color).

tex but more anteriorly than stimulus identity (x=30.89; y=-35.64; z=21.41; p<0.01). These mass-univariate effects did not survive correction for multiple comparisons (e.g. 210-320 ms: $\bar{t}$=1.79, p=0.21). Nonetheless, spatial decoders, which mitigate the trade-off between temporal specificity and the necessity to correct statistical estimates for multiple comparisons, showed that perceptual category was reliably decoded from a large set of brain areas ($\bar{t}$=4.82; p<0.001; 594 significant vertices) (Fig.2G).

Importantly, when training the classifier on all active trials to distinguish letters (E/H) and digits (4/6), we could significantly (max AUC= 0.55; SEM=0.011; p<.01) decode this contrast for different unambiguous tokens (A/C versus 9/8); suggesting that the response is tracking the abstract letter/digit contrast, abstracted from the specific pixel arrangement.

Fourth, trial difficulty (i.e. the distance between the presented stimulus and the closest unambiguous character) could be decoded between 270 and 1485 ms and peaked at 590 ms (l2-regularized regression fit across sensors, R=0.12; SEM=0.024; p<0.01). Difficulty signals were localized more anteriorly than those of stimulus category (x=12.58; y=-91.44; z=-1.23; p<0.01).

While spatial decoding led to significant clusters in the temporal, parietal and prefrontal areas (Fig.2B), the peak location of stimulus difficulty was highly variable across subjects and included the dorso-parietal cortex, the temporo-parietal junction and the anterior cingulate cortex (Fig.2E).

Finally, temporal decoders of subjects' button press (left versus right index fingers) were significant from 458 ms after stimulus onset and peaked at 604 ms (AUC=0.85; SEM=0.011; p<0.001). A significant cluster of motor signals could be detected around sensorimotor cortices between 590 and 840 ms ($\bar{t}$=4.98, p<0.001, Fig.2A). Response-lock analyses revealed qualitatively similar but stronger results. For example, temporal decoders were significant from 350 ms prior to the response and up to 500 ms after the response reaching an AUC of 94% at response time (p<0.001).

Overall, the time at which representations became maximally decodable correlated with their peak location along the postero-anterior axis (Fig.2D-E) (r=0.57, p<0.001). These results thus strengthen the classic notion that perceptual processes are hierarchically organized across space, time and function. Importantly however, this cascade of representations spreads over more than 600 ms and largely exceeds the time it takes the feedforward response to ignite the ventral and dorsal pathways (Fig.1G and Supplementary Video 1).

2.3. A hierarchy of recurrent layers explains the spatio-temporal dynamics of neural representations

To clarify how a cascade of representations can be generated over extended time periods, we propose to distinguish feedforward and/or recurrent architectures depending on (i) the spatial location, (ii) the timing and (iii) the spatio-temporal dynamics of their representations (Fig.3). This is done by simulating different architectures, and assessing their similarity to the MEG data. In these models, we assume that each 'layer' generates new hierarchical features, in order to account for the organization of spatial decoders (Fig.2E). Furthermore, we only discuss architectures which can code for all representations simultaneously, in order to account for the overlapping temporal decoding scores (Fig.2C). Finally, we only model discrete activations (i.e. a representation is either encoded or not) as any more subtle variation can be trivially accounted for by signal-to-noise ratio considerations.

Each model predicts (1) 'source' decoding time courses (i.e. what is decodable within each layer) and (2) 'temporal generalization' (TG) maps. TG is used to characterize the dynamics of neural representations and consists in assessing the extent to which a temporal decoder trained at a given time sample generalizes to other time samples [24] (Fig.3D).

Our spatial and temporal decoding results can be accounted for by a feedforward architecture that both (i) generates new representations at each layer and (ii) propagates low-level representations across layers (Fig.3 Model 1: 'broadcast'). This architecture predicts that representations would not be maintained within brain areas. This lack of maintenance is not supported by our data. First, the position of the stimulus was decodable in the early visual cortex between 80-320 ms ($\bar{t}$=5.18, p<.001) and thus longer than the stimulus presentation. Second, most temporal decoders significantly generalized over several hundreds of milliseconds (Fig.4A-B). For example, the temporal decoder trained to predict stimulus position from t=100ms could accurately generalize until ≈500 ms as assessed with spatio-temporal cluster tests across subjects (Fig.4A). Similarly, temporal decoders of perceptual category and button-press generalized, on average, for 287 ms (SEM=12.47; p<.001) and 689 ms (SEM=30.94; p<.001) respectively. Given that the neural activity underlying the decoded representations is partially stable over several hundreds of milliseconds, recurrent connections seem necessary to account our data (Fig.4 Model 2-4).

Consequently, we then considered a simple hierarchy of recurrent layers, where recurrence only maintains activated units (Fig.3 Model 2: 'maintain'). This architecture predicts strictly square TG matrices (i.e. temporal decoders would be equivalent to one another in terms of their performance) and is thus at odds with the largely diagonal TG matrices observed empirically (Fig.4A). Specifically, the duration of significant temporal decoding (fitting a new decoder at each time sample) was significantly longer than the generalization of a single decoder to subsequent time samples (e.g. 1,239 versus 287 ms for perceptual category (t=-61.39; p<.001) and 1,215 versus 689 ms for button-press (t=-16.26; p<.001), Fig.4B). These results thus suggest that the decoded representations depend on dynamically-changing activity: i.e. each feature is linearly coded by partially distinct brain activity patterns at different time samples.

