Reviewed Preprint
v1 • September 9, 2025
Not revised

Reviewed Preprint
v2 • May 28, 2026
Revised by authors

For correspondence:
j.castanheira@ucl.ac.uk

Funding: See page 22

Reviewing editor: Nathan Faivre, Centre National de la Recherche Scientifique, France

How attention simplifies mental representations for planning

Jason da Silva Castanheira[1], Chang (Christina) He[1], Nicholas Shea[2], Stephen M Fleming[1,3,4]

[1]Department of Experimental Psychology, University College London, London, United Kingdom • [2]Institute of Philosophy, School of Advanced Study, University of London, London, United Kingdom • [3]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, United Kingdom • [4]Canadian Institute for Advanced Research (CIFAR), Brain, Mind and Consciousness Program, Toronto, Canada

eLife Assessment

This **important** study utilizes behavioral data and computational modeling to show that spatial properties of visual attention affect human planning. The methodology and statistical analyses are **convincing**, though the way attention is conceptualized and modeled could be refined. The findings of this study will interest cognitive scientists studying attention, perception, and decision-making.

https://doi.org/10.7554/eLife.108034.2.sa4

Abstract

Human planning is efficient—it frugally deploys limited cognitive resources to accomplish difficult tasks—and flexible—adapting to novel problems and environments. Computational approaches suggest that people construct simplified mental representations of their environment, balancing the complexity of a task representation with its utility. These models imply a nested optimisation in which planning shapes perception, and perception shapes planning – but the perceptual and attentional mechanisms governing how this interaction unfolds remain unknown. Here, we harness virtual maze navigation to characterise how spatial attention controls which aspects of a task representation enter subjective awareness and are available for planning. We find that spatial proximity governs which aspects of a maze are available for planning, and that when task-relevant information follows natural (lateralized) contours of attention, people can more easily construct simplified and useful maze representations. This influence of attention varies considerably across individuals, explaining differences in people's task representations and behaviour. Inspired by the 'spotlight of attention' analogy, we incorporate the effects of visuospatial attention into existing computational accounts of value-guided construal. Together, our work bridges computational perspectives on perception and decision-making to better understand how individuals represent their environments in aid of planning.

Significance statement

Humans have an impressive ability to plan. Theoretical models in computer science propose that instead of using all the available information in a scene, a decision-maker should form a simplified mental representation of their environment over which they plan. However, little is known about how perceptual and attentional processes shape the planning process in humans. We find that people form simplified mental representations in line with the natural contours of spatial attention, whereby information limited to a visual hemifield is more readily available for planning, like a spotlight illuminating a part of the environment. We develop a novel computational model of the effects of attention on planning and characterise systematic variation between individuals in how they simplify their mental representations.

Introduction

Humans have an impressive ability to plan. We are able to model the world, simulate potential outcomes, and select among possible courses of action. Take, for example, your first trip to London. You want to visit Buckingham Palace despite being jet-lagged. Looking at a map of the underground, you're overwhelmed with information but need to make a plan. How do you solve this problem? Even simple decisions like this involve many potential actions and outcomes, making it impossible to systematically evaluate every possible option, especially given limited cognitive resources (Callaway et al., 2022; Daw et al., 2005; Dezfouli & Balleine, 2013; Griffiths et al., 2019; Huys et al., 2012; Newell & Simon, 1956). Explaining how people plan efficiently and flexibly under these constraints is a long-standing challenge in human and machine intelligence (Daw et al., 2005; Griffiths et al., 2019; Hassabis et al., 2017; Saxe et al., 2021).

Theories of human problem-solving conceptualize planning as a search through a 'decision tree' of all potential actions and their outcomes (Breslow & Aha, 1997; Callaway et al., 2022; Huys et al., 2012; Newell & Simon, 1956; Quinlan, 1986). In our example, an individual may first list all possible tube stations within walking distance and then evaluate which action sequence will get them closer to their destination. Previous work proposes different algorithmic strategies for how an agent efficiently searches over a complex decision tree. These strategies include ignoring low-value actions (i.e., pruning) (Huys et al., 2012, 2015; Knuth & Moore, 1975; Mingers, 1989), limiting how far in the future one might search (i.e., depth)(Callaway et al., 2022; Keramati et al., 2016; Korf, 1985; Snider et al., 2015), or relying on previously learnt strategies (i.e., habits) (Daw et al., 2005; Dezfouli & Balleine, 2013; Keramati et al., 2016; Kool et al., 2016).

This previous work, however, largely assumes that a decision-maker has a fixed representation of the problem. When planning involves constructing and evaluating multiple multi-step trajectories within a decision tree, the computational burden increases with the complexity of the representation of the problem space. Consider planning in a two-dimensional spatial grid, for example. A finer-grained grid presents many choice points about which way to turn. A coarser-grained grid presents fewer choice points. Since the number of branches is a multiplicative function of the number of choice points, a simplified representation of the task space, if chosen appropriately, can have a profound effect on reducing the computational demands of planning.

One elegant approach to forming such a simplified representation is to adaptively select the granularity of information required to complete the task (Ho et al., 2022), known as value-guided construal (VGC). Unlike previous accounts, which model human planning as a search over all items (e.g.., tube lines), the VGC model predicts that a cognitively limited decision-maker selects a manageable subset of information over which to plan— i.e., a task representation—balancing utility and complexity (Ho et al., 2022). In our example, the VGC algorithm would plan over a few relevant tube lines rather than planning over all possible stations. To select the representation that achieves the best balance between utility and complexity, the model searches across all possible combinations of tube lines, computing the value (i.e., the plan's utility minus its cost) of each representation for planning a specific journey. The algorithm then selects the representation with the highest value, which ensures that an ideal observer selects a representation which only includes the items (i.e., tube lines) that lead to successful planning while excluding as many items as possible to keep the plan as simple as possible. For our purposes, items included in the representation are considered taskrelevant, while items that are not represented are considered task-irrelevant. This algorithm, therefore, provides a normative standard of an efficient plan to which we can compare people's actual plans.

In previous work, Ho and colleagues discovered that people's awareness of, and memory for, obstacles in a maze varies in line with the predictions of a VGC model. The VGC model implies two nested optimisations – an outer loop of construal, and an inner loop that runs a plan conditional on a particular task representation. The VGC model is a normative model and remains agnostic as to the cognitive mechanisms controlling the construal. In particular, the perceptual and

attentional mechanisms governing *how* information is selected to become part of a task representation remain unknown. Initiating such a nested computation plausibly rests on inductive biases – general principles that a perceptual system can apply to select task-relevant information, before refining it as part of the planning process (Gershman, 2021). Selective attention is proposed as one general mechanism by which the brain selects relevant information, either voluntarily (endogenous) or reflexively (exogenous)(Carrasco, 2011; Carrasco et al., 2004; Chica et al., 2013; Landry et al., 2021, 2023; Nobre & Kastner, 2014).

Previous studies have demonstrated that attention guides the selection of particular features of the environment to support reinforcement learning(Niv, 2019). However, it remains unknown whether and how attention shapes value-guided construal "on the fly" during planning. For instance, one possibility is that forming a simplified task representation is a "late" passive side-effect of the planning process – a tendency to focus on what we are thinking about. Alternatively, VGC may reflect an "early" selection of perceptual information, perhaps based on a rapid feedforward sweep of perceptual input(V. A. Lamme & Roelfsema, 2000). These alternatives echo classic debates between early and late selection models of attention(Broadbent, 1958; Nelson et al., 2012), but now situated within the broader landscape of computational accounts of planning. More generally, despite the wealth of literature on attention, and pioneering efforts to incorporate attentional constraints into models of decision-making(Ho et al., 2022; Niv, 2019), we lack a basic understanding of how attention influences planning.

To make progress on this question, we examined the role of visuospatial attention on how people construct simplified task representations across three experiments in human participants. We build on previous work using maze navigation to provide a rich readout of people's current task representations. We predicted that if visuospatial attention is guiding the formation of task representations, the construal process will be constrained by inductive biases characteristic of attentional selection. For instance, previous work has illustrated how attentional selection is biased by the spatial context in which information is presented: presenting distractors alongside task-relevant stimuli makes attentional selection more challenging(He et al., 1996; Liu et al., 2009; Whitney & Levi, 2011). Attention, in this case, spills over to the neighbouring stimuli. These findings align with the metaphorical attentional spotlight or zoom lens, which stipulates that the focus of visual attention can move around the visual field, illuminating a limited spatial extent at a time (Norman, 1968; Posner et al., 1980). According to this model, individuals can, for example, orient their attention preferentially to a single hemifield—i.e., lateralizing—which is enabled by a hemispheric lateralization of alpha power over posterior cortex(Bagherzadeh et al., 2020; Jensen, 2024; Keefe & Störmer, 2021; Landry et al., 2024).

Our focus in this study was to examine how participants perceive and represent their environment (the maze stimulus). This is a distinct process to how participants orient their attention during navigation itself, which is not part of our current study. To do so, we harness classical signatures of attentional selection to characterise how visuospatial attention shapes awareness of maze obstacles during planning. First, we demonstrate "attentional overspill": participants preferentially incorporate task-irrelevant information into their task representation when it is presented in spatial proximity to task-relevant information. Second, we observe that attentional overspill is reduced when task-relevant information is lateralised to a single hemifield, allowing participants to more effectively form optimal task representations. Finally, we extend the VGC model to incorporate visuospatial attention as a key psychological mechanism for constructing simplified task representations. Together, our findings furnish a computational account of how attention and perception guide simplified representations in the service of planning.

Results

To examine the role of visuospatial attention in planning, we relied on a previously developed maze navigation paradigm in which participants solved 2-D mazes (Figures S1-6) (Ho et al., 2022), avoiding obstacles obstructing their path (Figure 1a, left panel). We operationalized planning using a maze navigation paradigm, akin to our tube-related example, where participants

were required to plan a route through the maze, avoiding obstacles that blocked their path. Obstacles predicted by the sVGC model to be included in the representation were considered task-relevant.

At the end of every trial, participants reported their awareness of specific obstacles (see Methods for details). The level of awareness reported for different obstacles provides a read-out of what features of the environment individuals were subjectively representing while solving a particular maze. While other markers of attention and awareness (for instance, behavioural or neurophysiological variables) could also be used, here we focused on direct awareness reports in order to relate our findings both to those of Ho and colleagues and to the subjective awareness reports used in consciousness science (e.g. the Perceptual Awareness Scale (Barnett et al., 2024; Overgaard & Sandberg, 2021; Ramsøy & Overgaard, 2004; Samaha et al., 2015)). Participants were instructed to maintain central fixation while planning (see dataset dSC 1), in line with previous empirical work using this task (Ho et al., 2022).

We first reanalyzed the data presented by Ho and colleagues (2022) (Ho et al., 2022) to examine the role of spatial attention in building task representations (datasets Ho 1 and 2). In a new experiment (dataset dSC 1), we designed novel mazes to test the effects of lateralization of attention in enabling efficient planning (see Methods & Table S1). In addition, we recorded the eye movements of participants during planning to ensure that any attentional effects observed were driven only by covert shifts in attention (see Methods).

We retained trials in which participants successfully navigated each maze (see Methods). Note that the successful navigation of the maze stimulus and the construal process represent two distinct processes. For example, we can imagine a trial in which a participant represents every single obstacle, whether relevant or irrelevant. We would predict that on this trial, the participant could successfully navigate the maze, yet their construal process would be suboptimal according to the normative sVGC model.

Our focus in the present study was to examine attentional effects on participants' perception of the maze stimulus. We did not quantify how individuals deploy their attention in the phase in which they were navigating through the maze.

A spotlight of attention influences task representations

We hypothesized that spatial attention would control which items are included in a task representation(He et al., 1996; Liu et al., 2009; Whitney & Levi, 2011). Specifically, we hypothesised that participants would deviate from the predictions of the VGC model and become distracted by task-irrelevant obstacles when they are presented in spatial proximity to task-relevant obstacles. Note that the task-relevance of obstacles is related to the maze's organization and computational model, and is not related to participants' subjective reports. To evaluate these predictions, we first computed the distance between a probe obstacle and every other obstacle in the maze. Second, we ranked the obstacles from the closest to the furthest from the probed item. Using the ranked obstacles, we trained a linear regression model to predict participants' awareness of the probed obstacle (in green) from their awareness of the remaining obstacles (in grey; Figure 1b).

Critically, we observed a significant effect of spatial context on task representations – an effect which is not predicted by the normative VGC model. Participants' awareness of a particular obstacle was positively predicted by the awareness of its close neighbours (β_1 = 0.26, SE = 0.01, 95% CI [0.25, 0.28]; β_2 = 0.29, SE = 0.01, 95% CI [0.27, 0.30]), whereas awareness of its furthest neighbours negatively predicted participant reports (β_5 = -0.13, SE = 0.01, 95% CI [-0.15, -0.12]; β_6 = -0.13, SE = 0.01, 95% CI [-0.15, -0.12]; see Table S2). In other words, the spatial context of an obstacle predicted whether it would be included in a simplified task representation – akin to a diffuse attentional spotlight which filters which aspects of the maze are available for planning. This effect remained significant for both task-relevant and task-irrelevant obstacles, and after controlling for the predictions of the VGC model (Figure S7 & Table S3, respectively). We observed the same effect in a separate experiment where participants planned their route upfront

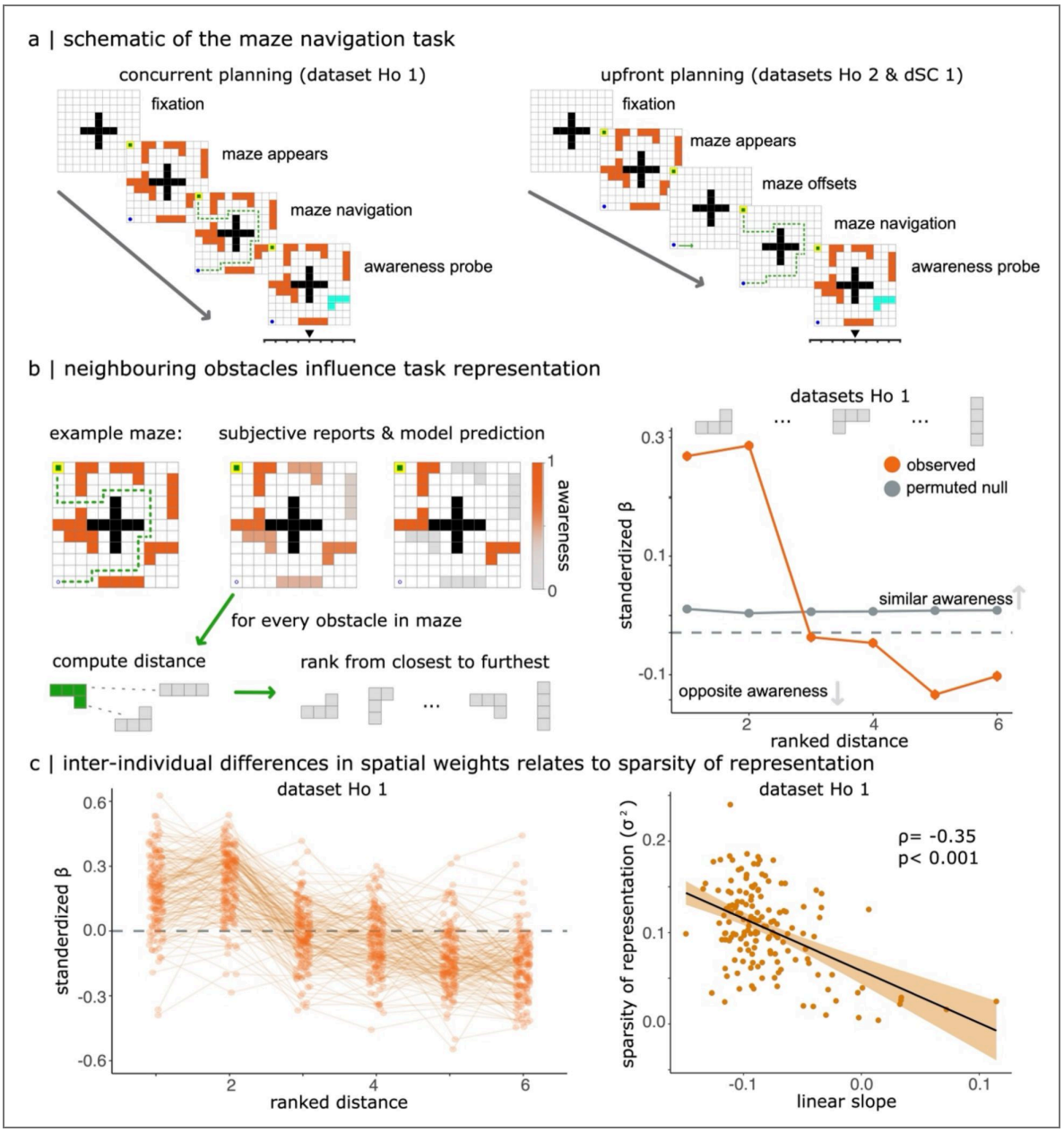

Fig. 1. Spatial attention shapes task representations.

(a) Schematic of the maze navigation task. Participants fixated at the start of each trial, after which a maze was presented, which they were asked to navigate. Maze stimuli either remained on the screen during navigation (left panel; *concurrent planning experiments*) or were removed before navigation (right panel; *upfront planning experiments*). Once participants finished navigating the maze, they were asked to report their awareness of every obstacle presented on a given trial in a random order. (b) Left panel: schematic of the analysis pipeline. An example maze is shown where seven obstacles (plotted in orange) are presented on every trial according to pre-defined mazes. Participants report their awareness of every obstacle at the end of each trial (middle maze). The VGC model predicts which obstacles in a maze will likely be included in participants' task representation (right maze). We use participants' awareness reports to test the influence of neighbouring obstacles on the probe obstacle (presented in green). We compute the influence of neighbouring obstacles (in grey) on participants' awareness of the probed obstacle (in green). Right panel: Results of the ranked regression model for dataset Ho 1. We observed that obstacles closest to the probed item (rank 1 & 2) positively impact awareness reports. In contrast, obstacles furthest from the probed item negatively impact awareness reports (rank 5 & 6). (c) Left panel: The effect of neighbouring obstacles on task representations varied across participants (each represented by a point). Right Panel: Inter-individual differences in the attentional effects correlate with the sparsity of participants' representations. Participants who showed the greatest influence of neighbouring obstacles (more negative slopes), showed the simplest representations (greatest variance in awareness reports).

before navigating the mazes (i.e., dataset Ho 2, see Figure S8 and Table S4 & S5). Finally, we replicated this pattern of results in our in-person experiment: closest neighbours positively predicted the awareness of an obstacle (β_1 = 0.19, SE = 0.007, 95% CI [0.18, 0.21]), whereas furthest neighbours negatively predicted participants' reports (β_3 = -0.10, SE = 0.01, 95% CI [-0.11, -0.08]; β_4 = -0.26, SE = 0.007, 95% CI [-0.27, -0.25]; β_5 = -0.29, SE = 0.007, 95% CI [-0.30, -0.27]; see Table S6 & S7 and Figure S9).

