nature portfolio

Peer Review File

Flexible hippocampal representation of abstract boundaries supports memory-guided choice.

Corresponding Author: Dr Raphael Kaplan

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Study overview:
Participants performed a task where they were trained on five stimuli, four 'boundary' and one 'central', each of which was associated two coordinates, price and expiration days. A 2AFC was then given where novel stimuli whose values fell within the four boundary stimuli, were given alongside the closest boundary stimulus plus the central stimulus, and tested on one of the two dimensions for which of the boundary/central stimuli the novel stimulus was closer to.

Next, participants were asked to position all stimuli on a 2D grid where one axis was labeled with price, and the other with freshness. The authors observed that subjects tended to place items within the shape defined by the boundary stimuli, though little is reported on how this relationship holds at the per-subject level.

The authors then ask whether RSA can explain encoding of stimuli in the hippocampus and mPFC. They fit an RSA GLM for each subject where the neural RDM is predicted by a combination of distance RDMs, relative choice distance between the center and boundary, and a contextual relevance (a binary factor for whether the item's location changed across contexts). They find that the distance to the nearest boundary node is significant in both regions, but not the other proposed RDMs, suggesting a representation reliant on the nearest boundary node.

The authors finally turn to MVPA classification to test whether context (square or distorted) is decodable. They train on N-1 subjects, and test on the remaining subject, where classification is a binary decision as to whether trial order was Distorted-Square-Distorted etc. or Square-Distorted-Square etc. They observe above-chance classification on the held-out subject, which they take as evidence of hippocampal representation of the geometry.

Major Comments / Critique
My biggest question in reading through this manuscript, and one that I think the importance of this study very much hinges on, is whether it's possible to show that participants truly represent the objects in a two-dimensional space, rather than simply learning two independent one-dimensional axes. E.g. I imagine the reported distance effects would be expected if subjects were treating price and date fully independently, and likewise, placement of stimuli on a 2D grid (especially given sufficient time to reason about both axes independently, as given here) might well give the placement results the authors find here. Have the authors performed a behavioral analysis as to whether choices are influenced by the non-relevant decision variable on each trial, or tested Euclidean against 2D distance for predicting behavior? The current study shows results suggestive of this (cued goods falling within a boundary, the ability to decode square vs. distorted shapes, the RDM factor of euclidan distance from boundary) but I think this could be quite a bit more explicit.

Behavior: Relatedly, in figure 2D, the authors present mean arrangements in space of the sixteen goods, and assert via a monte carlo simulation that the cued goods fell within the predefined shapes above chance. However, I can't find any information within the manuscript on the actual statistics of cued stimuli falling within the boundary on a per-subject level. I certainly hope the authors aren't comparing the statistics of the monte carlo test to the mean cued good locations (that is, the locations plotted in 2D), which seems somewhat fraught with issues of regression to the mean. The actual methods for the monte carlo simulation are somewhat sparse, with no indication of how the sixteen points are sampled for each iteration (moreover, if they're being sampled uniformly, there's no need for a simulation at all here – the distribution should be possible to calculate empirically, no?)

RSA Analysis: This seems like a useful place to perhaps look at questions of Euclidean vs. independent axes of distances – the authors could potentially compare fits between such metrics. In addition, I wonder if it would be worth testing for the influence of other boundary stimuli that are in theory defining the abstract space here – however, I realize this may be somewhat difficult, as the closest stimulus (and thus the one that always co-occurred) is the only unique one of the four, and there may be some difficulty overcoming the resulting symmetry between the other three for establishing an RDM. Nonetheless, it seems to me that a representation that depended on the anchoring of all four points might not be so surprising.

MVPA Analysis: This seems mildly suggestive that the classifier accuracy is really just picking up on something performance-related rather than anything about geometry. The authors test earlier for an explicit behavioral effect of block type, but I only see statistical results mentioned, and can't find the actual methodology used here. The fact that classification for under-performing subjects doesn't seem distributed around chance or 50% but rather below-chance is somewhat confusing otherwise.

Minor comments:

A schematic in figure 1 of the alternating square/distorted runs + the distinction between context-independent and context-dependent goods would be quite helpful.

Figure 2, C&E are swapped, and may want more informative labels than "x axis" and "y axis"

The statistical methods are missing quite a bit of information, and much of the crucial information about task structure only appears distributed throughout the methods.