It is difficult to determine, with MEG alone, whether such dynamic maintenance results from a change of neural activity within or across brain areas. Indeed, Model 1 and Model 3 can equally predict diagonal TG (Fig.3). However, these two models, and their combination (Model 4) diverge in terms of *where* information should be decodable. Specifically, source analyses revealed that both stimulus position and percep-

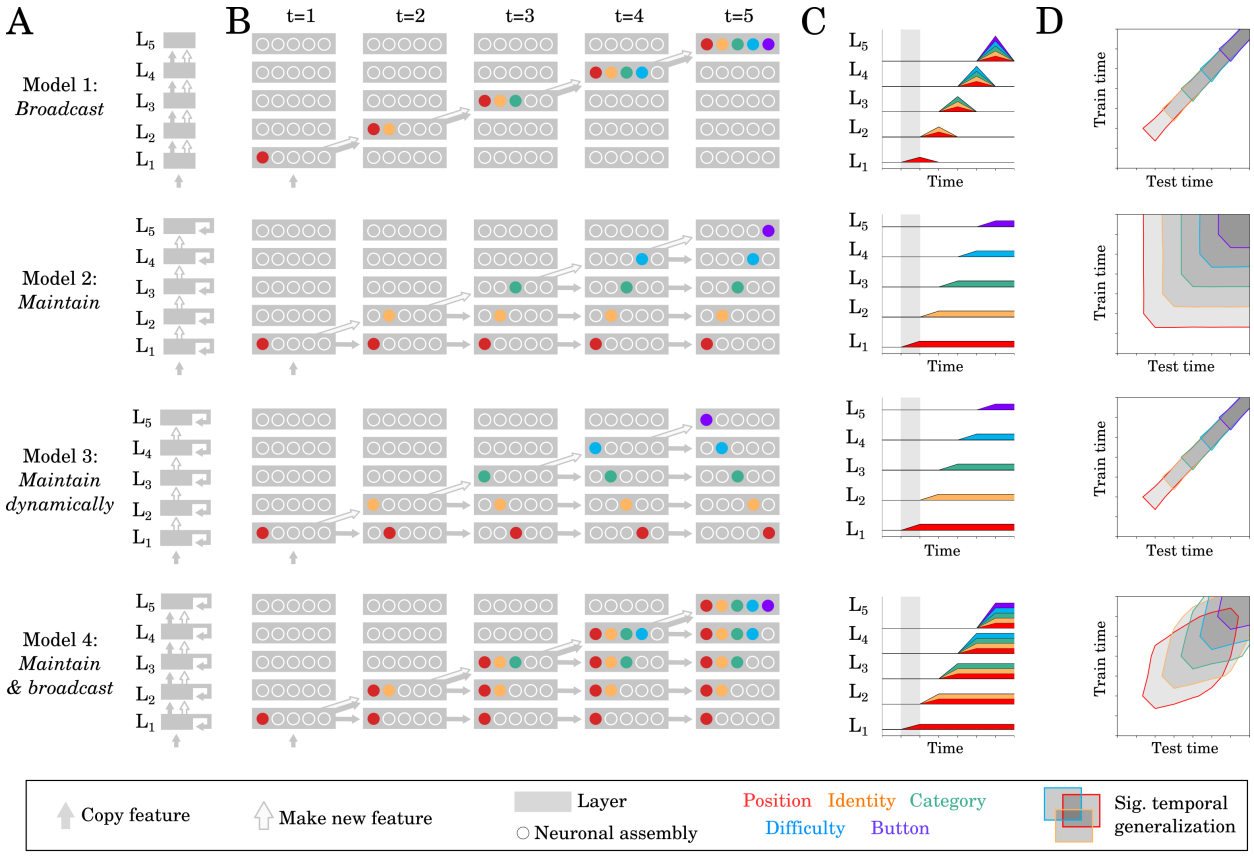

Figure 3: **Source and temporal generalization predictions for various neural architectures.** (A) Four increasingly complex neural architectures compatible with the spatial and temporal decoders of Fig.2. For each model (rows), the five layers (L1, L2 ... L5) generates new representations. The models differ in their ability to (i) propagate low-level representations across the hierarchy, (ii) maintain information with each layer in a stable or dynamic way. (B) Activations within each layer plotted at five distinct time samples. Dot slots indicate different neural assemblies within the same layer. Colors indicate which feature is linearly represented. For clarity purposes, only effective connections are plotted between different time samples. (C) Summary of the information represented within each layer across time. (D) Expected result for of the temporal generalization analyses, based on the processing dynamics of each model.

tual category can be decoded across a wide variety of partially-overlapping brain areas (Fig.2B, Supplementary video 2), similarly to Model 4. Nonetheless, our MEG study remains limited in assessing whether within brain regions dynamics also contribute to the diagonal TG, which would suggest a mixture between models 3 and 4.

Together, source and TG analyses thus suggest that the slow and sequential generation of increasingly abstract representations depends on a hierarchy of recurrent layers that generate, maintain and broadcast representations across the cortex.

2.4. *Hierarchical recurrence induces an accumulation of delays*

Can a hierarchy of recurrent processes account for single-trial dynamics? To address this issue, we hypothesized that recurrent processes would take variable amounts of time to converge to each intermediary representation. In this view, (i) each feature is predicted to propagate across brain areas at distinct moments, and (ii) the successive rise of decodable representations is thus predicted to incrementally correlate with reaction times (Fig.5A-E).

To test this hypothesis, we estimated how the peak of each temporal decoder varied with reaction times. For clarity purposes, we split reaction times into four quantiles, and averaged the time courses of temporal decoders relative to their training time. These analyses showed that the latencies of (i) perceptual category (r=0.35; p=0.006), (ii) stimulus difficulty (r=0.37; p=0.004) and (iii) button press (r=0.66; p<0.001) increasingly correlated with reaction times (Fig.5F-G).

