Imagining internal mental representations for planning

Jason da Silva Castanheira (j.castanheira@ucl.ac.uk)[1], Chang (Christina) He[1], Nicholas Shea[2], Stephen M. Fleming[1,3,4]

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

Abstract

Computational approaches to human planning propose that people construct simplified mental representations of their environment, known as value-guided construals (VGC), for computational efficiency. The VGC model successfully predicts patterns of external attention when people are required to navigate mazes externally. However, it remains unknown whether internal selection processes untethered to external stimuli (i.e., offline planning and imagination) engage similar VGCs. We developed a novel maze navigation task in which the start and goal locations are revealed after stimulus offset. Despite the physical absence of the maze stimulus, we find that participants remain capable of constructing efficient mental representations aligned with the VGC model. These internal plans demonstrate significant but weaker influences of spatial grouping, similar to the pattern demonstrated for external attention. Together, our work reveals how VGC extend to purely internal, imagined task representations. These findings will help guide future research on how human-like planning may be implemented in artificial systems.

Keywords: Planning; Attention; Representation; Cognitive neuroscience; Awareness; Psychology; Internal attention; Consciousness.

Background

Explaining the efficiency with which humans plan presents a long-standing challenge for cognitive science and those who wish to replicate it in artificial systems. Even seemingly simple everyday problems require a decision-maker to construct and evaluate multiple multi-step trajectories within a decision tree, which quickly becomes computationally intractable (Callaway et al., 2022; Huys et al., 2012; Newell & Simon, 1956).

One promising approach proposes that people build simplified mental representations of problems known as value-guided construals (VGC; (Ho et al., 2022). Unlike previous accounts which model human planning as a search over all available information, the VGC model predicts that a cognitively limited decision-maker selects a manageable subset of information over which to plan. This subset is termed a task representation and achieves an optimal balance of utility and complexity.

Take, for example, navigating the tube in London: the VGC algorithm would plan over a few relevant tube lines rather than planning over the entire network. To select this subset (i.e., representation) the model searches across all possible combinations of tube lines, computing each representation's value (i.e., utility minus cost) for planning a specific journey. The algorithm then returns the optimal (highest value) representation for that journey, only including the tube lines that lead to successful planning while excluding other lines. For our purposes, items included in the representation are considered task-relevant, while items that are not represented are considered task-irrelevant. The representation obtained from this model serves as a normative standard for an efficient task representation against which we can compare participants' percepts.

Recent findings propose that external visuospatial attention acts as an inductive bias for the VGC process, shaping the information included in an individual's mental representation (Castanheira et al., 2025). Individuals are more likely to incorporate information into their construal if presented in close spatial proximity to task-relevant information, regardless of its own task relevance (Castanheira et al., 2025). In addition, individuals form mental representations more closely aligned with the normative VGC model when task-relevant information is spatially confined to a single hemifield.

Notably, the experimental work to date has relied on maze navigation tasks in which people form online plans while the stimulus is being displayed (which we refer to as *external plans*). In contrast, much of human planning is done *offline*, using imagination and simulation (which we designate as *internal* plans). For instance, if we wish to figure out how to get home from work during a tube strike, we may call up a mental map of the bus routes to our house, rather than consulting a physical map. It remains unknown how or whether VGCs operate when task representations are formed internally.

A distinction between internal and external forms of attention has long been recognized: people can guide cognitive resources to prioritize not only certain external information, but also mental representations, memories, and thoughts (Griffin & Nobre, 2003; James, 1890; Nobre & Gresch, 2025). Individuals can prioritize information held in iconic sensory and working memory using retroactive cues to guide their behaviour (Sperling, 1960). Because flexible everyday behaviour requires people to often shift between internal and external forms of attention (Nobre & Gresch, 2025), both forms of selection are likely to influence our conscious experiences of the world. However, whether and how people construct simplified mental representations in the service of planning by harnessing internal attentional selection remains to be demonstrated.