Summary:
I believe this study could be a very nice addition to a growing body of work showing hippocampal representations of abstract spaces. In particular, existing studies have tended to focus on relative relationships between exchangeable stimuli, whereas here the authors have designed an experiment where subjects may plausibly learn certain stimuli as defining a geometric boundary and providing structure to others. However, at the moment, I have a number of reservations in drawing the same conclusions as the authors – first and foremost, I really would want to rule out that the current results might not depend on a 2D space at all, but only the univariate price and date relations, later translated through comparison to a 2D grid. Second, many of the methods are not fully elaborated on, and as such, I don't feel that I can comment much on the approach of the authors. Finally, I do worry (given the task design) that factors besides geometry may be underlying some of the MVPA results. If the authors can suitably address these factors, however, I think study as such would be quite reasonable.

Reviewer #2

(Remarks to the Author)
This study investigates how the hippocampus and medial prefrontal cortex (mPFC) represent abstract spatial boundaries in decision-making contexts. Using a novel memory-guided choice task involving produce items with price and freshness dimensions, the authors found that participants maintained 2D map-like representations of abstract spaces after the task. fMRI analyses revealed that Euclidean distances between cued and boundary goods modulated pattern similarity in both the hippocampus and mPFC. Notably, hippocampal activity patterns could distinguish between different abstract space configurations (square vs. distorted), with classification accuracy relating to task performance. These findings suggest that the hippocampus flexibly represents abstract boundary-defined contexts, potentially supporting adaptive decision-making in complex environments.
This is an innovative study with beautiful results. I have a few points to raise about this work, which could potentially improve it.

1. A potential confound in the RSA: The issue here stems from the design of the decision trials. In each trial, participants saw the cued good along with the landmark and the closest boundary good. This means that goods closer to each other in the abstract space would more often be presented with the same boundary good during decision trials. While the authors conducted a control analysis for the visual similarity of the cued goods themselves, they did not account for this task-induced visual similarity. This is problematic because their RSA analysis was based upon the decision trials. In these trials, the neural similarity between two cued goods could be influenced not just by their abstract spatial relationship, but also by whether they shared the same boundary good in their respective trials. This confound could explain why the authors found a significant effect of Euclidean distance to the closest boundary, but not to the landmark (which was present for all decision trials), on neural pattern similarity in the hippocampus and mPFC. To address this point, it might be beneficial to conduct a control analysis where neural similarity is compared between pairs of goods that share a boundary good versus those that don't, while controlling for abstract spatial distance (if there are enough trials to perform such an analysis). Alternatively, the authors can look for ways to explicitly model this boundary-item-based visual similarity as a separate model in the RSA.

2. The study used two features for the abstract space: price (ranging from 1-5.8 euros) and freshness (1-33 days). Looking at Figure 2D, there appears to be a wider range of responses on one axis (presumably the freshness axis?) compared to the other. This asymmetry in feature distributions could have significant implications for the results. Namly, I wonder whether the dominance of the distorted shape in the hippocampal findings and its stronger correlation with behavioral performance might be driven by this feature asymmetry rather than the shape distortion itself. If one feature (e.g., freshness) had a wider range

or was more salient to participants, it could make the distorted shape more distinctive and easier to represent. If this systematic bias in responses does exist – is there a way to quantify it and perhaps perform an individual differences analysis testing whether the bias predicts behavioral performance and neural activity in the hippocampus/mPFC?

3. Whether the dominance of the distorted shape across multiple analyses was a consequence of the features asymmetry or not, it could have important implications for understanding how abstract spaces are represented in the brain and how they influence behavior. It might reflect fundamental aspects of how irregular vs. regular spatial configurations is processed and remembered. It is worth explicitly acknowledging and discussing the pattern of stronger effects for the distorted shape across different analyses and propose potential explanations for this pattern in the discussion section.

4. In a similar way, I've noticed that in Figure 2C, which displays the post-fMRI mean boundary drag-and-rate placement, participants' reconstructions (pink dots) form more of a square shape than the original distorted shape (grey dots). However, the reported behavioral correlations and hippocampal findings show stronger effects for the distorted shape. How do the authors explain this discrepancy?

5. The drag-and-rate task, as shown in Figure 1C, included the landmark item (the apple) along with the boundary and cued goods. However, the results reported in Figure 2C and the associated text do not mention the positioning of this landmark item. Did people place the landmark? If so, it would be interesting to assess the fidelity of landmark representation and its relationship with performance and neural activity.

5. Reading the Methods, it occurred to me that the practice session was quite extensive – including 5 blocks with 12 trials each. I am wondering whether it could have cued participants to the task design and the understanding that they should learn an abstract shape during encoding to be able to perform the decisions task. It would be beneficial to explain more why the authors decided to opt for this extensive practice session and discuss its potential effects on the results.