Overall, these results show that we can track with

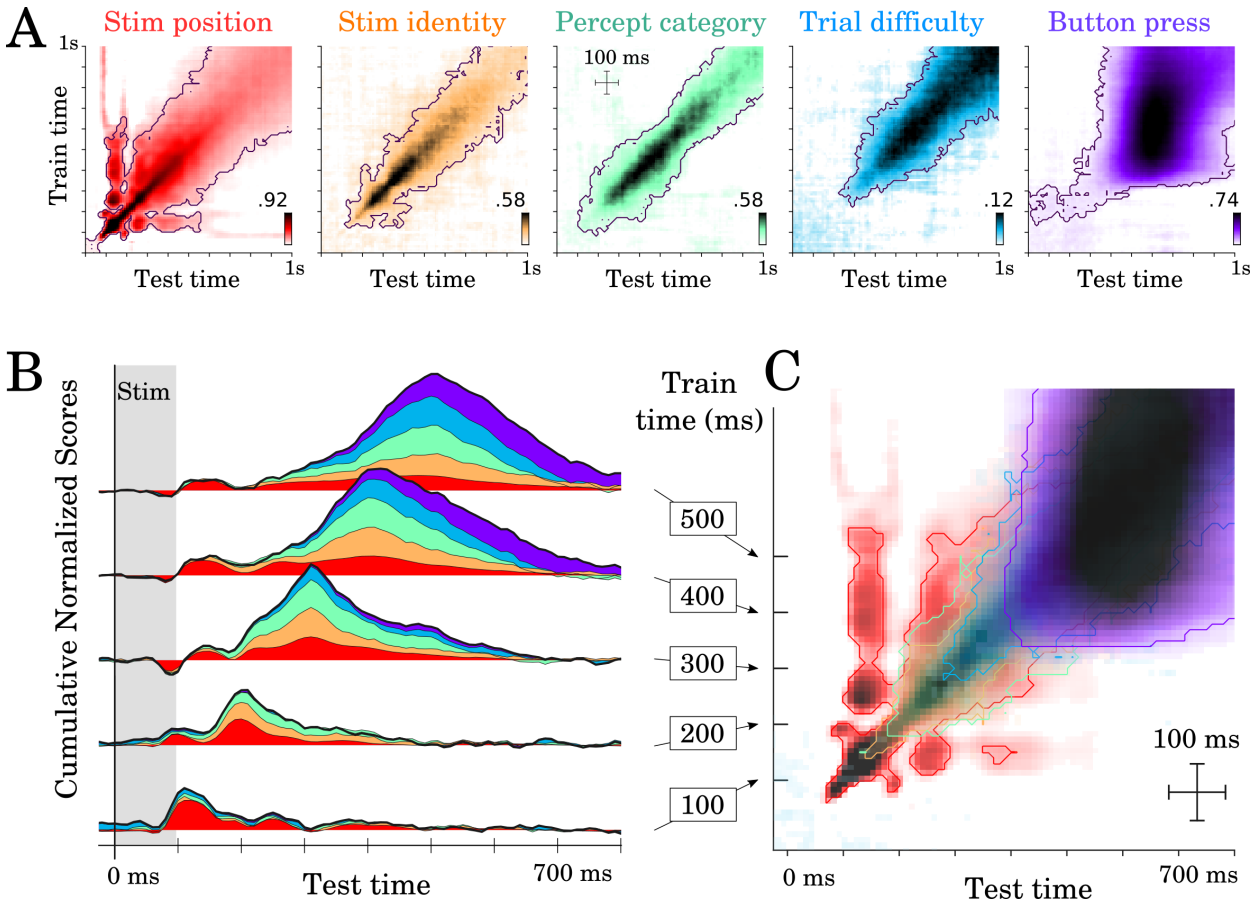

Figure 4: **Temporal generalization results.** (A) Temporal generalization for each of the five features orthogonally varying in our study. Color indicate decoding score (white=chance). Contours indicate significant decoding clusters across subjects. (B) Cumulative temporal generalization scores for the temporal decoders trained at 100, 200, 300, 400 and 500 ms respectively. These decoding scores are normalized by mean decoding peak for clarity purposes. (C) Same data as A but overlaid. For clarity purposes, contours highlight the 25th percentile of decoding performance.

MEG, a series of decisions generated by hierarchical recurrent processes. This neural architecture partially accounts for subjects' variable and relatively-slow reaction times.

2.5. *Hierarchical recurrence implements a series of all-or-none decisions*

An architecture based on successive decisions predicts a loss of ambiguous information akin to all-or-none categorization across successive processing stages (Fig.6A). To test this prediction, we quantified the extent to which the decoding of 'percept category' and of 'motor action' varied linearly or categorically with (i) categorical evidence and (ii) motor evidence respectively (i.e. the extent to which the stimulus (i) objectively looks like a letter or a digit and (ii) should have led to a left or right button press given its pixels).

The probabilistic decoding predictions of percept category correlated linearly with sensory evidence between 210 and 530 ms (r=0.38 +/-0.03, temporal-cluster p<0.001). The spatial decoders fit from 200 to 400 ms clustered around the VWFA ($\bar{t}$=4.6; p=.02; 224 vertices) (Fig 6H). These results suggest that this region first represents the stimulus objectively (i.e. in its full ambiguity).

Between 400 and 810 ms, the predictions of 'perceptual category' decoders were better accounted for by sigmoidal (r=0.77 +/-0.03,199 p<0.001) than by linear trends (r=0.77 +/-0.03, p<0.001). Spatial decoding analyses restricted to the 500-700 ms time window was more distributed ($\bar{t}$=4.4; p=.022; 110 vertices). Finally, ambiguous stimuli (steps 5 and 6 on the continuum) reached maximum decodability 205 ms later than unambiguous stimuli (steps 1 and 8) (p<0.001) (Fig.6J). The interaction between trend (linear or sigmoidal) and