In the present study, we investigate how simplified mental representations are constructed during offline planning using internal attention. Across two behavioural experiments, we

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

test how the mental representations of a problem differ between contexts of online and offline planning, relying on external and internal attention, respectively. We hypothesized that mental representations during *internal* planning would consist of only a subset of task-relevant information (Luck & Vogel, 1997), balancing the cost of keeping information in mind and its utility. In contrast to *external* planning, however, we predicted that spatial attention effects for *internal* plans would be significantly reduced. This prediction follows from the interplay between online planning and external perception, and previous work suggesting that the spatial tuning of internal representations is weaker (Favila et al., 2022). Together, our work aims to apply computational models of planning to internal selection processes to better understand the factors governing mental representation.

Methods

Experimental Task

We modified a previously established maze-navigation task to test how participants plan online and offline. Participants were asked to move a circular avatar from a starting location to the goal location using the arrow keys (Figure 1a). Each maze consisted of an 11x11 grid with 6 blue obstacles and black central walls in the shape of a fixation cross, which participants needed to navigate around. All trials began with a fixation cross, after which participants were prompted to navigate the maze and report their awareness of all obstacles presented on that trial.

In experiment 1, participants exclusively planned their route internally (i.e., in their imagination). Participants began each trial with a central fixation, and then were shown the obstacles without the goal location. The obstacles were hidden, and then the goal location and starting location were presented. Participants then had to plan their route to the goal location. Participants had thus first seen the maze obstacles prior to knowing where they needed to go (see Figure 1a), and therefore could only plan their specific route when given the goal and starting location, after the maze obstacles had disappeared. In half of the trials, participants were prompted to report their awareness of the different obstacles without explicitly navigating the maze. On the other half of the trials, the circular avatar would appear after a brief delay, prompting participants to begin to navigate the maze. Navigation and plan-only trials were randomly interleaved.

In experiment 2, participants solved maze trials that required either external or internal planning, in a randomized, interleaved fashion. For internal planning trials, participants were presented with both the maze obstacles and the goal location. In contrast, for external planning trials, participants were first shown the location of the maze obstacles without the goal location, as in experiment 1. I.e., for half of the trials, participants were not presented with the goal immediately alongside the maze obstacles.

Across all experiments, participants reported how aware they were of each obstacle while planning, reporting awareness on a nine-point scale for all 96 trials.

Participants

32 participants (mean age = 21.56, SD = 5.81; 8 male) and 41 participants (mean age = 25.39, SD = 5.43; 11 male) completed experiments 1 and 2, respectively. Testing was conducted in-person at University College London. We excluded trials where participants' reaction times were longer than 20 seconds, or where participants deviated more than nine moves from the optimal path. This resulted in 10% of trials being excluded on average per person. Two participants from experiment 1 and 8 participants from experiment 2 were excluded from the data analysis due to poor performance on the maze navigation task. These participants had 80% or fewer trials remaining after exclusions.

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 before the experiment.

Maze stimuli

We generated 12 unique maze stimuli, each of which had two potential goal locations, yielding two sets of mazes. These two sets of mazes were perceptually the same, except for the location of the goal. This resulted in 12 pairs of mazes which were perceptually the same but yielded different sVGC model predictions. In the offline planning task participants could not predict where the upcoming goal location would appear.

VGC model

We fit the normative value-guided construal model to our maze stimuli (Ho et al., 2022). In brief, this model computes the optimal simplified representation of a task by maximizing the utility of the representation while minimizing the computational cost of keeping the representation in mind. This model assumes that an optimal decision-maker combines a subset of cause-effect relationships to represent their environment when planning. The optimal construal maximizes the value of representation (VOR), defined for each possible construal as:

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

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

The optimal simplified mental representation is selected via a SoftMax decision rule. The modelled awareness of each obstacle is computed as the marginalized probability of each obstacle i being included within a construal. This marginalized probability, $P(\mathrm{Obstacle}_i)$, for every obstacle in each maze, is used to predict participants' awareness reports. For a detailed explanation of the computational model, see (Ho et al., 2022).

In line with our previous work on the effects of external visuospatial attention on task representations (Castanheira et

al., 2025), we focus our analyses in the present paper on the static version of the VGC model (i.e., sVGC). Here, task representations are assumed to remain stable across planning. We selected this model so that we could make a more direct comparison between internal selection and our previous finding on external attention.

The sparsity of task representations

We operationalized the sparsity of participants' mental representations as the variance of their awareness reports. An individual with a sparse representation shows a high variance in their awareness reports, being strongly aware of some obstacles and unaware of others. In contrast, a participant with low variance (low sparsity) is similarly aware (or unaware) of all obstacles for that maze.