6. It's worth clarifying some aspects of the methods –

* what was the structure of each run? (Encoding followed by decisions? How many trials each?)
* If I understand correctly, each shape included two runs – did the coordinates of the cued goods in the same shape repeat? If so, are there performance differences between the first and second run of the same shape?
* What trials are included when analyzing task performance? for example, in fig 3e -- was accuracy analyzed across the two shapes? or only in the distorted condition? do you find differences in correlation if context is included in the analysis? (same question about context regarding figure 4c).

7. In Figure 2 there appears to be a mismatch between the labels on the figure itself and their descriptions in the figure legend (C and E swapped?).

Reviewer #3

(Remarks to the Author)
In this paper, the authors investigated cognitive maps of abstract knowledge spaces and whether the hippocampus and mPFC represent changes to the boundaries of 2-D knowledge spaces. Participants were scanned while they completed a decision-making task in which they learned about grocery store goods along two dimensions (price and freshness). There were two different shapes of knowledge spaces that were implicitly learned by participants (square and distorted). Participants then completed a surprise drag-and-drop task outside of the scanner where they placed the goods on a grid according to their price and freshness. The authors report that the precision of placement on the drag-and-drop task was related to their choice accuracy during the decision-making task. Using multivariate fMRI analyses, the authors also found that the hippocampus and mPFC represented Euclidean distance between decisions cues and the closest boundary item during decision-making. Based on a multivariate classification analysis, the authors found that the hippocampus flexibly represents changes in abstract boundaries.

While the study is certainly interesting to the field of learning and cognitive maps, the paper was very difficult to read. The introduction lacked a clear research question and clear motivation for the study based on the previous literature. It was not clear what the gap in the literature is and how the current study will address the gap. The description of the methods and paradigm were confusing, and potentially lacking important details, which makes it difficult to determine whether the methods are well-suited for the research question and makes it difficult to interpret the results. Our detailed comments and concerned are outlined below:

Introduction:
In the introduction the authors set up the hypothesis, however the specific question and predictions are unclear. For example, the authors' hypothesis states that "map-like representations of knowledge in the human hippocampus and mPFC are sensitive to abstract boundary changes." It is unclear what the specific predictions would be for behavior and neural activity if the authors were to test this hypothesis. Is map-like knowledge defined by the neural representation or is there a behavioral construct (e.g., errors/error rate) that is a signature of map-like learning? How is the paradigm designed to specifically test this prediction?

Also, the authors cite previous papers that have demonstrated that the hippocampus does represent cognitive maps of abstract knowledge space, but I think the novelty of the current study and how it diverges from past work to add something

new to the literature got a little bit lost.

Paradigm:
The design and analyses are not well described or motivated by previous literature. Specifically, more work is needed to explicitly explain why participants completed alternating blocks of distorted and square knowledge spaces. The surprise drag-and-drop task at the end is conducted outside of the scanner, after participants had exposure to both the distorted and square knowledge spaces. How did the authors predict the participants would respond on this drag-and-drop task? Should they be responding based on one of the shape spaces that they had learned? Or demonstrate some morphed/in-between version of both square and distorted?

Additionally, what were participants instructed to do during encoding? Were they watching the screen and trying to learn the price/freshness of each item? During the decision-making task it wasn't clear whether participants were given feedback on their responses (correct/incorrect) to reinforce the learning of the abstract spaces. Were participants continuing to learn during the decision-making trials? Or did they do all the learning of food + price + freshness during encoding? How was this evaluated? Were trials in which participants were early in learning excluded from the fMRI analyses? Or were all trials included in the analyses? Also after each block of trials did the participants know they were switching to a new "space" with different foods/characteristics?

I'm not sure why the contextual relevance manipulation was included. What does this tell us? Why was this manipulation needed?

Statistical/fMRI methods:
I have a few questions as well about the specific methods used.