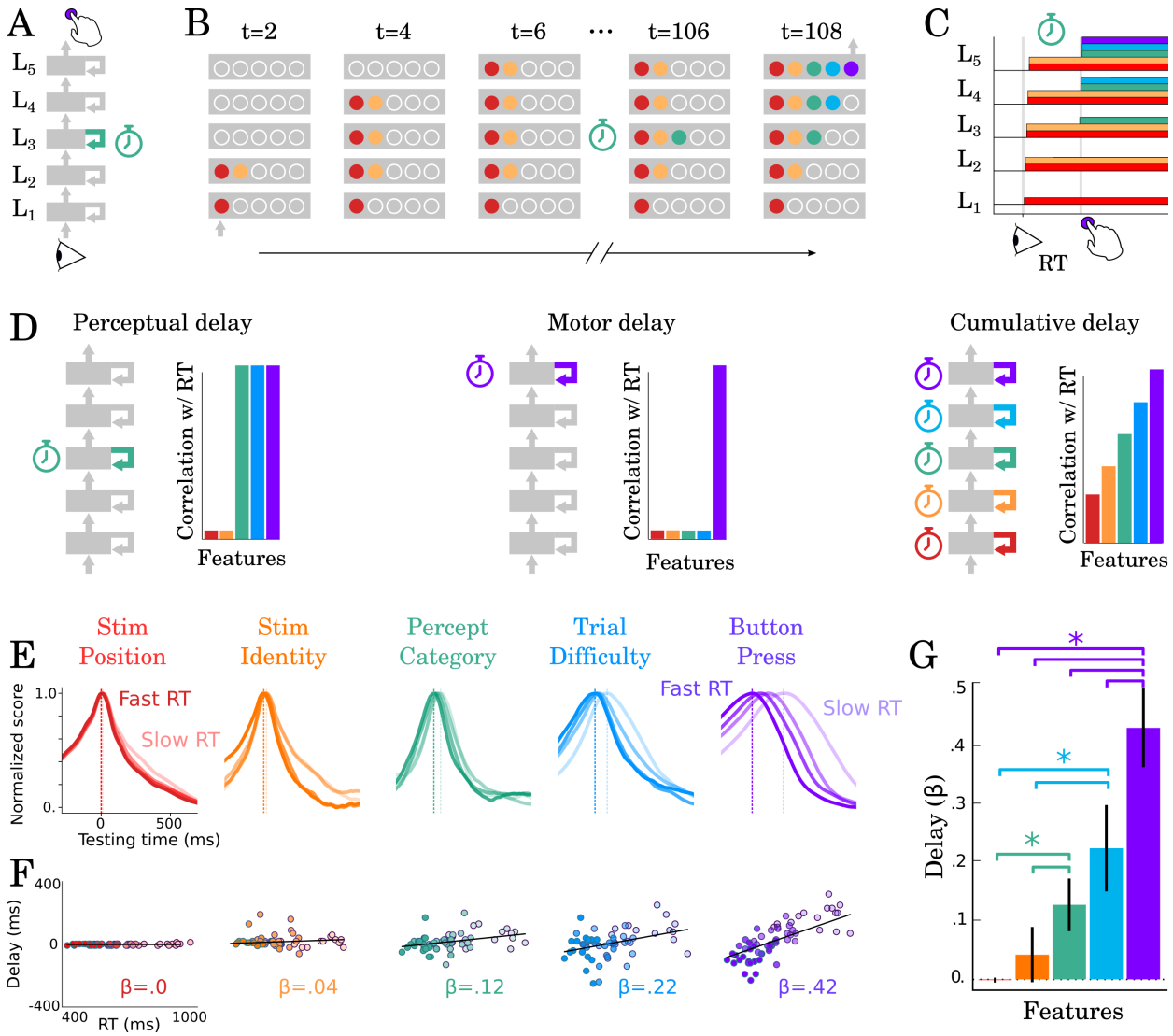

Figure 5: **Correlation between TG peaks and reaction times.** (A, B) Recurrent processing at a given processing stage is hypothesized to take a variable amount of time to generate adequate representations. (C) According to this hypothesis, the rise of the corresponding and subsequent representations would correlate with reaction times. (D, left) Predictions when delays are only induced by the perceptual stage of processing. (D, middle) Predictions when delays are only induced by the motor processing processing stage. (D, right) Predictions when delays are induced by all processing stages. (E) TG scores aligned to training time, split into trials within the fastest and slowest reaction-time quantile and averaged across reaction times bins. Dark and light lines indicate the average decoding performance for trials with fastest and slowest reaction times respectively. (F) Each subject (dot) mean peak decoding time (y-axis) as a function of reaction time (x-axis) color-coded from dark (fastest) to light (slowest). The beta coefficients indicate the average delay estimate. (G) The average slope between processing delay and reaction time for each feature. Error-bars indicate the SEM.

window latency was significant across subjects (r=0.07; SEM=0.01; p=0.002).

This progressive categorization of the letter/digit representations contrasts with the all-or-none pattern of motor signals. Specifically, the probabilistic predictions of button-press decoders varied categorically with response evidence from 440 to 1,290 ms (sigmoid > linear cluster, $\bar{t}$=3.17; p<0.001). There was also a more transient linear trend from 410 to 580 ms ($\bar{t}$=3.69; p<0.001). This suggests that, unlike perceptual category, motor signals largely derive from categorical inputs.

Together, delay (Fig.5) and categorization (Fig. 6) analyses thus show that perceptual representations slowly become categorical and are subsequently followed by all-or-none motor representations.