Representations over multiple goals

Participants not presented with the goal location could nevertheless start making hypothetical plans while the maze is on the screen, based on where the goal might appear. To evaluate the influence of various potential goals on task representations, we derived sVGC model predictions for 12 potential goal locations for each of the mazes. The 12 goal locations for each maze were chosen pseudo-randomly to cover a large range of possible locations. We used the mean prediction of the sVGC model across these 12 potential mazes as a proxy for representations across multiple goals in the offline planning task. We used the mean sVGC model prediction across goals in a hierarchical regression model to predict participants' awareness reports alongside the standard sVGC model predictions for the particular maze and goal locations presented to participants on each trial.

Spatial proximity effects

To examine the effects of spatial context on participants' task representations, we first rank-ordered all obstacles based on spatial proximity to every obstacle in all mazes. Participants' awareness ratings for neighbouring obstacles were then used as predictors for their awareness of the target obstacle. For example, the participant's awareness report of the closest neighbour to obstacle$_p$ on a trial was used as a predictor of the participant's report of obstacle$_p$. This yielded a hierarchical regression model with 5 regression coefficients predicting participants' awareness reports based on spatial proximity:

$$\text{report}_{\text{obs } p} = \beta_0 + \beta_1 * \text{report}_{\text{obs i}} + \beta_2 * \text{report}_{\text{obs ii}} \ldots + \beta_5 * \text{report}_{\text{obs v}} + \beta_6 * \text{sVGC} + (1 \mid \text{MazeID}) + (1 \mid \text{ParticipantID})$$

where (1| MazeID) and (1| ParticipantID) are random intercepts of each maze and participant, respectively, and β_1 reflects the spatial contribution of the closest neighbouring obstacle to participants' awareness of obstacle$_i$. We interpret any significant predictors in this model as a significant effect of spatial context above and beyond that of the sVGC model predictions (β_6). We additionally tested whether the task (external vs internal) planning moderated the effect of spatial context.

Note that the VGC model predicts that obstacles may be excluded from a mental representation regardless of their spatial proximity to other task-relevant obstacles and their proximity to the optimal path (Castanheira et al., 2025; Ho et al., 2022). Any effect observed in the above regression model would indicate a significant spatial context effect not predicted by the VGC model.

To ensure that the above spatial proximity effects were not driven by spatial smoothness of our data, we conducted 1000 spatial null permutations. For every iteration, we permuted the mapping between each obstacle in a given maze and their spatial location, maintaining the number of neighbouring obstacles for every trial. We then fit the hierarchical linear regression model described above to the permuted data to build a null distribution of coefficients. We fit a linear model to these permuted beta coefficients to assess the strength of the null spatial attention effects relative to our observed effect.

Statistical analyses

All analyses were run in R. We relied on the lme4 package to run all of our hierarchical linear regression models, and RVAideMemoire for our permutation t-test analyses.

Results

We modified a previously developed maze navigation paradigm to examine whether participants build simplified mental representations in a value-guided fashion. Participants were instructed to navigate a series of 2-D mazes and told to avoid obstacles obscuring their path (Figure 1a). Participants were first shown the obstacles of each maze without the goal or start location (experiment 1). These obstacles were then hidden, and after a fixed delay, the goal and start location were shown. In this way, participants were required to formulate a plan without the aid of the concurrent visual presentation of the obstacles (Figure 1a).

People construct simplified representations offline

We hypothesized that individuals would construct simplified mental representations for the internal planning task. To test this hypothesis, we first computed the sparsity of participants' mental representation for each maze for each task. We operationalized the sparsity of mental representations as the variance in participants' awareness reports of obstacles for a maze, where a larger variance indicates that participants had a sparser mental representation—being aware of a subset of obstacles, and unaware of others. We observed that participants in experiment 1 had a mean sparsity of mental representation of 2.05 for plan-only trials, and 2.12 for the execute trials. We replicated these findings in a second experiment.

