Neuron

Joint representation of working memory and uncertainty in human cortex

Highlights

- Humans know the uncertainty of their working memory and use it to make decisions
- The content and the uncertainty of working memory can be decoded from BOLD signals
- Decoding errors predict memory errors at the single-trial level
- Decoded uncertainty correlates with behavioral reports of working memory uncertainty

Authors

Hsin-Hung Li, Thomas C. Sprague, Aspen H. Yoo, Wei Ji Ma, Clayton E. Curtis

Correspondence

clayton.curtis@nyu.edu

In brief

Li et al. demonstrate that the content and the uncertainty of working memory decoded from BOLD signals in human cortex predict behavioral memory errors and uncertainty reports. The results support the theory that neural populations represent the content and uncertainty of working memory jointly by a probabilistic code.

Li et al., 2021, Neuron *109*, 1–14
November 17, 2021
https://doi.org/10.1016/j.neuron.2021.08.022

CellPress

Neuron

CellPress

Article

Joint representation of working memory and uncertainty in human cortex

Hsin-Hung Li,[1,4] Thomas C. Sprague,[1,3,4] Aspen H. Yoo,[1] Wei Ji Ma,[1,2,5,6] and Clayton E. Curtis[1,2,5,6,7,*]
[1]Department of Psychology, New York University, New York, NY 10003, USA
[2]Center for Neural Science, New York University, New York, NY 10003, USA
[3]Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106, USA
[4]These authors contributed equally
[5]These authors contributed equally
[6]Senior author
[7]Lead contact
*Correspondence: clayton.curtis@nyu.edu
https://doi.org/10.1016/j.neuron.2021.08.022

SUMMARY

Neural representations of visual working memory (VWM) are noisy, and thus, decisions based on VWM are inevitably subject to uncertainty. However, the mechanisms by which the brain simultaneously represents the content and uncertainty of memory remain largely unknown. Here, inspired by the theory of probabilistic population codes, we test the hypothesis that the human brain represents an item maintained in VWM as a probability distribution over stimulus feature space, thereby capturing both its content and uncertainty. We used a neural generative model to decode probability distributions over memorized locations from fMRI activation patterns. We found that the mean of the probability distribution decoded from retinotopic cortical areas predicted memory reports on a trial-by-trial basis. Moreover, in several of the same mid-dorsal stream areas, the spread of the distribution predicted subjective trial-by-trial uncertainty judgments. These results provide evidence that VWM content and uncertainty are jointly represented by probabilistic neural codes.

INTRODUCTION

Working memory (WM) extends the duration over which neural representations are available to guide purposeful behaviors and supports a wide range of cognitive functions, such as learning and decision making (Collins and Frank, 2012; Curtis and Lee, 2010; Wagner, 1999). Although WM is a fundamental building block of cognition, the neural activity that supports WM is noisy and resource limited (reviewed in Ma et al., 2014). Thus, decisions based on WM are inevitably subject to uncertainty (Fougnie et al., 2012; Keshvari et al., 2012; Ma et al., 2014). Access to the uncertainty in our WM enables us to use the extent to which we "trust" our memory to make better decisions. Indeed, people's reported confidence in their WM performance correlates with the magnitude of memory errors, reflecting their ability to track the quality of their memory (van den Berg et al., 2017; Fougnie et al., 2012; Rademaker et al., 2012; Samaha and Postle, 2017). Moreover, people incorporate knowledge of WM uncertainty to improve their decisions in change detection tasks (Devkar et al., 2017; Keshvari et al., 2012; Yoo et al., 2020) and post-memory wagers (Honig et al., 2020; Yoo et al., 2018).

Even though uncertainty plays a key role in supporting WM-guided behavior, we know little about how WM uncertainty is represented in the brain. Previous studies have established that the contents of visual WM (VWM; e.g., the specific remembered orientation, color, motion direction, or spatial location) can be decoded from activation patterns in visual, parietal, frontal cortex, and subcortical regions (Albers et al., 2013; Brissenden et al., 2021; Christophel et al., 2017, 2018; Emrich et al., 2013; Ester et al., 2013, 2015; Harrison and Tong, 2009; Jerde et al., 2012; Lee et al., 2013; Lorenc et al., 2018; Rademaker et al., 2019; Rahmati et al., 2018, 2020; Riggall and Postle, 2012; Serences et al., 2009; Sprague et al., 2014, 2016; Xing et al., 2013; Yu and Shim, 2017). However, these previous studies decoded VWM representations assuming a single point estimate of the memorized stimulus averaged over many trials. As we motivate next, ignoring both the distribution of decoded estimates and their trial-by-trial variability limits our ability to test theories of how neural populations encode VWM content and uncertainty, especially when it comes to links to memory behavior.

Neural population activity is noisy (Faisal et al., 2008; Tolhurst et al., 1983; Tomko and Crapper, 1974). According to the theory of probabilistic population codes, the brain knows the generative model that describes neural population activity as a function of stimulus features (e.g., location or orientation), including the distribution of the noise. Using this knowledge would make it possible to assess the appropriate level of uncertainty associated with a stimulus feature (Foldiak, 1993; Jazayeri and Movshon, 2006; Ma et al., 2006; Sanger, 1996; Zemel et al., 1998), a process known as "inverting" the generative model. Under this theory, a population of neurons contains a joint representation of a

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

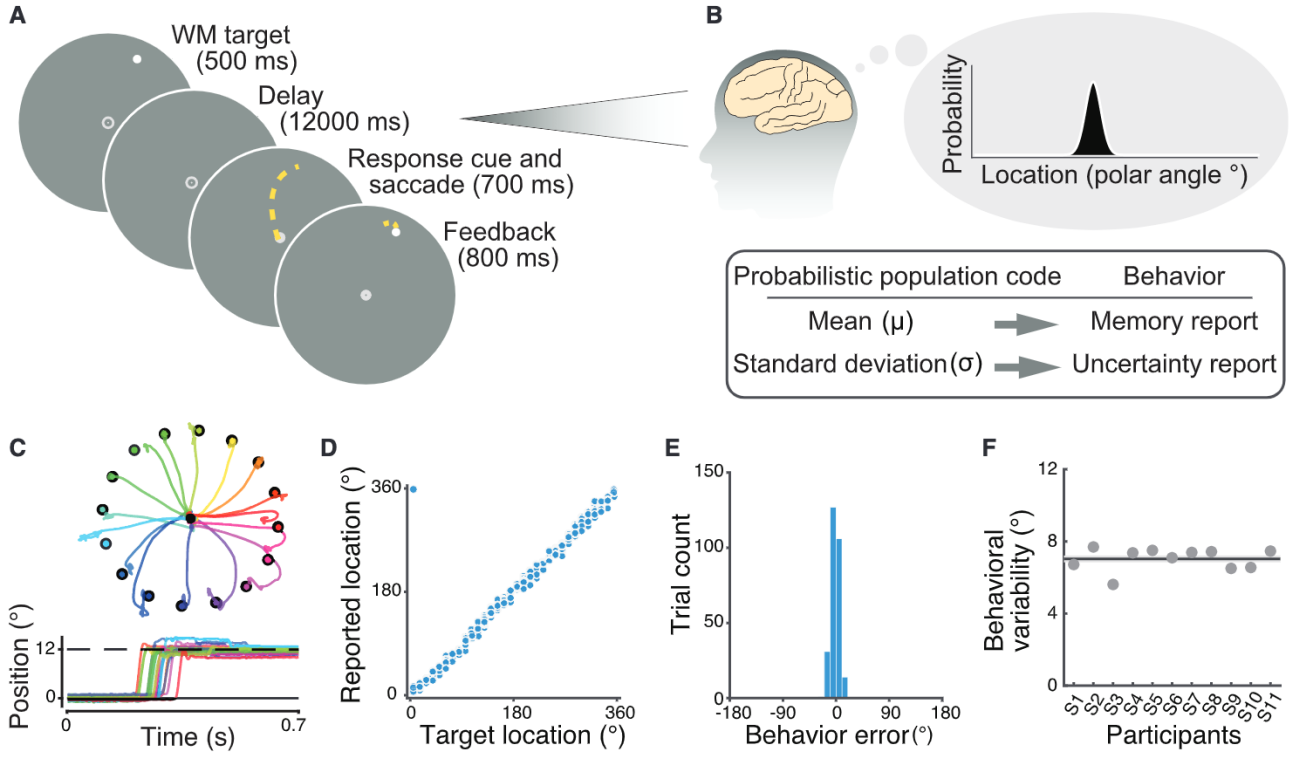

Figure 1. Procedures and working memory performance in experiment 1

(A) Procedures. Participants maintained fixation while remembering the location of the target, presented at a pseudorandom position 12° from fixation. Thereafter, participants generated a memory-guided saccade to the remembered location. Feedback, in the form of a white dot presented at the actual target location, permitted comparison with the landing spot following the memory-guided saccade.

(B) We hypothesized that VWM is represented by a probabilistic population code. Per this hypothesis, populations of neurons represent the remembered target as a probability distribution over stimulus feature values (polar angle of the target in this case). This probability distribution allows a joint representation of the estimate of the memorized target (mean of the distribution) and the uncertainty of memory (SD of the distribution). Two key predictions stem from this hypothesis: readout of the mean of the maintained probability distribution guides memory reports, and the SD of the probability distribution forms one's memory uncertainty.

(C) Example traces of memory-guided saccades for different locations across one scanning run (16 trials). The colored dots evenly spaced on an imaginary circle represent the target locations.

(D) Memory reports from an example participant plotted against the target location.

(E) Memory error distribution from the example participant in (D).

(F) The variability of memory reports for individual participants (dots), quantified by the SD of the memory error distribution. The black horizontal line shows mean across participants, and gray shaded interval shows ± SEM.

stimulus along with uncertainty about the stimulus, and potentially even an entire Bayesian posterior probability distribution over the stimulus. Probabilistic population coding thus provides a testable hypothesis for how neural populations jointly represent a stimulus estimate and the associated uncertainty. In support of this hypothesis, previous studies reported that the probability distributions decoded from neural activity measured in visual cortex predict aspects of visual behavior (van Bergen and Jehee, 2019; van Bergen et al., 2015; Walker et al., 2020). Here, we ask whether higher cognitive processes, such as the items maintained by VWM, are also encoded as probability distributions by neural populations. We hypothesized that similar computational principles explain how neural populations maintain VWM representations. Specifically, we predicted that an item maintained in VWM is represented as a posterior probability distribution over the feature space (e.g., location). In this scenario, access to the content of VWM (e.g., remembered location) would involve a readout of the mean of the distribution, while memory uncertainty would be reflected in the width of the distribution (Figure 1B). The critical direct test of this hypothesis hinges on whether the parameters of the probability distribution actually predict the quality and uncertainty of measured memory behavior.

We measured fMRI blood-oxygen-level-dependent (BOLD) activity in humans performing two experiments to test each prediction from the above hypothesis. In the experimental design, we set out to ensure that we measured the uncertainty of VWM instead of sensory-evoked responses as in previous studies (van Bergen et al., 2015; Walker et al., 2020). We used a long memory delay to help isolate our measurements from the stimulus epoch, and we analyzed the BOLD signal corresponding to a late time window during the delay. Moreover, we conducted a passive viewing control experiment to make sure that the decodable neural signals we observed represent VWM content and not merely sensory responses. For decoding analysis, we adapted and inverted a generative model for the BOLD activity

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

CellPress

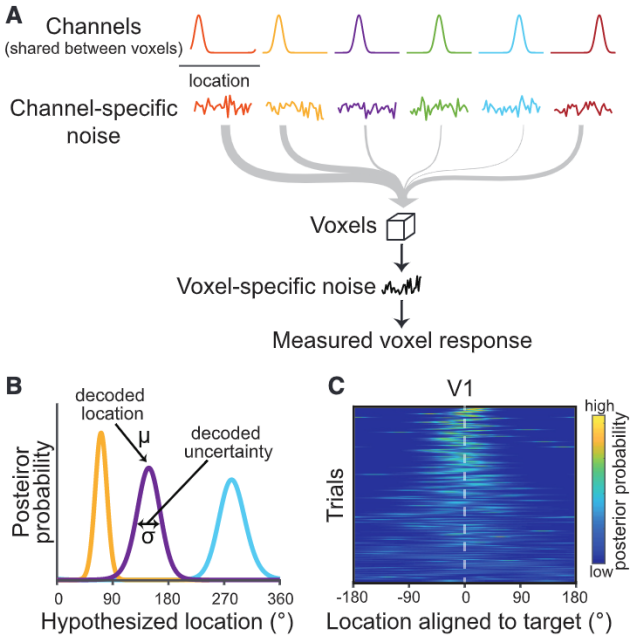

Figure 2. Generative model used to estimate and decode working memory representations

(A) Schematic of the generative model for BOLD response for spatial VWM (van Bergen and Jehee; 2021). The tuning function (mean response amplitude as a function of remembered target location) of each voxel is modeled as a weighted sum of eight basis functions evenly spanning the entire location space (0°–360°; note that six are shown here in cartoon depiction). Two sources of noise are considered: noise arising from each channel, which is shared across voxels, and noise arising from each voxel independently.

(B) Posterior probability distributions decoded from the memory delay of three example trials. The decoded memorized location is derived from the circular mean of the posterior distribution. The decoded uncertainty in memory is derived from the circular SD of the posterior distribution.