First, the use of the Monte Carlo simulation is slightly opaque and needs a bit more unpacking. For example, the defined geometric shape seems to comprise almost all of the screen (based on Figure 2C). Providing the proportion of the screen that is within the shape would be useful. Additionally, when reading that Monte Carlo was used, we thought that the authors were going to define a baseline, accounting for the participant bias for placing items close together, or account for some other noise. That is, the Monte Carlo approach assumes that all points are independently sampled, but we know that there are biases in participants fixating towards the center of the screen as well as placing items clustered around each other at random. For instance, if one were to drag and drop all of the items in the center, and add some noise, you would still get a very low p-value on the Monte Carlo comparison, but it would not mean much for your data. These sorts of biases could be integrated into your Monte Carlo sampling in order to create a meaningful baseline to compare to the alternative null. In the case of centroid bias + noise, the distances that comprise the behavioral RDMs would show some distribution around the null as opposed to being consistently positive or negative. Showing the distribution of average participant distances that informed each behavioral RDM would therefore be useful in addition to the other behavioral metrics shown in Fig 2. I think the work could be strengthened by thinking through different null scenarios in the placement other than simple independent random points assumption.

Second, we are a bit confused by the classifier results. The distorted shape has points that are confounded with the extreme ends of each dimension (e.g., the most expensive food and freshest food + the cheapest foods and least fresh food). Instead of getting a representation of the shape of the space per se, the classifier might instead be picking up on extremity of options. This leads to why we are concerned with Figure 4C. Our original expectation was that the magnitude of performance difference would be positively correlated with classification accuracy, regardless of which task participants performed better on. It is hard to estimate visually, but we are not certain that the absolute value of performance difference would correlate with the classification accuracy. Instead, we are wondering why worse performance on the distorted shape leads to worse classifier accuracy. If the participants were not tracking the extremum of the distorted shape properly and the classifier was simply picking up on extremum (instead of abstract knowledge shape), then you would expect this sort of correlation.

Relevant literature:
There are a couple important papers that missing from the paper. We have copied them below.

In relating event boundaries and hippocampal cognitive maps, a relevant paper is Sherrill, K. R., Molitor, R. J., Karagoz, A. B., Atyam, M., Mack, M. L., & Preston, A. R. (2023). Generalization of cognitive maps across space and time. Cerebral Cortex, bhad092. https://doi.org/10.1093/cercor/bhad092

A very relevant preprint in our minds is this one by Park et al. following up on work you have cited.
Park, S. A., Zolfaghar, M., Russin, J., Miller, D. S., O'Reilly, R. C., & Boorman, E. D. (2023). The representational geometry of cognitive maps under dynamic cognitive control [Preprint]. Neuroscience. https://doi.org/10.1101/2023.02.04.527142

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I appreciate the authors' thorough efforts addressing the robustness of the 2D spatial encoding. The fact that this metric (as opposed to the one-dimensional distances) shows a learning effect is quite interesting, and perhaps speaks to different timeframes of integrating simpler scalar distances into an abstract metric space. Likewise, exploring this in the space of mPFC RSA is a solid addition, and significantly reinforces the notion that participants are relying on boundary effects to form a cognitive map of this space versus cognitive approaches that treat the two axes as entirely independent. I believe this to be a much improved manuscript as a result.

Overall, I am quite satisfied with the changes here.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have addressed all of my concerns—thank you.
I would, however, suggest including a detailed description of all behavioral analyses and the coefficients used in these models (including t-tests and ANOVAs) in the Methods section. This addition would enhance the clarity of the results.

(Remarks on code availability)
code was not provided

Reviewer #3

(Remarks to the Author)
The authors have done a good job of addressing our concerns. In the Introduction, the motivation and predictions/hypotheses are now clearly outlined. The authors also have added additional rationale for the paradigm and have made clear how the present study is distinct from past work. The authors have clarified all our questions and confusions about the details of the paradigm and have updated the manuscript with more detail about the task and methods (including an updated task figure). Notably, the authors have now added a new analysis of the choice accuracy data that is a great addition to the paper and significantly strengthens the results.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Response to Reviewers

Reviewer #1 (Remarks to the Author):

Study overview:

Participants performed a task where they were trained on five stimuli, four 'boundary' and one 'central', each of which was associated two coordinates, price and expiration days. A 2AFC was then given where novel stimuli whose values fell within the four boundary stimuli, were given alongside the closest boundary stimulus plus the central stimulus, and tested on one of the two dimensions for which of the boundary/central stimuli the novel stimulus was closer to.

Next, participants were asked to position all stimuli on a 2D grid where one axis was labeled with price, and the other with freshness. The authors observed that subjects tended to place items within the shape defined by the boundary stimuli, though little is reported on how this relationship holds at the per-subject level.

The authors then ask whether RSA can explain encoding of stimuli in the hippocampus and mPFC. They fit an RSA GLM for each subject where the neural RDM is predicted by a combination of distance RDMs, relative choice distance between the center and boundary, and a contextual relevance (a binary factor for whether the item's location changed across contexts). They find that the distance to the nearest boundary node is significant in both regions, but not the other proposed RDMs, suggesting a representation reliant on the nearest boundary node.