Next, we explored whether the influence of neighbouring obstacles on task representations varied across individuals. To do so, we fit the regression model described above to quantify each participant's attentional spillover, and quantified the linear slope of the resulting beta coefficients. Negative slopes indicate a significant effect of attentional spillover on task representation. The influence of attention varied considerably across participants: while on average, participants' task representations were influenced by attention (mean effect = -0.08; s.d. = 0.04), a subset of participants showed minimal influence of attention on their task representation (i.e., flat slopes; Figure 1c).

We hypothesized that participants with the largest attention effects (i.e., most negative slopes) would also show sparser task representation (i.e., a "spotlight of attention" which is focused only on a subset of obstacles). To test this, we computed the sparsity of participants' task representations by estimating the variance of their awareness reports, with higher variance indexing those participants who report being very aware of some obstacles and unaware of others. In line with our hypothesis, we observed that participants who were most influenced by neighbouring obstacles also showed sparser task representations (dataset Ho 1: ρ = -0.35, $p < 0.001$; dataset Ho 2: ρ = -0.49, $p < 0.001$; dataset dSC 1: ρ = -0.51, $p < 0.01$; see Figure S10). To address concerns of overfitting, we tested whether the spatial attention effects observed in a lateralized set of mazes generalized to task representations of non-lateralized mazes and vice versa (dataset dSC 1). We observed that inter-individual differences in spatial attention effects in one condition predicted the sparsity of task representations in the other (ρ = -0.48, $p < 0.01$; ρ = -0.42, $p < 0.05$).

Attentional limits constrain the optimality of task representations

Prior psychological research indicates that attention can be efficiently allocated to a "hemifield" of visual space – with information being preferentially processed when presented in the attended hemifield(Eriksen & St. James, 1986; Posner, 1980; James, 1890). Building on this work, we hypothesized that participants would select task-relevant information with greater ease – constructing task representations more closely aligned with the VGC model – when task-relevant information is spatially confined to a visual hemifield (i.e., presented unilaterally).

To test this hypothesis, we derived a measure of task-relevant lateralization inspired by the attention literature (Ghafari et al., 2024; Keefe & Störmer, 2021; Vollebregt et al., 2015) (Figure 2a). Specifically, we separated maze stimuli across the vertical meridian and computed the ratio of task-relevant information presented on the left versus right side derived from the sVGC model. For example, the maze shown in Figure 2a has twice the amount of task-relevant information presented in the left hemifield than in the right (lat. Index= 1/3). A lateralization index of 0.0 indicates that both hemifields contain equal amounts of task-relevant information (i.e., non-lateralized). The lateralization index was computed using the continuous VGC predictions for each obstacle (see Methods). We used this task-relevance lateralization index as a moderator in a hierarchical linear regression model to test whether participants' awareness reports were better predicted by the original VGC model in mazes showing the greatest lateralization of task-relevant information. In addition, we monitored participants' eye movements in dataset dSC 1 to ensure that attention shifts would be covert as opposed to overt—a distinction which could not be determined in the online samples of datasets Ho 1 and 2.

In line with our hypothesis, we observed a significant moderation effect whereby the greater the lateralization of task-relevant information across the vertical meridian, the better the original VGC model was at predicting participants' awareness reports ($\beta_{interaction}$ = 0.01, SE = $2.65*10^{-3}$, 95% CI [0.01, 0.02], $p_{FDR} < 0.001$; Figure 2c left panel & Table S8). We replicated these findings with the

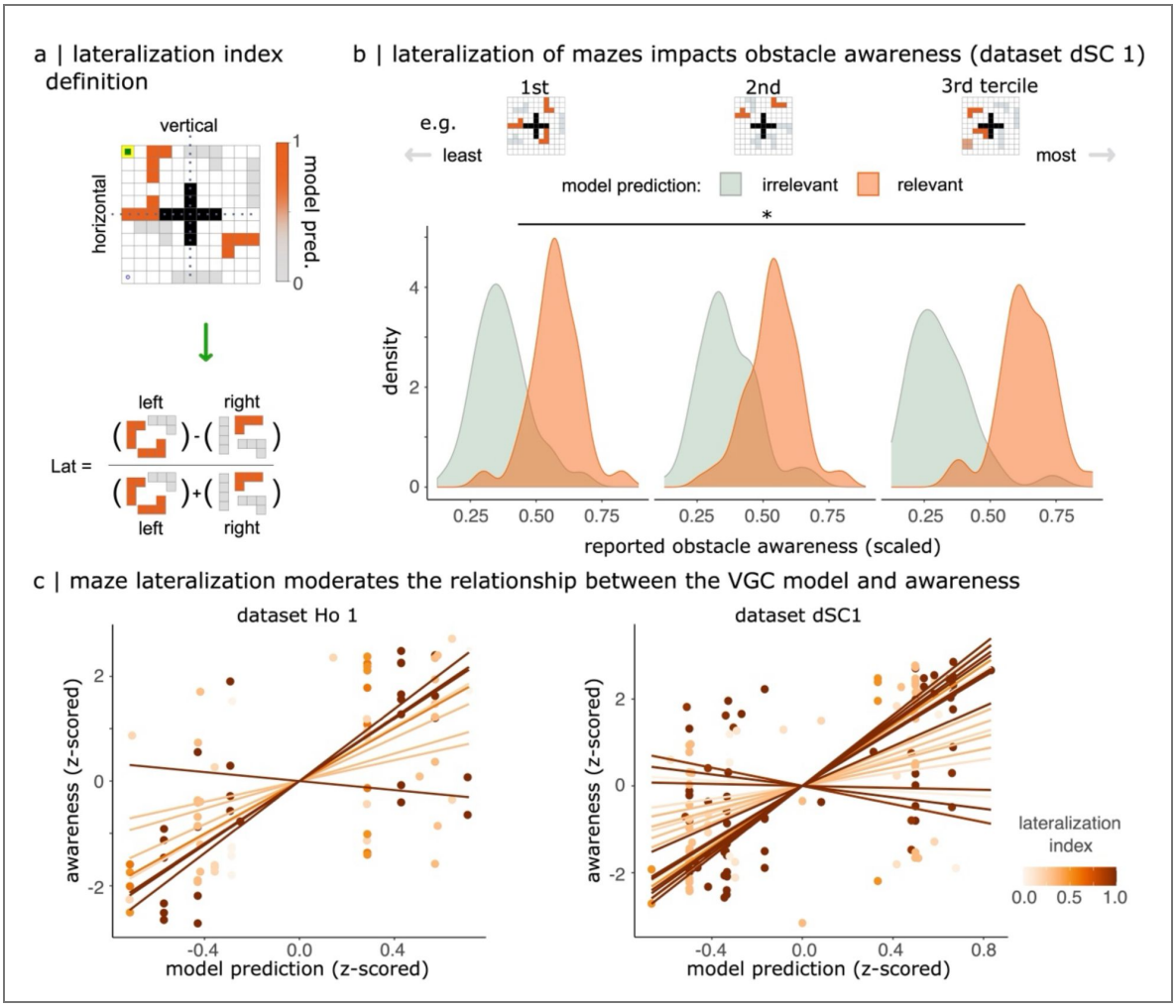

Fig. 2. Lateralization of task-relevant information affects task representations.

(a) For each maze, we computed a vertical meridian lateralization index. This index reflects whether task-relevant information is lateralized to a hemifield. In the example plotted, there is more task-relevant information presented on the left than on the right of the maze, therefore this would correspond to a moderate level of vertical meridian (i.e., left vs right) lateralization. We similarly computed an attention index for the horizontal meridian (i.e., above vs below). (b) Density plots of the reported awareness of obstacles on the basis of whether the value-guided construal (VGC) model predicted them to be task-relevant (≥0.5; in orange) or task-irrelevant (< 0.5; in grey). Note sVGC model predictions for each obstacle were binarized for visualization purposes only. Participants were more likely to be aware of obstacles predicted as task-relevant. We split maze stimuli based into terciles based on the degree to which task-relevant information was presented preferentially to one hemifield (x-axis). The leftmost plots are mazes where task-relevant information is presented on both hemifields. In contrast, the rightmost plot depicts mazes with the largest lateralization. We observed that the awareness reports of participants become increasingly aligned to the VGC model's predictions as lateralization increases. (c) Scatter plot of the effect of maze lateralization on the relationship between the value-guided model and participants' awareness of obstacles. We observed a significant vertical meridian lateralization effect whereby participants' awareness reports were more strongly aligned with the VGC model's predictions when task-relevant information was presented unilaterally in all datasets. Each point represents an obstacle in a maze, and each line represents the model fit for that specific maze stimulus.

data collected in dataset Ho 2(Ho et al., 2022) (p_{FDR}< 0.01; see Table S9). These results indicate that participants' task representations are more closely aligned with the ideal observer (i.e., the original-VGC model) when taskrelevant information is presented unilaterally.

In our new dataset (dSC 1), we designed novel maze stimuli to validate these lateralised effects of attention while addressing some limitations of previous experiments (see Methods). We again observed that lateralization of task-relevant information impacted participants' awareness reports. Participants were less aware of task-irrelevant stimuli on trials where the lateralization of task-relevant information was larger (Figure 2b) and we replicated the moderation effect of information lateralization on the extent to which the original VGC model captured participants' awareness reports ($\beta_{interaction}$ = 0.01, SE = 2.65*10^{-3}, 95% CI [0.01, 0.02], p< 0.001; Figure 2c & Table S10). This effect did not vary significantly as a function of the specific hemifield (i.e., left vs right) in which task-relevant information was presented (β= 0.01, SE= 0.02, 95% CI [-0.03, 0.04], p = 0.738; ΔBIC = 58.30 in favour of the null effect; see table S14).

We note that for three maze stimuli whose model predictions were lateralized there was nevertheless a poor fit to the sVGC model (see Figure 2c, right panel). These outliers correspond to maze stimuli where participants, on average, lateralized their attention to the incorrect hemifield (i.e., the opposite hemifield to that predicted by the sVGC model). In contrast with our observations of consistent and strong attentional effects relative to the vertical meridian, effects relative to the horizontal meridian (superior vs. inferior) were inconsistent across experiments. Specifically, we observed a significant moderation effect in dataset Ho 2 ($\beta_{interaction}$ = 0.01, SE = 2.85*10^{-3}, 95% CI [0.00, 0.01], p_{FDR}< 0.05; see Table S9), but not in dataset Ho 1, and the moderation effect was negative rather than positive in dataset dSC 1 ($\beta_{interaction}$ = -0.01, SE = 2.22*10^{-3}, 95% CI [-0.01, 0.00], p< 0.05). The effect in the Ho 2, but not the dSC1, dataset became insignificant after accounting for nuisance covariates (see Table S15 & S16).

Inter-individual variation in lateralization of attention

Next, we investigated participants' tendency to pay attention to obstacles within a single hemifield (left vs right) regardless of the sVGC model predictions. To do so, we computed an awareness lateralization index (ALI) based on participants' self-reported awareness reports of obstacles on each trial (Figure 3a). Large positive values indicate that participants were preferentially aware of the right hemifield, whereas negative values indicate preferential awareness of the left hemifield. Values close to zero indicate that participants paid attention to both hemifields equally (see Methods for details). We observed that participants' tendency to lateralize their awareness varied greatly across the Ho datasets 1 and 2 (Figure 3b); some participants preferentially paid attention to a single hemifield, regardless of whether the sVGC model predictions were lateralized. For the dSC1 dataset, we observed that on some trials, participants significantly lateralized their awareness (|ALI| > 0.5; Figure 3c) even though the sVGC model predictions were non-lateralized. These findings suggest that participants' tendency to pay attention to a single hemifield may represent an observable inter-individual difference in how they allocate their awareness to form task construals.

To further explore these inter-individual differences, we tested whether participants' tendencies to lateralize their attention to a single hemifield was consistent across trials and maze stimuli. We observed that participants' tendency to lateralize their attention to a single hemifield was similar for left and right lateralized maze stimuli (Spearman ρ= 0.72, Figure 3d). This suggests that participants who preferentially attended to a single hemifield did so regardless of which hemifield they should attend to. More consequentially, the tendency for participants to lateralize their awareness on maze stimuli whose model predictions were also lateralized linearly correlated with participants' tendency to lateralize their attention on non-lateralized maze stimuli (Spearman ρ= 0.88, Figure 3d). Taken together, these findings emphasize that some individuals tend to preferentially attend to a single hemifield when planning. This tendency, importantly, represents an inter-individual difference in how participants allocate their attention across various maze types.

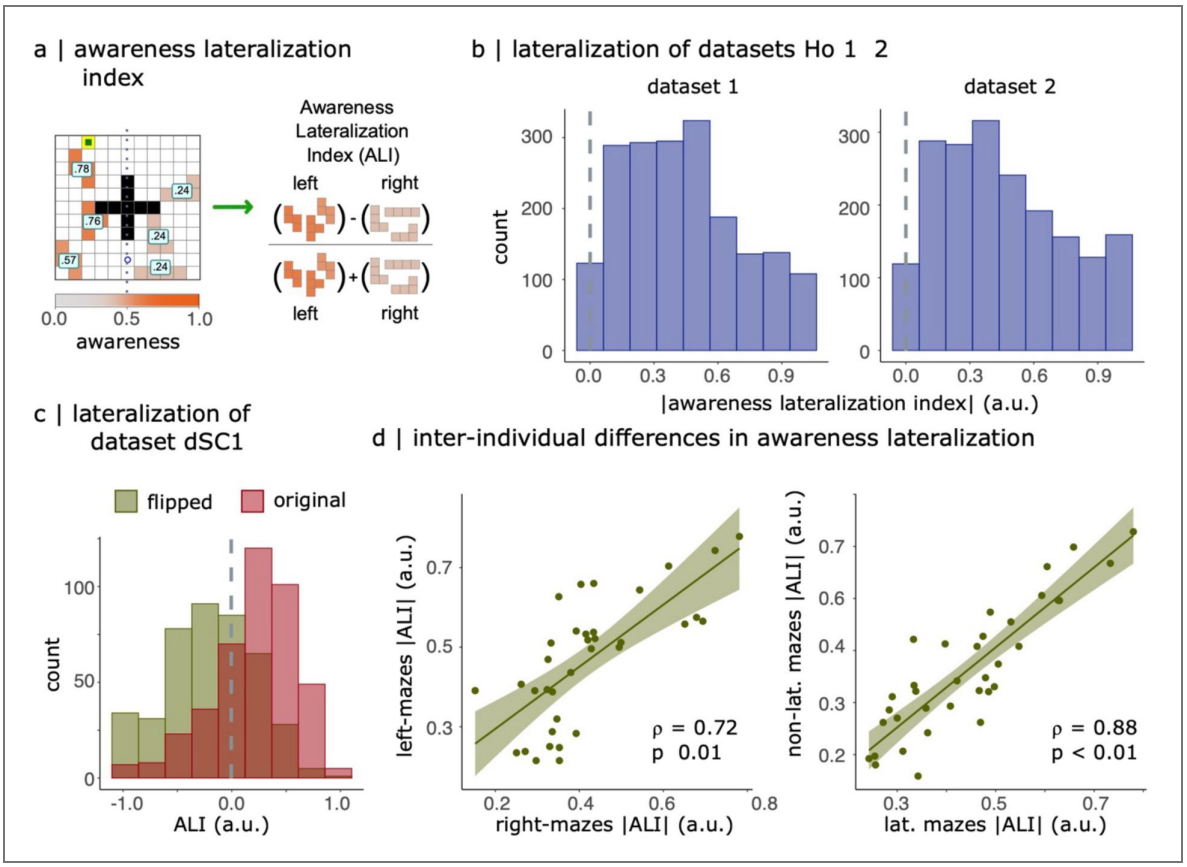

Fig. 3. Inter-individual variation in lateralization of awareness.

(a) For each maze and participant, we computed an awareness lateralization index (ALI). This index reflects the degree to which participants tended to pay attention to obstacles in a single hemifield. In the example plotted, the participant preferentially paid attention to the obstacles presented to the left hemifield regardless of whether they were taskrelevant or task-irrelevant. Note that this lateralization index is based on participants' selfreports, unlike the lateralization index presented in Figure 2 which is based on the sVGC model predictions. (b) Histogram of the ALI of participants across the Ho 1 & 2 datasets. In these experiments, some participants showed substantial lateralization of awareness (ALI > 0.5), despite the maze stimuli for these experiments being—on average—non-lateralized in their sVGC model predictions. (c) Histogram of the ALI of participants for maze stimuli with non-lateralized VGC model predictions. We plot ALI values separately for the original non-lateralized mazes, and the left-right reversed (flipped) mazes separately. Participants on average did not lateralize their awareness. We note, however, that on some trials participants' awareness reports were strongly lateralized, which contrasts with the sVGC model predictions. (d) Scatter plot of participants' tendency to lateralize their attention to either hemifield (i.e., absolute value of ALI). We plot this for mazes with left and right lateralized model predictions (left panel) and for mazes with non-lateralized and lateralized model predictions (right panel). The large linear relationships indicate that participants' tendency to lateralize their awareness is a stable inter-individual difference.

Attentional spotlight model of task representations

Taken together, our results corroborate a critical role for visuospatial attention in constructing task representations. Notably, these filtering effects of attention on value-guided construal are not currently part of the original VGC framework proposed by Ho and colleagues. In what follows we explicitly incorporate the influence of a spotlight of attention into the original VGC model to formulate the spotlight-VGC model(Eriksen & St. James, 1986; Posner, 1980).

To achieve this, we computed the predictions of the existing VGC model for each obstacle's task relevance in a given maze, and averaged these predictions within an attentional spotlight of 3 squares (Figure 4a & S8, see Methods for details). This process yielded novel model predictions, whereby some obstacles which were once predicted as task-irrelevant by the normative sVGC model are now predicted as task-relevant by the attentional spotlight model. We depict the effects of this spatial spotlight in Figure 4a: task-irrelevant stimuli (plotted in grey; see middle left obstacle) neighbouring task-relevant obstacles (plotted in orange) become more task-relevant, whereas task-relevant information becomes less relevant when surrounded by task-irrelevant information (see bottom right orange obstacle). This deviation in model predictions from the normative sVGC model was used to predict participants' awareness reports. We hypothesized that this spotlight-VGC model would predict participants' reports better than the original VGC model, which does not account for spatial attention.

In line with this hypothesis, we observed that the spotlight-VGC model predicted participants' awareness reports better than the original VGC model in all three datasets (dataset Ho 1: ΔBIC= 84.63; Ho 2: ΔBIC= 203.43; dSC 1: ΔBIC= 70.72; see Figure 4b right panel). For dataset dSC 1, we observed a significant improvement in model fit for non-lateralized maze stimuli (ΔBIC= 161.93) but failed to find any improvement when maze stimuli were lateralized (ΔBIC= -42.02; see Figure 4c). These findings dovetail with the previously discussed moderation effects, and suggest that the spotlight-VGC model is particularly useful in improving our ability to explain human behaviour in situations when attentional filtering is more complex.