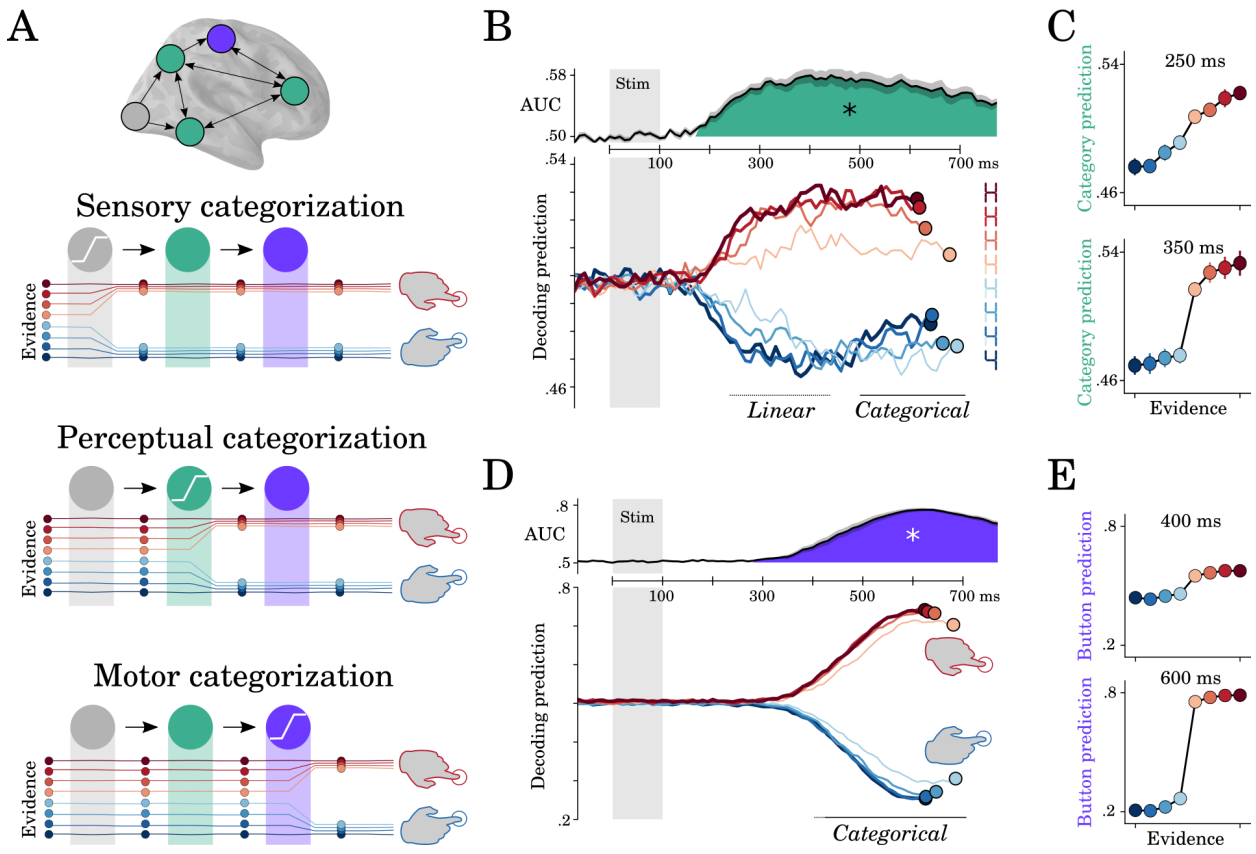

Figure 6: **Motor and perceptual decisions**. (A) Hypothesis space for when responses become categorical: during sensory, perceptual or motor processing. (B, top) Time course of decoding the perceptual decision. (B, bottom) Classifier predictions split into different levels of sensory evidence. (C) Averaging probabilities in different time-windows shows the linear-categorical shift in how information is represented. (D, top) Time course of decoding the motor decision. (D, bottom) Splitting classifier predictions into different levels of difficulty. (E) Different windows of classifier predictions, showing the categorical responses throughout processing.

3. Discussion

Our results show that briefly-flashed stimuli elicit a cascade of representations that spread well beyond the initial feedforward recruitment of the visual pathways. A hierarchy of recurrent processes that incrementally build representations best explains this cascade, its accumulated delays and its all-or-none categorizations.

While macroscopic MEG signals advantageously give a birds-eye view of the cortical correlates of perceptual decision making, it should be stressed that their source reconstruction remains a coarse approximation. Consequently, identifying (1) the role of subcortical areas and (2) the extent to which representations dynamically change within each brain area will necessitate invasive brain recordings.

Nonetheless, our results bridge three important lines of research on the neural and computational bases of visual processing.

First, core-object recognition research, generally based on ≈100 ms-long image presentations has repeatedly shown that the spiking responses of the inferotemporal cortex is better explained by recurrent models than by feedforward ones [9, 14]. In particular, Kar et al have recently shown that images that are challenging to recognize, lead to delayed content-specific spiking activity in the macaque's infero-temporal cortex [14]. Our findings, based on simpler but highly-controlled stimuli, are consistent with these results and further highlight that perceptual representations are not confined to the inferotemporal cortices, but also reach a large variety of parietal and prefrontal areas [25].

Second, the present study makes important contributions to the perceptual decision making literature [11, 13]. With some notable exceptions (e.g. [26]), this line of research primarily aims to isolate motor and supra-modal decision signals in the presence of sustained visual inputs: i.e. neural responses ramping to-

wards a virtual decision threshold, independently of the representation on which this decision is based [13]. The present study complements this approach by tracking the representation-specific signals that slowly emerge after a brief stimulus. Our results thus open an exciting avenue for querying the gating mechanisms of successive decisions and clarifying the role of the prefrontal areas in the coordination multiple perceptual and supramodal modules [27].

Finally, our results constitute an important confirmation of modern theories of perception. In particular, the Global Neuronal Workspace Theory predicts that perceptual representations need to be broadcast to associative cortices via the fronto-parietal areas to lead to subjective reports [8]. Yet, at some notable exceptions [28, 29], previous studies often fail to dissociate perceptual contents and perceptual reports (e.g. [30, 31]). By contrast, the present experimental design allows an unprecedented dissection of the distinct processing stages that transform sensory input into perceptual representations and, ultimately, actions. The generation of letter and digit representations in the dedicated brain areas [32, 17] and their subsequent broadcast to the cortex reinforce the notion that subjective perception relate to the global sharing of content-specific representations across brain areas [8, 33].