(C) Posterior probability distributions decoded from an example participant's primary visual cortex. Each row presents the posterior probability distribution decoded from the delay period of a single trial, where trials are sorted (from top to bottom) on the basis of the decoded uncertainty of each trial (from lowest to highest uncertainty). The posterior distributions are circularly shifted to align to the target position of each trial (0°).

(van Bergen and Jehee, 2021; van Bergen et al., 2015). This yielded, on a trial-to-trial basis, a probability distribution over a memorized stimulus location from an activation pattern measured from retinotopic visual, parietal, and frontal cortex. Although fMRI BOLD activity is subject to measurement noise, we still predicted that the decoded probability distribution would bear a resemblance to the one that the brain might use for its decision making. In experiment 1, we demonstrated that we can reliably decode the content of spatial VWM from BOLD activation. Moreover, trial-by-trial errors in the decoded positions predicted behavioral recall errors, revealing a close relationship between the decoded memory content and memory recall. In experiment 2, we further established that the decoded uncertainty predicted explicit uncertainty reports when participants introspected the quality of their VWM in a wager task. Our results support the theory that the brain uses knowledge of the generative process of neural activity to represent memorized items probabilistically; in other words, that neural activity multiplexes the content of VWM and its uncertainty.

RESULTS

Experiment 1

In experiment 1, we used a Bayesian generative model to decode single-trial VWM representations from neural activation patterns and assessed how the decoded memory content related to memory reports. We studied spatial VWM in a memory-guided saccade task. In each trial, we presented participants with a brief (500 ms) target dot, followed by a 12 s delay period (Figure 1A). The polar angle of the target was chosen pseudo-randomly from 1 of 32 positions that spanned the full circle. Participants were asked to remember the location of the target while maintaining central fixation throughout the delay period. After the delay period, the empty fixation dot was replaced by a filled dot serving as the response cue. Upon the response cue onset, participants reported the remembered position by making a saccadic eye movement (e.g., Funahashi et al., 1989; Hikosaka and Wurtz, 1983; Figures 1A and 1B). Behavioral memory reports were measured as the polar angle of the saccade endpoint.

We adapted a generative model (van Bergen and Jehee, 2021; van Bergen et al., 2015) to decode a probability distribution over the stimulus location (polar angle) from the delay period brain activity for each single trial. To focus on VWM maintenance activity, we used the averaged BOLD response for each voxel from 5.25 to 12.00 s after the delay period onset as the input to the model. The generative model described the multivariate voxel response given a stimulus location by a multivariate normal distribution. To estimate the mean of this distribution, the model approximated each voxel's spatial tuning curve by a weighted sum of eight basis functions (channels) that evenly tiled visual (polar angle) space (Figure 2A). For the covariance of the multivariate normal distribution, the model incorporated the empirical noise covariance estimated by the data and a theoretical noise covariance matrix that considered two sources of variability: the noise of each location channel and the noise specific to each voxel (see STAR Methods). For each trial, we used the circular mean of the decoded probability distribution to represent the decoded remembered location.

We first demonstrated that we can decode VWM content from delay period fMRI signals. We defined four retinotopic visual (V1, V2, V3, and V3AB), four parietal (IPS0, IPS1, IPS2, and IPS3), and two frontal (iPCS and sPCS) areas as regions of interest (ROIs) using population receptive field mapping techniques (Dumoulin and Wandell, 2008; Mackey et al., 2017). Similar to previous studies using other decoding methods (Hallenbeck et al., 2021; Jerde et al., 2012; Rahmati et al., 2018; Sprague et al., 2014, 2016), we found that the remembered stimulus location could be decoded from the delay period BOLD responses in retinotopic visual, parietal, and frontal cortex. First, we plotted a distribution of the trial-wise decoding error (decoded location minus target location; Figure 3A) for each ROI. These decoding error distributions reliably exhibited a single peak centered near 0°, indicating the robustness of our decoder (Table S1). We quantified the existence of decodable VWM information by comparing

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

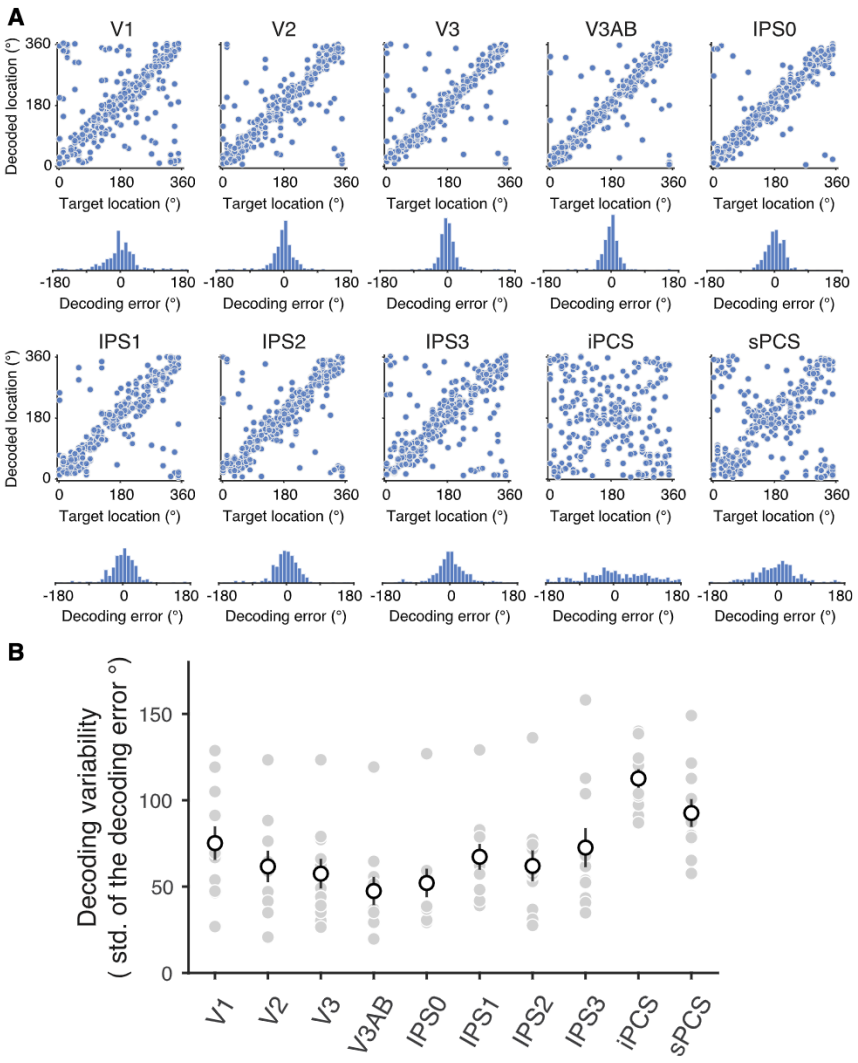

Figure 3. Working memory content can be precisely decoded

(A) Decoding performance of an example participant. For each ROI, the top figure represents the decoded location as a function of the memorized target location. The bottom figure is the distribution of signed decoding error (decoded location minus the memory target location).

(B) Decoding performance quantified as decoding variability, the SD of the decoding error distribution. The filled gray dots represent individual participants. The empty white dots represent the group average. The error bars represent ± SEM. Decoding performance varied significantly across ROIs (permutation one-way repeated-measures ANOVA, $F[9, 90] = 32.82$, $p < 0.001$, $\eta_p^2 = 0.77$).

the SD of the decoding error distributions with a null distribution obtained by a permutation procedure (see STAR Methods). At the individual participant level, target locations were robustly decoded in most ROIs (Table S2). At the group level, VWM contents were decoded in all ROIs ($p < 0.001$; unless otherwise noted, we report p values corrected for multiple comparisons across ROIs via false discovery rate [FDR] with $q = 0.05$). The SD of the decoding error distributions varied significantly across ROIs (permutation one-way repeated-measures ANOVA, $p < 0.001$, $\eta_p^2 = 0.77$; Figure 3B), with smaller SDs in extrastriate cortex regions V3AB and V3, and IPS0 in intraparietal sulcus, indicating a more precise decoding performance in these regions. Two regions in the prefrontal cortex, iPCS and sPCS (the putative human homolog of macaque frontal eye field [FEF]), had the largest SDs, indicating lower decoding performance in these areas. In another analysis, we obtained similar results when using the circular correlation between decoded location and target location to quantify the decoding performance (Figure S1A; Table S3).

As we are interested in WM, we established that the signals we decoded cannot be attributed to sensory responses to the target stimuli. First, we are modeling the BOLD responses well into the delay period. Second, in a passive viewing experiment, a subset of participants (n = 3) performed a discrimination task at central fixation without the requirement to remember peripheral targets. Instead, we presented a highly salient, but irrelevant, flickering checkerboard at the same locations used in the WM task for the same duration as the VWM target stimulus (500 ms). Compared with the VWM experiment, the SD of the decoding error distribution (averaged across subjects and ROIs) on the basis of the same delay period time points increased from 71° to 130° in the passive viewing experiment (Figures S2A and S2B). The near doubling of variability of the decoding error was only barely distinguishable from the null distribution in a subset of ROIs and participants (Table S2). Moreover, the circular correlation between the decoded location and the target location was at zero under passive viewing for most participants and ROIs (Figure S2C; Table S3). When we instead decoded stimulus location from earlier time points with strong evoked sensory signals (0.75–5.25 s from the delay deploy period onset), we were able to accurately decode stimulus location (Figures S2D–S2F). Together, these results indicated that the neural representations of the target only persisted through the late delayed period when they were actively maintained in VWM.

So far, we have shown that we can decode the location of the memorized target. However, if the decoded VWM representation obtained from the BOLD signal drives behavioral performance, the decoded VWM representation should contain information relevant for behavioral memory reports beyond the physical location of the target. To investigate this issue, we leveraged our single-trial decoded locations, and we tested the prediction that (signed) memory error and (signed) decoding errors correlate at the trial-by-trial level. That is, we tested whether the direction of errors in memory and errors in decoding are the same

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

CellPress

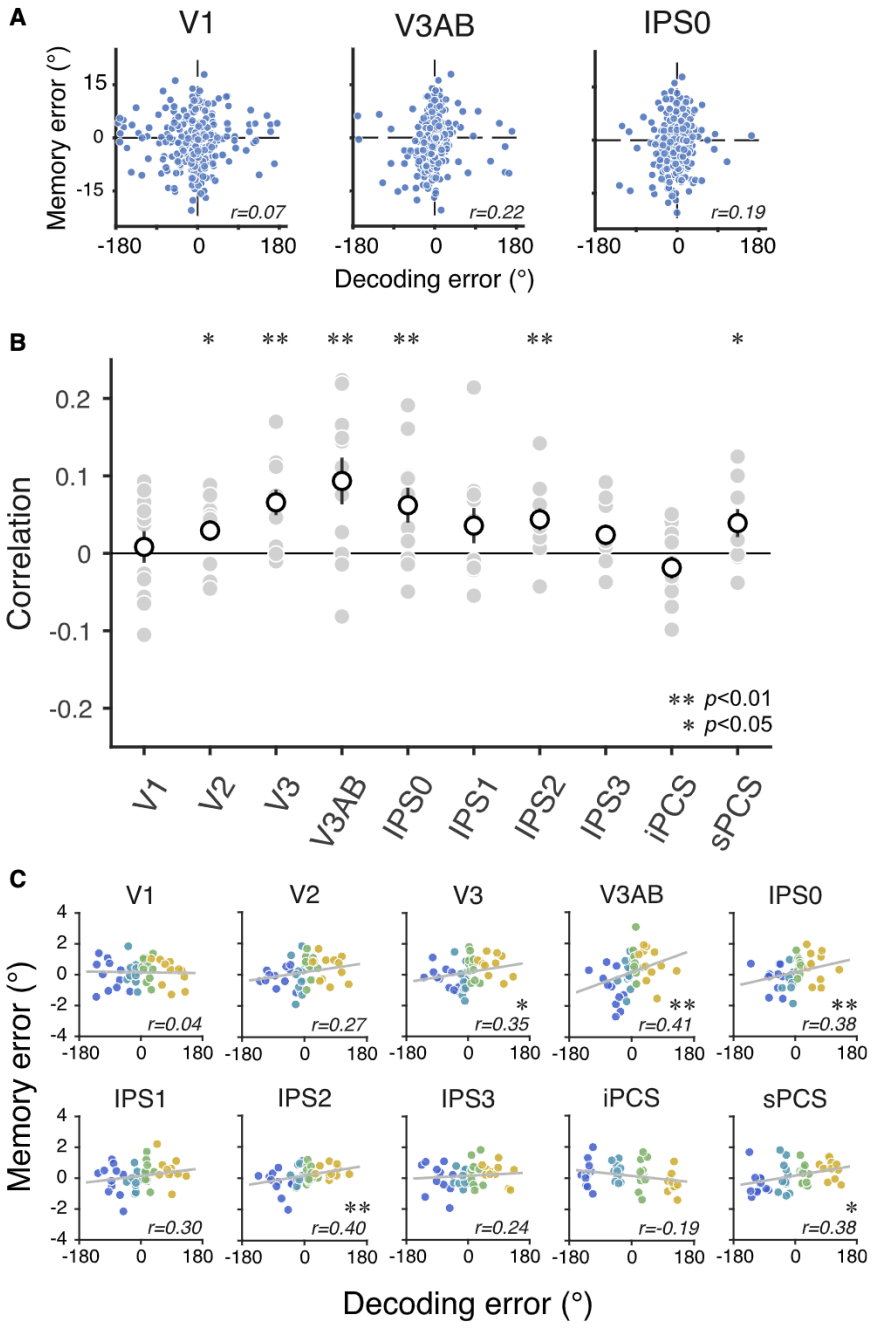

Figure 4. Errors in neural decoding of working memory predict behavioral memory errors in experiment 1

(A) Memory errors plotted against neural decoding errors of an example participant for three ROIs.