To further explore inter-individual differences in task construal, we tested whether adjusting the attentional spotlight width to each participant's awareness reports improved the predictions of the attentional spotlight model. To do so, we first determined the width attentional spotlight of each individual in the dSC1 dataset based on lateralized maze stimuli. We then generated person-specific attentional spotlight model predictions for the non-lateralized maze stimuli to avoid overfitting the data (Figure S11). We note that 7 participants had either flat attentional slopes or negative beta coefficients, which prevented the selection of an appropriate attentional spotlight width (see Methods for details). We observed a significant improvement in model fit for the person-specific attentional spotlight model relative to both the group-level attentional spotlight model (ΔBIC= -1487.39) and the normative sVGC model (ΔBIC= -1655.29). While the limited trial numbers per participant in our current dataset warrants caution in interpreting these findings, these findings do encourage further research on inter-individual differences in attentional deployment during planning.

Maze navigation performance

The previous analyses focused on participants' task representations during planning. We next sought to explore links between participants' task representations and maze navigation performance. Participants performed the maze navigation task near-ceiling: they solved 95% of maze stimuli in under 20 seconds, with minimal deviation from the optimal path (i.e., 9 moves or fewer). Notwithstanding this limited variance in task performance, we explored whether participants' task construals may have impacted their navigation speed. To do so, we first regressed out the effects of the sVGC model from participants' awareness reports and used the mean squared residuals for each trial to predict response times (see Methods for details). Surprisingly, we observed a negative relationship between mean squared residual variance and response times (β = -0.31, SE = 0.05, 95% CI [-0.41, -0.21], p< 0.001), indicating that participants were faster on trials where the sVGC model explained less variance in their awareness reports. In other

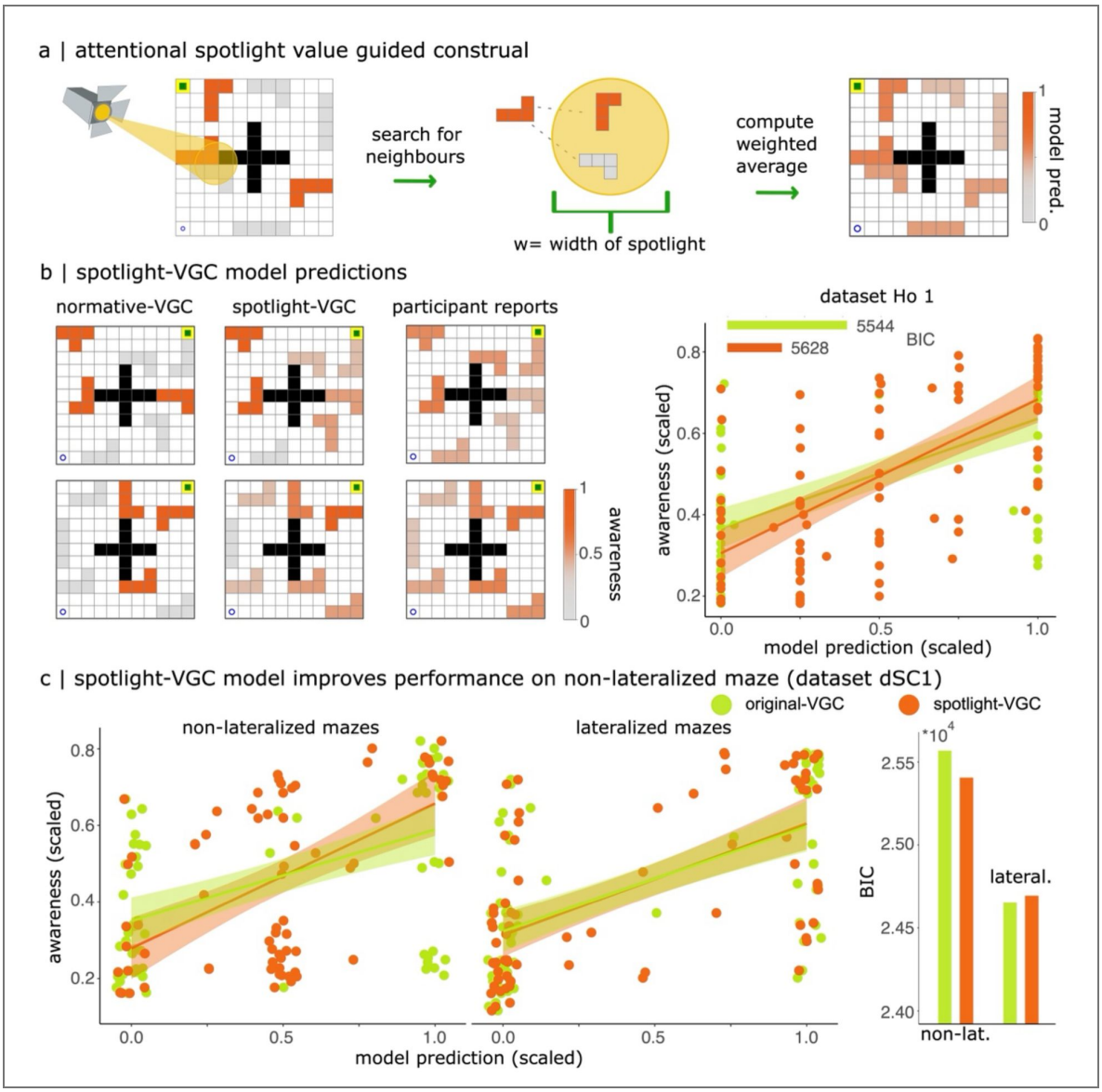

Fig. 4. A VGC model augmented with an attentional spotlight model predicts participants' task representations.

(a) Schematic of the attentional spotlight model. Inspired by the spotlight of attention analogy, we recompute an obstacle's probability of being included in a task representation as the weighted average of its neighbours. We first search for all neighbours of obstacle$_i$ that are w squares away. We then compute $P(\text{Obstacle}_i)$ as the weighted average of obstacle$_i$ and its neighbours. This generates more graded model predictions (far right panel). (b) Left panel: Each row represents a different example maze stimulus. The left column depicts the original VGC model prediction $P(Obstacle_i)$ for every obstacle in the example maze. The middle column shows the attentional-spotlight model prediction for every obstacle. Obstacles that were considered task-relevant (deep orange) in the original model become less important when surrounded by task-irrelevant information (grey obstacles). The right column shows the participants' average awareness of each obstacle in the example mazes. Right panel: Scatter plot of the linear relationship between participants' awareness reports of obstacles and model predictions (original-VGC in green and the spotlight-VGC model in orange) for dataset Ho 1. The latter fits participants' reports better than the original VGC model. (c) Scatter plot of the linear relationship between participants' awareness reports of obstacles and model predictions (original VGC in green and the spotlight-VGC model in orange) for dataset dSC 1 separately for non-lateralized (left panel) and lateralized mazes (right panel). Although both models fit participants' awareness reports better for lateralized mazes, the advantage of the spotlight model over the original model (better model fit / lower BIC) was observed only in non-lateralized mazes.

words, trials in which participants deviated *more* from the sVGC model predictions were solved faster. We note that one reason for this may be the strong influence of the lateralisation effect on navigation performance (see paragraph below), which itself is not part of the sVGC model prediction.

We then explored whether participant performance differed between lateralised and non-lateralised mazes. Here, we reasoned that the initial phase of lateralised attentional selection would lead to lateralised mazes being easier to navigate than non-lateralised ones. Consistent with this hypothesis, participants were faster (β = -0.04, SE = $5.91*10^{-3}$, 95% CI [-0.06, -0.03], $p < 0.001$) and followed the optimal path more closely (β = -0.59, SE = 0.09, 95% CI [-0.78, -0.40], $p< 0.001$) when maze stimuli were more lateralized.

Sensitivity analyses

We conducted a series of control analyses to verify the robustness of our experimental results. First, we verified that the spatial proximity effect (Figure 1b) was not driven solely by the spatial smoothness of participants' awareness reports by conducting null permutation tests (grey line, Figure 1b). For each maze stimulus, we permuted the rank of the neighbouring obstacles. We then fit the same linear model to assess the effect of spatial context on task representations. This procedure was repeated 1000 times to generate a null distribution of beta coefficients. The resulting null distribution showed no discernible effect of spatial context. Second, we used null permutation tests to verify that the improved fit of the spotlight-VGC model was not driven by greater spatial smoothness of the model predictions (see Supplemental Figure S12). Third, we assessed whether nuisance covariates could explain the moderation effects we observed. Specifically, we added the distance from the goal, starting location, center walls, and fixation as nuisance covariates in our hierarchical regression models. Maze lateralization remained a significant moderator of the relationship between the original VGC model and participants' awareness reports after controlling for these covariates (see Table S11-13). This was not the case, however, for lateralization effects along the horizontal meridian (see Table S15- 16).

Fourth, we sought to verify that the lateralization effects we observed were not driven by a change in eye movement patterns. For dataset dSC 1, we continuously tracked the position of participants' gaze. We explicitly instructed participants to maintain central fixation while planning (see Methods for details) and removed the obstacles from the screen after 6 seconds. This allowed us to verify that greater awareness of obstacles was not driven by longer fixation times. We confirmed that participants maintained central fixation on both lateralized and non-lateralized maze stimuli in most trials (see Figure S13). Excluding trials where participants exhibited excessive eye movements during planning, we continued to observe qualitatively similar lateralization effects (see Figure S15 and Table S17).

Finally, we examined the convergent validity of participants' awareness reports by reanalyzing the memory recall data reported in Ho and colleagues' experiment (Ho et al., 2022). We reasoned that participants should demonstrate similar task representations regardless of the measure used to probe the construal. In line with this prediction, we observed that the obstacle awareness reports and memory/hover measures were strikingly correlated within three independent samples of participants (Spearman ρ = 0.86 between memory accuracy and awareness; ρ = 0.86 between confidence in memory and awareness; ρ = 0.76 between the probability of hovering over the obstacle and awareness; ρ = 0.65 between the duration of the mouse hovering and awareness; see Tables S18 and S19).

Discussion

Searching for a solution in a complex multi-step task is challenging. Recent computational work suggests that humans overcome this challenge by constructing simplified perceptual representations of their environment. In the present study, we reveal a role for visuospatial attention in constructing these simplified perceptual representations.

Participants' task-representations are informed by visuospatial attention

We provide several lines of evidence for the critical role of visuospatial attention in constructing task representations. First, we observed a significant effect of the spatial context in which information is presented. Participants were less likely to incorporate taskrelevant information into their construal when it was surrounded by task-irrelevant information. These effects mirror perceptual crowding effects(Liu et al., 2009; Whitney & Levi, 2011) which reveal that attention spills over to distractors presented alongside taskrelevant stimuli when presented in close spatial proximity. Second, we observed that participants incorporated task-relevant information into their task representations more frequently when relevant obstacles were grouped together within the same hemifield (Figure 2). Participants' task representations in such settings were more closely aligned with an ideal observer model – suggesting that the natural contours of visuospatial attention interact with the capacity of observers to form efficient task representations.

Incorporating attention into a model of value-guided construal

The VGC model articulates how an ideal decision-maker should represent an environment while balancing complexity and utility(Ho et al., 2022). We developed an extension of this model that accounts for the effects of spatial attention on planning. Our model, inspired by the analogy of a spotlight of attention(Carrasco, 2011; Posner, 1980), provides a better fit to participants' awareness reports than the original VGC model (Figure 4). This improved model fit was most evident for mazes where task-relevant information was presented to both hemifields (Figure 3c), suggesting the augmented model is helpful in explaining behaviour in contexts where attentional selection is more complex. These deviations from the original VGC model, therefore provide a useful benchmark to compare human performance and offer insights into natural constraints on human cognition. For instance, we demonstrate that spatial context biases whether information is to be included or excluded from a representation of the environment. These effects may reflect inductive biases in humans who have learned and evolved to select information from real-world environments where obstacles tend to be grouped together in visual scenes(Kaiser et al., 2019; Peelen & Kastner, 2014). However, it is plausible that these inductive biases on value-guided construal may themselves be learnt, and vary according to other environmental demands and contexts which impose systematic regularities on useful task representations (e.g., attending preferentially to intersections when planning on the Tube). Future research can explore the flexibility of participants' task representations across environmental contexts, and ask how these inductive biases are acquired.

Inter-individual differences in attention

We also observed considerable inter-individual differences in attentional effects across participants (Figure 1c). While some participants were strongly influenced by the spatial context of neighbouring stimuli, others showed more limited evidence for an attentional effect (Figure 1b). Inter-individual differences in attention predicted the sparsity of participants' simplified representations: participants with larger attention effects exhibited sparser representations. Moreover, these inter-individual differences in effects of spatial proximity could be incorporated into the attentional spotlight model by varying the width of the spotlight, resulting in better model predictions.

Beyond these spatial proximity effects, we also observed that participants varied in their tendency to lateralize their attention to a single hemifield (Figure 3). This tendency was observed across all three datasets, including on maze stimuli whose value-guided model predictions were not lateralized. This suggests that although a strategy of allocating attention is sub-optimal for these maze stimuli, some individuals preferentially attend to a single hemifield in a heuristic-like fashion. This tendency to attend to a single hemifield was a robust inter-individual difference across maze stimuli (Figure 3d), and dovetails with individual-level variation in spatial proximity effects. Taken together, these findings offer novel insights into how people vary in the

ways they allocate spatial attention to solve complex problems. Future research could explore how these individual differences constrain performance on other tasks that require planning and search in highdimensional spaces.

Mental representations and task performance

We observed that participants were faster and deviated less from the optimal path on maze stimuli that were lateralized. This effect is not predicted by the original sVGC model but dovetails with the interpretation that early visuospatial attention operates as an inductive bias to guide the formation of simplified task representations. Surprisingly, we also observed that participants were faster to navigate mazes on trials where their simplified task representation deviated from the sVGC model prediction. We interpret this seemingly contradictory finding in the following way: there are several factors beyond the sVGC model – including, for instance, maze lateralisation – that predict both construal and performance on the maze navigation task. Further work is needed to understand how inductive biases such as lateralisation shape both construal and performance, and the real-world benefits that such strategies might afford for naturalistic stimuli.

Planning and Consciousness

Our experiments investigate the connection between planning and participants' reports of their awareness of features of the task environment. The results may therefore be relevant to understanding the functions of conscious experience. While an intimate connection between attention and consciousness is widely recognized(Cohen et al., 2012; Koch & Tsuchiya, 2007; V. A. F. Lamme, 2003; Tsuchiya & Koch, 2014), there is less work explicitly considering the connection between planning and consciousness(Fleming & Michel, n.d.; MacIver & Finlay, 2022). However, there are several reasons why the kind of planning at work in our experiments is likely to require the task to be represented consciously.

As we mentioned at the outset, simplified representations reduce the computational burden of planning in a branching multi-step task space. The same consideration suggests that planning should be based on conscious, rather than unconscious, representations(Shea & Frith, 2016). Initial stages of perceptual processing can carry information about a range of different and incompatible possibilities at once, for example, a probability distribution across a range of possible orientations of a line. The probabilistic representation attaches some probability (or probability density) to many different possibilities. There is, of course, a certain burden in integrating and weighing probabilistic information of this kind, for which the brain is thought to deploy various solutions (e.g. approximate Bayesian inference). These initial stages of perceptual inference are typically thought to be unconscious. However, forward planning from multiple possibilities in a branching task space rapidly becomes intractable as the combinatorial possibilities explode. Consciousness, by contrast, provides a much sharper representation of the current state, from which planning can proceed forward(Block, 2018; Stocker & Simoncelli, 2007).

Given the computational cost of running through and comparing many potential multi-step action sequences, it makes sense to base that process on a reliable estimate of the current world state. While it is doubtless useful to produce some kinds of unlearned and habitual action very rapidly at the first hint of information, for example of the presence of a predator, with multi-step forward planning it makes sense to integrate information from more sensory modalities and across a longer timescale before then committing to using a representation as the basis for planning. This again suggests that conscious representations, which are known to integrate information across modalities and time(Bekinschtein et al., 2009; Deroy et al., 2016; Herzog et al., 2020; Mudrik et al., 2014; Strauss et al., 2015), are perfectly suited to the functional needs of this kind of planning task. Furthermore, planning depends on both facts and values. Potential actions are assessed based on the expected value of outcomes. The role of value was captured, in our studies, by an extension of the VGC (value-guided construal) model. Consciousness is thought to facilitate the integration of different sources of value(Braver et al., 2014; Dickinson & Balleine, 2010; Dung, 2022).

The task and associated computational model thus offer a flexible tool for characterising the computations by which conscious representations influence decisions and actions. Future work could tell us more about the way bottom-up attention-driven inputs and taskbased value jointly influence what information reaches conscious experience. This provides a novel ecologically-valid probe of the connections between attention, consciousness, and decision-making that does not require the explicit labelling of taskrelevant stimuli. Neural (e.g. M/EEG) data collected while participants plan could help understand the timescale and computational steps that lead to the formation of a conscious task representation. Modifications of the paradigm would also be suited to exposing the way non-consciously-presented cues do and do not influence the way participants plan.

Methodological Considerations and Future Directions

We close by reflecting on opportunities for further work in this area. First, an important next step is to explore the process by which task representations are formed, and how inductive biases might affect the process of task construal. The sVGC model is a normative model of the optimal task representation. Since it's construction involves an exhaustive calculation over possible paths, it is not a plausible basis for a model of the psychological process by which participants actually construct task representations. More recently a process model of task construal has been proposed, the Just in Time model (JIT). The hypothesis of the JIT model is that participants' task representations are built up over time by iteratively simulating possible paths through the maze, affording insight into the construal process (Chen et al., 2026). In future work, it would be of interest to ask whether the attentional effects we observe in our experiments could be meshed with a dynamic JIT account of construal. We speculate that visuospatial attention may operate as an early filter, limiting the space of potential construals based on coarse spatial features of the environment, constraining a dynamic selection of obstacles. Brain imaging techniques with high time resolution, such as M/EEG, may be able to shed further light on how task representations are formed as participants plan.

Second, in the current work we were unable to distinguish whether these attentional effects are driven by a fixed spotlight of attention, or whether attention operates akin to a zoom lens, shifting the 'width' of the focus of attention according to the task demands (Eriksen & St. James, 1986; Müller et al., 2003; Schad & Engbert, 2012). The latter view would be consistent with growth-cone models of attention in which the focus of attention expands and contracts in accordance with task demands, mirroring the various receptive field sizes in the visual hierarchy (Pooresmaeili et al., 2014; Pooresmaeili & Roelfsema, 2014). In partial support of this idea, we found significant inter-individual differences in the width of participants' attentional spotlight (Figure S11). It is also possible that attention is deployed within or along parts of obstacles, rather than on entire obstacles. Future work using naturalistic measures of eye movements may be able to address these questions.