4. Method

4.1. Target stimuli

Using the font designed in [15], the stimuli were made from 0, 4, 5, 6, 8, 9, A, C, E, H, O, S, or from a linear combination of two of these characters varying in a single black bar (hereafter 'pixel'). The corresponding 'morphs' were created by adjusting the contrast of the remaining pixel along eight equally spaced steps between 0 (no bar) and 1 (black bar).

4.2. Experiment 1

Eight subjects with normal or corrected vision, seated ≈60 cm from a 19" CRT monitor (60Hz refresh rate, resolution: 1024x768), performed a stimulus identification task with continuous judgements across 28 variably ambiguous stimuli generated from digit stimuli. Ten euros were provided in compensation for this 1-hour experiment.

Subjects performed four blocks of 50 trials, each organized in the following way. After a 200 ms fixation, a target stimulus, randomly selected from one of the 28 stimuli, was flashed for 83 ms on a 50% gray background to the left or to the right of fixation. The orientation of the reporting disk (e.g 5-6-8-9 versus 5-9-8-6) was counterbalanced across subjects. Subjects had then up to 10 seconds to move a cursor on a large disk to report their percepts. The radius on the disk indicated subjective visibility (center=did not see the stimulus, disk border=max visibility). The angle on the disk indicated subjective identity (e.g. 5, 6, 8, 9 for the top left, top right, bottom right, and bottom left 'corners' respectively). Inter-trial interval was 500 ms. To verify that subjects provided meaningful reports, the target stimulus was absent 15% of the trials. Absent trials were rated with a low visibility (i.e radius below 5% of the disk radius) in most cases. Absent trials and trials reported with a low visibility were excluded from subsequent analyses. The report distribution plotted in Fig.1B were generated with Seaborn's bivariate Gaussian kernel density estimate function with default parameters.

Modeling categorical reports. To test whether subjective reports of stimulus identity varied linearly or categorically with sensory evidence, we analyzed how reports' angle (i.e. subjective identity) varied with the expected angle given the stimulus (i.e. sensory evidence).

For each morph (5-6, 5-8, 9-8 and 6-8) separately, we fit a linear model:

$$\hat{y} \leftarrow \beta_1 x + \beta_0 \tag{1}$$

and a sigmoidal model:

$$\hat{y} \leftarrow \frac{1}{1+\exp(\beta_1 x + \beta_2)} + \beta_0 \tag{2}$$

where $\hat{y}$ is the report angle predicted by the model, x is expected angle given the stimulus pixels and β_0 is a free bias parameter.

To minimize the effects of noise, behavioral reports were first averaged within each level of evidence, sorted from the stimulus with the least pixels (e.g. 5, in 5-6 morph) to the stimulus with the most pixels (e.g. 6 in the 5-6 morph). The resulting averages were normalized to range between 0 and 1 within each subject. The β parameters were fit with Scipy's 'curve_fit' function [34] to minimize a mean squared error across trials i:

$$\underset{\beta}{\operatorname{argmin}} \sum_i (y_i - \hat{y}_i)^2 \tag{3}$$

Because the linear and sigmoidal models have distinct numbers of free parameters, we compared them within a 5-split cross-validation. Specifically, the two models were repeatedly fit and tested on independent

trials. A Pearson correlation coefficient r summarised the ability of each model to accurately predict y_{test} given x_{train}, y_{train} and x_{test}. Finally, a Wilcoxon test was applied across subjects to test whether the two models were consistently above chance ($r > 0$) and consistently different from one another ($r_{sigmoid} > r_{linear}$).

Experiment 2. This experiment was performed at Neurospin, Gif usr Yvette, thanks to the support of Stanislas Dehaene. Seventeen subjects performed a discrete identification task across 22 variably ambiguous stimuli generated from letters and digits inside an Elekta Neuromag MEG scanner (204 planar gradiometers and 102 magnetometers). Seventy euros were provided in compensation to the 1-hour experiment and 30 minutes of preparation.

Participants' head shape was digitized along with five fiducial points on the forehead and on each aural canal. Five head-position coils were placed on subjects head and localized at the beginning of each block.

The trial structure was as follows. A black fixation cross was displayed on a 50% gray background for 300 ms followed by a 100ms-long target stimulus presented on the left or on the right of fixation. Two task-irrelevant flankers (e.g. stimulus can be read as an S or a 5) were displayed on the side of this target stimulus to increase our chances of eliciting recurrent processing via crowding [18]. Subjects were given two seconds to report the identity of the stimulus. Reports of stimulus identity were given by pressing a button with the left and right index fingers respectively. The identity-button mapping changed on every block to orthogonalize the neural correlates of stimulus identity and the neural correlates of motor actions. For example, in block 1, perceiving an E or a 4 should have been reported with a left button press, whereas in block 2, E and 4 should have been reported with a right button press. The identity-button was explicitly reminded before each block. In addition, a visual feedback was displayed after non-ambiguous trials. Specifically, the fixation turned green for 100ms or red for 300 ms in correct and incorrect trials respectively. The brain responses to these feedback stimulations are not analyzed in the present study. Inter-trial interval was 1 second. Subjects were provided a short training to ensure they understood the task, and identified non-ambiguous targets at least 80% of the time.

A total of 1920 trials, grouped into 40 blocks, were performed by each subject, 320 of which were presented passively at the end of each block – subjects were not required to provide a response. The trial structure was generated by (i) permuting all combinations of stimulus features (e.g. position, identity, response mapping, difficulty), and (ii) shuffling the order of presentation for each subject. The experiment was presented using Psychtoolbox [35].

All experiments were approved by the local ethics committee. All subjects signed an informed consent form.