Having established that participants report being aware of only a subset of obstacles during internal planning, we sought to examine whether these mental representations were aligned with the static value-guided construal (sVGC) model. We fit separate hierarchical linear models and observed that the sVGC model significantly predicted participants'

awareness reports for both plan-only (β = 0.22, SE = 0.01, 95% CI [0.20, 0.24]) and execute (β = 0.27, SE = 0.01, 95% CI [0.24, 0.29]) trials. We then fit a hierarchical linear model to test the interaction between the sVGC model predictions and the trial type (plan-only vs execute). We observed a significant two-way interaction, where the sVGC predictions were better for execute trials (β = 0.04, SE = 0.02, 95% CI [0.01, 0.07]). We replicated the finding that the sVGC model predicted participants' awareness reports in a second experiment where participants were required to plan externally and internally. Taken together, these results indicate that participants form simplified mental representations during offline planning in accordance with the sVGC model.

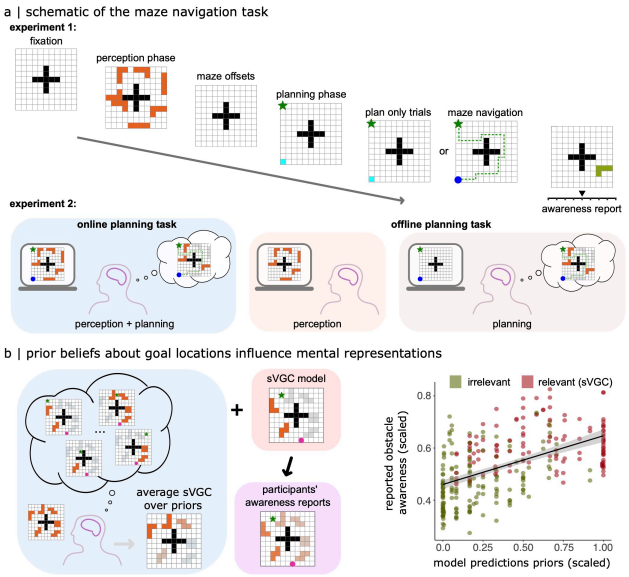

Figure 1: People construct simplified mental representations offline.

(a) Schematic of the internal planning maze task (experiment 1). Participants fixated at the start of each trial, after which a maze was presented. The goal location appeared after a brief delay. Participants were therefore required to plan while untethered to the presentation of the maze stimulus. On every trial, participants reported their mental representation using a nine-point awareness scale. On half of the trials, participants executed their plans (execute), whereas on the other half participants were not required to navigate the maze and just reported their plan (plan-only).

(b) Left panel: We tested whether participants' mental representations were predicted by potential goal locations during the internal planning task. To test this, we generated the mean VGC model prediction over 12 possible goal locations (left panel, blue box) to predict participants' awareness reports. Right panel: Scatter plot depicting the relationship between sVGC model predictions over potential goal locations and participants' awareness reports. Each dot represents an obstacle. Model predictions over potential goal locations (x-axis) significantly predicted participants' awareness reports above and beyond the normative sVGC model. Points are coloured based on whether that obstacle was predicted as task-relevant (sVGC >0.5) or irrelevant (sVGC < 0.5) based on the normative sVGC model.

Potential goals influence simplified representations

We then sought to explore whether potential goal locations influence participants' mental representations during internal planning. We reasoned that participants may form predictions of where the goal location may appear when first presented with the obstacles for each maze (Figure 1b, left panel), which in turn may bias their subsequent simplified mental representation. To test this hypothesis, we first derived predictions based on the sVGC model across multiple potential goal locations for a given maze. We used the average of these sVGC model predictions for a maze as a proxy of representations across multiple goals (see Methods for details). We used these marginalized model predictions as regressors in hierarchical linear models in addition to the normative sVGC model (based on the ground truth goal and start location) to predict participants' awareness reports. We observed that representations marginalized across potential goals positively predicted participants' awareness reports (β = 0.12, SE = 0.01, 95% CI [0.10, 0.14]; sVGC: (β = 0.17, SE = 0.01, 95% CI [0.15, 0.19]; Figure 1b) beyond the normative sVGC model. Comparing the Bayesian Information Criteria between a model which included the marginalized model predictions across possible goals against just the normative sVGC model corroborated these effects (ΔBIC= 65.05, in favour of potential goal effect). We replicated these findings in a second experiment where participants were tasked with planning both externally and internally. In experiment 2, we observed that, for internal plans, representations are governed by a value-guided process based on the present goal (sVGC; (β = 0.26, SE = 0.01, 95% CI [0.23, 0.28]) and other potential goal locations (β = 0.07, SE = 0.01, 95% CI [0.05, 0.10]). We did not explore the effect of potential goals for external planning trials.