Third, while we observed clear lateralization effects along the vertical meridian (i.e., left vs right hemifield), effects along the horizontal meridian were less clear (i.e., above vs below; see Table S15-16). One potential explanation of this asymmetry is the retinotopic organization of the cortex, in which spatially adjacent stimuli can be retinotopically distant if presented on the opposite side of the vertical (but not horizontal) meridian, facilitating distractor inhibition. Importantly, while the visuospatial attention effects observed in the Ho 1 and 2 datasets are likely driven by both covert and overt shifts in attention, the findings presented in experiment 3 (i.e., dSC1 dataset) rule out the contribution of overt shifts in attention through the use of eye tracking (see Figure S13-14)(Carrasco, 2011; Pooresmaeili & Roelfsema, 2014).

Fourth, it will also be necessary to elaborate on how bottom-up and top-down aspects of attentional selection are combined to guide complex task representations and plans. Foundational questions remain unanswered, for instance: can multiple spatial locations be preferentially selected at once, i.e. are there multiple spotlights (Awh & Pashler, 2000; McMains & Somers, 2004; Pylyshyn & Storm, 1988; Shaw & Shaw, 1977)? There is also discourse on how spatial attention may move from one location to another: are the intervening visual regions between

attended locations similarly selected (Dubois et al., 2009; Kr & Np, 1999; McMains & Somers, 2004, 2005)? Our findings tentatively suggest that individuals are able to attend to disparate spatial regions to form sparse task representations, yet there is substantial variability in how individuals orient their attention during the task. The present paradigm and computational modelling, in conjunction with carefully designed stimuli, may help resolve these outstanding questions.

Finally, our present study focused on studying mental representations for planning in the context of a navigation task. Whether these effects hold across other forms of planning, including planning over abstract spaces, remains to be demonstrated. An important next step to further our understanding of task representations would be to extend the current paradigm to other forms of planning and more naturalistic tasks, such as navigating immersive virtual reality (VR) environments, planning over cognitive rather than perceptual representations (e.g.. planning over an abstract space), or internally-guided planning based on working memory. In this spirit, recent work has applied the VGC model to a physical reasoning task in which participants were asked to predict the trajectory of a blue ball (Chen et al., 2026). Future work could also profitably examine the relevance of visuospatial attention for the navigation process itself in this task. While our present findings speak to how individuals perceive the maze while planning, it remains unclear how attention is deployed during navigation along a path, such as how object-based attention progressively spreads along trajectories in time and space (Pooresmaeili & Roelfsema, 2014; Wong & Scholl, 2024).

Conclusions

Complex daily decisions require a decision-maker to arbitrate over countless potential multi-step actions and their outcomes, making searching for a solution difficult. We shed light on how this is achieved by clarifying the role of visuospatial attention in forming simplified perceptual representations to aid in planning. We build on previous work on the effect of task relevance and develop a computational model which explicitly incorporates the role of attention in value-guided construal. Our model bridges the literature on perception, attention and computational models of planning to provide a more complete computational account of human cognition. We believe the results of this paper can inform future research on a comprehensive theory of human cognition and inspire novel biologically-informed intelligent algorithms.

Methods

Experimental Task

To test our hypotheses we relied on a previously established mazenavigation task where participants are asked to move a circle avatar from a starting location to a goal using the arrow keys(Ho et al., 2022). Each maze consisted of an 11×11 grid with blue obstacles (7 obstacles in datasets Ho 1 & 2, and 6 obstacles in dataset dSC 1), and black central walls arranged in the shape of a fixation cross. Each trial began with a fixation cross (center walls), after which participants were prompted to navigate to the goal. Experiments differed in terms of i) the mazes participants navigated, ii) whether the obstacles were presented before or during the execution of the plan, and iii) what the participant reported.

We reanalyzed the data of Ho and colleagues' experiments 1 and 2 for the present study. In experiment 1 (i.e., dataset Ho 1), participants were presented with the obstacles throughout the trial. At the end of each trial, participants were asked to rate "How aware of the highlighted obstacle were you at any point?" using a nine-point scale. In experiment 2 (i.e., dataset Ho 2), participants were similarly asked to rate their awareness of the various obstacles but were required to plan their solution before they began to solve the maze. See (Ho et al., 2022) for details concerning the experimental procedures.

We did not reanalyze the results of the fourth experiment by Ho and colleagues. In this experiment, participants were not presented with all the information (i.e., obstacles) at once to solve the maze. Instead, they discovered obstacles by hovering over them with a cursor.

To further test the effects of attention on task representations, we designed a novel set of maze stimuli. This consisted of 12 mazes with task-relevant obstacles lateralized to a hemifield (left or right) and 12 non-lateralized stimuli. Each maze consisted of six obstacles, three on each hemifield, none of which crossed the veridical meridian. This ensured that there were an equal number of obstacles for computing the lateralization index (see below). Maze stimuli of both sets were equated on several nuisance covariates (see Supplemental Table S1). Maze stimuli were vertically and horizontally reversed (i.e., left-right flipped) such that participants could not predict the location of the start or goal location. This resulted in four potential orientations of each maze across all 24 mazes, 96 trials in total.

The design of the in-person experiment (i.e., dataset dSC 1) closely followed the second experiment of Ho and colleagues(Ho et al., 2022). On every trial, participants were presented with a maze stimulus for 6 seconds, over which they were required to plan. The maze stimulus was offset, and participants were required to solve the maze after a one-second delay. On every trial, participants reported on their task representations using a nine-point awareness scale.

Participants

For datasets Ho 1 & 2, participants completed the task online on Prolific. In dataset Ho 1, 194 participants completed submissions, 161 of whom were included in the final sample after exclusions. In dataset Ho 2, 188 participants completed submissions, 162 of whom were included. Participants were excluded from analyses based on preregistered exclusion criteria as detailed in (Ho et al., 2022). In short, participants were excluded if 20% or more of their trials were removed based on reaction times, or if they failed 2 out of 3 comprehension checks.

For dataset dSC 1, 35 participants (mean age = 23.14, SD = 5.35; 12 male) completed an in-person eye-tracking experiment (see Eye-tracking acquisition). None of the participants were excluded from the data analysis. We excluded trials where participants' reaction times were longer than 20 seconds, or where participants deviated more than nine moves from the optimal path (which reflected 3SD above the mean).

Ethics

All procedures were approved by the University College London ethics committee and adhered to the Declaration of Helsinki. Informed consent was obtained from each participant prior to each experiment.

VGC model

We fit the previously described VGC model to our maze stimuli(Ho et al., 2022). Briefly, this model computes the optimal simplified task representation such that it maximizes the utility of the representation while also minimizing the cognitive cost of keeping information in mind. This model assumes that a decision-maker combines a subset of cause-effect relationships to represent their environment in aid of planning. For every possible construal, the model computes the value of a representation:

$$\mathrm{VOR}\left(c\right) = U\left(\pi_c\right) - C\left(c\right).$$

where U(π_c is the utility of a construed plan π_c, and *C(c)* represents the cost of keeping that information in mind.

A task representation is selected according to a SoftMax decision rule. We then compute a marginalized probability for each obstacle being included within a construal,

$$P\left(\text{Obstacle}_i\right) = \sum \|\varphi_{Obstaclei} \in c\| \, P\left(c\right),$$

where $\phi_{obstaclei}$ is the cause-effect relationship for obstacle$_i$, P(c) is the probability that the task representation is selected, and ‖X‖ is a statement which evaluates to 1 if X is true, and 0 when X is false. We use the values of $P(\text{Obstacle}_i)$ for every obstacle in a maze to predict participants' awareness reports. See (Ho et al., 2022) for a detailed explanation of the computational model.

We focused our analyses on the *static* version of the VGC model (i.e., sVGC), whereby task representations are assumed to remain stable across planning. Our choice was informed by the design of the experiment where participants were required to plan over all obstacles at once.

Spatial proximity effects

To examine how the spatial context of information influences participants' awareness reports, we ran a hierarchical linear regression model. First, for every obstacle in every maze, we rank-ordered all other obstacles based on spatial proximity. That is, the participant's awareness report of the closest item to obstacle$_i$ on the trial was used as a predictor of the participant's report of obstacle$_i$ in a hierarchical linear regression model. This yielded a regression model with 6 regression coefficients predicting participants' awareness reports based on spatial proximity:

$$\begin{aligned}\text{report}_{\text{obstacle i}} &= \beta_0 + \beta_1 * \text{report}_{\text{obstacle 1}} + \beta_2 \\ * \text{report}_{\text{obstacle 2}} \ldots &+ \beta_{6*}\text{report}_{\text{obstacle 6}} \\ &+ (1|\text{MazeID}) + (1|\text{ParticipantID}) + \varepsilon\end{aligned}$$

where *(1 | MazeID)* and *(1 | ParticipantID)* are random intercepts of each maze and participant, respectively, and β_1 reflects the contribution of the closest obstacle to obstacle$_i$. We interpret any significant effects in this model as the influence of neighbouring stimuli on participants' representations. We also fit the above hierarchical linear regression model for each participant separately. We report these individual beta coefficients in Figure 1b.

To ensure that the above spatial proximity effects were not driven by the VGC model predictions, we regressed out the effects of VCG model predictions from participants' awareness reports, and used the residuals of the model as the dependent variable in a second regression where we similarly predicted the effects of neighbouring stimuli on representations.

We verified that these effects were not explained by the spatial smoothness of our data by conducting 1000 spatial null permutations. For every iteration, we permuted the mapping between each obstacle in a maze and their spatial location maintaining the number of neighbouring obstacles for every trial. We fit a hierarchical linear regression model using this permuted data and built a distribution of null beta coefficients to compare to our observed effects.

The sparsity of task representations

We sought to test the relationship between i) inter-individual differences in attention effects and ii) the sparsity of task representations. First, we estimated the magnitude of each person's attention effect by fitting a linear slope to the beta coefficients obtained (see Spatial proximity effects). A participant with a large negative slope, therefore, showed a larger effect of neighbouring obstacles on their representation. Second, we operationalized the sparsity of participants' simplified representation as the variance of their awareness reports. A participant with a sparse representation shows a high variance in their awareness of different obstacles in a given maze. Last, we tested the linear monotonic relationship between the sparsity of participants' representations and the attention effects using Spearman correlation.

Lateralization index

To test the effects of lateralization of task-relevant stimuli on participants' awareness reports, we developed a lateralized index of task-relevance inspired by the alpha-power attention literature(Ghafari et al., 2024; Keefe & Störmer, 2021; Vollebregt et al., 2015). We divided each maze into a right and left hemifield and computed the ratio of task-relevant obstacles on both sides:

$$Lat.index = \frac{\sum sVGC_{left} - \sum sVGC_{right}}{\sum sVGC_{left} + \sum sVGC_{right}}$$

where sVGC is the model's prediction of each obstacle task-relevance for that maze. Note obstacles only with a majority of its blocks within a single hemifield were considered (3 or more squares). This yielded an index of task-relevance lateralization for each maze stimulus. We repeated the above procedure to obtain an index of task-relevance lateralization for the horizontal meridian (superior vs inferior hemifield).

We tested whether the lateralization index moderated the relationship between the value-guided model predictions (sVGC) and participants' awareness reports using a hierarchical linear regression model.

$$report = \beta_0 + \beta_1 * sVGC + \beta_2 * Lat + \beta_3 * Lat * VGC + (1|Maze_{ID} + Participant_{ID})$$

where β_3 represents the interaction between the VGC model predictions and the lateralization index.

Inter-individual differences in lateralization of awareness reports

To examine inter individual differences in participants tendency to lateralize their attention to one henifiled (i.e., left vs right) while planning, we computed an awareness lateralization index (ALI) based on participants reports. For each trial, we compute the ratio between participant's awareness of obstacles presented on the right vs left hemifield for every trial:

$$Awareness\ Lat.index = \frac{\sum Awareness\ Report_{left} - \sum wareness\ Report_{right}}{\sum wareness\ Report_{left} + \sum wareness\ Report_{right}}$$

where negative values of ALI indicate that participants preferentially paid attention to obstacles presented on the left hemifield. We report the average absolute value of the ALI across participants for the Ho datasets 1 and 2. For the dSC1 dataset we computed the ALI for every participant for lateralized and non-lateralized maze stimuli. We compute the Spearman correlation between participants' tendency to lateralize their attention on non-lateralized mazes and lateralized mazes. A large correlation indicates that participants' tendency to lateralize their attention to a hemifield was consistent across both maze types.

Spotlight-VGC model

Inspired by previous literature comparing visuospatial attention to a spotlight that moves across the visual field, we developed an extension of the VGC model to account for the effects of attentional selection in forming task representations.

To do this, we recomputed the *P*(Obstacle$_i$) as a weighted average of its neighbours. We computed the distance between every obstacle in the maze, and searched for obstacles with neighbours within 3 squares (Manhattan distance) away from obstacle$_i$. We fixed the 'width' of the attentional spotlight to a distance of 3 squares based on the observation that the two neighbouring obstacles positively predicted the awareness of a probe. We observed that the mean and median distance

between neighbouring obstacles of the 2nd rank (i.e., second closest) was 3 squares away for all mazes (Figure S15). We therefore opted to fix the value of the attention spotlight to 3 squares based on these observations. Future work utilizing this model should consider the statistics of their maze stimuli when deciding on the 'width' of the attentional spotlight. Neighbouring obstacles that fell within the attention spotlight were averaged as follows:

$$P\left(\text{Obstacle}_i\right) = \frac{P\left(\text{Obstacle}_i\right) + mean\left(P\left(\text{Obstacle}_n\right)\right)}{2}$$

where n is the number of obstacles that fall within the width of the attentional *spotlight* (i.e., neighbouring items). We repeat this procedure for all obstacles within each maze. If an obstacle did not have any neighbours, then the value of $P(\text{Obstacle}_i)$ remained identical to the value of original VGC model.

We used the outputs of the attention spotlight model in a hierarchical linear regression to predict participants' awareness reports, where we included participant and maze random intercepts:

$$report = \beta_0 + \beta_1 * At.Sp.Model + (1|Maze_{ID} + Participant_{ID})$$

All linear regression models were fit with the *lmer* package in R.

Personalized spotlight model

To assess inter-individual differences in the width of the attentional spotlight, we aimed to test whether person-specific attentional spotlight model predictions outperformed a constant a model where we held the width of the attentional spotlight constant. To do so, we first used the beta coefficients obtained for each participant from the spatial proximity effects model. We then thresholded the betas and recorded for each participant at which rank their beta coefficient dropped below 0.05 (i.e., a small effect). We then used the median distance between obstacles (i.e., se Figure S15) of this rank as the width of the attentional spotlight for each participant. This resulted in a majority of participants with an attentional spotlight value of 3 and 4 (Figure S11). We note that 7 participants were excluded from these analyses: these reflect participants with flat spatial proximity slopes or a negative beta coefficient for the first rank. We then adjusted the attentional spotlight model predictions according to each participant's width, and used these model predictions in a hierarchical linear model to predict participants' awareness reports.

Null permutations

To ensure that the improved model fit of the attentional spotlight model was not driven by the spatial smoothness of our data, we conducted a series of control analyses where we permuted the model predictions within mazes.

To do so, we re-assigned the $P(\text{Obstacle}_i)$ of each obstacle in a given maze to a random item such that each obstacle was given a new model prediction. This permutation procedure maintains the distribution of $P(\text{Obstacle}_i)$ across obstacles for each maze, while randomizing the location of task-relevant information. We repeated this procedure for each maze separately. We then used these random model predictions to predict participants' reports using the same hierarchical linear model described in Spotlight-VGC model. We repeated this procedure 1000 times to generate a null distribution of beta coefficients. We compared the observed beta value for the spotlight-VGC model against this distribution. We note that averaging neighbouring obstacles before or after the permutation of the model predictions qualitatively yielded the same result.

Eye-tracking acquisition

For dataset dSC1, participants completed the computer task while their eye-position and pupil size were monitored using an EyeLink 1000 Plus eye tracker at 1000Hz (SR Research, Osgoode, ON). Participants were seated comfortably in a dimly lit room in front of a 24-inch monitor set to the resolution of 1,920 × 1,080 pixels at 60 Hz. Participants were positioned 60 centimetres away from

the screen and rested their heads on a mount. Stimuli were presented on MATLAB 2019a using psychtoolbox (3.0.16), synchronized with the eye tracker. Before the start of the experiment, participants completed a standard 5-point calibration procedure. Drift correction was applied after every block.

Eye-tracking preprocessing & analysis

Eye-tracking data were preprocessed with the PuPL toolbox in MATLAB(Kinley & Levy, 2022). Impossible data points (i.e., gaze outside the screen's bounds) were removed, in addition to data 50ms before and 150 ms after eye blinks (identified by pupillometry noise(Hershman et al., 2018)). Segments of missing gaze position, up to 400ms long, were interpolated using cubic splines. We analyzed eye position data between -1000 ms and 6000 ms around the presentation of the maze, which corresponds to the planning window and the one second prior to planning. To verify that participants did not move their eyes more frequently during planning for lateralized mazes, we computed the standard deviation of eye position along the X-axis for each trial. We compared the fluctuations across lateralized and non-lateralized trials with a two-sample t-test. To verify the robustness of our behavioural effects, we identified and removed from further analysis trials where participants' eye position exceeded two squares away from fixation.

Navigation performance & mental representations

To examine how mental representations related to the navigation task we examined whether the lateralization of maze stimuli related to the time it took participants to navigate each maze (i.e., their response time). We ran a hierarchical linear regression model where we predicted the response time of each trial from the optimal number of moves it takes to solve that maze and the lateralization index as fixed effects, and participant IDs as random effects. We repeated this analyses to predict the deviation from the optimal path (i.e., the difference between the optimal number of moves and the total moves for a given trial).

We then explored whether participants were faster at navigating mazes in which the sVGC model more closely aligned with participants' awareness reports. To do so, we first regressed out the effects of the sVGC model predictions from the awareness reports of participants using a hierarchical linear regression model. We then took the mean squared residuals of each trial from this regression model and used that as a predictor in a second regression model. The mean squared residual of each trial represents the unexplained variance in participants awareness reports after accounting for the sVGC model, where larger numbers indicate more unexplained variance. Here, we predicted participants' response times from the optimal number of moves and the mean squared residuals as fixed effects, and participants' IDs as random effects.

Data availability

All in-house code used for data analysis and visualization is available on GitHub *https://github.com/jasondsc/ConsciousDetour*. The reanalyzed data presented herein are available from *https://www.nature.com/articles/s41586-022-04743-9*. The data from experiment 3 are available from *https://osf.io/sa6vf/*.