4.3. Structural MRI

For each subject, an anatomical MRI with a resolution of 1×1×1.1 mm was acquired after the MEG experiment with a 3T Siemens scanner. Gray and white matter were segmented with Freesurfer 'recon-all' pipeline [20] and coregistered with each subject's digitized head shapes along with fiducial points.

4.4. Preprocessing

The continuous MEG recording was noise-reduced using Maxfilter's SSS correction on the raw data, bandpass-filtered between 0.5 and 40 Hz using MNE-Python's default parameters with firwin design [21] and downsampled to 250 Hz. Epochs were then segmented between -300 ms and +1500 ms relative to stimulus onsets.

After coregistering the MEG sensor data with subjects' structural MRI and the head position coils, we computed the forward model using a 1-layer (inner skull) boundary element model, for each subject separately and fit a minimum-norm inverse model (signal to noise ratio: 3, loose dipole fitting: 0.2, with normal orientation of the dipole relative to the cortical sheet) using the noise covariance across sensors averaged over the pre-stimulus baseline across trials. Finally, the inverse model was applied to single-trial data resulting in a dynamic Statistical Parameter Map (dSPM) [19] value for each source at each time sample.

4.5. Modeled features

We investigated whether single-trial source and sensor evoked responses varied as a function of five features: (1) the position of the stimulus on the computer screen (left versus right of fixation), (2) the morph from which the stimulus is generated (E-6 versus H-4), (3) the category of the stimulus (letter versus digit), (4) the difficulty of the trial (maximum difficulty = stimuli with pixel at 50% contrast; minimum difficult stimuli with pixels at 0% or 100% contrast), and (5) the response button used to report the stimulus (left versus right button). By design, these five features are independent of one another.

It is challenging to dissociate brain responses that represent objective sensory information from those that

represent perceptual decisions as the two are generally collinear. To address this issue, we first fit univariate and multivariate models to predict perceptual category: i.e. whether the button press indicated a character that belongs to the digit or to the letter category. This feature is independent of the button press (e.g. the letter E and the digit 4 can be reported with the same button in a given block). Furthermore, this feature is not necessary to perform the task (i.e. knowing whether E and H are letters is unnecessary to discriminate them). We reasoned that if subjects automatically generates letter/digit representations during perceptual categorization, then we should be able to track the generation of this abstract feature from brain activity.

4.6. Mass univariate statistics

To estimate whether brain responses correlated with each of these five features, we first fit, within each subject, mass univariate analyses at each source location and for each time sample with a linear regression:

$$\beta = (X^T X)^{-1} X y \tag{4}$$

where $X \in \mathbb{R}^{n,f}$ is a design matrix of n epochs by $f = 5$ features and $y \in \mathbb{R}^n$ is the univariate brain response at a given source and at given time. The effect sizes β were then passed to second-level statistics across subjects corrected for multiple comparisons using non-parametric spatio-temporal cluster testing (see below).

4.7. Decoding

Decoding analyses consists in predicting each feature from multivariate brain responses. Decoding analyses were performed within a 5-split stratified K-Fold cross-validation using l2-regularized linear models. Classifiers consisted of logistic regressions (with scikit-learn [36]'s default parameters: $C = 1$):

$$\underset{\beta}{\operatorname{argmin}} \sum_i log(1 + exp(-y_i \beta^T \vec{x}_i)) + C||\beta||^2 \tag{5}$$

where $y_i \in \{\pm 1\}$ is the feature to be decoded at trial i and x_i is the multivariate brain response.

Regressors consisted of ridge regression (with scikit-learn [36]'s default parameters: $\alpha = 1$).

$$\underset{\beta}{\operatorname{argmin}} \sum_i (y_i \beta^T x_i)^2 + \alpha||\beta||^2 \tag{6}$$

For each subject independently, decoding performance was summarized across trials, with an area under the curve (AUC) and a Spearman r correlation score for classifiers and regressors respectively.

All decoders were provided with data normalized by the mean and the standard deviation in the training set.

Spatial decoding consists in fitting a series of decoders at each brain source independently, across all 1,500 time samples relative to stimulus onset. This analysis results in a decoding brain map that indicates where a feature can be linearly decoded in the brain. These decoding maps were then passed to cluster-corrected second-level statistics across subjects.

Temporal decoding consists in fitting a series of decoders at each time sample independently, across all 306 MEG sensors. This analysis results in a decoding time course that indicates when a feature can be linearly decoded from MEG signals. These decoding time courses were then passed to cluster-corrected second-level statistics across subjects.

Temporal generalization (TG) consists in testing whether a temporal decoder fit on a training set at time t can decode a testing set at time t' [24]. TG can be summarized with a square training time × testing time decoding matrix. To quantify the stability of neural representations, we measured the duration of above-chance generalization of each temporal decoder. To quantify the dynamics of neural representations, we compared the mean duration of above-chance generalization across temporal decoders to the duration of above-chance temporal decoding (i.e. the diagonal of the matrix versus its rows). These two metrics were assessed within each subject and tested with second-level statistics across subjects.

4.8. Linear versus Categorical

To test whether neural representations varied as a function of (i) reaction times (RTs, split into 4 quantiles), (ii) sensory evidence (i.e. the extent to the stimulus objectively corresponds to a letter) and (iii) motor evidence (i.e. whether the stimulus should have led to the left button press), we analyzed the extent to which decoders' predictions covaried with each of these three variables z:

$$f(z, \beta^T X) \tag{7}$$

where f is a linear or a sigmoidal model, X is the multivariate brain response and β is the decoder's coefficient fit with cross-validation.

4.9. Statistics

Univariate, decoding and TG models were fit within subjects, and tested across subjects. In case of repeated estimates (e.g. temporal decoding is repeated at each time sample), statistics derived from non-parametric cluster-testing with 10,000 permutations across subjects with MNE-Python's default parameters [21].