Attentional spillover during internal planning

We were interested in exploring how the mental representations built during external versus internal planning differed from one another. To test this question, we conducted a second experiment (experiment 2) in which participants solved maze stimuli either externally (i.e., the maze obstacles were presented alongside the goal) or internally. Participants completed both external and internal planning trials in an interleaved fashion.

Our previous work demonstrates that visuospatial attention guides simplified mental representations and influences which information is included within a mental representation. Here, we tested the hypothesis that there would be spatial attention effects in the context of internal planning, but less strongly than when planning externally. If spatial attention effects were present, task-irrelevant obstacles would be more likely to be included in a participant's mental representation when presented in close spatial proximity to task-relevant obstacles. We first computed the distance between a probe obstacle and every other obstacle in the maze. Second, we

ranked the obstacles from closest to furthest from the probe and used these ranks to train a linear regression model to predict participants' awareness of the probed obstacle (green obstacle, Figure 2, left panel) from the remaining neighbouring obstacles. We ran these analyses separately for the external and internal planning tasks.

For the external planning task, we observed a significant effect of spatial context on task representations, an effect not predicted by the normative VGC model but which replicates our previous findings (Castanheira et al., 2025). Participants' awareness of an obstacle was positively predicted by the awareness of its close neighbours (β1 = 0.25, SE = 0.01,95% CI [0.24, 0.27]; β2 = 0.04, SE = 0.01,95% CI [0.02, 0.06]), whereas awareness of its furthest neighbours negatively predicted participant reports (β5 = −0.19, SE = 0.01, 95% CI [-0.21, −0.17]; β6 = −0.24, SE = 0.01, 95% CI [-0.26, −0.22]). For example, if participants reported being unaware of an obstacle, they would similarly report being unaware of its close neighbours and aware of obstacles further away. This effect of spatial context goes beyond that predicted by the sVGC model.

Critically, we also observed a similar spatial context effect in the internal planning task. Obstacles closest to a probe item positively predicted participants' awareness reports (β1 = 0.13, SE = 0.01,95% CI [0.11, 0.16]; β2 = 0.03, SE = 0.01,95% CI [0.01, 0.05]), whereas distant neighbours negatively predicted participants' awareness reports (β5 = −0.08, SE = 0.01, 95% CI [-0.10, −0.06]; β6 = −0.15, SE = 0.01, 95% CI [-0.17, −0.13])

We found that the task moderated the influence of neighbouring obstacles of the first (β = -0.13, SE = 0.01,95% CI [-0.15, -0.10]), fourth (β = 0.10, SE = 0.01,95% CI [0.07, 0.10]), and fifth (β = 0.07, SE = 0.01,95% CI [0.04, 0.10]) rank, such that the influence of spatial inductive biases on task representation was weaker in the internal planning task. Taken together, these results extend the influence of spatial context on the formation of task representations to instances of internal planning. The spatial attention effects observed in the internal planning task were robust to null permutations ($p < .001$; Figure 2 right panel).

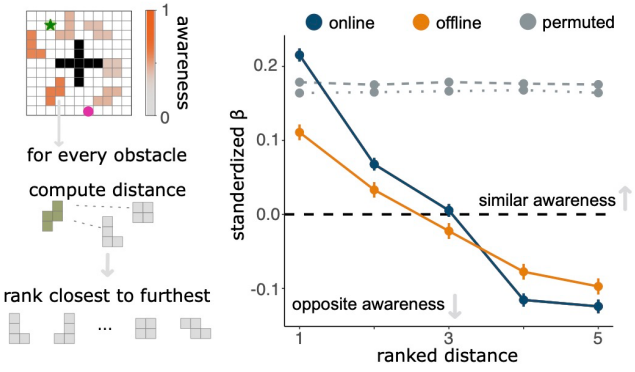

Figure 2: Spatial attention effects observed for internal planning

Left panel: Schematic of the analysis pipeline. We tested whether participants' awareness reports of an obstacle were predicted by its neighbours (i.e., a spatial attention effect). To do so, we first rank-ordered obstacles (in grey) in the maze based on their proximity to a probe obstacle (in green). We used the awareness reports of the ranked obstacles to predict the awareness of the probed obstacle in a hierarchical linear regression model (see Methods).