Acknowledgements

J.d.S.C. is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) postdoctoral fellowship program. S.M.F. is a CIFAR Fellow in the Brain, Mind and Consciousness Program and is supported by a Wellcome/Royal Society Sir Henry Dale Fellowship [206648/Z/17/Z] and UKRI under the UK government's Horizon Europe funding guarantee (selected as ERC Consolidator, grant number 101043666).

Additional files

Supplementary materials.

Additional information

Funding

Funder	Grant reference number	Author
Wellcome	https://doi.org/10.35802/206648	Stephen M Fleming
EC | European Research Council (ERC)	101043666	Stephen M Fleming

Author ORCID iDs

Jason da Silva Castanheira: https://orcid.org/0000-0002-7022-9009
Stephen M Fleming: https://orcid.org/0000-0003-0233-4891

References

Awh E., Pashler H. (2000) Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception and Performance* **26**:834-846 https://doi.org/10.1037/0096-1523.26.2.834

Bagherzadeh Y., Baldauf D., Pantazis D., Desimone R. (2020) Alpha Synchrony and the Neurofeedback Control of Spatial Attention. *Neuron* **105**:577–587.e5 https://doi.org/10.1016/j.neuron.2019.11.001 | PubMed

Barnett B., Andersen L. M., Fleming S. M., Dijkstra N. (2024) Identifying content invariant neural signatures of perceptual vividness. *PNAS Nexus* **3**:pgae061 https://doi.org/10.1093/pnasnexus/pgae061 | PubMed

Bekinschtein T. A., Shalom D. E., Forcato C., Herrera M., Coleman M. R., Manes F. F., Sigman M. (2009) Classical conditioning in the vegetative and minimally conscious state. *Nature Neuroscience* **12**:1343-1349 https://doi.org/10.1038/nn.2391 | PubMed

Block N. (2018) If perception is probabilistic, why does it not seem probabilistic?. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**:20170341 https://doi.org/10.1098/rstb.2017.0341 | PubMed

Braver T. S., Krug M. K., Chiew K. S., Kool W., Westbrook J. A., Clement N. J., Adcock R. A., Barch D. M., Botvinick M. M., Carver C. S., *et al.* (2014) Mechanisms of motivation-cognition interaction: Challenges and opportunities. *Cognitive, Affective & Behavioral Neuroscience* **14**:443-472 https://doi.org/10.3758/s13415-014-0300-0 | PubMed

Breslow L. A., Aha D. W. (1997) Simplifying decision trees: A survey. *The Knowledge Engineering Review* **12**:1-40 https://doi.org/10.1017/S0269888997000015

Broadbent D. E. (1958) *Perception and Communication* Elsevier Science & Technology.

Callaway F., van Opheusden B., Gul S., Das P., Krueger P. M., Griffiths T. L., Lieder F. (2022) Rational use of cognitive resources in human planning. *Nature Human Behaviour* **6**:1112-1125 https://doi.org/10.1038/s41562-022-01332-8 | PubMed

Carrasco M. (2011) Visual attention: The past 25 years. *Vision Research, Vision Research 50th Anniversary Issue: Part 2* **51**:1484-1525 https://doi.org/10.1016/j.visres.2011.04.012 | PubMed

Carrasco M., Ling S., Read S. (2004) Attention alters appearance. *Nature Neuroscience* **7**:308-313 https://doi.org/10.1038/nn1194 | PubMed

Chen T., Cheyette S., Allen K., Tenenbaum J., Smith K. (2026) "Just in Time" World Modeling Supports Human Planning and Reasoning. *arXiv* https://doi.org/10.48550/ARXIV.2601.14514

Chica A. B., Bartolomeo P., Lupianez J. (2013) Two cognitive and neural systems for endogenous and exogenous spatial attention. *Behavioural Brain Research* **237**:107-123 https://doi.org/10.1016/j.bbr.2012.09.027 | PubMed

Cohen M. A., Cavanagh P., Chun M. M., Nakayama K. (2012) The attentional requirements of consciousness. *Trends in Cognitive Sciences* **16**:411-417 https://doi.org/10.1016/j.tics.2012.06.013 | PubMed

Daw N. D., Niv Y., Dayan P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* **8**:17041711 https://doi.org/10.1038/nn1560 | PubMed

Deroy O., Faivre N., Lunghi C., Spence C., Aller M., Noppeney U. (2016) The Complex Interplay Between Multisensory Integration and Perceptual Awareness. *Multisensory Research* **29**:585-606 https://doi.org/10.1163/22134808-00002529 | PubMed

Dezfouli A., Balleine B. W. (2013) Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLOS Computational Biology* **9**:e1003364 https://doi.org/10.1371/journal.pcbi.1003364 | PubMed

Dickinson A., Balleine B. (2010) Hedonics: The cognitive-motivational interface. In: Kringelbach M. L., Berridge K. C. (Eds). *Pleasures of the brain* Oxford University Press. pp. 74-84 https://doi.org/10.1093/oso/9780195331028.003.0006

Dubois J., Hamker F. H., VanRullen R. (2009) Attentional selection of noncontiguous locations: The spotlight is only transiently "split". *Journal of Vision* **9**:3 https://doi.org/10.1167/9.5.3 | PubMed

Dung L. (2022) Assessing tests of animal consciousness. *Consciousness and Cognition* **105**:103410 https://doi.org/10.1016/j.concog.2022.103410 | PubMed

Eriksen C. W., James J. D. (1986) Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics* **40**:225-240 https://doi.org/10.3758/BF03211502 | PubMed

Fleming S., Michel M. (no date) Sensory Horizons and the Functions of Conscious Vision.

Gershman S. (2021) *What makes us smart: The computational logic of human cognition* Princeton University Press. pp. vii-205

Ghafari T., Mazzetti C., Garner K., Gutteling T., Jensen O. (2024) Modulation of alpha oscillations by attention is predicted by hemispheric asymmetry of subcortical regions. *eLife* **12**:RP91650 https://doi.org/10.7554/eLife.91650 | PubMed

Griffiths T. L., Callaway F., Chang M. B., Grant E., Krueger P. M., Lieder F. (2019) Doing more with less: Meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences, Artificial Intelligence* **29**:24-30 https://doi.org/10.1016/j.cobeha.2019.01.005

Hassabis D., Kumaran D., Summerfield C., Botvinick M. (2017) Neuroscience-Inspired Artificial Intelligence. *Neuron* **95**:245-258 https://doi.org/10.1016/j.neuron.2017.06.011 | PubMed

He S., Cavanagh P., Intriligator J. (1996) Attentional resolution and the locus of visual awareness. *Nature* **383**:334-337 https://doi.org/10.1038/383334a0 | PubMed

Hershman R., Henik A., Cohen N. (2018) A novel blink detection method based on pupillometry noise. *Behavior Research Methods* **50**:107-114 https://doi.org/10.3758/s13428-017-1008-1 | PubMed

Herzog M. H., Drissi-Daoudi L., Doerig A. (2020) All in Good Time: Long-Lasting Postdictive Effects Reveal Discrete Perception. *Trends in Cognitive Sciences* **24**:826-837 https://doi.org/10.1016/j.tics.2020.07.001 | PubMed

Ho M. K., Abel D., Correa C. G., Littman M. L., Cohen J. D., Griffiths T. L. (2022) People construct simplified mental representations to plan. *Nature* **606**:129-136 https://doi.org/10.1038/s41586-022-04743-9 | PubMed

Huys Q. J. M., Eshel N., O'Nions E., Sheridan L., Dayan P., Roiser J. P. (2012) Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees. *PLOS Computational Biology* **8**:e1002410 https://doi.org/10.1371/journal.pcbi.1002410 | PubMed

Huys Q. J. M., Lally N., Faulkner P., Eshel N., Seifritz E., Gershman S. J., Dayan P., Roiser J. P. (2015) Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences* **112**:3098-3103 https://doi.org/10.1073/pnas.1414219112 | PubMed

Jensen O. (2024) Distractor inhibition by alpha oscillations is controlled by an indirect mechanism governed by goal-relevant information. *Communications Psychology* **2**:1-11 https://doi.org/10.1038/s44271-024-00081-w | PubMed

Kaiser D., Quek G. L., Cichy R. M., Peelen M. V. (2019) Object Vision in a Structured World. *Trends in Cognitive Sciences* **23**:672-685 https://doi.org/10.1016/j.tics.2019.04.013 | PubMed

Keefe J. M., Stormer V. S. (2021) Lateralized alpha activity and slow potential shifts over visual cortex track the time course of both endogenous and exogenous orienting of attention. *NeuroImage* **225**:117495 https://doi.org/10.1016/j.neuroimage.2020.117495 | PubMed

Keramati M., Smittenaar P., Dolan R. J., Dayan P. (2016) Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences* **113**:12868-12873 https://doi.org/10.1073/pnas.1609094113 | PubMed

Kinley I., Levy Y. (2022) PuPl: An open-source tool for processing pupillometry data. *Behavior Research Methods* **54**:2046-2069 https://doi.org/10.3758/s13428-021-01717-z | PubMed

Knuth D. E., Moore R. W. (1975) An analysis of alpha-beta pruning. *Artificial Intelligence* **6**:293-326 https://doi.org/10.1016/0004-3702(75)90019-3

Koch C., Tsuchiya N. (2007) Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences* **11**:16-22 https://doi.org/10.1016/j.tics.2006.10.012 | PubMed

Kool W., Cushman F. A., Gershman S. J. (2016) When Does Model-Based Control Pay Off?. *PLOS Computational Biology* **12**:e1005090 https://doi.org/10.1371/journal.pcbi.1005090 | PubMed

Korf R. E. (1985) Depth-first iterative-deepening: An optimal admissible tree search. *Artificial Intelligence* **27**:97-109 https://doi.org/10.1016/0004-3702(85)90084-0

Kr C., Np B. (1999) Visuospatial attention: Beyond a spotlight model. *Psychonomic Bulletin & Review* **6** https://doi.org/10.3758/bf03212327 | PubMed

Lamme V. A. F. (2003) Why visual attention and awareness are different. *Trends in Cognitive Sciences* **7**:12-18 https://doi.org/10.1016/S1364-6613(02)00013-X | PubMed

Lamme V. A., Roelfsema P. R. (2000) The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences* **23**:571-579 https://doi.org/10.1016/s0166-2236(00)01657-x | PubMed

Landry M., da Silva Castanheira J., Jerbi K. (2023) Differential and Overlapping Effects between Exogenous and Endogenous Attention Shape Perceptual Facilitation during Visual Processing. *Journal of Cognitive Neuroscience* **35**:1279-1300 https://doi.org/10.1162/jocn_a_02015 | PubMed

Landry M., da Silva Castanheira J., Raz A., Baillet S., Sackur J. (2024) A lateralized alpha-band marker of the interference of exogenous attention over endogenous attention. *Cerebral Cortex* **34**:bhad457 https://doi.org/10.1093/cercor/bhad457 | PubMed

Landry M., Da Silva Castanheira J., Sackur J., Raz A. (2021) Investigating how the modularity of visuospatial attention shapes conscious perception using type I and type II signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance* **47**:402-422 https://doi.org/10.1037/xhp0000810 | PubMed

Liu T., Jiang Y., Sun X., He S. (2009) Reduction of the Crowding Effect in Spatially Adjacent but Cortically Remote Visual Stimuli. *Current Biology* **19**:127-132 https://doi.org/10.1016/j.cub.2008.11.065 | PubMed

MacIver M. A., Finlay B. L. (2022) The neuroecology of the water-to-land transition and the evolution of the vertebrate brain. *Philosophical Transactions of the Royal Society of London Series B Biological Sciences* **377**:20200523 https://doi.org/10.1098/rstb.2020.0523 | PubMed

McMains S. A., Somers D. C. (2004) Multiple Spotlights of Attentional Selection in Human Visual Cortex. *Neuron* **42**:677-686 https://doi.org/10.1016/S0896-6273(04)00263-6 | PubMed

McMains S. A., Somers D. C. (2005) Processing Efficiency of Divided Spatial Attention Mechanisms in Human Visual Cortex. *The Journal of Neuroscience* **25**:9444-9448 https://doi.org/10.1523/JNEUROSCI.2647-05.2005 | PubMed

Mingers J. (1989) An Empirical Comparison of Pruning Methods for Decision Tree Induction. *Machine Learning* **4**:227-243 https://doi.org/10.1023/A:1022604100933

Mudrik L., Faivre N., Koch C. (2014) Information integration without awareness. *Trends in Cognitive Sciences* **18**:488-496 https://doi.org/10.1016/j.tics.2014.04.009 | PubMed

Müller N. G., Bartelt O. A., Donner T. H., Villringer A., Brandt S. A. (2003) A Physiological Correlate of the "Zoom Lens" of Visual Attention. *The Journal of Neuroscience* **23**:3561-3565 https://doi.org/10.1523/JNEUROSCI.23-09-03561.2003 | PubMed

Nelson M. D., Crisostomo M., Khericha A., Russo F., Thorne G. L. (2012) Classic Debates in Selective Attention: Early vs Late, Perceptual Load vs Dilution, Mean RT vs Measures of Capacity. *Perception* **41**:997-1000 https://doi.org/10.1068/p7309 | PubMed

Newell A., Simon H. (1956) The logic theory machine-A complex information processing system. In: *IRE Transactions on Information Theory* IRE Transactions on Information Theory. **2** pp. 61-79 https://doi.org/10.1109/TIT.1956.1056797

Niv Y. (2019) Learning task-state representations. *Nature Neuroscience* **22**:1544-1553 https://doi.org/10.1038/s41593-019-0470-8 | PubMed

Nobre K., Kastner S. (2014) *The Oxford Handbook of Attention* OUP Oxford.

Norman D. A. (1968) Toward a theory of memory and attention. *Psychological Review* **75**:522-536 https://doi.org/10.1037/h0026699

Overgaard M., Sandberg K. (2021) The Perceptual Awareness Scale—Recent controversies and debates. *Neuroscience of Consciousness* **2021**:niab044 https://doi.org/10.1093/nc/niab044 | PubMed

Peelen M. V., Kastner S. (2014) Attention in the real world: Toward understanding its neural basis. *Trends in Cognitive Sciences* **18**:242-250 https://doi.org/10.1016/j.tics.2014.02.004 | PubMed

Pooresmaeili A., Poort J., Roelfsema P. R. (2014) Simultaneous selection by object-based attention in visual and frontal cortex. *Proceedings of the National Academy of Sciences* **111**:6467-6472 https://doi.org/10.1073/pnas.1316181111 | PubMed

Pooresmaeili A., Roelfsema P. R. (2014) A Growth-Cone Model for the Spread of Object Based Attention during Contour Grouping. *Current Biology* **24**:2869-2877 https://doi.org/10.1016/j.cub.2014.10.007 | PubMed

Posner M. I. (1980) Orienting of attention. *The Quarterly Journal of Experimental Psychology* **32**:3-25 https://doi.org/10.1080/00335558008248231 | PubMed

Posner M., Snyder C., Davidson B. (1980) Attention and the detection of signals. *Journal of Experimental Psychology: General* **109**:160-174 https://doi.org/10.1037/0096-3445.109.2.160 | PubMed

Pylyshyn Z. W., Storm R. W. (1988) Tracking multiple independent targets: Evidence for a parallel tracking mechanism*. *Spatial Vision* **3**:179-197 https://doi.org/10.1163/156856888X00122 | PubMed

Quinlan J. R. (1986) Induction of decision trees. *Machine Learning* **1**:81-106 https://doi.org/10.1007/BF00116251

Ramsöy T. Z., Overgaard M. (2004) Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences* **3**:1-23 https://doi.org/10.1023/B:PHEN.0000041900.30172.e8

Samaha J., Bauer P., Cimaroli S., Postle B. R. (2015) Top-down control of the phase of alpha-band oscillations as a mechanism for temporal prediction. *Proceedings of the National Academy of Sciences* **112**:8439-8444 https://doi.org/10.1073/pnas.1503686112 | PubMed

Saxe A., Nelli S., Summerfield C. (2021) If deep learning is the answer, what is the question?. *Nature Reviews Neuroscience* **22**:55-67 https://doi.org/10.1038/s41583-020-00395-8 | PubMed

Schad D. J., Engbert R. (2012) The zoom lens of attention: Simulating shuffled versus normal text reading using the SWIFT model. *Visual Cognition* **20**:391-421 https://doi.org/10.1080/13506285.2012.670143 | PubMed

Shaw M. L., Shaw P. (1977) Optimal allocation of cognitive resources to spatial locations. *Journal of Experimental Psychology: Human Perception and Performance* **3**:201-211 https://doi.org/10.1037/0096-1523.3.2.201

Shea N., Frith C. D. (2016) Dual-process theories and consciousness: The case for 'Type Zero' cognition. *Neuroscience of Consciousness* **2016**:niw005 https://doi.org/10.1093/nc/niw005 | PubMed

Snider J., Lee D., Poizner H., Gepshtein S. (2015) Prospective Optimization with Limited Resources. *PLOS Computational Biology* **11**:e1004501 https://doi.org/10.1371/journal.pcbi.1004501 | PubMed

Stocker A. A., Simoncelli E. P. (2007) A Bayesian Model of Conditioned Perception. *Advances in Neural Information Processing Systems* **2007**:1409-1416 PubMed

Strauss M., Sitt J. D., King J.-R., Elbaz M., Azizi L., Buiatti M., Naccache L., van Wassenhove V., Dehaene S. (2015) Disruption of hierarchical predictive coding during sleep. *Proceedings of the National Academy of Sciences of the United States of America* **112**:E1353–1362 https://doi.org/10.1073/pnas.1501026112 | PubMed

James W. (1890) *The principles of psychology, Vol I.* APA PsycNET. https://doi.org/10.1037/10538-000

Tsuchiya N., Koch C. (2014) On the Relationship Between Consciousness and Attention. In: Gazzaniga M. S., Mangun G. R. (Eds). *The Cognitive Neurosciences* (5th) The MIT Press. https://doi.org/10.7551/mitpress/9504.003.0092

Vollebregt M. A., Zumer J. M., ter Huurne N., Castricum J., Buitelaar J. K., Jensen O. (2015) Lateralized modulation of posterior alpha oscillations in children. *NeuroImage* **123**:245-252 https://doi.org/10.1016/j.neuroimage.2015.06.054 | PubMed

Whitney D., Levi D. M. (2011) Visual Crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences* **15**:160-168 https://doi.org/10.1016/j.tics.2011.02.005 | PubMed

Wong K. W., Scholl B. J. (2024) Spontaneous path tracing in task-irrelevant mazes: Spatial affordances trigger dynamic visual routines. *Journal of Experimental Psychology: General* **153**:2230-2238 https://doi.org/10.1037/xge0001618 | PubMed

da Silva Castanheira J, et al (2025) How Attention Simplifies Mental Representations For Planning. Open Science Framework. https://doi.org/10.17605/OSF.IO/SA6VF

Ho MK, et al (2020) Value-Guided Construal. Open Science Framework. https://doi.org/10.17605/OSF.IO/ZPQ69

Peer reviews

Reviewer #1 (Public review):

Summary:

This study investigated how visuospatial attention influences the way people build simplified mental representations to support planning and decision-making. Using computational modeling and virtual maze navigation, the authors examined whether spatial proximity and the spatial arrangement of obstacles determine which elements are included in participants' internal models of a task. The study developed and tested an extension of the value-guided construal (VGC) model that incorporates features of spatial attention for selecting simpler task mental representation.