(B) Correlations are computed as circular correlations between the behavioral memory errors and neural decoding errors. The filled gray dots represent individual participants. The unfilled white dots represent the group average. The error bars represent ± SEM.

(C) Memory errors plotted against decoding errors. The four colors indicate four bins (within each of 11 participants) sorted by decoding error. The gray line in each panel represents the best linear fit. The value at the lower right of each panel is the Pearson correlation coefficient.

(e.g., clockwise) with respect to the target. Accordingly, we computed the trial-wise circular correlation between memory error and decoding error for each participant and ROI. We found that the strength of this correlation varied across ROIs (permutation one-way repeated-measures ANOVA, $p < 0.001$, $\eta_p^2 = 0.30$). For individual ROIs, we found significant positive correlations in multiple regions, including V2, V3, V3AB, IPS0, IPS2, and sPCS (bootstrapping test, $p < 0.05$; Figure 4B). Following previous studies using a similar Bayesian decoding approach (van Bergen and Jehee, 2021; van Bergen et al., 2015), we quantified the correlations by binning the trials on the basis of their decoding errors, computing the memory error of each bin, and pooling the data across participants. We observed similar patterns, as significant positive correlations were observed in multiple ROIs, including V3, V3AB, IPS0, IPS2, and sPCS (permutation test, $p < 0.05$; Figure 4C). Overall, we found that memory behavior was linked to the neural representations we decoded, supporting our prediction that access to the content of VWM involves a readout of the mean of the population-encoded probability distribution.

Experiment 2

In experiment 1, participants reported the remembered location using a memory-guided saccade, and we quantified performance on the basis of saccade landing position. We found not only that the population activity encoded the VWM target location but additionally that errors in the decoder predicted errors in memory. Next, we tested the hypothesis that the population activity encodes a joint representation of both memory content and the uncertainty of their memory. Indeed, we can reflect on and directly report the confidence we have in our memory (Fougnie et al., 2012; Honig et al., 2020; Rademaker et al., 2012; Yoo et al., 2018). Do these introspective reports reflect the uncertainty associated with the neural representation, quantified on the basis of the posterior distribution decoded from neural activation patterns? To test this, in experiment 2, we adapted our task so that participants were required to explicitly report the uncertainty of their memory with a wager.

The experimental procedures were similar to experiment 1, with a few modifications. In addition to the filled dot at central fixation, the response cue contained a circular annulus with a radius matching the eccentricity of the target (Figure 5A).

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

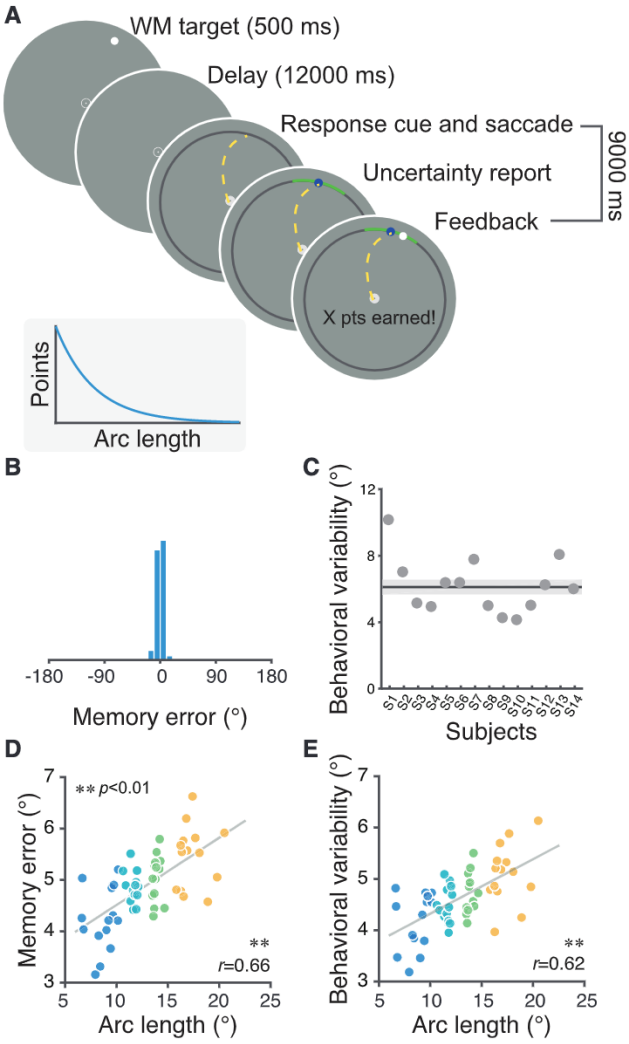

Figure 5. Procedures and working memory performance in experiment 2

(A) The procedures were similar to those of experiment 1 except for a few modifications. To report the remembered location, the participants generated memory-guided saccades onto a ring and then reported their memory uncertainty by adjusting the length of an arc centered at the reported location. The trial ended with the onset of the feedback stimulus, a white dot presented at the target location. Participants earned points only if the target location was within the arc, and the points they earned decreased with the arc length. To earn a high score, participants should set shorter arcs when less uncertain.

(B) The distribution of memory error from one example participant.

(C) The variability of memory reports for individual participants, quantified by the SD of the memory error distribution. The black horizontal line shows mean across participants, and gray shaded interval shows ± SEM.

(D) Memory error as a function of reported arc length, binned. Four colors represent four bins (within each of 14 participants) with increasing arc length.

(E) Behavioral variability as a function of reported arc length. In trials in which participants reported longer arc lengths, behavioral recall of remembered positions had larger errors (D; permutation test, $p < 0.001$) and was more variable (E; permutation test, $p < 0.001$).

In (D) and (E), the gray lines represent the best linear fits.

Participants reported their memory by making a saccade to the position on the annulus that matched the eccentricity of the remembered target. Then, participants placed a wager by adjusting the length of an arc attached to the reported location (Honig et al., 2020; Yoo et al., 2018). Participants were instructed to use the length of the arc to reflect the uncertainty of their memory. At the end of a trial, a feedback dot was presented at the true target location. Participants gained points if the target location was within the arc or otherwise gained zero points. The points they could gain decreased with the length of the arc, so observers were motivated to reflect their uncertainty using the arc length. In order to obtain the highest points, an optimal observer would increase the length of the arc with higher VWM uncertainty (Honig et al., 2020; Yoo et al., 2018).

Behaviorally, participants were able to monitor the quality of their VWM. Both the magnitude of memory errors (Figure 5D; permutation test, $p < 0.001$) and the variability of memory reports (Figure 5E; permutation test, $p < 0.001$) increased with the reported arc length. For some participants, the reported arc length varied as a function of target location, with shorter arc lengths and smaller errors at the cardinal angles (Figure S3A). To evaluate whether the participants could track VWM uncertainty independent of the target location, we regressed out the effect of target location (the polar angle between the target to the nearest cardinal angles) from the arc length. Still, the arc length correlated with the magnitude of memory error and the variability of memory reports (Figures S3B and S3C), indicating that the participants' ability to track the uncertainty across trials was not driven solely by the physical locations of the target. In sum, participants not only were aware of their memory uncertainty but used these estimates to inform their wagers.

Next, we tested the hypothesis that these subjective estimates of memory uncertainty are jointly represented in the neural population activity that encodes the memory itself. To test this hypothesis, we correlated decoded uncertainty (SD of the decoded posterior probability distribution) with behaviorally reported memory uncertainty (arc length, with the effect of target angle regressed out) at a single-trial level for each participant and each ROI. Decoded uncertainty correlated with the reported arc length significantly in V3AB and IPS1 (bootstrapping test, $p < 0.05$; Figure 6B). Additionally, we binned each participant's trials on the basis of decoded uncertainty and pooled the data across participants. We found that participants reported larger arc length in trials with higher decoded uncertainty in V2, V3, V3AB, IPS0, IPS1, and IPS2 (permutation test, $p < 0.05$; Figure 6C). These results support the notion that the uncertainty of VWM can be represented along with the memorized location by a probabilistic population code, and the uncertainty encoded in the neural population is used for explicit uncertainty reports.

In perceptual decision making, people use their knowledge of their own reaction times when making uncertainty judgments (Kiani et al., 2014). Thereby, saccade reaction time might implicitly track VWM uncertainty in both experiments. Behaviorally, reported arc length increased with saccade reaction time, indicating an impact of reaction time on uncertainty judgement (Figure S4A). In terms of fMRI BOLD activity, saccade reaction time correlated with decoded uncertainty in V3AB and IPS0, when

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

CellPress

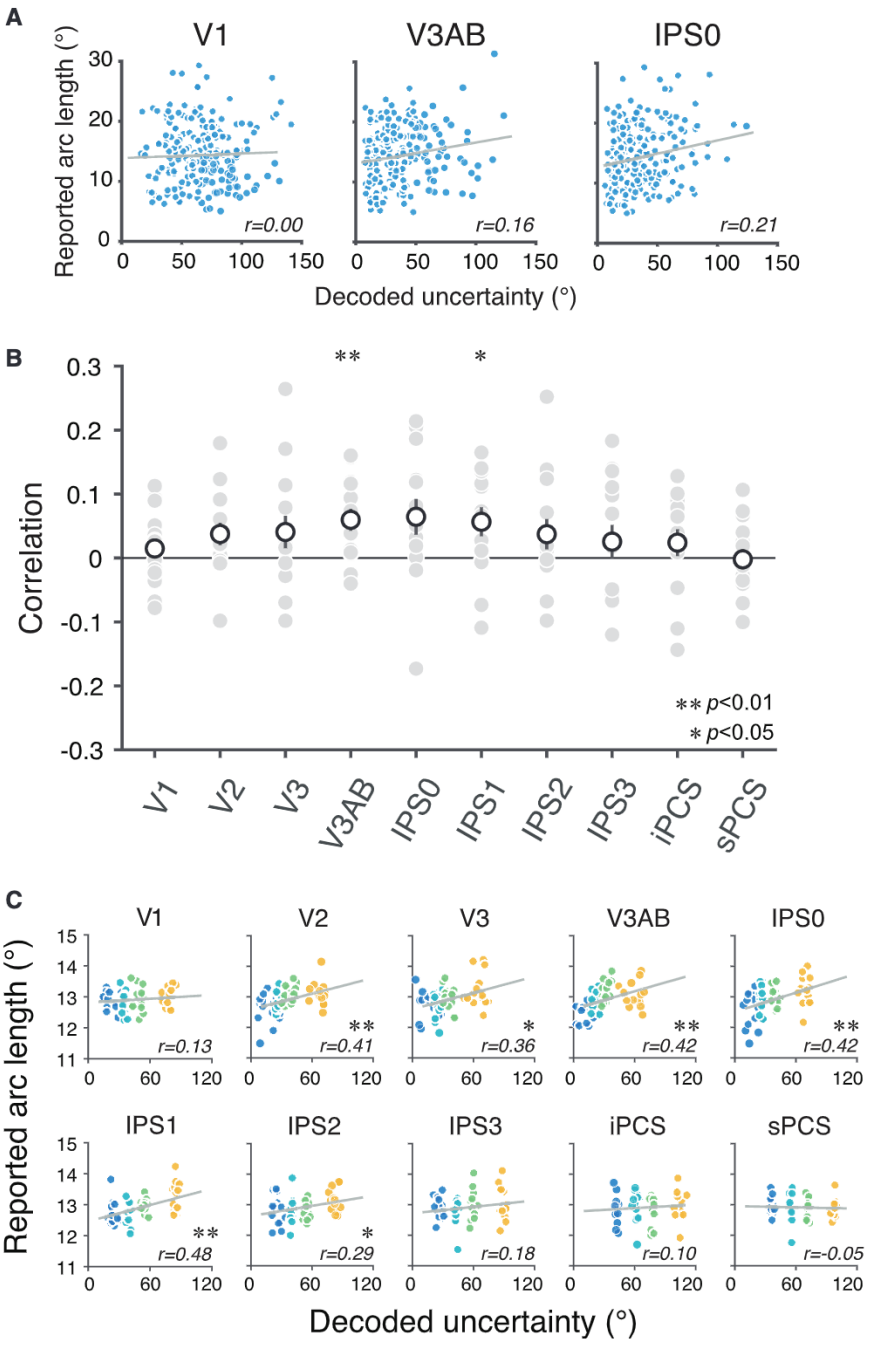

Figure 6. Decoded memory uncertainty predicts subjective memory uncertainty

(A) Reported arc length (subjective VWM uncertainty report) plotted against decoded uncertainty of an example participant for three ROIs.

(B) Correlations between reported arc length and decoded uncertainty. The filled gray dots represent individual participants. The empty white dots represent the group average. The error bars represent ±SEM.