Simulations. To formalize how distinct neural architectures lead to distinct spatio-temporal dynamics, we modeled discrete linear dynamical systems forced with a transient input U. Specifically:

$$X_{t+1} = AX_t + BU_t \tag{8}$$

where X is a multidimensional times series (i.e. neurons x time), A is the architecture, and corresponds to square connectivity matrix (i.e. neurons x neurons), B is an input connectivity matrix (i.e. inputs x neurons), and U is the input vector.

Distinct architectures differ in the way units are connected with one another. For simplicity purposes, we order units in the A matrix such that their row index correspond to their hierarchical levels.

In this view, the recurrent, feedforward and skip connections of the architecture A were modeled as a binary diagonal matrix R, a shift matrix F and a matrix S with 1 entries in the last column respectively. These three matrices were modulated by specific weights, as detailed below. The input U was only connected to the first "processing stage", i.e. to the first unit(s) of A, via a matrix B constant across architectures, and consisted of a transient square-wave input, that mimics the transient flash of the stimulus onto subjects' retina.

To model multiple features, we adopted the same procedure with multiple units per layer. Each unit within each layer was forced to encode a specific feature.

Each architecture was fed an input at t=1, and simulated for 8 time steps. Finally, temporal generalization analyses based on the architectures' activations were applied for each of the features.

5. Acknowledgement

This project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 660086, the Bettencourt-Schueller Foundation, the Fondation Roger de Spoelberch, the Philippe Foundation and the Abu Dhabi Institute G1001. We are infinitely grateful to Stanislas Dehaene for his support.

[1] Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology* **160**, 106–154 (1962).

[2] Maunsell, J. & van Essen, D. C. The connections of the middle temporal visual area (mt) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience* **3**, 2563–2586 (1983).

[3] Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nature neuroscience* **2**, 1019 (1999).

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

[5] Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111**, 8619–8624 (2014).

[6] Schrimpf, M. *et al.* Using brain-score to evaluate and build neural networks for brain-like object recognition. In *Cosyne 19, Date: 2019/02/28-2019/03/03, Location: Lisbon, Portugal* (2019).

[7] Ratcliff, R. & Smith, P. L. A comparison of sequential sampling models for two-choice reaction time. *Psychological review* **111**, 333 (2004).

[8] Dehaene, S. & Changeux, J.-P. Experimental and theoretical approaches to conscious processing. *Neuron* **70**, 200–227 (2011).

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

[10] Gray, C. M., König, P., Engel, A. K. & Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**, 334 (1989).

[11] Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annual review of neuroscience* **30** (2007).

[12] Shadlen, M. N. & Newsome, W. T. Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. *Journal of neurophysiology* **86**, 1916–1936 (2001).

[13] O'connell, R. G., Dockree, P. M. & Kelly, S. P. A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature neuroscience* **15**, 1729 (2012).

[14] Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience* **22**, 974 (2019).

[15] King, J.-R. & Dehaene, S. A model of subjective report and objective discrimination as categorical decisions in a vast representational space. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130204 (2014).

[16] Dehaene, S. & Cohen, L. The unique role of the visual word form area in reading. *Trends in cognitive sciences* **15**, 254–262 (2011).

[17] Shum, J. *et al.* A brain area for visual numerals. *Journal of Neuroscience* **33**, 6709–6715 (2013).

[18] Strasburger, H., Rentschler, I. & Jüttner, M. Peripheral vision and pattern recognition: A review. *Journal of vision* **11**, 13–13 (2011).

[19] Dale, A. M. *et al.* Dynamic statistical parametric mapping: combining fmri and meg for high-resolution imaging of cortical activity. *Neuron* **26**, 55–67 (2000).

[20] Fischl, B. Freesurfer. *Neuroimage* **62**, 774–781 (2012).

[21] Gramfort, A. *et al.* Mne software for processing meg and eeg data. *Neuroimage* **86**, 446–460 (2014).

[22] Hagler Jr, D. J. & Sereno, M. I. Spatial maps in frontal and prefrontal cortex. *Neuroimage* **29**, 567–577 (2006).

[23] Wandell, B. A., Dumoulin, S. O. & Brewer, A. A. Visual field maps in human cortex. *Neuron* **56**, 366–383 (2007).

[24] King, J. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences* **18**, 203–210 (2014).

[25] Freedman, D. J. & Miller, E. K. Neural mechanisms of visual categorization: insights from neurophysiology. *Neuroscience & Biobehavioral Reviews* **32**, 311–329 (2008).

[26] Philiastides, M. G. & Sajda, P. Eeg-informed fmri reveals spatiotemporal characteristics of perceptual decision making. *Journal of Neuroscience* **27**, 13082–13091 (2007).

[27] Sarafyazd, M. & Jazayeri, M. Hierarchical reasoning by neural circuits in the frontal cortex. *Science* **364**, eaav8911 (2019).

[28] Tong, F., Nakayama, K., Vaughan, J. T. & Kanwisher, N. Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* **21**, 753–759 (1998).

[29] King, J.-R., Pescetelli, N. & Dehaene, S. Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron* **92**, 1122–1134 (2016).

[30] Sergent, C., Baillet, S. & Dehaene, S. Timing of the brain events underlying access to consciousness during the attentional blink. *Nature neuroscience* **8**, 1391 (2005).

[31] Van Vugt, B. *et al.* The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science* **360**, 537–542 (2018).

[32] Cohen, L. *et al.* The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain* **123**, 291–307 (2000).

[33] Lamme, V. A. Why visual attention and awareness are different. *Trends in cognitive sciences* **7**, 12–18 (2003).

[34] Jones, E., Oliphant, T., Peterson, P. *et al.* SciPy: Open source scientific tools for Python (2001–). URL `http://www.scipy.org/`. [Online; accessed ¡today¿].

[35] Kleiner, M. *et al.* What's new in psychtoolbox-3. *Perception* **36**, 1 (2007).

[36] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830 (2011).