Strengths:

(1) Original Perspective: The study introduces an explicit attentional component to established models of planning, offering an approach that bridges perception, attention, and decision-making.

(2) Methodological Approach: The combination of computational modeling, behavioral data, and eye-tracking provides converging measures to assess the relationship between attention and planning representations.

(3) Cross-validated data: The study relies on the analysis of three separate datasets, two already published and an additional novel one. This allows for cross-validation of the findings and enhances the robustness of the evidence.

(4) Focus on Individual Differences: Reports of how individual variability in attentional "spillover" correlates with the sparsity of task representations and spatial proximity add depth to the analysis.

Appraisal of Aims and Results:

The study sets out to determine how spatial attention shapes the construction of task representations in planning contexts. The authors provide evidence that spatial proximity and arrangement influence which environmental features are incorporated into internal models used for navigation, and that accounting for these effects improves model predictions. There is clear documentation of individual variation, with some participants showing greater attentional spillover and more sparse awareness profiles.

Comments on revised version:

The authors did a great job and I am very happy with the revised manuscript.

https://doi.org/10.7554/eLife.108034.2.sa3

Reviewer #2 (Public review):

Summary:

Castanheira et al. investigate the role of spatial attention for planning during three maze navigation experiments (one new experiment and two existing datasets). Effective planning in complex situations requires the construction of simplified representations of the task at hand. The authors find that these mental representations (as assessed by conscious awareness) of a given stimulus are influenced by (spatially) surrounding stimuli. Individual participants varied in the degree to which attention influenced their task representations, and this attentional effect correlated with the sparsity of representations (as measured by the range of awareness reports across all stimuli). Spatially grouping task-relevant information on either the left or right side of the maze led to mental representations more similar to optimal representations predicted by the value-guided construal (VGC) model - a normative model describing a theoretical approach to simplifying complex task information. Finally, the authors propose an update to this model, incorporating an attentional spotlight component; the revised descriptive model predicts empirical task representations better than the original (normative) VGC model.

Strengths:

The novelty of this study lies in the proposal and investigation of a cognitive mechanism through which a normative model like value-guided construal can enable human planning. After proposing attention as this mechanism, the authors make concrete hypotheses about mismatches between the VGC predictions and real human behavior, which are experimentally validated. Thus, not only does this study describe a possible mechanism for simplification of task information for planning, but the authors also propose a descriptive model, revising VGC to incorporate this attentional component.

A strength of this paper is the variety of investigative approaches: analysis of existing data, novel experiment, and a computational approach to predict experimental findings from a theoretical model. Analyzing pre-existing datasets increases the size of the participant cohort and strengthens the authors' conclusions. Meanwhile, comparing the predictions of the existing normative model and the authors' own refined model is a clever approach to

substantiate their claims. In addition, the authors describe several crucial controls, which are key to the interpretability of their results. In particular, the eye tracking results were critical.

In summary, this paper constitutes an important step toward a more complete understanding of the human ability to plan.

Comments on revised version:

I am overall happy with the revision and agree that the authors have addressed most of the comments.

https://doi.org/10.7554/eLife.108034.2.sa2

Reviewer #3 (Public review):

Summary:

The authors build on a recent computational model of planning, the "value-guided construal" framework by Ho et al. (2022), which proposes that people plan by constructing simple models of a task, such as by attending to a subset of obstacles in a maze. They analyze both published experimental data and new experimental data from a task in which participants report attention to objects in mazes. The authors find that attention to objects is affected by spatial proximity to other objects (i.e., attentional overspill) as well as whether relevant objects are lateralized to the same hemifield. To account for these results, the authors propose a "spotlight-VGC" model, in which, after calculating attention scores based on the original VGC model, attention to objects is enhanced based on distance. They find that this model better explains participant responses when objects are lateralized to different hemifields. These results demonstrate complex interactions between filtering of task-relevant information and more classical signatures of attentional selection.

Strengths:

(1) The paper builds on existing modeling work in a novel manner and integrates classic results on attention into the computational framework.

(2) The authors report new and extensive analyses of existing data that shed light on additional sources of systematic variability in responses related to attentional spillover effects

(3) They collect new data using new stimuli in the original paradigm that directly test predictions related to the lateralization of task-relevant information, including eye tracking data that allows them to control for possible confounds.

(4) The extended model (spotlight-VGC) provides a formal account of these new results.

Comments on revised version:

I also agree that the authors addressed our comments and the manuscript is much stronger now.

https://doi.org/10.7554/eLife.108034.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Reviewer #1 (Public review):***

> *Summary:*
>
> *This study investigated how visuospatial attention influences the way people build simplified mental representations to support planning and decision-making. Using computational modeling and virtual maze navigation, the authors examined whether spatial proximity and the spatial arrangement of obstacles determine which elements are included in participants' internal models of a task. The study developed and tested an extension of the value-guided construal (VGC) model that incorporates features of spatial attention for selecting simpler task mental representation.*
>
> *Strengths:*
>
> *(1) Original Perspective:*
>
> *The study introduces an explicit attentional component to established models of planning, offering an approach that bridges perception, attention, and decisionmaking.*
>
> *(2) Methodological Approach:*
>
> *The combination of computational modeling, behavioral data, and eye-tracking provides converging measures to assess the relationship between attention and planning representations.*
>
> *(3) Cross-validated data:*
>
> *The study relies on the analysis of three separate datasets, two already published and an additional novel one. This allows for cross-validation of the findings and enhances the robustness of the evidence.*
>
> *(4) Focus on Individual Differences:*
>
> *Reports of how individual variability in attentional "spillover" correlates with the sparsity of task representations and spatial proximity add depth to the analysis.*

We thank the Reviewer for their overall positive assessment of our work and their helpful comments. We have addressed each point below.

> *Weaknesses:*
>
> *(1) Clarity of the VGC model and behavioral task:*
>
> *The exposition of the VGC model lacks sufficient detail for non-expert readers. It is not clear how this model infers which maze obstacles are relevant or irrelevant for planning, nor how the maze tasks specifically operationalize "planning" versus other cognitive processes.*
>
> *The method for classifying obstacles as relevant or irrelevant to the task and connecting metacognitive awareness (i.e., participants' reports of noticing obstacles) to attentional capture is not well justified. The rationale for why awareness serves as a valid attention proxy, as opposed to behavioral or neurophysiological markers, should be clearer.*

We thank the reviewer for urging further clarity here. Our work builds closely on the previous maze navigation paradigm and VGC model developed and reported by Ho et al. Nature (2022). We directly adopted variants of their maze stimuli, computational model and obstacle awareness measures, and married these with an investigation of the role of visuospatial attention. We agree that it would be useful for the reader to have a more in-depth description of the paradigm and model, and how it operationalises planning, without

needing to refer back to the original Ho et al. paper. We have now added additional explanatory sections to the Introduction and Methods as follows:

On page 4:

“One elegant approach to forming such a simplified representation is to adaptively select the granularity of information required to complete the task (Ho et al., 2022), known as value-guided construal (VGC). Unlike previous accounts, which model human planning as a search over all items (e.g.., tube lines), the VGC model predicts that a cognitively limited decision-maker selects a manageable subset of information over which to plan— i.e., a task representation—balancing utility and complexity (Ho et al., 2022). In our example, the VGC algorithm would plan over a few relevant tube lines rather than planning over all possible stations. To select the representation that achieves the best balance between utility and complexity, the model searches across all possible combinations of tube lines, computing the value (i.e., the plan’s utility minus its cost) of each representation for planning a specific journey. The algorithm then selects the representation with the highest value, which ensures that an ideal observer selects a representation which only includes the items (i.e., tube lines) that lead to successful planning while excluding as many items as possible to keep the plan as simple as possible. For our purposes, items included in the representation are considered taskrelevant, while items that are not represented are considered task-irrelevant. This algorithm, therefore, provides a normative standard of an efficient plan to which we can compare people’s actual plans.”

On page 6:

“We operationalized planning using a maze navigation paradigm, akin to our tube-related example, where participants were required to plan a route through the maze, avoiding obstacles that blocked their path. Obstacles predicted by the sVGC model to be included in the representation were considered task-relevant.”

“At the end of every trial, participants reported their awareness of specific obstacles (see Methods for details). The level of awareness reported for different obstacles provides a read-out of what features of the environment individuals were subjectively representing while solving a particular maze. While other markers of attention and awareness (for instance, behavioural or neurophysiological variables) could also be used, here we focused on direct awareness reports in order to relate our findings both to those of Ho and colleagues and to the subjective awareness reports used in consciousness science (e.g. the Perceptual Awareness Scale (Barnett et al., 2024; Overgaard & Sandberg, 2021; Ramsøy & Overgaard, 2004; Samaha et al., 2015)). Participants were instructed to maintain central fixation while planning (see dataset dSC 1), in line with previous empirical work using this task (Ho et al., 2022).”

To visualize our effects, we binarized the predictions of the sVGC model such that obstacles with a marginalized probability greater than 0.5 were considered taskrelevant, while other obstacles were considered task-irrelevant (e.g., Figure 2b). We now clarify this point in the caption of Figure 2.

> *(2) Attention framework:*
>
> *The account of attention is largely limited to the "spotlight" model. When solving a maze, participants trace the correct trail, following it mentally with their overt or covert attention. In this perspective, relevant concepts are also rooted in attention literature pertaining to object-based attention using tasks like curve tracing (e.g., Pooresmaeili & Roelfsema, 2014) and to mental maze solving (e.g., Wong & Scholl, 2024), which may be highly relevant and add nuance to the current work. This view of attention may be more pertinent to the task than models of simultaneously tracking multiple objects cited here.*

> *Prior work (notably from the Roelfsema group) indicates that attentional engagement in curve-tracing tasks may be a continuous, bottom-up process that progressively spreads along a trajectory, in time and space, rather than a "spotlight" that simply travels along the path. The spread of attention depends on the spatial proximity to distractors - a point that could also be pertinent to the findings here.*
>
> *Moreover, the tracing of a "solution" trail in a maze may be spontaneous and not only a top-down voluntary operation (Wong & Scholl, 2024), a finding that requires a more careful framing of the link to conscious perception discussed in the manuscript.*
>
> *Conceptualizing attention as a spatial spotlight may therefore oversimplify its role in navigation and planning. Perhaps the observed attentional modulation reflects a perceptual stage of building the trail in the maze rather than a filter for a later representation for more efficient decision making and planning. A fuller discussion of whether the current model and data can distinguish between these frameworks would benefit readers.*

We thank the reviewer for highlighting relevant findings in the attention literature that were missing from our discussion. We fully agree that a complete account of the interplay of planning, navigation, and attention is likely to recruit the kind of curvetracing processes highlighted by the reviewer. However, we emphasise that our current focus is not on the process of navigation through a maze, but on the process of construing the maze itself. In other words, we are focused not on how people represent their path from A to B, but how they represent the maze itself, which they then use as a basis for planning between A and B. The VGC model predicts that a subset of obstacles will be included in this construal. We think that a spotlight model is a good starting point for this work, because attention is being deployed across the whole maze stimulus, and then becomes attached to particular objects located in particular positions. This is a distinct process from that involved in navigating the path itself. Accordingly, our stimuli were designed such that task-relevant obstacles could be presented either proximally or distally to the optimal path (e.g., Figure 1a and Supplemental Figures S1-6). An obstacle that blocks any possible path on one side of the maze is task-relevant but located a long way from the optimal path. The results of Ho and colleagues' (2022) third experiment demonstrate how task-relevant yet distal obstacles are better remembered than task-irrelevant proximal obstacles (see Figure 4 of Ho et al., 2022). We also observed that obstacles further away from the navigation path were often represented by participants (see Figures S1-6), which cannot be explained by curve tracing alone.

While these results cannot definitively rule out the possibility that participants automatically trace the path while also construing the maze, they suggest that the value-guided construal process is an independent predictor of participants' representations beyond proximity to the navigated path. To make this distinction clearer, we now cite the papers alluded to by the reviewer, in the Discussion on pages 28-29, while also acknowledging the potential for investigating attention during the navigation process itself:

"Future work may also wish to examine the relevance of visuospatial attention for the navigation process itself in this task. While our present findings speak to how individuals perceive the maze while planning, it remains unclear how attention is deployed during navigation along a path, such as how object-based attention progressively spreads along trajectories in time and space(Pooresmaeili & Roelfsema, 2014; Wong & Scholl, 2024)."

There is also one additional nuance to the current spotlight model that we were inspired to consider by the reviewer's comment. This is the idea that attentional effects may spread within or along the obstacles themselves. We cannot explore this in the current data because we asked for awareness of the entire obstacles, not parts of obstacles, but it may be possible to explore this in future work, for instance, with eye tracking measures.

More generally, the growth-cone (i.e., zoom lens) model of attention for curve tracing proposed by Roelfsema and colleagues shares considerable similarities with the spotlight of attention model. Both models argue for the grouping of spatially proximal items based on attention. While the growth-cone model argues for varying sizes of zoom lenses (i.e., receptive fields of neurons) that facilitate the tracing of proximal items, both models predict that spatially proximal items are preferentially processed together because of attention. Indeed, the spotlight model could model these varying zoom lenses by altering the width of the attentional spotlight dynamically across the visual scene based on the spatial proximity of obstacles. Following related comments by Reviewer 2, we now investigate inter-individual differences in the attentional spotlight of participants and observed that these differences significantly predict participants' mental representations (see Attentional spotlight model of task representations). We have now updated the Discussion to include consideration of these alternative model frameworks:

On page 27:

"Second, in the current work we were unable to distinguish whether these attentional effects are driven by a fixed spotlight of attention, or whether attention operates akin to a zoom lens, shifting the 'width' of the focus of attention according to the task demands (Eriksen & St. James, 1986; Müller et al., 2003; Schad & Engbert, 2012). The latter view would be consistent with growth-cone models of attention in which the focus of attention expands and contracts in accordance with task demands, mirroring the various receptive field sizes in the visual hierarchy (Pooresmaeili et al., 2014; Pooresmaeili & Roelfsema, 2014). In partial support of this idea, we found significant inter-individual differences in the width of participants' attentional spotlight (Figure S11). It is also possible that attention is deployed within or along parts of obstacles, rather than on entire obstacles. Future work using naturalistic measures of eye movements may be able to address these questions."

> *(3) Lateralization of attention:*
>
> *The analysis considers whether relevant information is distributed bilaterally or unilaterally across the visual display, but does not sufficiently address evidence for attentional asymmetries across the left and right visual fields due to hemispheric specialization (e.g., Bartolomeo & Seidel Malkinson, 2019). Whether effects differ for left versus right hemifield arrangements is not made explicit in the presented findings.*

We thank the reviewer for this suggestion. To address this point, we fitted a three-way interaction model between VGC model prediction, lateralization index, and side (left vs right hemifield). We did not find evidence for the three-way effect (β= 0.01, SE= 0.02, 95% CI [-0.03, 0.04], p = 0.738; ΔBIC = 58.30 in favour of the null effect; see table below), suggesting that the side to which participants lateralized their attention did not influence their task representations. This result is now reported on page 12:

"This effect did not vary significantly as a function of the specific hemifield (i.e., left vs right) in which task-relevant information was presented (β= 0.01, SE= 0.02, 95% CI [-0.03, 0.04], p = 0.738; ΔBIC = 58.30 in favour of the null effect; see table S14)."

We also explored inter-individual differences in participants' tendency to lateralize their attention (see also the next point). We observed that participants tended to lateralize their attention slightly more to the right-hand side for non-lateralized maze stimuli, despite the normative sVGC model predicting that participants should not lateralize their attention for these stimuli (Figure 3c). These results may speak to potential asymmetries in lateralization, but given the exploratory nature of these analyses, they should be verified and replicated in future work.

> *(4) Individual differences:*

> *Individual differences in attentional modulation are a strength of the work, but similar analyses exploring individual variation in lateralization effects could provide further insight, and the lack of such analyses may mask important effects.*

Thank you for this suggestion. In new analyses, we explored whether i) participants exhibited differences in their tendency to lateralize their awareness reports, and ii) whether the degree to which they tended to lateralize their awareness predicted their performance on a separate set of maze stimuli. In short, we observed substantial variation in participants' tendency to lateralize their awareness (Figure S11) and found that this tendency reflected an inter-individual difference which was stable across maze types. We report these new findings on pages 14-16.

"Inter-individual variation in lateralization of attention

Next, we investigated participants' tendency to pay attention to obstacles within a single hemifield (left vs right) regardless of the sVGC model predictions. To do so, we computed an awareness lateralization index (ALI) based on participants' self-reported awareness reports of obstacles on each trial (Figure 3a). Large positive values indicate that participants were preferentially aware of the right hemifield, whereas negative values indicate preferential awareness of the left hemifield. Values close to zero indicate that participants paid attention to both hemifields equally (see Methods for details). We observed that participants' tendency to lateralize their awareness varied greatly across the Ho datasets 1 and 2 (Figure 3b); some participants preferentially paid attention to a single hemifield, regardless of whether the sVGC model predictions were lateralized. For the dSC1 dataset, we observed that on some trials, participants significantly lateralized their awareness ($|ALI| > 0.5$; Figure 3c) even though the sVGC model predictions were non-lateralized. These findings suggest that participants' tendency to pay attention to a single hemifield may represent an observable inter-individual difference in how they allocate their awareness to form task construals."

"To further explore these inter-individual differences, we tested whether participants' tendencies to lateralize their attention to a single hemifield was consistent across trials and maze stimuli. We observed that participants' tendency to lateralize their attention to a single hemifield was similar for left and right lateralized maze stimuli (Spearman $\rho = 0.72$, Figure 3d). This suggests that participants who preferentially attended to a single hemifield did so regardless of which hemifield they should attend to. More consequentially, the tendency for participants to lateralize their awareness on maze stimuli whose model predictions were also lateralized linearly correlated with participants' tendency to lateralize their attention on non-lateralized maze stimuli (Spearman $\rho = 0.88$, Figure 3d). Taken together, these findings emphasize that some individuals tend to preferentially attend to a single hemifield when planning. This tendency, importantly, represents an inter-individual difference in how participants allocate their attention across various maze types."