(C) Reported arc length plotted against decoded uncertainty. The four colors indicate four bins (within each of 14 participants) with increasing decoded uncertainty. The gray line in each panel represents the best linear fit. The value at the lower right of each panel is the Pearson correlation coefficient.

binning trials on the basis of decoded uncertainty in experiment 2 (Figures S4B–S4E).

Regarding the decoded memorized location, the results of experiment 2 replicate those of experiment 1. VWM contents were decodable in all the ROIs (permutation test, $p < 0.001$ for all ROIs). The precision of the neural decoding error distribution varied across ROIs (permutation one-way repeated-measures ANOVA, $p < 0.001$, $\eta_p^2 = 0.81$; Figures S1C and S1D; also see Tables S4–S6), with the highest precision observed in V3AB and the lowest precision in iPCS and sPCS (Figure S1D). To further evaluate the behavioral relevancy of the decodable information, we correlated (signed) memory error with (signed) decoding error. The main effect of ROI on this correlation was significant (permuted one-way repeated-measures ANOVA, $p < 0.01$, $\eta_p^2 = 0.18$). Memory error correlated with the neural error in all the ROIs, except iPCS (bootstrapping test, $p < 0.05$; Figure 7B). We obtained similar results when binning each participant's trials on the basis of decoding error and pooled the data across participants (Figure 7C).

For the analyses presented thus far, we selected the 750 voxels from each ROI and participant that showed the strongest location selectivity (see STAR Methods), thereby equalizing the number of voxels across ROIs while maximizing the number of voxels included in analyses (Figures S5A and S5B). To ensure that our findings did not depend on the specific number of voxels, we selected different numbers of voxels ranging from 32 to 1,250 per ROI, or without any voxel selection (using all the voxels in each ROI and participant) and conducted the same analyses. We found that our results (decoding performance in Figure S6; relationships between the outputs of the decoder and behaviors in Figures S5C–S5E) were robust with respect to the number of voxels selected.

Although our model made direct predictions that the decoded location and decoded uncertainty are reflected in memory reports and uncertainty judgments, respectively, there could additionally exist a relationship between the decoded uncertainty and the variability of memory reports (van Bergen et al., 2015). Within the context of spatial VWM, decoded uncertainty did not correlate with the magnitude of memory error or with the variability of

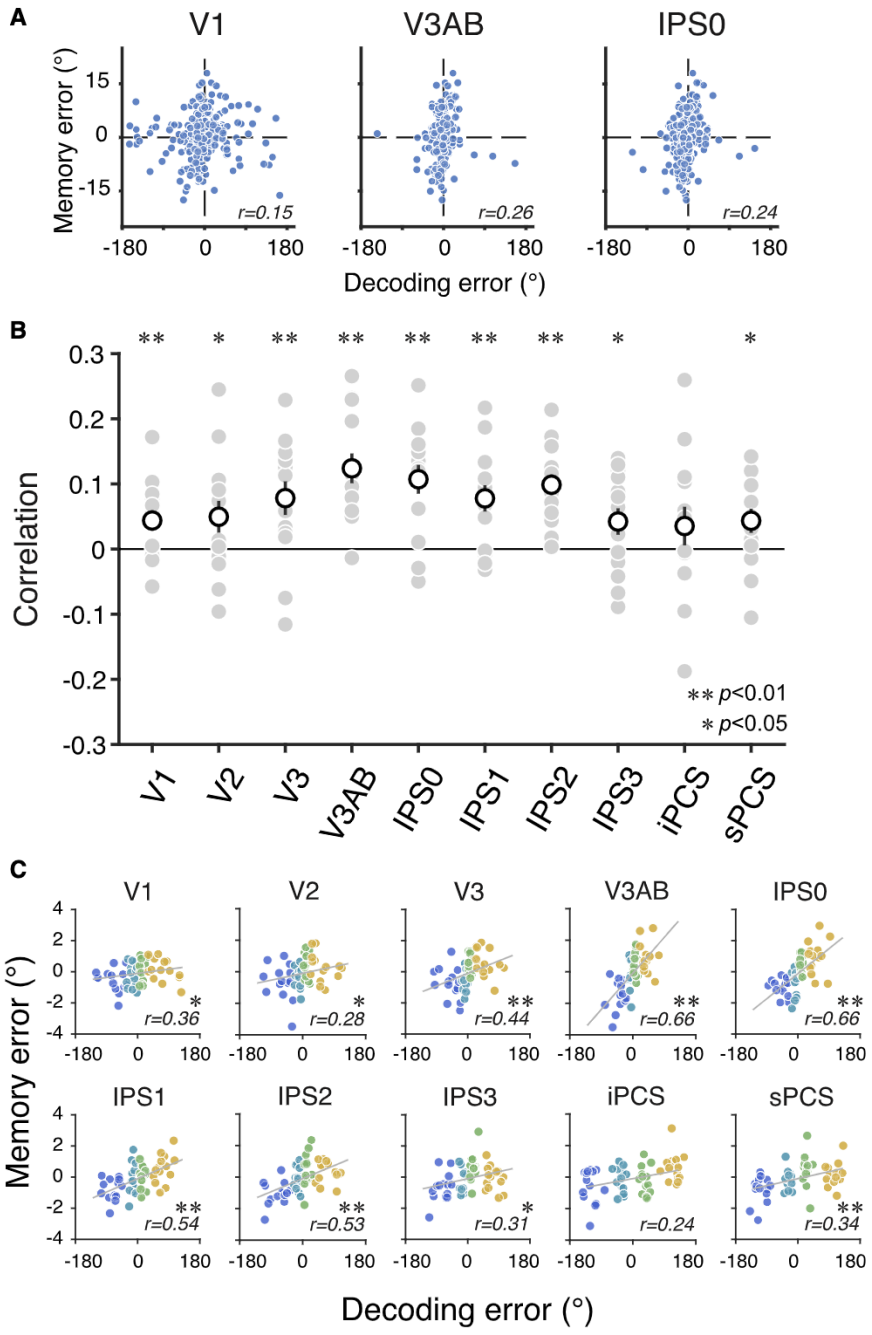

Figure 7. Errors in neural decoding of working memory predict behavioral memory errors in experiment 2

(A) Behavioral memory error plotted against neural decoding error of an example participant.

(B) Correlations are computed as circular correlations between the memory error and neural decoding error. The filled gray dots represent individual participants. The empty white dots represent the group average. The error bars represent ±1 SEM.

(C) Memory error plotted against decoding error. The four colors indicate four bins (within each of 14 participants) sorted by the decoding error. The gray line in each panel represents the best linear fit. The value at the lower right of each panel is the Pearson correlation coefficient. Overall, these results replicate those reported in experiment 1 (Figure 4).

DISCUSSION

Although it is well established that the contents of WM can be decoded from human brain activity, it remains unknown whether and how memory uncertainty is represented in the brain. Here, inspired by the theory of probabilistic population codes (Foldiak, 1993; Jazayeri and Movshon, 2006; Ma et al., 2006; Sanger, 1996), we tested the hypothesis that the human brain encodes VWM as a probability distribution over the remembered feature space. In two independent experiments using a generative model of neural activity combined with a multivariate Bayesian decoder, we tested two central predictions that stem from this hypothesis. First, after validating that our procedures could decode the precise contents of VWM during a memory retention interval, we discovered that errors in our neural decoder predicted the direction and amplitude of memory errors made later in the trial. Second, we discovered that the uncertainty in our neural decoder predicted the memory uncertainty explicitly reported by our participants. Together, these results provide strong evidence that the content of our WM is a readout of a noisy probability distribution encoded in the population activity of neurons whose distribution width conveys information about memory uncertainty.

The theory of probabilistic population codes was originally proposed to explain how the brain can jointly represent the estimate and the uncertainty of sensory stimuli used during perception. Per this theory, the brain possesses knowledge about the generative process of neural activity (Beck et al., 2008; Foldiak, 1993; Jazayeri and Movshon, 2006; Ma et al., 2006; Sanger, 1996; Zemel et al., 1998). This knowledge and

memory reports (Figure S7). To investigate whether such relationships between decoded uncertainty and memory errors exist at a cross-subject level, for each participant, we averaged the decoded uncertainty across trials. Widespread across multiple ROIs in visual cortex and IPS, we found that participants with larger averaged decoded uncertainty performed worse in the behavioral memory reports when quantified as their averaged magnitude of behavioral memory error (Figure 8) or as the SD of their behavioral error distribution (Figure S8). These results demonstrate a linkage between the precision of VWM neural representation in these brain regions and the precision of behavioral memory reports.

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

CellPress

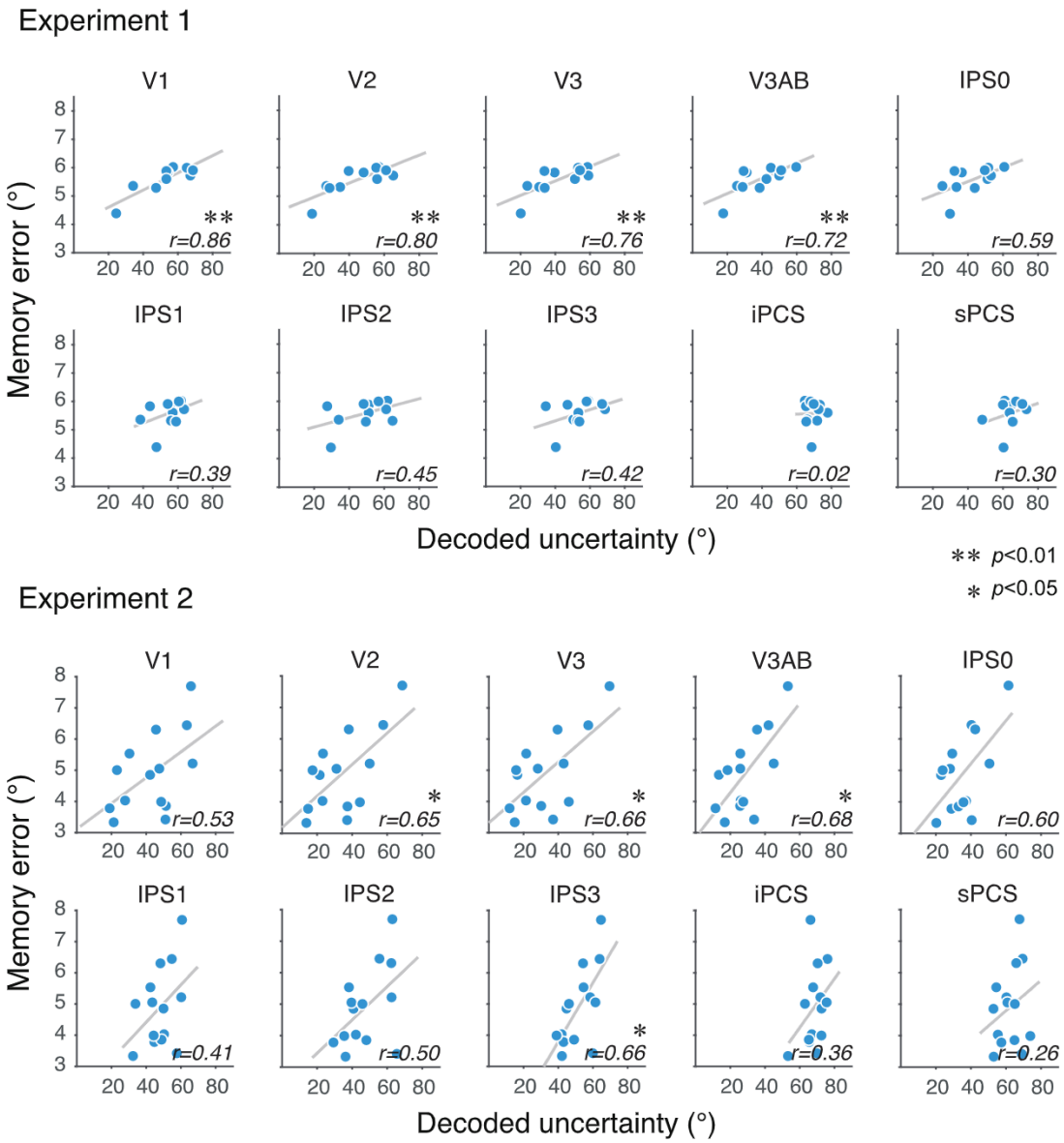

Figure 8. Participants with overall greater decoded uncertainty have less precise working memory

(A) Experiment 1. Each dot represents one participant. The decoded uncertainty (x axis) and the absolute value of memory error (y axis) was averaged across trials for each participant. The gray lines represent the best linear fit.

(B) Experiment 2.

the variability in the response of cortical neurons naturally allow the neural population to represent the probability distribution over the perceptual stimulus space for any pattern of neural activity. Neurophysiological experiments measuring population-level neural responses have found evidence supporting the predictions of probabilistic population codes for visual perception in non-human primates (Berens et al., 2012; Graf et al., 2011; Walker et al., 2020). Here, we demonstrate that probabilistic population codes are not limited to visual perception but are also used to represent information actively maintained in WM in support of higher cognition. Our computational neuroimaging approach—using the knowledge of a generative model and applying Bayesian decoding on the observed fMRI activation—mimics how the "decision maker" in the brain performs inference on the basis of its knowledge of its own generative model and the observed neural activity during the VWM delay period.