Right panel: We observed that obstacles closest to the probed item (ranks 1 & 2) positively impact awareness reports. In contrast, obstacles furthest from the probed item negatively impact awareness reports (rank 5 & 6). This effect is significantly different from the null for both external and internal planning, but weaker for internal plans (i.e., flatter slope).

External plans are more closely aligned with a value-guided construal model

Having demonstrated that individuals form simplified mental representations in accordance with a value-guided model during internal planning, we then explored whether internal plans were as equally well captured by the value-guided construal model as external plans. To test this, we relied on data from experiment 2, where participants solved either external (i.e., perceptually guided) or internal (unattended to perception) planning trials in an interleaved fashion. Here, we hypothesized that participants' mental representations during internal planning would be less closely aligned to the sVGC model predictions. Our rationale was that internal planning, untethered to perception, requires individuals to maintain all obstacles in working memory, which is cognitively more demanding. We observed that the sVGC model significantly predicted participants' awareness reports for both external (β = 0.58, SE = 0.01, 95% CI [0.56, 0.59]) and internal (β = 0.30, SE = 0.01, 95% CI [0.28, 0.32]) planning separately. Trial type (i.e., external vs internal planning) moderated the relationship between the sVGC model and participants' awareness reports (β = -0.26, SE = 0.01, 95% CI [-0.29, -0.23]), such that internal representations were less well predicted by the sVGC model. Taken together, these results indicate that participants form simplified mental representations during internal planning in accordance with the sVGC model, albeit to a weaker degree than those formed during external planning.

Discussion

The capacity to simplify complex problems by selectively representing goal-relevant information is critical to humans' impressive ability to plan. Here, we investigated the unexplored question of how people construct simplified mental representations when a physical stimulus is unavailable. We showed that people hold onto a representation of a problem internally, in their imagination, and plan over it. These mental representations are constructed in imagination for planning, based on an inference about task-relevant features. Mental representations associated with internal plans demonstrate similar, but weaker, spatial proximity effects. This is consistent with classical work showing that the brain uses perceptual resources for

imagination and simulation (Bisiach & Luzzatti, 1978; Block, 2023; Dijkstra et al., 2019), but now extends these results to computational models of planning. Our findings extend this previous work by showing that mental representations in imagination can be constructed on the basis of a retro-cue through inference (i.e., based on task-relevant information).

People selectively represented a subset of obstacles during internal planning (i.e., ~3 obstacles), and these representations were well predicted by the VGC model. Even when the perceptual inputs were presented beforehand and not available during goal presentation, participants were still able to form efficient representations offline by directing internal attention towards information that was relevant and filtering out that which was irrelevant. This also aligns with the considerations of the VGC model and previous work on working memory capacity: we can only retain/represent a limited amount of information from our environment at any one time. Models of working memory developed over the past 30 years similarly emphasize a capacity-limited system with several components, including the so-called visuospatial sketchpad (A. Baddeley, 2010; A. D. Baddeley & Hitch, 1974), which individuals can use to manipulate information in mind. These two lines of work dovetail with one another, emphasizing the critical limits of the brain for holding information over short periods of time and using this information in aid of achieving a goal. A limited working memory capacity may represent an adaptive computational bottleneck to facilitate planning over a subset of information. Indeed, future work may build upon the sVGC model to predict what information, and how much, will be retained in working memory for subsequent planning.

Mental representations of potential goals

In addition to the normative sVGC model significantly predicting participants' mental representations, we also observed a significant effect of potential goal contexts on task representations during internal planning.

Recent empirical work corroborates that humans can compute the value of items to achieve a specific goal across shifting contexts—the usefulness of a chair is different if you need to start a fire or cut vegetables (Castegnetti et al., 2021). This work corroborates that higher-order brain regions track the value of an item for the current goal, unrelated to its monetary value, and that activity in the vmPFC correlated with an item's usefulness across goals (Castegnetti et al., 2021). Taken together with our present findings, these results suggest that to construct a flexible plan, humans may simulate the usefulness (i.e., value) of an item across goal contexts. We speculate that this value, computed over potential goals, is then used to inform the simplified mental representation of the problem at hand. Future research can clarify the computations that govern the extraction of value across potential goals and how the brain arbitrates between value computed from potential and current goals.