> *(5) Distinction between overt and covert attention:*
>
> *The current report at times equates eye movement patterns with the locus of attention. However, attention can be covertly shifted without corresponding gaze changes (see, for example, Pooresmaeili & Roelfsema, 2014).*

We fully agree, and thank the reviewer for prompting further reflection on this distinction. In the online experiments run by Ho and colleagues (i.e., datasets Ho1 and Ho2), participants' eye movements were not tracked, and therefore, they could not disambiguate whether participants were engaging in covert or overt attention to sample maze obstacles. In our third experiment (i.e., dataset dSC1), we both recorded eye movements and explicitly instructed participants to fixate centrally while viewing the maze. This ensured that participants oriented their attention only covertly during planning (see Figure S13-14).

We now elaborate on this important distinction in the Results section of the manuscript, page 12:

“In addition, we monitored participants’ eye movements in dataset dSC 1 to ensure that attention shifts would be covert as opposed to overt—a distinction which could not be determined in the online samples of datasets Ho 1 and 2.”

On page 28:

“Importantly, while the visuospatial attention effects observed in the Ho 1 and 2 datasets are likely driven by both covert and overt shifts in attention, the findings presented in experiment 3 (i.e., dSC1 dataset) rule out the contribution of overt shifts in attention through the use of eye tracking (see Figure S13-14)(Carrasco, 2011; Pooresmaeili & Roelfsema, 2014).”

> *The implications for interpreting the relationship between eye movement, memory, and attention in this setting are not fully addressed. The potential dynamics of attention along a maze trajectory and their impact on lateralization analysis would benefit from further clarification.*

We thank the reviewer for urging more clarity here. The attentional dynamics we document in our study concern how people perceive / construe the maze itself, rather than how they deploy their attention to guide active navigation. We have now sought to make this distinction clear at a number of points in the paper. The core idea is that attention acts as an early filter to select which obstacles are part of a task construal, which then affects both awareness and memory.

We have now clarified the focus of our study in the introduction on pages 5-7:

“Our focus in this study was to examine how participants perceive and represent their environment (the maze stimulus). This is a distinct process to how participants orient their attention during navigation itself, which is not part of our current study. To do so, we harness classical signatures of attentional selection to characterise how visuospatial attention shapes awareness of maze obstacles during planning.” … “Our focus in the present study was to examine attentional effects on participants’ perception of the maze stimulus. We did not quantify how individuals deploy their attention in the phase in which they were navigating through the maze.”

We did not explicitly test for memory effects in our new experiments, but Ho and colleagues demonstrated that the sVGC model predicted not only awareness reports, but also participants’ memory of obstacles (see Ho et al., 2022). Indeed, task representations computed from memory or awareness reports were strikingly similar in their experiments (Spearman $\rho = 0.86$ between memory accuracy and awareness; $\rho = 0.86$ between confidence in memory and awareness). In relation to eye movements, we refer the reviewer back to our previous response, which details how eye movements were measured and controlled during maze construal.

> *Figure 1 legend (b) --> (c)*

We have corrected this typo in the figure caption.

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *Castanheira et al. investigate the role of spatial attention for planning during three maze navigation experiments (one new experiment and two existing datasets). Effective planning in complex situations requires the construction of simplified representations of*

> *the task at hand. The authors find that these mental representations (as assessed by conscious awareness) of a given stimulus are influenced by (spatially) surrounding stimuli. Individual participants varied in the degree to which attention influenced their task representations, and this attentional effect correlated with the sparsity of representations (as measured by the range of awareness reports across all stimuli). Spatially grouping taskrelevant information on either the left or right side of the maze led to mental representations more similar to optimal representations predicted by the valueguided construal (VGC) model - a normative model describing a theoretical approach to simplifying complex task information. Finally, the authors propose an update to this model, incorporating an attentional spotlight component; the revised descriptive model predicts empirical task representations better than the original (normative) VGC model.*
>
> *Strengths:*
>
> *The novelty of this study lies in the proposal and investigation of a cognitive mechanism through which a normative model like value-guided construal can enable human planning. After proposing attention as this mechanism, the authors make concrete hypotheses about mismatches between the VGC predictions and real human behavior, which are experimentally validated. Thus, not only does this study describe a possible mechanism for simplification of task information for planning, but the authors also propose a descriptive model, revising VGC to incorporate this attentional component.*
>
> *A strength of this paper is the variety of investigative approaches: analysis of existing data, novel experiment, and a computational approach to predict experimental findings from a theoretical model. Analyzing pre-existing datasets increases the size of the participant cohort and strengthens the authors' conclusions. Meanwhile, comparing the predictions of the existing normative model and the authors' own refined model is a clever approach to substantiate their claims. In addition, the authors describe several crucial controls, which are key to the interpretability of their results. In particular, the eye tracking results were critical.*
>
> *In summary, this paper constitutes an important step toward a more complete understanding of the human ability to plan.*

We thank the Reviewer for their thoughtful and positive assessment of our findings. We also appreciate the constructive feedback on our methodology, which we believe has substantially improved our manuscript.

> *Weaknesses:*
>
> *(1) There is a critical conceptual gap in the study and its interpretation, mainly due to the reliance on a self-report metric of awareness (rather than an objective measure of behavioral performance).*
>
> *a. Awareness is tested by a 9-point self-report scale. It is currently unclear why awareness of task-irrelevant obstacles in this task would necessarily compromise optimal planning. There is no indication of whether self-reported awareness affects performance (e.g., navigation path distance, time to complete the maze, number of errors). Such behavioral evidence of planning would be more compelling.*

We thank the reviewer for prompting further reflection on the connection between construal and navigation performance. We wish to emphasise that the primary focus of our study was on measuring and modeling participants' task construals using perceptual awareness judgments, building on the methods developed by Ho and colleagues, rather than on navigation performance itself. However, as the reviewer points out, there is a natural

relationship between construal and performance – if you represent the wrong obstacles, plans may be disrupted.

To explore the relationship between task construals and performance on the navigation task we first regressed out the effects of the sVGC model on participants' awareness reports and computed the mean squared residuals for each trial. We then used these values to predict participants' navigation response times on each trial. We observed a significant negative relationship, suggesting that on trials where participants' representations showed greater deviations from the normative model, they were in fact faster at navigating the mazes. This relationship was surprising, and at odds with the initial idea that adhering to normative VGC aids in task performance. However, we think that this direction of effect may make sense if one considers that a large part of the actual construal (rather than the normative prediction) in our data was in fact driven by effects such as lateralisation which are not accounted for by the sVGC model. If one is faster at harnessing inductive biases such as lateralisation, then one may be faster to complete the maze but also show a greater deviation from the predictions of the original model.

To further explore these effects, we next focused on the distinction between lateralised and non-lateralised mazes. Here, we reasoned that the initial phase of lateralised attentional selection would lead to lateralised mazes being easier to navigate than nonlateralised ones. We conducted new analyses to determine whether participants navigated lateralized maze stimuli faster and with fewer moves than maze stimuli with non-lateralized model predictions. As detailed in Methods, we excluded trials in which participants significantly deviated from the optimal number of moves (9 or more moves) and took longer than 20 seconds to solve the maze. In line with our interpretation that attention operates as an inductive bias, participants were faster and deviated less from the optimal path on lateralized compared to non-lateralized mazes.

We now report these new results on navigation performance on pages 20-21:

"Maze navigation performance

The previous analyses focused on participants' task representations during planning. We next sought to explore links between participants' task representations and maze navigation performance. Participants performed the maze navigation task near-ceiling: they solved 95% of maze stimuli in under 20 seconds, with minimal deviation from the optimal path (i.e., 9 moves or fewer). Notwithstanding this limited variance in task performance, we explored whether participants' task construals may have impacted their navigation speed. To do so, we first regressed out the effects of the sVGC model from participants' awareness reports and used the mean squared residuals for each trial to predict response times (see Methods for details). Surprisingly, we observed a negative relationship between mean squared residual variance and response times (β = -0.31, SE = 0.05, 95% CI [-0.41, -0.21], $p < 0.001$), indicating that participants were faster on trials where the sVGC model explained less variance in their awareness reports. In other words, trials in which participants deviated more from the sVGC model predictions were solved faster. We note that one reason for this may be the strong influence of the lateralisation effect on navigation performance (see paragraph below), which itself is not part of the sVGC model prediction."

"We then explored whether participant performance differed between lateralised and nonlateralised mazes. Here, we reasoned that the initial phase of lateralised attentional selection would lead to lateralised mazes being easier to navigate than non-lateralised ones. Consistent with this hypothesis, participants were faster (β = -0.04, SE = $5.91*10^3$, 95% CI [-0.06, -0.03], $p< 0.001$) and followed the optimal path more closely (β = -0.59, SE = 0.09, 95% CI [-0.78, -0.40], $p< 0.001$) when maze stimuli were more lateralized."

And in the Discussion section, on page 23:

“Mental representations and task performance

We observed that participants were faster and deviated less from the optimal path on maze stimuli that were lateralized. This effect is not predicted by the original sVGC model but dovetails with the interpretation that early visuospatial attention operates as an inductive bias to guide the formation of simplified task representations. Surprisingly, we also observed that participants were faster to navigate mazes on trials where their simplified task representation deviated from the sVGC model prediction. We interpret this seemingly contradictory finding in the following way: there are several factors beyond the sVGC model – including, for instance, maze lateralisation – that predict both construal and performance on the maze navigation task. Further work is needed to understand how inductive biases such as lateralisation shape both construal and performance, and the real-world benefits that such strategies might afford for naturalistic stimuli.”

> *b. Relatedly, it would have been more convincing to have an objective measure of awareness, for instance, how the presence or absence of a "task-irrelevant" obstacle affects performance (e.g., change navigation path distance or time to complete the maze), or whether participants can accurately recall the location of obstacles.*

We thank the reviewer for prompting further reflection on the validity and robustness of our awareness measures. We emphasise however that our focus is not (primarily) on maze navigation performance, but on task construal, which as noted in our previous response may come apart from navigation performance for a variety of reasons. Our primary goal is to measure participants’ subjective awareness of the maze as a marker of their idiosyncratic (conscious) mental representation on each trial. In doing so, we build on a rich tradition of measuring subjective awareness in consciousness and perception science (for instance, work using the Perceptual Awareness Scale, or detection judgments). In this sense, we think our awareness scale (following Ho et al.) represents a valid and straightforward way of assessing our target psychological construct. However, we also agree with the Reviewer that convergent evidence from other measures is always valuable. In Ho and colleagues’ original paper, they developed a variant of the maze task where participants had to recall the location of obstacles, as well as rate their awareness (Exp 3) and a variant in which participants could hover their mouse over hidden obstacles in the maze to reveal their location – an online metric of attentional deployment (Exp 4). These data afforded us the opportunity to validate the awareness reports against an objective measure of recall, as suggested by the Reviewer. In reanalysing these data, we observed that the obstacle awareness and memory/hover measures were strikingly correlated within two independent samples of participants (Spearman $\rho = 0.86$ between memory accuracy and awareness; $\rho = 0.86$ between confidence in memory and awareness; $\rho = 0.76$ between the probability of hovering over the obstacle and awareness; $\rho = 0.65$ between the duration of the mouse hovering and awareness). These re-analyses are now reported on page 22 of our manuscript, to highlight the convergent validity of the awareness metric:

“Finally, we examined the convergent validity of participants’ awareness reports by reanalyzing the memory recall data reported in Ho and colleagues’ experiment(Ho et al., 2022). We reasoned that participants should demonstrate similar task representations regardless of the measure used to probe the construal. In line with this prediction, we observed that the obstacle awareness reports and memory/hover measures were strikingly correlated within three independent samples of participants (Spearman $\rho = 0.86$ between memory accuracy and awareness; $\rho = 0.86$ between confidence in memory and awareness; $\rho = 0.76$ between the probability of hovering over the obstacle and awareness; $\rho = 0.65$ between the duration of the mouse hovering and awareness; see Tables S18 and S19).”

> *c. Consequently, I'm not sure that we can conclude that the spatial context does impact participants' ability to plan spatial navigation or to "incorporate taskrelevant*

> *information into their construal". We know that the spatial context affects subjective (self-reported) awareness, but the authors do not present evidence that spatial context affects behavioral performance.*

Following the line of argument above, we think it's important to separate out task construal (the simplified representation of the maze, measured by awareness reports), and the impact of this on navigation and other aspects of behaviour. The awareness reports (and other convergent measures) show that task-relevant information (as predicted by the VGC) is incorporated into the construal, a process which is modulated by spatial context. These are the key targets of our modeling. Whether this impacts performance is a distinct question, and one that we now address in our response to point a above.

> *d. Another concern that may complicate interpretation is the following: Figure 3c shows improved VGC model predictions (steeper slope) for mazes with greater lateralization. However, there are notable outliers in these plots, where a high lateralization index does not correspond to good model performance. There is currently no discussion/explanation of these cases.*

The Reviewer astutely points out some outliers in our analysis. While on average lateralized maze stimuli are represented more closely to the sVGC model, there are indeed some noticeable outlier mazes. These mazes represent stimuli in which participants tended to lateralize their attention to the 'wrong hemifield'—e.g., participants were more aware of obstacles in the right hemifield despite sVGC model predicting that obstacles on the left hemifield were task-relevant. We believe this explains the poor sVGC model fits on these trials. We note, however, that on average participants were capable of attending to the correct hemifield without explicit instructions (i.e., 9 out of 12 mazes).

We have now included a discussion of these outliers in the results section of the paper on page 12:

"We note that for three maze stimuli whose model predictions were lateralized there was nevertheless a poor fit to the sVGC model (see Figure 2c, right panel). These outliers correspond to maze stimuli where participants, on average, lateralized their attention to the incorrect hemifield (i.e., the opposite hemifield to that predicted by the sVGC model)."

> *(2) I noticed an issue with clarity regarding task-relevance. It is currently not fully clear which obstacles are "task irrelevant". Also, the term is used inconsistently, sometimes conflating with "awareness". For example, in the "Attentional spotlight model of task representations" section, the authors state that "taskrelevant information becomes less relevant when surrounded by task-irrelevant information". But they really mean that participants become less aware of those task-relevant obstacles. I assume task-relevance is an objective characteristic related to maze organization, not to a participant's construal. Indeed, the following paragraph provides evidence of model predictions of awareness.*

We apologize for any confusion regarding the terminology of our manuscript. We indeed use the terms task-relevant and task-irrelevant to refer to obstacles that are objectively predicted by the normative sVGC model or the attentional spotlight model to be included in (>0.5) or excluded from (<0.5) task construals, respectively. This designation reflects the predictions from the computational model and does not reflect participants' reported awareness. We then ran linear hierarchical models to predict participants' awareness reports from these model predictions. The Reviewer is correct that the task-relevance of obstacles is indeed related to the maze's organization, and not related to participants' subjective reports of awareness. We have now clarified this point throughout the manuscript to better emphasize the difference between the model predictions of taskrelevance and participants' subjective reports.

On page 17:

"To achieve this, we computed the predictions of the existing VGC model for each obstacle's task relevance in a given maze, and averaged these predictions within an attentional spotlight of 3 squares (Figure 4a & S8, see Methods for details). This process yielded novel model predictions, whereby some obstacles which were once predicted as task-irrelevant by the normative sVGC are now predicted as task-relevant by the attentional spotlight model. We depict the effects of this spatial spotlight in Figure 4a: task-irrelevant stimuli (plotted in grey; see middle left obstacle) neighbouring taskrelevant obstacles (plotted in orange) become more task-relevant, whereas taskrelevant information becomes less relevant when surrounded by task-irrelevant information (see bottom right orange obstacle). This deviation in model predictions from the normative sVGC model was used to predict participants' awareness reports. We hypothesized that this spotlight-VGC model would predict participants' reports better than the original VGC model, which does not account for spatial attention."

> *(3) The behavioral paradigm has some distinct disadvantages, and the validity of the task is not backed up by behavioral data.*
>
> *a. I understand the need for central fixation, but it also makes the task less naturalistic.*

The fixation cross was required on every trial such that participants could maintain central fixation for our eye tracking experiment. While this design is less naturalistic, it allows us to examine the eye movements of participants. Requiring participants to fixate during the 'planning' phase of the experiment allowed us to isolate the effects of covert attention from changes in awareness due to overt shifts in attention. In other words, differences in participants' awareness reports in the 3rd experiment cannot be explained by longer fixation times to specific obstacles.

> *b. The task with its top-down grid view does not seem to mimic real human navigation. Though this grid may be similar to mental maps we form for navigation, the sensory stimuli corresponding to possible paths and to spatial context during real-life navigation are very different.*

We agree with the reviewer that while our task is engaging for participants and simple to follow, it does not mimic naturalistic navigation in humans. There is a natural tension in computational / experimental work in cognitive science in wanting to build closely on previous results and paradigms, while ensuring that results can generalise to real-world contexts. Here, our choice of paradigm and measures was closely built on previous papers using this task from Ho and colleagues (2022, 2023). While preparing this response, we learnt that the MIT group had also harnessed this same task to develop a novel dynamic variant of the VGC model (Chen et al., 2026) called the Just in Time model (JIT). The advantage of building on this prior work is that we are able to iteratively refine and expand the VGC approach, and (in our case) bring it into closer contact with work on modeling the deployment of spatial attention in human vision. The top-down aspect of the maze notably facilitated the study of the spatial deployment of attention. We now discuss the novel dynamic variant of the VGC model in our paper on page 27:

"We close by reflecting on opportunities for further work in this area. First, an important next step is to explore the process by which task representations are formed, and how inductive biases might affect the process of task construal. The sVGC model is a normative model of the optimal task representation. Since it's construction involves an exhaustive calculation over possible paths, it is not a plausible basis for a model of the psychological process by which participants actually construct task representations. More recently a process model of task construal has been proposed, the Just in Time model (JIT). The

hypothesis of the JIT model is that participants' task representations are built up over time by iteratively simulating possible paths through the maze, affording insight into the construal process (Chen et al., 2026). In future work, it would be of interest to ask whether the attentional effects we observe in our experiments could be meshed with a dynamic JIT account of construal. We speculate that visuospatial attention may operate as an early filter, limiting the space of potential construals based on coarse spatial features of the environment, constraining a dynamic selection of obstacles. Brain imaging techniques with high time resolution, such as M/EEG, may be able to shed further light on how task representations are formed as participants plan."

> *c. Behavioral performance is not reported, so it is unknown whether participants are able to properly complete the task. The task seems pretty difficult to navigate, especially when the obstacles disappear, and in combination with the central fixation.*

Behavioural performance is now reported in response to point 1a above.

> *d. There is no discussion of whether/how this navigation task generalizes to other forms of planning.*

We fully agree that an important next step would be to generalise our results on construal to naturalistic forms of planning – for instance, using immersive VR mazes, and or investigating cognitive rather than perceptual construals. We have now added a line to this effect to the Discussion on page 28.

"An important next step to further our understanding of task representations would be to extend the current paradigm to other forms of planning and more naturalistic tasks, such as navigating immersive virtual reality (VR) environments, planning over cognitive rather than perceptual representations (e.g. planning over an abstract space), or internallyguided planning based on working memory."