Under the rubric of probabilistic population codes, the fluctuations of both the content and the uncertainty of VWM arise from the noise in the neural population response that encodes the memorized stimulus. Thus, it is critical that we can decode VWM content and its uncertainty on a trial-by-trial basis in order to study their relationships with behavior. Previous neuroimaging studies have mostly reported VWM decoding accuracy (or

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

fidelity) per condition, averaged across all trials in the experiment (e.g., Albers et al., 2013; Christophel et al., 2012, 2018; Emrich et al., 2013; Ester et al., 2013, 2015; Harrison and Tong, 2009; Jerde et al., 2012; Lee et al., 2013; Lorenc et al., 2018; Rahmati et al., 2018; Riggall and Postle, 2012; Serences et al., 2009; Sprague et al., 2014, 2016; Xing et al., 2013; Yu and Shim, 2017). These indices represent decoding quality aggregated across many trials, and thus they are inadequate to explain or to estimate how VWM content and its uncertainty fluctuate across individual trials.

Unlike previous studies that used simpler linear encoding models (Brouwer and Heeger, 2009) to decode the content of spatial VWM (Hallenbeck et al., 2021; Jerde et al., 2012; Sprague et al., 2014, 2016), we used a generative model that improves the precision of decoding by estimating sources of measurement and neural noise (van Bergen and Jehee, 2021). In both experiments, we found remarkably precise representations of the memorized target locations in a widely distributed network of brain regions, including visual, parietal, and frontal cortex. Encouraged by the robustness of the decoding, we asked whether these population responses in fact encoded the memory, including small spatial errors in memory, rather than the veridical target locations. On a trial-by-trial basis, we found that errors in our neural decoder predicted the direction and amplitude of memory errors (Figure 4). These results indicate that the neural representations we decoded from the late delay period preceding the participants' memory-guided responses, contain information that affects behaviors beyond that present in the physical stimulus. Specifically, it strongly suggests that one's memory depends on the readout of these population-encoded representations. In neurophysiological studies, this type of correlation (between neural noise and behavioral choices) has sometimes been used to infer a causal link between neural responses and behaviors (reviewed in Nienborg et al., 2012) We observed the strongest correlations in V3AB in dorsal extrastriate cortex, followed by neighboring regions V3 and IPS0. These results are generally consistent with neurophysiological studies of perceptual decisions reporting that choice-related activity is weak at best in early sensory cortex, and stronger in higher tier sensory cortices (Britten et al., 1996; Camarillo et al., 2012; Dodd et al., 2001; Goris et al., 2017; Nienborg and Cumming, 2009; Nienborg et al., 2012; Yu et al., 2015).

Next, we leveraged the signal-trial decoding to investigate the neural basis of VWM uncertainty. As VWM uncertainty is defined as the width of one's belief distribution over possible stimulus values (memorized location) or the subjective sense of the quality of one's own memory, uncertainty can fluctuate across trials even when remembering the exact same stimulus. In line with previous behavioral studies (Fougnie et al., 2012; Honig et al., 2020; Rademaker et al., 2012; Samaha and Postle, 2017; Yoo et al., 2018), we found that uncertainty judgements tracked the quality of VWM on a trial-by-trial basis (Figures 5D and 5E). Importantly, this demonstrates that participants were aware of the quality of their memory and adjusted their uncertainty reports in step with their memory fidelity. We also found that the measured population-encoded responses, when analyzed with our generative model, stored memory uncertainty. On a trial-by-trial basis, memory uncertainty decoded from the retention interval in V3AB and IPS1 predicted the uncertainty explicitly reported and used by the participants in the wagers made later in the trial (Figure 6). Recall that we also observed strong correlations between decoding error and memory error in V3AB. Theoretically, an estimate of an item and the uncertainty of the estimate can be jointly encoded as a single probability distribution by the same population of neurons. Our findings suggest that such an efficient mechanism exists in V3AB to support VWM.

Our theory-guided approach of decoding uncertainty from the width of a modeled probability distribution is a departure from previous fMRI studies investigating the neural correlates of uncertainty or confidence in perception (Bang and Fleming, 2018) and decision making (De Martino et al., 2013; Lebreton et al., 2015). These previous studies used linear regression to identify brain regions whose activity increased (or decreased) with uncertainty report or confidence rating, thereby identifying the brain regions that represent uncertainty by a "rate code" (i.e., increasing or decreasing averaged response amplitude with uncertainty or confidence). Perhaps such regions act as a downstream decoder and extract the uncertainty information represented by the neural populations that encode the stimulus features, in a way similar to how we decode uncertainty from voxel activity. How the regions with different coding schemes for uncertainty—the probabilistic population code reported here and the rate code described in these previous studies—interact is still an open question.

The decodable VWM signals across nearly all ROIs in both experiments are generally consistent with the notion that the storage of VWM content involves a widespread cortical network (Christophel et al., 2017; Ester et al., 2015; Hallenbeck et al., 2021; Lee and Baker, 2016). For spatial VWM, the decodable signals may hinge on the retinotopic organization of visual, parietal, and frontal cortex. During the WM delay, participants most likely attended to the target location covertly (Awh and Jonides, 2001; Awh et al., 1999; Jerde et al., 2012), increasing the response of the voxels selective for the memorized and attended location (Gandhi et al., 1999; Itthipuripat et al., 2019; Kastner et al., 1999). This is consistent with the findings that the locus of spatial attention can be decoded by using voxels' preferred locations (Brefczynski-Lewis et al., 2009; Datta and DeYoe, 2009). It remains an open question whether, how, and in what contexts the neural representations of memorized locations and attended locations are distinguishable.

The quality of VWM representations varies greatly across different brain regions. In dorsal extrastriate cortex, V3AB and its neighboring regions IPS0 and V3 showed the highest performance in decoding the memorized target locations. The SDs of decoding error distributions were about one-third of that of the region with the lowest decoding performance. Moreover, decoding error and decoded uncertainty from V3AB and IPS0 exhibited the strongest correlations with behavioral memory error and uncertainty judgement respectively. Thus, these regions could be most critical for maintaining the content of spatial VWM. These results converge with recent studies on mental imagery and episodic memory: Breedlove et al. (2020) built an encoding model to predict the brain activity corresponding to different "imagined" images. They found higher prediction accuracy in

higher level visual areas (V3AB and IPS) than in early visual cortex. Similarly, Favila et al. (2020) found that during retrieval of spatial positions from episodic memory, the spatially localized memory-evoked responses in extrastriate regions V3AB and hV4 were more precise than those observed in early visual cortex. Together, their and our results highlight the importance of the dorsal mid- and high-level visual cortex in maintaining behaviorally relevant information in the absence of bottom-up inputs. We cannot exclude the possibility that the type of task and stimulus affects which regions exhibit higher decoding performance or higher correlations with behaviors. The high decoding performance in dorsal extrastriate and posterior parietal cortex might be specific to tasks involving maintenance of spatial information. For other tasks requiring maintenance of different feature values, other regions may be more critically involved in supporting behavior (Christophel et al., 2017; Lee and Baker, 2016; Lee et al., 2013). Indeed, some previous studies investigating WM for orientation have found decoding accuracy in V1 to be as high as mid- or high-level visual cortices (Harrison and Tong, 2009; Pratte and Tong, 2014), though orientation representations have also be found in frontal and parietal regions (Ester et al., 2015; Yu and Shim, 2017). To understand the generalization of our results, future studies should directly compare which regions carry WM representations for different types of stimuli under the same procedures and with the same decoding algorithms.

Despite the centrality of the prefrontal cortex to WM theory (Curtis and Sprague, 2021; Sreenivasan et al., 2014), the inferior and superior branches of the precentral sulcus (iPCS and sPCS) had the lowest decoding performance, and only in sPCS did decoding error correlate with memory error. In addition, across participants the decoding quality from frontal cortex did not predict how well a participant performed in the VWM tasks (Figure 8; Figure S8). We choose sPCS and iPCS as ROIs because in the frontal cortex, they exhibit the clearest retinotopic organization (Mackey et al., 2017) and the strongest decodable spatial VWM signals in previous fMRI studies (Jerde et al., 2012; Sprague et al., 2014, 2016). The sPCS is believed to be the human homolog of monkey FEF (Blanke et al., 1999; Curtis and Connolly, 2008), a macaque region known to be critically involved in spatial VWM (Armstrong et al., 2009; Bruce and Goldberg, 1985; Sommer and Wurtz, 2001), covert attention orienting (Moore and Armstrong, 2003; Moore and Fallah, 2001), and saccadic eye movement (Bruce et al., 1985; Tehovnik et al., 2000). Perhaps surprisingly, our results indicate that compared with dorsal high-level visual cortex and IPS0, sPCS contained a quite coarse representation of memorized locations.

In a previous study, van Bergen et al. (2015) developed and applied the Bayesian decoding method used here to quantify sensory uncertainty from early visual cortex activation patterns (voxels pooled across V1, V2, and V3) evoked by visual stimuli. They found that the uncertainty decoded from early visual cortex correlated with participants' behavioral variability and errors in an orientation estimation task. Here, we did not observe a correlation between the decoded uncertainty and the variability (or error) of memory reports within individuals (when the individual participant means were removed) (Figure S7). This discrepancy might reflect differences in how locations and orientations are encoded. Perhaps the correlation between decoded uncertainty and behavioral variability reported by van Bergen et al. (2015) indicates that their observers did not solely use the posterior mean when reporting orientation. For example, the use of a "posterior probability matching" strategy (reporting an orientation by drawing a sample from the posterior distribution; e.g., Wozny et al., 2010) would increase the correlation between uncertainty and behavioral variability. Uncertainty and error (or bias) would correlate if an observer weighted the prior information more when the uncertainty was high. It is well documented that in orientation estimation, observers use a prior reflecting the statistics of the orientations in the natural environment (e.g., more cardinal than oblique orientations; Girshick et al., 2011; Wei and Stocker, 2015), which is different from the statistics of the orientations used in van Bergen et al. (2015) (e.g., uniform distribution). In the case of spatial VWM used here, it is unlikely that observers used a prior for encoding locations and instead assumed that objects appeared uniformly at all possible locations (polar angles).

Contrary to our results, van Bergen et al. (2015) did not observe a significant correlation between signed decoding error and signed behavioral error during a perceptual task involving orientation estimation. They only analyzed ROIs V1, V2, and V3 in early visual cortex. However, we found these early regions have weaker correlations than higher level regions V3AB and IPS0, and sometimes the correlations were non-significant (V1 in experiment 1, Figures 4B and 4C; V2 in experiment 1, Figure 4C). In addition, van Bergen et al. (2015) used an earlier version of the Bayesian decoder that has a lower decoding performance when directly compared with TAFKAP (van Bergen and Jehee, 2021). Both the choice of the ROI and the version of the decoder could contribute to their null results, as well as differences in stimuli (orientation versus space) and task (perception versus memory).

In a number of cortical areas, we observed strong correlations between participants' average decoded uncertainty (across all trials) and their average memory error (across all trials; Figure 8). Participants who on average represented remembered locations more precisely in their neural activation patterns (i.e., with lower decoded uncertainty) were those whose WM was more precise. This result was consistent across both experiments, and the result stands when we used an alternative index, the SD of the distribution of memory errors, to quantify memory precision (Figure S8). These findings support previous studies that identified cross-subject correlations between average decoding performance and average behavioral performance (Albers et al., 2013; Christophel et al., 2018; Ester et al., 2013). Overall, the strong cross-subject correlations we observed demonstrated that our model-based decoding approach not only provided unprecedented accuracy of decoding single-trial spatial VWM content but also extracted features of individuals' neural circuitry that constrained individual WM performance. WM abilities predict a number of cognitive and intellectual functions, suggesting that it might be a core component upon which many high-level cognitive abilities depend (Daneman and Carpenter, 1980; Süß et al., 2002). Although the neural sources of these individual differences in WM remain elusive, our results suggest that the noise in the population encoding may be an important neural source of the individual differences in VWM quality.

Overall, across two computational neuroimaging experiments, we demonstrated that humans encode WM representations as a probability distribution maintained via the activity patterns in posterior parietal and extrastriate visual cortex. These results extend previous studies identifying probabilistic sensory representations during perceptual processing and establish that probabilistic population codes are an efficient and general neural coding principle used to support higher cognitive behaviors such as WM.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Procedures
  - Setup and stimuli
  - Eyetracking
  - Behavioral data analysis
  - Retinotopic mapping and the identification of region of interest (ROI)
  - MRI acquisition
  - MRI data preprocessing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Generative model
  - Model fitting and decoding
  - Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2021.08.022.

ACKNOWLEDGMENTS

We thank New York University's Center for Brain Imaging for technical support. This research was supported by the National Eye Institute (NEI) (R01 EY-016407 to C.E.C., R01 EY-027925 to C.E.C. and W.J.M., and F32 EY-028438), a Sloan Research Fellowship to T.C.S., the NEI Visual Neuroscience Training Program (T32 EY-007136 to T.C.S.), and Nvidia Hardware Grants to C.E.C. and T.C.S.

AUTHOR CONTRIBUTIONS

H.-H.L., T.C.S., A.H.Y., W.J.M., and C.E.C. designed the experiments. H.-H.L. and T.C.S. conducted the experiments and analyzed the data. H.-H.L., T.C.S., A.H.Y., W.J.M., and C.E.C. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 14, 2021
Revised: July 9, 2021
Accepted: August 17, 2021
Published: September 14, 2021

REFERENCES

Albers, A.M., Kok, P., Toni, I., Dijkerman, H.C., and de Lange, F.P. (2013). Shared representations for working memory and mental imagery in early visual cortex. Curr. Biol. *23*, 1427–1431.