Spatial organization of mental representations

We observed that although internal planning demonstrates similar spatial attention effects as observed in external planning, these effects were significantly weaker (Figure 2a). We replicated these spatial attention effects on trials where participants were not required to navigate the maze, but simply had to report their awareness. It is striking that the direction of internal attention, in the offline case, exhibits spatial effects. These findings support the view that, despite internal planning being untethered to perception, the mental representations used in imagination demonstrate considerable spatial organization—an effect not predicted by the normative sVGC model. An ideal decision-maker, in contrast, would show minimal, if any, spatial bias in their mental representations. For visually representable problems, our findings suggest that mental representations and planning occur in a space constrained by perceptual features.

We speculate that these spatial effects may reflect the neural similarity of representations during online and offline planning. This interpretation is aligned with previous brain imaging work suggesting that perception and memory systems activate similar sensory brain regions, which are spatially organized (Block, 2023; Dijkstra et al., 2019; Favila et al., 2022). These authors note, however, that the spatial organization observed in memory is less fine-grained (Favila et al., 2022)—a result that aligns with our observation of a weaker spatial attention effect for offline planning.

In the present paper, we investigated how individuals harness internal attention to build simplified mental representations to aid planning. We observed that akin to external (i.e., perceptually guided) planning, participants represented a subset of spatially organized, task-relevant information, despite no physical stimulus being present. Together, our findings illustrate how humans can deploy perceptual resources and internal attention to select information "in mind" for the purposes of planning, efficiently sculpting a task representation via imagination.

Acknowledgments

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 UKRI under the UK government's Horizon Europe funding guarantee (selected as ERC Consolidator, grant number 101043666).

References

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), R136–R140. https://doi.org/10.1016/j.cub.2009.12.014

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In *Psychology of Learning and Motivation* (Vol. 8, pp.

47–89). Elsevier. https://doi.org/10.1016/S0079-7421(08)60452-1

Bisiach, E., & Luzzatti, C. (1978). Unilateral Neglect of Representational Space. *Cortex*, *14*(1), 129–133. https://doi.org/10.1016/S0010-9452(78)80016-1

Block, N. (2023). *The Border Between Seeing and Thinking* (1st ed.). Oxford University PressNew York. https://doi.org/10.1093/oso/9780197622223.001.0001

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*(8), 1112–1125. https://doi.org/10.1038/s41562-022-01332-8

Castanheira, J. D. S., Shea, N., & Fleming, S. M. (2025). *How attention simplifies mental representations for planning*. https://doi.org/10.7554/eLife.108034.1

Castegnetti, G., Zurita, M., & De Martino, B. (2021). How usefulness shapes neural representations during goal-directed behavior. *Science Advances*, *7*(15), eabd5363. https://doi.org/10.1126/sciadv.abd5363

Dijkstra, N., Bosch, S. E., & Van Gerven, M. A. J. (2019). Shared Neural Mechanisms of Visual Perception and Imagery. *Trends in Cognitive Sciences*, *23*(5), 423–434. https://doi.org/10.1016/j.tics.2019.02.004

Favila, S. E., Kuhl, B. A., & Winawer, J. (2022). Perception and memory have distinct spatial tuning properties in human visual cortex. *Nature Communications*, *13*(1), 5864. https://doi.org/10.1038/s41467-022-33161-8

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194. https://doi.org/10.1162/089892903322598139

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*(7912), Article 7912. https://doi.org/10.1038/s41586-022-04743-9

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*(3), e1002410. https://doi.org/10.1371/journal.pcbi.1002410

James W. (1890). *The principles of psychology, Vol I.* Henry Holt and Co. https://doi.org/10.1037/10538-000

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Newell, A., & Simon, H. (1956). The logic theory machine–A complex information processing system. *IRE Transactions on Information Theory*, *2*(3), 61–79. IRE Transactions on Information Theory. https://doi.org/10.1109/TIT.1956.1056797

Nobre, A. C., & Gresch, D. (2025). How the brain shifts between external and internal attention. *Neuron*, *113*(15), 2382–2398. https://doi.org/10.1016/j.neuron.2025.06.013

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29. https://doi.org/10.1037/h0093759