> ***Reviewer #2 (Recommendations for the authors):***
>
> *(1) There are, of course, benefits to simple tasks like the ones described, but it would be interesting to compare the results to a possible experiment in which a top-down grid/map is used for planning, but then task execution is carried out in a simulated environment corresponding to the map. Also, perhaps beyond the scope of the questions addressed in this paper, but I am curious how unexpected obstacles affect representations. For instance, if participants plan based on a topdown map and then begin "real" navigation but encounter an unexpected obstacle that was not indicated on the map, does this modulate representations/awareness of future obstacles (near vs. far)?*

We fully agree that all of these lines of investigation would be super interesting to pursue in future studies, and we have added a line to the discussion to that effect on page 28:

"An important next step to further our understanding of task representations would be to extend the current paradigm to other forms of planning and more naturalistic tasks, such as navigating immersive virtual reality (VR) environments, planning over cognitive rather than perceptual representations (e.g.. planning over an abstract space), or internallyguided planning based on working memory."

> *(2) Regarding self-reported awareness as a metric, an additional experiment could ask participants to recreate the maze (identify locations of obstacles after they disappear). This would be a more objective measure of awareness.*

Yes indeed, and as described above, this was a metric used by Ho and colleagues in their previous experiment. As we describe in more detail above, the task representations obtained

via memory or awareness reports demonstrated striking similarity (ρ = 0.86).

> *(3) What is meant by "all possible orientations of the maze" in this Methods sentence: "For dataset dSC 1, participants solved each of these 24 mazes four times (i.e., all possible orientations of the maze)"?*

We thank the Reviewer for prompting more clarity here. We vertically and horizontally reversed mazes (i.e., left-right flipped) such that participants could not predict the location of the goal or start location. In this way, each maze stimulus had four potential orientations. This resulted in 96 trials of 24 unique mazes. We have clarified this point in the Methods section on page 30:

Maze stimuli were vertically and horizontally reversed (i.e., left-right flipped) such that participants could not predict the location of the start or goal location. This resulted in four potential orientations of each maze across all 24 mazes, 96 trials in total.

> *(4) For lateralization, it was unclear until reading the Methods that the lateralization index was calculated using the VGC-predicted level of taskrelevance. From the main text and Figure 2, I assumed you were just counting the number of task-relevant obstacles on each side, rather than also quantifying relevance. I understood after reading the Methods, but this could be clarified further.*

We agree with the Reviewer that this was not evident from the text. We have now updated the Results section of the manuscript to clarify this point on page 11:

“To test this hypothesis, we derived a measure of task-relevant lateralization inspired by the attention literature (Ghafari et al., 2024; Keefe & Störmer, 2021; Vollebregt et al., 2015) (Figure 2a). Specifically, we separated maze stimuli across the vertical meridian and computed the ratio of task-relevant information presented on the left versus right side derived from the sVGC model. For example, the maze shown in Figure 2a has twice the amount of task-relevant information presented in the left hemifield than in the right (lat. Index= 1/3). A lateralization index of 0.0 indicates that both hemifields contain equal amounts of task-relevant information (i.e., non-lateralized). The lateralization index was computed using the continuous VGC predictions for each obstacle (see Methods).”

> *(5) The explanation in the Methods of how the width of the attentional spotlight was chosen references Figure 1b and Supplementary Figure S2, but it seems that Supplementary Figure S8 explains this more in the caption. Also, I don't see how Figure S2 supports this.*

We apologize for this typo. The explanation of how we selected the width of the attentional spotlight should indeed reference supplemental Figure 15 (previously Figure S8). We have now corrected this and elaborated on this choice in the Methods section on page 35:

“We fixed the ‘width’ of the attentional spotlight to a distance of 3 squares based on the observation that the two neighbouring obstacles positively predicted the awareness of a probe. We observed that the mean and median distance between neighbouring obstacles of the 2nd rank (i.e., second closest) was 3 squares away for all mazes (Figure S15). We therefore opted to fix the value of the attention spotlight to 3 squares based on these observations. Future work utilizing this model should consider the statistics of their maze stimuli when deciding on the ‘width’ of the attentional spotlight.”

> *(6) The attentional spotlight width was assumed to be 3 squares, based on the linear regression predictions of the effect of neighboring obstacles on stimulus awareness. Given the individual differences across participants, it would be interesting to choose a different attentional spotlight size for each participant. Would a participant-specific attentional spotlight width improve the predictions of the spotlight-VGC model?*

The Reviewer highlights a very interesting question: do individuals vary in terms of their attentional spotlight? To test this hypothesis, we first estimated the size of the attentional spotlight for each individual based on lateralized maze stimuli, and then used this to generate personalized attentional spotlight model predictions for each subject based on these values (Figure S11). We restricted this analysis to the dSC1 dataset, where we had substantially more trials (96 in total).

In brief, we observed that indeed the personalized spotlight model fit participants' awareness reports better than both a normative sVGC model and a group-level attentional spotlight model. We interpret these findings with some caution as i) a subset of individuals had flat attentional slopes and therefore were excluded from these analyses, and ii) we believe we require additional trials to ensure a robust model fit at the individual level. While our results are encouraging, we hope future investigations into inter-individual differences will extend these findings.

We have included these additional analyses in the main text.

On page 18:

"To further explore inter-individual differences in task construal, we tested whether adjusting the attentional spotlight width to each participant's awareness reports improved the predictions of the attentional spotlight model. To do so, we first determined the width attentional spotlight of each individual in the dSC1 dataset based on lateralized maze stimuli. We then generated person-specific attentional spotlight model predictions for the non-lateralized maze stimuli to avoid overfitting the data (Figure S11). We note that 7 participants had either flat attentional slopes or negative beta coefficients, which prevented the selection of an appropriate attentional spotlight width (see Methods for details). We observed a significant improvement in model fit for the person-specific attentional spotlight model relative to both the group-level attentional spotlight model (ΔBIC= -1487.39) and the normative sVGC model (ΔBIC= -1655.29). While the limited trial numbers per participant in our current dataset warrants caution in interpreting these findings, these findings do encourage further research on inter-individual differences in attentional deployment during planning."

On pages 23-24:

"Inter-individual differences in attention

We also observed considerable inter-individual differences in attentional effects across participants (Figure 1c). While some participants were strongly influenced by the spatial context of neighbouring stimuli, others showed more limited evidence for an attentional effect (Figure 1b). Inter-individual differences in attention predicted the sparsity of participants' simplified representations: participants with larger attention effects exhibited sparser representations. Moreover, these inter-individual differences in effects of spatial proximity could be incorporated into the attentional spotlight model by varying the width of the spotlight, resulting in better model predictions."

"Beyond these spatial proximity effects, we also observed that participants varied in their tendency to lateralize their attention to a single hemifield (Figure 3). This tendency was observed across all three datasets, including on maze stimuli whose value-guided model predictions were not lateralized. This suggests that although a strategy of allocating attention is sub-optimal for these maze stimuli, some individuals preferentially attend to a single hemifield in a heuristic-like fashion. This tendency to attend to a single hemifield was a robust inter-individual difference across maze stimuli (Figure 3d), and dovetails with individual-level variation in spatial proximity effects. Taken together, these findings offer novel insights into how people vary in the ways they allocate spatial attention to solve

complex problems. Future research could explore how these individual differences constrain performance on other tasks that require planning and search in highdimensional spaces."

On page 17 of the Supplemental Materials:

> *(7) The supplementary text about lateralization effects, above Supplementary Table S8, references Table S6, but it is Table S6 does not seem to display lateralization results.*

We thank the Reviewer for pointing out this typo: we now refer to the correct supplementary table (S9).

> *(8) Why does it matter that "the maze stimuli were not designed to test horizontalmeridian lateralization effects"? What is the effect on power? Is it because there is not a good enough range in lateralization indices? It would be good to clarify, or just remove that explanation, since the cortical retinotopy explanation seems more convincing.*

We did not specifically design the maze stimuli such that there is an equal number of obstacles above and below the horizontal meridian. As such, the lateralization index derived along the horizontal meridian does not control for the number of obstacles in each hemifield, which may influence participants' awareness reports. In contrast, we designed maze stimuli such that this would not be a concern for the vertical meridian. We have clarified this point in the discussion on page 27.

"Third, while we observed clear lateralization effects along the vertical meridian (i.e., left vs right hemifield), effects along the horizontal meridian were less clear (i.e., above vs below; see Table S15-16). One potential explanation of this asymmetry is the retinotopic organization of the cortex, in which spatially adjacent stimuli can be retinotopically distant if presented on the opposite side of the vertical (but not horizontal) meridian, facilitating distractor inhibition. Importantly, while the visuospatial attention effects observed in the Ho 1 and 2 datasets are likely driven by both covert and overt shifts in attention, the findings presented in experiment 3 (i.e., dSC1 dataset) rule out the contribution of overt shifts in attention through the use of eye tracking (see Figure S13-14)(Carrasco, 2011; Pooresmaeili & Roelfsema, 2014)."

> *(9) For Figure 2c, it would be helpful to directly state what each dot and line mean.*

We updated the caption of Figure 2c to clarify what we are plotting: each point represents an obstacle, and each line the linear fit for a maze stimulus.

"Each point represents an obstacle in a maze, and each line represents the model fit for that specific maze stimulus."

> *(10) Figures and wording imply there is only a single probe obstacle per trial, but methods and model imply that participants are asked to report awareness for every obstacle. This should be clarified.*

We apologize for any confusion regarding the methodology of our study. The Reviewer is correct that participants reported their awareness of every obstacle presented on a given trial. We have clarified this in the Results section of the manuscript on page 7:

"Note, participants reported their awareness of every obstacle presented on a given trial."

We have also updated the caption of Figure 1 to clarify this point:

"Once participants finished navigating the maze, they were asked to report their awareness of every obstacle presented on a given trial in a random order."

> *(11) What is the reason for the exclusion of participants (33 for experiment 1 and 26 for experiment 2)?*

Participants were excluded from the Ho et al. datasets 1 and 2 based on their preregistered exclusion criteria, as detailed in the Methods section of their paper. In short, trials were excluded if participants took longer than 20 seconds to complete the trial, or if they spent longer than 5 seconds in the initial state. Participants were excluded if less than 80% of trials remained after reaction time exclusions or if they failed 2 out of 3 comprehension checks. We have elaborated on this point in the Methods section on page 31.

"Participants were excluded from analyses based on pre-registered exclusion criteria as detailed in (Ho et al., 2022). In short, participants were excluded if 20% or more of their trials were removed based on reaction times, or if they failed 2 out of 3 comprehension checks."

> *(12) The supplemental figures are not referenced in order, and some are not referenced at all; this should be fixed.*

We thank the Reviewer for pointing this out and have reorganized our Supplementary materials accordingly.

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *The authors build on a recent computational model of planning, the "value-guided construal" framework by Ho et al. (2022), which proposes that people plan by constructing simple models of a task, such as by attending to a subset of obstacles in a maze. They analyze both published experimental data and new experimental data from a task in which participants report attention to objects in mazes. The authors find that attention to objects is affected by spatial proximity to other objects (i.e., attentional overspill) as well as whether relevant objects are lateralized to the same hemifield. To account for these results, the authors propose a "spotlight-VGC" model, in which, after calculating attention scores based on the original VGC model, attention to objects is enhanced based on distance. They find that this model better explains participant responses when objects are lateralized to different hemifields. These results demonstrate complex interactions between filtering of task-relevant information and more classical signatures of attentional selection.*
>
> *Strengths:*
>
> *(1) The paper builds on existing modeling work in a novel manner and integrates classic results on attention into the computational framework.*
>
> *(2) The authors report new and extensive analyses of existing data that shed light on additional sources of systematic variability in responses related to attentional spillover effects*
>
> *(3) They collect new data using new stimuli in the original paradigm that directly test predictions related to the lateralization of task-relevant information, including eye tracking data that allows them to control for possible confounds.*
>
> *(4) The extended model (spotlight-VGC) provides a formal account of these new results.*

We thank the Reviewer for their positive assessment of our manuscript and their insightful comments, which has improved the clarity of our findings.

> *Weaknesses:*

> *(1) The spotlight-VGC model has a free parameter - the "width" of the attentional spotlight. This seems to have been fixed to be 3 squares. It would be good if the authors could describe a more principled procedure for selecting the width so that others can use the model in other contexts.*

Our choice for this parameter was informed by the spatial effects reported in Figure 1b. We observed that the two closest neighbouring obstacles to a probe had similar awareness (i.e., positive beta weights). We therefore compute the mean and median distances between obstacle pairs that were the second closest obstacle to a probe. This distance was 3 squares away, as depicted in Figure S15. We fixed the width of the attentional spotlight across all studies based on this observation. We agree that future research utilizing this model may need to tune this hyperparameter depending on the mean distance between a probe and its neighbours.

We have clarified this point in the methods section on page 35:

“We fixed the ‘width’ of the attentional spotlight to a distance of 3 squares based on the observation that the two neighbouring obstacles positively predicted the awareness of a probe. We observed that the mean and median distance between neighbouring obstacles of the 2nd rank (i.e., second closest) was 3 squares away for all mazes (Figure S15). We therefore opted to fix the value of the attention spotlight to 3 squares based on these observations. Future work utilizing this model should consider the statistics of their maze stimuli when deciding on the ‘width’ of the attentional spotlight.”

Following the suggestion of Reviewer 2 point 6, we now also explored inter-individual differences in this parameter. To do so, we first used the lateralized mazes in the dSC1 dataset to determine the optimal width of the attentional spotlight for each individual.

Then, we used this spotlight to derive model predictions for each person. We observed that these personalized attentional spotlight model predictions fit participants’ awareness reports on non-lateralized mazes better than the fixed-width spotlight model. We believe this preliminary result suggests the importance of modelling inter-individual differences in attentional deployment during planning. We report these effects on page 17.

> *(2) Have the authors considered other ways in which factors such as attentional spillover and lateralization could be incorporated into the model? The spotlightVGC model, as presented, involves first computing VGC predictions and only afterwards computing spillover. This seems psychologically implausible, since it supposes that the "optimal" representation is first formed and then it gets corrupted. Is there a way to integrate these biases directly into the VGC framework, perhaps as a prior on construals? The authors gesture towards this when they talk about "inductive biases", but this is not formalized.*

We thank the reviewer for bringing up this very important point. We think that a full computational treatment of the inductive bias would be a distinct project, but now seek to expand our discussion on the mechanisms by which representations could be formed. In this context, we specifically highlight novel computational work from the MIT group that was published as a preprint in the time since we submitted our paper, and which proposes a new process account of construal, the “Just in Time” (JIT) model. We also elaborate on a possible mechanism by which visuospatial attention may aid the dynamics of the construal process. In short, we agree with the reviewer that spatial attention may bias individuals to search over a subset of potential representations based on low-level spatial characteristics of the obstacles (e.g., their spatial spread in the visual field), prior to (or in concert with) a dynamic JIT-like selection process. We now elaborate on these possibilities on pages 27-28:

“We close by reflecting on opportunities for further work in this area. First, an important next step is to explore the process by which task representations are formed, and how

inductive biases might affect the process of task construal. The sVGC model is a normative model of the optimal task representation. Since it's construction involves an exhaustive calculation over possible paths, it is not a plausible basis for a model of the psychological process by which participants actually construct task representations. More recently a process model of task construal has been proposed, the Just in Time model (JIT). The hypothesis of the JIT model is that participants' task representations are built up over time by iteratively simulating possible paths through the maze, affording insight into the construal process (Chen et al., 2026). In future work, it would be of interest to ask whether the attentional effects we observe in our experiments could be meshed with a dynamic JIT account of construal. We speculate that visuospatial attention may operate as an early filter, limiting the space of potential construals based on coarse spatial features of the environment, constraining a dynamic selection of obstacles. Brain imaging techniques with high time resolution, such as M/EEG, may be able to shed further light on how task representations are formed as participants plan."

[…]

"Fourth, it will also be necessary to elaborate on how bottom-up and top-down aspects of attentional selection are combined to guide complex task representations and plans. Foundational questions remain unanswered, for instance: can multiple spatial locations be preferentially selected at once, i.e. are there multiple spotlights (Awh & Pashler, 2000; McMains & Somers, 2004; Pylyshyn & Storm, 1988; Shaw & Shaw, 1977)? There is also discourse on how spatial attention may move from one location to another: are the intervening visual regions between attended locations similarly selected (Dubois et al., 2009; Kr & Np, 1999; McMains & Somers, 2004, 2005)? Our findings tentatively suggest that individuals are able to attend to disparate spatial regions to form sparse task representations, yet there is substantial variability in how individuals orient their attention during the task. The present paradigm and computational modelling, in conjunction with carefully designed stimuli, may help resolve these outstanding questions."

> *(3) Can the authors rule out that the lateralization effects are the result of memory biases since the main measure used is a self-report of attention?*

We thank the reviewer for bringing up this important point. In our experiments, we sought to measure participants' subjective awareness of the maze stimuli as a readout of their conscious task representation on each trial. This approach marries an extensive literature on measures of perceptual awareness in consciousness science (e.g., using the Perceptual Awareness Scale) with computational models of planning. Participants' memory of (their awareness of) the obstacles is inherent to this approach, but just as with similar approaches in consciousness science (e.g. measures of iconic memory in the Sperling paradigm), we think it provides a reasonably "online" measure of awareness. It's important of course to ensure that results obtained with awareness reports are not idiosyncratic, and generalise to other approaches to quantifying task representations.

To further bolster the convergent validity of our awareness measure, we reanalyzed the data from Ho and colleagues. In their original paper, they developed a variant of the maze-navigation task where participants were asked to recall the location of obstacles as well as report their awareness (Exp 3) and a third variant of the task where participants could hover their cursors over hidden obstacles to reveal their locations (Exp 4). These data allowed us to validate the awareness reports against objective measures of recall and mouse-tracking data. We observed that the subjective awareness reports of participants were strikingly correlated with recall/hover measures across two independent samples of participants (Spearman ρ = 0.86 between memory accuracy and awareness; ρ = 0.86 between confidence in memory and awareness; ρ = 0.76 between the probability of hovering over the obstacle and awareness; ρ = 0.65 between the duration of the mouse hovering and awareness). We believe these findings

validate participants' awareness reports. These findings are now reported on page 22 of the manuscript.

"Finally, we examined the convergent validity of participants' awareness reports by reanalyzing the memory recall data reported in Ho and colleagues' experiment (Ho et al., 2022). We reasoned that participants should demonstrate similar task representations regardless of the measure used to probe the construal. In line with this prediction, we observed that the obstacle awareness reports and memory/hover measures were strikingly correlated within three independent samples of participants (Spearman $\rho = 0.86$ between memory accuracy and awareness; $\rho = 0.86$ between confidence in memory and awareness; $\rho = 0.76$ between the probability of hovering over the obstacle and awareness; $\rho = 0.65$ between the duration of the mouse hovering and awareness; see Tables S18 and S19)."

https://doi.org/10.7554/eLife.108034.2.sa0