Armstrong, K.M., Chang, M.H., and Moore, T. (2009). Selection and maintenance of spatial information by frontal eye field neurons. J. Neurosci. *29*, 15621–15629.

Awh, E., and Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. Trends Cogn. Sci. *5*, 119–126.

Awh, E., Jonides, J., Smith, E.E., Buxton, R.B., Frank, L.R., Love, T., Wong, E.C., and Gmeindl, L. (1999). Rehearsal in spatial working memory: evidence from neuroimaging. Psychol. Sci. *10*, 433–437.

Bang, D., and Fleming, S.M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. Proc. Natl. Acad. Sci. U S A *115*, 6082–6087.

Beck, J.M., Ma, W.J., Kiani, R., Hanks, T., Churchland, A.K., Roitman, J., Shadlen, M.N., Latham, P.E., and Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. Neuron *60*, 1142–1152.

Berens, P. (2009). CircStat: a MATLAB toolbox for circular statistics. J. Stat. Softw. *31*, 1–21.

Berens, P., Ecker, A.S., Cotton, R.J., Ma, W.J., Bethge, M., and Tolias, A.S. (2012). A fast and simple population code for orientation in primate V1. J. Neurosci. *32*, 10618–10626.

Blanke, O., Morand, S., Thut, G., Michel, C.M., Spinelli, L., Landis, T., and Seeck, M. (1999). Visual activity in the human frontal eye field. Neuroreport *10*, 925–930.

Breedlove, J.L., St-Yves, G., Olman, C.A., and Naselaris, T. (2020). Generative feedback explains distinct brain activity codes for seen and mental images. Curr. Biol. *30*, 2211–2224.e6.

Brefczynski-Lewis, J.A., Datta, R., Lewis, J.W., and DeYoe, E.A. (2009). The topography of visuospatial attention as revealed by a novel visual field mapping technique. J. Cogn. Neurosci. *21*, 1447–1460.

Brissenden, J.A., Tobyne, S.M., Halko, M.A., and Somers, D.C. (2021). Stimulus-specific visual working memory representations in human cerebellar lobule VIIb/VIIIa. J. Neurosci. *41*, 1033–1045.

Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S., and Movshon, J.A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis. Neurosci. *13*, 87–100.

Brouwer, G.J., and Heeger, D.J. (2009). Decoding and reconstructing color from responses in human visual cortex. J. Neurosci. *29*, 13992–14003.

Bruce, C.J., and Goldberg, M.E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. J. Neurophysiol. *53*, 603–635.

Bruce, C.J., Goldberg, M.E., Bushnell, M.C., and Stanton, G.B. (1985). Primate frontal eye fields. II. Physiological and anatomical correlates of electrically evoked eye movements. J. Neurophysiol. *54*, 714–734.

Camarillo, L., Luna, R., Nácher, V., and Romo, R. (2012). Coding perceptual discrimination in the somatosensory thalamus. Proc. Natl. Acad. Sci. U S A *109*, 21093–21098.

Christophel, T.B., Hebart, M.N., and Haynes, J.-D. (2012). Decoding the contents of visual short-term memory from human visual and parietal cortex. J. Neurosci. *32*, 12983–12989.

Christophel, T.B., Klink, P.C., Spitzer, B., Roelfsema, P.R., and Haynes, J.-D. (2017). The distributed nature of working memory. Trends Cogn. Sci. *21*, 111–124.

Christophel, T.B., Iamshchinina, P., Yan, C., Allefeld, C., and Haynes, J.-D. (2018). Cortical specialization for attended versus unattended working memory. Nat. Neurosci. *21*, 494–496.

Collins, A.G.E., and Frank, M.J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur. J. Neurosci. *35*, 1024–1035.

Curtis, C.E., and Connolly, J.D. (2008). Saccade preparation signals in the human frontal and parietal cortices. J. Neurophysiol. *99*, 133–145.

Curtis, C.E., and Lee, D. (2010). Beyond working memory: the role of persistent activity in decision making. Trends Cogn. Sci. *14*, 216–222.

Curtis, C.E., and Sprague, T.C. (2021). Persistent activity during working memory from front to back. Front. Neural Circuits *15*, 696060.

Daneman, M., and Carpenter, P.A. (1980). Individual differences in working memory and reading. J. Verbal Learn. Verbal Behav. *19*, 450–466.

Datta, R., and DeYoe, E.A. (2009). I know where you are secretly attending! The topography of human visual attention revealed with fMRI. Vision Res. *49*, 1037–1044.

De Martino, B., Fleming, S.M., Garrett, N., and Dolan, R.J. (2013). Confidence in value-based choice. Nat. Neurosci. *16*, 105–110.

Devkar, D., Wright, A.A., and Ma, W.J. (2017). Monkeys and humans take local uncertainty into account when localizing a change. J. Vis. *17*, 4.

Dodd, J.V., Krug, K., Cumming, B.G., and Parker, A.J. (2001). Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. J. Neurosci. *21*, 4809–4821.

Dumoulin, S.O., and Wandell, B.A. (2008). Population receptive field estimates in human visual cortex. Neuroimage *39*, 647–660.

Emrich, S.M., Riggall, A.C., Larocque, J.J., and Postle, B.R. (2013). Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. J. Neurosci. *33*, 6516–6523.

Ester, E.F., Anderson, D.E., Serences, J.T., and Awh, E. (2013). A neural measure of precision in visual working memory. J. Cogn. Neurosci. *25*, 754–761.

Ester, E.F., Sprague, T.C., and Serences, J.T. (2015). Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. Neuron *87*, 893–905.

Faisal, A.A., Selen, L.P.J., and Wolpert, D.M. (2008). Noise in the nervous system. Nat. Rev. Neurosci. *9*, 292–303.

Favila, S.E., Kuhl, B.A., and Winawer, J. (2020). Perception and memory have distinct spatial tuning properties in human visual cortex. bioRxiv. https://doi.org/10.1101/811331.

Foldiak, P. (1993). Computation and Neural Systems, F.H. Eeckman and J. Bower, eds. (New York: Springer Science & Business Media).

Fougnie, D., Suchow, J.W., and Alvarez, G.A. (2012). Variability in the quality of visual working memory. Nat. Commun. *3*, 1229.

Funahashi, S., Bruce, C.J., and Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J. Neurophysiol. *61*, 331–349.

Gandhi, S.P., Heeger, D.J., and Boynton, G.M. (1999). Spatial attention affects brain activity in human primary visual cortex. Proc. Natl. Acad. Sci. U S A *96*, 3314–3319.

Girshick, A.R., Landy, M.S., and Simoncelli, E.P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nat. Neurosci. *14*, 926–932.

Goris, R.L.T., Ziemba, C.M., Stine, G.M., Simoncelli, E.P., and Movshon, J.A. (2017). Dissociation of choice formation and choice-correlated activity in macaque visual cortex. J. Neurosci. *37*, 5195–5203.

Graf, A.B.A., Kohn, A., Jazayeri, M., and Movshon, J.A. (2011). Decoding the activity of neuronal populations in macaque primary visual cortex. Nat. Neurosci. *14*, 239–245.

Hallenbeck, G.E., Sprague, T.C., Rahmati, M., Sreenivasan, K.K., and Curtis, C.E. (2021). Working memory representations in visual cortex mediate distraction effects. Nat. Commun. *12*, 4714.

Harrison, S.A., and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature *458*, 632–635.

Hikosaka, O., and Wurtz, R.H. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. III. Memory-contingent visual and saccade responses. J. Neurophysiol. *49*, 1268–1284.

Honig, M., Ma, W.J., and Fougnie, D. (2020). Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. Proc. Natl. Acad. Sci. U S A *117*, 8391–8397.

Itthipuripat, S., Sprague, T.C., and Serences, J.T. (2019). Functional MRI and EEG index complementary attentional modulations. J. Neurosci. *39*, 6162–6179.

Jazayeri, M., and Movshon, J.A. (2006). Optimal representation of sensory information by neural populations. Nat. Neurosci. *9*, 690–696.

Jerde, T.A., Merriam, E.P., Riggall, A.C., Hedges, J.H., and Curtis, C.E. (2012). Prioritized maps of space in human frontoparietal cortex. J. Neurosci. *32*, 17382–17390.

Kastner, S., Pinsk, M.A., De Weerd, P., Desimone, R., and Ungerleider, L.G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron *22*, 751–761.

Kay, K.N., Winawer, J., Mezer, A., and Wandell, B.A. (2013). Compressive spatial summation in human visual cortex. J. Neurophysiol. *110*, 481–494.

Keshvari, S., van den Berg, R., and Ma, W.J. (2012). Probabilistic computation in human perception under variability in encoding precision. PLoS ONE *7*, e40216.

Kiani, R., Corthell, L., and Shadlen, M.N. (2014). Choice certainty is informed by both evidence and decision time. Neuron *84*, 1329–1342.

Lebreton, M., Abitbol, R., Daunizeau, J., and Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. Nat. Neurosci. *18*, 1159–1167.

Ledoit, O., and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. *88*, 365–411.

Lee, S.-H., and Baker, C.I. (2016). Multi-voxel decoding and the topography of maintained information during visual working memory. Front. Syst. Neurosci. *10*, 2.

Lee, S.-H., Kravitz, D.J., and Baker, C.I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. Nat. Neurosci. *16*, 997–999.

Lorenc, E.S., Sreenivasan, K.K., Nee, D.E., Vandenbroucke, A.R.E., and D'Esposito, M. (2018). Flexible coding of visual working memory representations during distraction. J. Neurosci. *38*, 5267–5276.

Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat. Neurosci. *9*, 1432–1438.

Ma, W.J., Husain, M., and Bays, P.M. (2014). Changing concepts of working memory. Nat. Neurosci. *17*, 347–356.

Mackey, W.E., and Curtis, C.E. (2017). Distinct contributions by frontal and parietal cortices support working memory. Sci. Rep. *7*, 6188.

Mackey, W.E., Winawer, J., and Curtis, C.E. (2017). Visual field map clusters in human frontoparietal cortex. eLife *6*, 6.

Moore, T., and Armstrong, K.M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. Nature *421*, 370–373.

Moore, T., and Fallah, M. (2001). Control of eye movements and spatial attention. Proc. Natl. Acad. Sci. U S A *98*, 1273–1276.

Nienborg, H., and Cumming, B.G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature *459*, 89–92.

Nienborg, H., Cohen, M.R., and Cumming, B.G. (2012). Decision-related activity in sensory neurons: correlations among neurons and with behavior. Annu. Rev. Neurosci. *35*, 463–483.

Pratte, M.S., and Tong, F. (2014). Spatial specificity of working memory representations in the early visual cortex. J. Vis. *14*, 22.

Rademaker, R.L., Tredway, C.H., and Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. J. Vis. *12*, 21.

Rademaker, R.L., Chunharas, C., and Serences, J.T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. Nat. Neurosci. *22*, 1336–1344.

Rahmati, M., Saber, G.T., and Curtis, C.E. (2018). Population dynamics of early visual cortex during working memory. J. Cogn. Neurosci. *30*, 219–233.

Rahmati, M., DeSimone, K., Curtis, C.E., and Sreenivasan, K.K. (2020). Spatially specific working memory activity in the human superior colliculus. J. Neurosci. *40*, 9487–9495.

Riggall, A.C., and Postle, B.R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. J. Neurosci. *32*, 12990–12998.

Samaha, J., and Postle, B.R. (2017). Correlated individual differences suggest a common mechanism underlying metacognition in visual perception and visual short-term memory. Proc. Biol. Sci. *284*, 20172035.

Sanger, T.D. (1996). Probability density estimation for the interpretation of neural population codes. J. Neurophysiol. *76*, 2790–2793.

Serences, J.T., Ester, E.F., Vogel, E.K., and Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. Psychol. Sci. *20*, 207–214.

Sommer, M.A., and Wurtz, R.H. (2001). Frontal eye field sends delay activity related to movement, memory, and vision to the superior colliculus. J. Neurophysiol. *85*, 1673–1685.

Sprague, T.C., Ester, E.F., and Serences, J.T. (2014). Reconstructions of information in visual spatial working memory degrade with memory load. Curr. Biol. *24*, 2174–2180.

Sprague, T.C., Ester, E.F., and Serences, J.T. (2016). Restoring latent visual working memory representations in human cortex. Neuron *91*, 694–707.

Sprague, T.C., Adam, K.C.S., Foster, J.J., Rahmati, M., Sutterer, D.W., and Vo, V.A. (2018). Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. eNeuro *5*, ENEURO.0098-18.2018.

Sreenivasan, K.K., Curtis, C.E., and D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. Trends Cogn. Sci. *18*, 82–89.

Süß, H.-M., Oberauer, K., Wittmann, W.W., Wilhelm, O., and Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. Intelligence *30*, 261–288.

Tehovnik, E.J., Sommer, M.A., Chou, I.H., Slocum, W.M., and Schiller, P.H. (2000). Eye fields in the frontal lobes of primates. Brain Res. Brain Res. Rev. *32*, 413–448.

Tolhurst, D.J., Movshon, J.A., and Dean, A.F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. Vision Res. *23*, 775–785.

Tomko, G.J., and Crapper, D.R. (1974). Neuronal variability: non-stationary responses to identical visual stimuli. Brain Res. *79*, 405–418.

van Bergen, R.S., and Jehee, J.F.M. (2019). Probabilistic representation in human visual cortex reflects uncertainty in serial decisions. J. Neurosci. *39*, 8164–8176.

van Bergen, R.S., and Jehee, J.F.M. (2021). TAFKAP: an improved method for probabilistic decoding of cortical activity. bioRxiv. https://doi.org/10.1101/2021.03.04.433946.

van Bergen, R.S., Ma, W.J., Pratte, M.S., and Jehee, J.F.M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. Nat. Neurosci. *18*, 1728–1730.

van den Berg, R., Yoo, A.H., and Ma, W.J. (2017). Fechner's law in metacognition: a quantitative model of visual working memory confidence. Psychol. Rev. *124*, 197–214.

Wagner, A.D. (1999). Working memory contributions to human learning and remembering. Neuron *22*, 19–22.

Walker, E.Y., Cotton, R.J., Ma, W.J., and Tolias, A.S. (2020). A neural basis of probabilistic computation in visual cortex. Nat. Neurosci. *23*, 122–129.

Wei, X.-X., and Stocker, A.A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. Nat. Neurosci. *18*, 1509–1517.

Wozny, D.R., Beierholm, U.R., and Shams, L. (2010). Probability matching as a computational strategy used in perception. PLoS Comput. Biol. *6*, e1000871.

Xing, Y., Ledgeway, T., McGraw, P.V., and Schluppeck, D. (2013). Decoding working memory of stimulus contrast in early visual cortex. J. Neurosci. *33*, 10301–10311.

Yoo, A.H., Klyszejko, Z., Curtis, C.E., and Ma, W.J. (2018). Strategic allocation of working memory resource. Sci. Rep. *8*, 16162.

Yoo, A.H., Acerbi, L., and Ma, W.J. (2020). Uncertainty is maintained and used in working memory. bioRxiv. https://doi.org/10.1101/2020.10.06.328310.

Yu, Q., and Shim, W.M. (2017). Occipital, parietal, and frontal cortices selectively maintain task-relevant features of multi-feature objects in visual working memory. Neuroimage *157*, 97–107.

Yu, X.-J., Dickman, J.D., DeAngelis, G.C., and Angelaki, D.E. (2015). Neuronal thresholds and choice-related activity of otolith afferent fibers during heading perception. Proc. Natl. Acad. Sci. U S A *112*, 6467–6472.

Zemel, R.S., Dayan, P., and Pouget, A. (1998). Probabilistic interpretation of population codes. Neural Comput. *10*, 403–430.

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

CellPress

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
fMRI data	This paper	https://osf.io/WDJRV
Behavioral data	This paper	https://osf.io/WDJRV
Software and algorithms		
MATLAB	MathWorks	https://www.mathworks.com/products/matlab.html
Custom code and algorithm	This paper	https://osf.io/WDJRV
TAFKAP	van Bergen and Jehee (2021)	https://github.com/jeheelab/TAFKAP

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Clayton Curtis (clayton.curtis@nyu.edu).

Materials availability

The study did not produce new materials.

Data and code availability

The processed fMRI data and raw behavioral data generated in this study have been deposited in the Open Science Framework https://osf.io/WDJRV (https://doi.org/10.17605/OSF.IO/WDJRV). Processed fMRI data contains extracted timeseries from each voxel of each ROI. The raw fMRI data are available under restricted access to ensure participant privacy; access can be obtained by contacting the corresponding authors. The data used to plot figures in this paper (participant means) are provided in the Source Data file.

All code for data analysis has been deposited in the Open Science Framework https://osf.io/WDJRV (https://doi.org/10.17605/OSF.IO/WDJRV) and is publicly available as of the date of publication.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Thirteen participants took part in Experiment 1 (two authors). Data of two participants were excluded because the eye tracking data were too noisy for extracting gaze positions reliably. Nine participants from Experiment 1 and five additional participants joined in Experiment 2. All participants had normal or corrected-to-normal vision. The experiments were conducted with the written, informed consent of each participant. The experimental protocols were approved by the University Committee on Activities involving Human Subjects at New York University, and participants received monetary compensation ($30/hr).

METHOD DETAILS

Procedures

Experiment 1

Participants performed a memory-guided saccade task in the fMRI scanner. Each trial started with the onset of the working memory target (light gray dot) with a duration of 500 ms followed by a delay period of 12000 ms. Participants were required to remember the location of the target and hold their gaze at the fixation point at the screen center until the end of the delay period. After the delay period, the response cue, the fixation point changing from a light gray circle outline into a filled light gray circle, instructing participants to make a saccadic eye movement to the remembered location. 700 ms after the onset of the response cue, a feedback stimulus (a white dot) was presented at the target location for 800 ms. Participants made a saccade to the feedback dot before moving their eyes back to the screen center. The intertrial interval was pseudo-randomly chosen to be 6, 9, or 12 s. Each participant completed 304 to 496 trials (346 trials per participant on average) in 2 to 3 1.5-hr scanning sessions on separate days. Each session consisted of 9 to 10 runs, each with 16 trials whose target had locations evenly spanning the circular space. Participants were allowed to take a break between runs.

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

Experiment 2

The procedures of Experiment 2 were the same as Experiment 1 except the following: In addition to the filled black dot at the screen center, the response cue contained a dark ring of which radius matching the eccentricity of the target. Participants made a saccadic eye movement onto the ring when reporting the remembered locations. Upon the detection of the saccade offset, a dot was presented at participants' saccade landing location. Participants held a dial in their dominant hands and were allowed to use the dial to manually adjust the location of the dot if they felt that its location did not match the location they intended to report (e.g., due to the noisy online gaze position readout). On average, only in 14% of the trials, participants' final reported location was the same as the location initially marked by the dot. To finalize the memory report, the participants pressed the button on a button box on the other hand. Upon the button press, an arc centered at the dot (reported location) appeared on the ring. In a post-estimation wager, the participants used the dial to adjust the length of the arc. Spinning the dial clockwise increased the length of the arc along the ring and vice versa. Participants were instructed to reflect the uncertainty of their memory on the length of the arc, the longer the arc, the more uncertain. Participants finalized the arc length by pressing a button and then the feedback stimulus (a white dot) appeared at the true target location. The number of points earned by participants for each trial was displayed on the screen along with the feedback stimulus. Participants were rewarded with some points only if the true target location fell within the arc. The number of points was $100e^{-0.08d}$, in which d was the length of the arc in polar angle (°). That is, the number of points they could gain decreased exponentially with the length of the arc. To gain more points, an optimal observer should increase the length of the arc with their uncertainty. Participants were well-informed regarding the structure of the betting game and the policy of reward. Each participant completed 180 to 270 trials (227 trials per participant on average) in 2 to 3 2-hr scanning sessions on separate days.

Passive viewing control experiment

We scanned a subset of participants (n = 3) on an additional control experiment in which we presented a high-contrast, salient flickering checkerboard stimulus at the same locations as the WM target stimuli while participants performed a demanding discrimination task at fixation. Trial timing was identical to that used in Experiments 1 and 2. Instead of a dim target stimulus, we presented a full-contrast flickering checkerboard (0.875 deg radius; 1 cycle/deg spatial frequency; 8 Hz flicker) for 500 ms, followed by a 12 s 'delay' period. Throughout the trial, including during the stimulus presentation period, participants carefully attended a rapidly-flashing "+" stimulus at fixation (4 Hz) to detect targets defined by a widening or heightening of the "+" and responding with one button for each target type. We adjusted the aspect ratio of the fixation discrimination stimulus across scanning runs to maintain performance ~75%. At the end of the 12 s "delay" period, the fixation task concluded, and participants received feedback about their detection performance via green/red/yellow dots (for correct/incorrect/missed responses) presented around fixation. Each of the 3 participants performed 2 sessions of this task, totaling 20-24 runs per participant.

Setup and stimuli

Visual stimuli were presented by an LCD (VPixx ProPix) projector located behind the scanner bore and were viewed by participants through an angled mirror with a field of view of 52° by 31°. A gray circular aperture with a diameter of 30° was presented on the screen throughout the experiments. The working memory target was a light gray dot with a diameter of 0.65°. It had an eccentricity at 12° from the central fixation point and its polar angle was pseudo-randomly chosen from 1 of 32 locations that evenly tiled the full circle within each run

Eyetracking

For all imaging sessions, we measured eye position using an EyeLink 1000 Plus infrared video-based eye tracker (SR Research) mounted beneath the screen inside the scanner bore operating at 500 Hz. The camera always tracked the participant's right eye, and we calibrated using either a 13-point (Experiment 1, Experiment 2, and the passive viewing experiment) or 5-point (retinotopy) calibration routine at the beginning of the session and as necessary between runs. We monitored gaze data and adjusted pupil/corneal reflection detection parameters as necessary during and/or between each run.

Behavioral data analysis

For Experiment 1, we used gaze position estimated from eye position traces as our measurement of VWM performance. We preprocessed raw gaze data using fully-automated procedures implemented within iEye_ts (https://github.com/tommysprague/iEye_ts) to remove blinks, adjust for drift over the course of a run, recalibrate gaze data trial-by-trial, automatically identify memory-guided saccades, and flag trials for rejection (for behavioral analyses).

We defined blinks as 200 ms before and after periods when pupil size fell below the 1.5th percentile of the distribution across all pupil size samples of the entire run (396 s). We computed velocity based on smoothed gaze time courses (5 ms standard deviation Gaussian kernel). We defined saccades based on a velocity threshold of 30 deg/s and a minimum duration of 0.0075 s and 0.25 deg amplitude. We defined periods between saccades as fixations. We drift-corrected each trial based on the modal fixation position during the trial period before the go cue appeared. To recalibrate gaze traces on each trial, we found the nearest fixation to the known target position during the feedback period (800 ms during which target was re-presented and participants were instructed to fixate this position) and fit a 3rd-order polynomial for each coordinate (X,Y) to map between actual WM position and measured gaze coordinate. We used this polynomial to recalibrate the X and Y traces across all trials within each run. We used trials for which measured

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

gaze position was within 2.5 deg visual angle of the feedback target location for fitting the polynomial, but all trials were subjected to the resulting recalibration.

We quantified WM error based on the endpoint of the large saccadic eye movement toward the remembered position (> 5 deg amplitude, < 150 ms saccade duration), which we call the 'primary saccade', and the final eye position before the feedback stimulus appeared ('final saccade'). On trials in which a subsequent corrective saccade is not made before the feedback stimulus appeared, these positions were considered identical. For both primary and final saccades, the saccade must have both begun and ended during the response period. Moreover, we exclude trials in which participants initiate a saccade faster than 100 ms after the response cue appeared (Mackey and Curtis, 2017).

We flagged trials for exclusion based on: (1) failures of automatic drift correction and/or excessive necessary drift correction (beyond 2.5 deg), (2) fixation outside a 2.5 deg aperture around fixation during the delay period, (3) ill-defined primary saccade, or (4) excessive error for primary saccade (> 5 degree visual angle). We included all trials for fMRI data analyses regardless of behavioral exclusion criteria during model estimation to ensure a balanced sampling of spatial positions, but only included trials with reliable behavioral estimates for all subsequent analyses including quantifying the decoding performance and correlating decoded results with behaviors.

The analysis for Experiment 2 was similar to that of Experiment 1, with a few exceptions. As participants were allowed to use the dial to manually adjust the remembered location, we used the final dot location after the manual adjustment as the participants' memory reports. Different from the definition of excessive error (> 5 degree visual angle) used in Experiment 1, we computed the memory error and the reported arc length in units of degree polar angle and excluded the trials with memory error exceeding the mean error plus three standard deviations. The same exclusion criterion was applied to the trials with excessive reported arc length.

For both experiments, when quantifying participants' behavioral memory error (Figures 1, 4, 5, and 7), we computed the error as the (signed) difference between the reported location and WM target position in polar angle.

Retinotopic mapping and the identification of region of interest (ROI)

Each participant was scanned for one 1.5-2 hour fMRI session for retinotopic mapping. The experimental procedures followed those reported by Mackey et al. (Mackey et al., 2017). Participants maintained fixation at the screen center while covertly tracking a bar aperture sweeping across the screen in discrete steps and in four directions: a vertical aperture moving from the left to the right, or from the right to the left of the screen; a horizontal aperture moving from the top to the bottom, or from the bottom to the top of the screen. The bar aperture was divided into three rectangular segments (defined as a central segment and two flanking segments) with equal sizes, each containing a random dot kinematogram (RDK). Participants' task was to discriminate in which one of the two flanking segments, the motion direction of the RDK was in the same direction as the one within the central segment. The dot motions of all the three segments changed with each discrete step. Participants reported their answer by a button press before the bar moved into the next step. The coherence of the random dot motion was staircased in order to keep the difficulty of the task at about 75% accuracy. Each session contained eight to nine runs. In each run, the bar aperture swept across the screen 12 times, and each swept consisted of 12 discrete steps. The four sweeping directions were interleaved and randomized within each run. While Mackey et al. (Mackey et al., 2017) presented different bar widths in different scanning runs, here we interleaved 3 different bar widths during the same run.

We fit a population receptive field (pRF) model with compressive spatial summation to the BOLD time series of the retinotopic mapping data for each participant (Dumoulin and Wandell, 2008; Kay et al., 2013) after smoothing on the surface (5 mm FWHM Gaussian kernel). We visualized on the cortical surface the voxels' preferred phase angle and eccentricity estimated by the pRF model. To define the ROIs, we set a threshold to only include voxels with greater than 10% response variability explained by the pRF model. We then drew ROIs by visual inspection, primarily by identifying reversals of the voxels' preferred phase angle on the cortical surface. We define bilateral dorsal visual ROIs V1, V2, V3, V3AB, IPS0, IPS1, IPS2, IPS3, iPCS and sPCS, each with a full visual field representation.

MRI acquisition

MRI data were acquired on a Siemens Prisma 3T scanner with a 64-channel head/neck coil. We collected functional imaging for the working memory experiments and the passive viewing experiment with 44 slices and a voxel size of 2.5^3 mm (4x simultaneous-multi-slice acceleration; FoV 200 × 200 mm, no in-plane acceleration, TE/TR: 30/750 ms, flip angle: 50 deg, Bandwidth: 2290 Hz/pixel; 0.56 ms echo spacing; P→ A phase encoding). Intermittently throughout each scanning session we acquired pairs of spin-echo images in the forward and reverse phase-encoding direction with identical slice prescription and no simultaneous-multi-slice acceleration (TE/TR: 45.6/3537 ms; 3 volumes per phase encode direction). These pairs are used to estimate a field map used to correct for local spatial distortions. The slice prescription was approximately parallel to the calcarine sulcus and covered most of the occipital lobe and the parietal lobe, with the exception of ventral temporal poles and ventral orbitofrontal cortex in some participants. The functional imaging data for retinotopic mapping was acquired in a separate session at a higher resolution, with a slice prescription spanning 56 slices (4x simultaneous multislice acceleration) and a voxel size of 2^3 mm (FoV 208 × 208 mm, no in-plane acceleration, TE/TR: 36/1200 ms, flip angle: 66 deg, Bandwidth: 2604 Hz/pixel (0.51 ms echo spacing), P→A phase encoding).

For each participant, in the retinotopic mapping session, we also collected 2 or 3 T1 weighted whole-brain anatomical scans (MPRAGE sequence; 0.8 mm^3).

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

MRI data preprocessing

T1-weighted anatomical images were segmented, and cortical surfaces were constructed using Freesurfer (v6.0). Functional data (EPI time series) of both the retinotopic mapping experiment and the VWM experiments were preprocessed by customized scripts using functions provided by AFNI. We applied B0 field map correction and reverse-polarity phase-encoding (reverse blip) correction to the functional data. Spatial smoothing (5 mm FWHM on the cortical surface) was only applied to the retinotopic mapping data. All the functional data were motion-corrected (6-parameter affine transform), aligned to the anatomical images, projected onto the cortical surface, then re-projected into volume space. This process incurs a small amount of smoothing along vectors perpendicular to the cortical surface, but no additional smoothing was applied. When possible, all linear and nonlinear spatial transformations were concatenated into a single transform operation to minimize additional smoothing. Linear trends were removed from the time series. For the VWM experiments, the time series of each voxel was first converted into percentage signal change for each run, and then normalized (z-score) across time points within each run.

QUANTIFICATION AND STATISTICAL ANALYSIS

Generative model

We decoded the content of WM using a generative model proposed by van Bergen et al. (2015) and van Bergen and Jehee (2021). Specifically, we used the method named TAFKAP described in (van Bergen and Jehee, 2021). Note that this Bayesian decoding approach allows us to concurrently read out memorized location and memory uncertainty from the same decoded probability distribution. Other decoding methods (e.g., linear regression, like inverted encoding models IEM; Sprague et al., 2018) can decode memorized locations but do not concurrently provide a theoretically-grounded representation of uncertainty. Here, in this and the next section we briefly describe the critical components of the model and the model-fitting procedures. In the generative model, the multivariate voxel response given the stimulus location (polar angle) was modeled as a multivariate normal distribution. The average response (mean) of each voxel given a stimulus was determined by its tuning function (voxel response as a function of polar angle). The voxel tuning function was approximated by a weighted sum of eight basis functions that evenly tiled the location space. The basis functions are raised sinusoidal functions

$$f(s)_k = \lfloor \cos(s - \phi_k) \rfloor^8$$

where $\lfloor\rfloor$ represents half-wave rectification and ϕ_k is the center of the kth channel. The response of ith voxel b_i given a stimulus s is then modeled as

$$b_i(s) = \sum_{k=1}^{8} W_{ik}(f_k(s) + \eta_k) + \upsilon_i$$

where $\mathbf{W}$ is a weighting matrix that determines the weights of each basis function for each voxel. Here, two sources of variability are considered. First, η is the noise specific to each basis function. This noise was carried over into each voxel by the weighting matrix $\mathbf{W}$. It modeled the noise shared across voxels with similar voxel tuning functions. The model assumed that η follows a zero-mean normal distribution whose covariance matrix is a constant noise magnitude multiplied with an identity matrix $\eta \sim N(0, \sigma^2\mathbf{I})$. Second, υ represents the noise specific to each voxel. The model assumed that the voxel-wise noise follows a zero-mean normal distribution $\upsilon \sim N(0, \Sigma)$. The covariance matrix of this distribution is approximated by a rank-one covariance matrix plus a diagonal matrix

$$\Sigma = \rho\tau\tau^{\mathrm{T}} + (1 - \rho)\mathbf{I} \circ \tau\tau^{\mathrm{T}}$$

where $\circ$ represents Hadamard product, element-wise product between two matrices. Thus, based on this generative process, the theoretical covariance matrix of the multivariate response of the voxels given a stimulus s is

$$\Omega_0 = \rho\tau\tau^{\mathrm{T}} + (1 - \rho)\mathbf{I} \circ \tau\tau^{\mathrm{T}} + \sigma^2\mathbf{W}\mathbf{W}^{\mathrm{T}}$$

The first two terms of the theoretical covariance matrix consider a simple form of covariance as a weighted sum between a diagonal matrix (where τ is a vector representing the standard deviation of the noise of each voxel) and a rank-one covariance matrix. The last term captures the covariance depending on the tuning functions of the voxels (i.e., the voxels selective for similar locations have higher covariance; see derivations in (van Bergen et al., 2015)).

In addition to the theoretical covariance matrix, the model also considered the empirical sample covariance

$$\Omega_{\text{sample}} = \frac{1}{N_{\text{train}}}\left(\mathbf{B} - \widehat{\mathbf{W}}\mathbf{G}\right)\left(\mathbf{B} - \widehat{\mathbf{W}}\mathbf{G}\right)^{\mathrm{T}}$$

where $\mathbf{B}$ is the training data and $\mathbf{G}$ is the response of the basis functions given the training set stimuli. Thus, for each training dataset, we assumed that the voxel activity pattern followed a multivariate normal distribution.

$$p(\mathbf{b}|s) \sim N(\mathbf{W}f(s), \Omega)$$
$$\Omega = \lambda\Omega_0 + (1 - \lambda)\Omega_{\text{sample}}$$

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

When the number of variables (voxels) is larger than the number of observations (trials), the sample covariance is not invertible. To ensure an invertible and stable estimation of the covariance matrix, here the covariance matrix was modeled as the sample covariance matrix "shrunk" (Ledoit and Wolf, 2004) to a target covariance matrix, the theoretical covariance matrix Ω_0. The degree of shrinkage is determined by a free parameter λ (see details in (van Bergen and Jehee, 2021)).

Model fitting and decoding

For each voxel, we averaged the z-normalized percentage signal change of the BOLD time-series over a time window at 5.25 to 12.00 s from the delay onset. This (time-averaged) voxel response corresponding to the delay period was the input to the model. To let the ROIs we compared have the same number of voxels, for each ROI we first selected the voxels that exhibited the strongest location selectivity. For each voxel within an ROI, we performed a one-way ANOVA on the training dataset (so the testset was not used for voxel selection) using the 32 target locations as a categorical independent variable and the voxel response as the dependent variable. For each ROI, we selected 750 voxels with the largest F value. These voxels were used for training the model, and later for decoding the data in the testset. The results we reported were robust with respect to the number of voxels selected: We varied the number of voxels selected (per ROI) across a wide range (from 32 to 1250 voxels), and found that the patterns of all the critical indices were robust with respect to the number of voxels selected (Figures S5 and S6).

For each participant and each ROI, after selecting the voxels, we trained a Bayesian decoding model using TAFKAP (van Bergen and Jehee, 2021), and decoded spatial positions using a leave-one-run-out cross-validation procedure. During the training phase the model used all the trials, except those from one held out run, to estimate the free parameters of the generative model. TAFKAP used a method called "bootstrap aggregating" or "bagging" to take the uncertainty of model parameters into account. Bagging is a special case of model averaging. By bootstrapping, the trials in the training dataset were resampled with replacement for multiple times to generate many bootstrap resampled datasets. Each resampled dataset had the same number of trials as the training dataset. For each resampled training dataset j, a set of free parameters θ_j was estimated by ordinary least-squares. Each trial in the testset (the held out run) was then decoded based on Bayes rule. For each trial in the testset, the posterior probability of the stimulus given the multivariate voxel response $\mathbf{b}$ was computed as

$$p(s|\mathbf{b};\theta_j) = \frac{p(\mathbf{b}|s;\theta_j)p(s)}{\int p(\mathbf{b}|s;\theta_j)p(s)ds}$$

We assumed the prior $p(s)$ to be a uniform distribution and we approximated the continuous posterior probability function by sampling 1000 steps evenly spanning the location space. The normalization factor in the posterior was computed by numerical integration. Note that for each trial in the testset, decoding was performed multiple times based on the parameters estimated using each resampled training dataset. The decoding results were averaged across all resampled training datasets to obtain one decoded posterior probability distribution

$$p_{\text{bag}}(s|\mathbf{b}) = \frac{1}{N_{\text{bootstrap}}}\sum_j p(s|\mathbf{b};\theta_j)$$

We then numerically estimated the circular mean of the posterior to represent the decoded location, and the circular standard deviation of the posterior to represent the uncertainty of the remembered location. The number of bootstrap resampled dataset generated (N_{boostrap}) was determined by a stopping criterion based on Jensen-Shannon divergence (see details in (van Bergen and Jehee, 2021)).

Statistical analysis

We tested whether the decoding error distributions (Figure 3A; Figure S1C) departed from a uniform distribution by a (permutation-based) V test (Berens, 2009), which is a circular variable equivalent of the Rayleigh test with the alternative hypothesis that the decoding error distributions had means centered at zero degree (polar angle). For each participant and each ROI, we computed the V statistics and compared it with the null distribution, obtained by randomly permuting the target location and then recomputing the decoding errors and their V statistics for 2000 times. The results of the V test are reported in Tables S1 and S4.

Decoding performance was quantified by two indices: the standard deviation of the decoding error (the decoded location minus the target location in polar angle) distribution (Tables S2 and S5), and the circular correlation between the decoded location and the target location (Tables S3 and S6). For each subject and ROI, we computed the standard deviation of the decoding error distribution and compared it with the null distribution. We obtained the null distribution by randomly permuting the target location and then recomputing the standard deviation of the decoding error for 2000 times. At the group level, we conducted the same permutation procedure to obtain the null distribution of the group-averaged standard deviation of the decoding error distribution. The same statistical procedures were applied to the circular correlation.

To relate decoding outputs to behaviors we conducted two sets of statistical tests in parallel (1) We conducted non-parametric bootstrapping to test the significance of the single-trial correlations reported in both experiments, including the circular correlation between the decoding error and memory error (Figures 4B and 7B), the correlation between decoded uncertainty and reported arc length (Figure 6B), the correlation between decoded uncertainty and saccade reaction time (Figures S4B and S4D), and the

Please cite this article in press as: Li et al., Joint representation of working memory and uncertainty in human cortex, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.08.022

correlation between decoded uncertainty and the magnitude of memory error (Figures S7A and S7D). For each ROI, we computed the correlation (or circular correlation) between the two variables in interest for each participant and averaged the correlation coefficients across participants. We then resampled the correlation coefficients (with replacement) and computed the averaged correlation coefficients. We repeated this procedure for 2000 iterations to obtain a bootstrapped distribution of the averaged correlation coefficients. The percentage of the iterations in this distribution that was higher or lower than zero was used to compute (two-tailed) *p-value*s. (2) Following the statistical tests conducted in the previous studies applying the same Bayesian decoding method (van Bergen and Jehee, 2021; van Bergen et al., 2015), we also computed binned-correlation for statistical analysis (Figures 4C, 6C, and 7C; Figures S4C, S4E, S7B, S7C, S7E, and S7F). For each participant, the trials were sorted into four bins with increasing decoding error or decoded uncertainty. The memory error or reported arc length was then computed for each bin. We then pooled data points across participants (four data points per participant) after removing the mean of each participant. For visualization, we added the grand means back to the data when plotting binned correlations. Pearson correlation coefficients were then computed based on the pooled data. We compared the correlation coefficients to the null distribution obtained by permuting the data points in the pooled dataset.

We conducted permutation ANOVA to test the effect of ROI on decoding performance (Figure 3B; Figure S1) and error correlations (Figures 4B and 7B). The *F*-statistic computed from the original data was compared to the null distribution of *F*-statistics, which was obtained by randomly permuting the ROI labels and calculating the *F*-statistic for 2000 times. We used a false-discovery rate (Benjamini–Hochberg procedure) for correction of multiple comparisons (the number of ROIs) with q = 0.05. We reported adjusted p values unless otherwise specified.