The authors finally turn to MVPA classification to test whether context (square or distorted) is decodable. They train on N-1 subjects, and test on the remaining subject, where classification is a binary decision as to whether trial order was Distorted-Square-Distorted etc. or Square-Distorted-Square etc. They observe above-chance classification on the held-out subject, which they take as evidence of hippocampal representation of the geometry.

Major Comments / Critique

My biggest question in reading through this manuscript, and one that I think the importance of this study very much hinges on, is whether it's possible to show that participants truly represent the objects in a two-dimensional space, rather than simply learning two independent one-dimensional axes. E.g. I imagine the reported distance effects would be expected if subjects were treating price and date fully independently, and likewise, placement of stimuli on a 2D grid (especially given sufficient time to reason about both axes independently, as given here) might well give the placement results the authors find here. Have the authors performed a behavioral analysis as to whether choices are influenced by the non-relevant decision variable on each trial, or tested Euclidean against 2D distance for predicting behavior? The current study shows results suggestive of this (cued goods falling within a boundary, the ability to decode square vs. distorted shapes, the RDM factor of euclidan distance from boundary) but I think this could be quite a bit more explicit.

1. To address the reviewer's comment, we added a general linear model(GLM) to investigate whether participants' choice accuracy on a given trial is more strongly predicted by the Euclidean distance to the closest boundary over the course of the experiment(as participants build up a map-like representation of the decision space). Testing whether the contribution to choice accuracy of the relative 1D distance, the 2D distance to the landmark good, or the 2D distance to the closest boundary good, changes over the course experiment, we observe that only the 2D Euclidean distance to the closest boundary both predicts accuracy throughout the task (t(28)=-4.03; p<.001) and significantly improves as the experiment progresses(t(28)=2.60; p=.015). In contrast, the relative 1D distance strongly relates to choice accuracy(t(28)=4.65; p<.001), but unlike 2D distance its predictability doesn't improve over the course of the experiment(t(28)=0.961; p=.345). Notably, the lack of improved relative 1D distance prediction of accuracy as the experiment progressed couldn't be explained by ceiling effects since accuracy increased over the course of the experiment(rho: 0.15;p<0.001).

We then tried to connect the 2D distance performance effects to our RSA effects. In a between-subjects analysis, we tested whether the progression of individual Euclidean distance choice accuracy predictability from our GLM relates to hippocampal and mPFC RSA values. We observe that individual mPFC RSA 2D distance effects correlate with the 2D distance-related improvements in choice accuracy(rho:0.180; p=0.0011), which didn't significantly relate to our hippocampal RSA effect(rho:-0.087, p=0.117).

We add the new behavioral results in a paragraph in the Results section on p.6 lines 121-35 and Table 1.

Investigating if participants exhibited context-independent learning-related improvements in the task, we investigated the effect of 1D RD versus 2D Euclidean distance on choice accuracy. More specifically, we used a general linear model(GLM) to test whether trial-by-trial accuracy was significantly predicted by the relative 1D distance, the 2D Euclidean distance to the landmark good, or the 2D Euclidean distance to the closest boundary good. Crucially, this GLM also permitted us to measure whether the predictability of any of the three aforementioned variables on accuracy changed over the course experiment (Eq.1). Consequently, this allowed us to distinguish general performance effects not linked to learning from ones related to learning. Only the Euclidean distance to the closest boundary both predicted accuracy throughout the task(t(28)=-4.03, p<.001) and significantly improved as the experiment progressed(t(28)=2.60, p=.015). The relative 1D distance also strongly related to choice accuracy(t(28)=4.65, p<.001), but unlike 2D distance its predictability didn't improve as the experiment progressed(t(28)=0.961, p=.345; see Table 1 for group level statistics of all predictors). Notably, the lack of improved 1D RD prediction of accuracy as the experiment progressed couldn't be explained by ceiling effects (Fig. 2B). These results are consistent with the hypothesis that participants build up map-like representations of multidimensional decision spaces that inform their choices, even when they aren't needed.

Predictor	T-statistic (28)	p value
Choice Distance	4.647	< .001***
ED from Boundary	-4.027	< .001***
ED from Landmark	-0.296	0.769
Choice Distance × Run	0.961	0.345
ED from Boundary × Run	2.596	0.015*
ED from Landmark × Run	1.984	0.057

and add the new fMRI results in the RSA Results section on p. lines.

To explore how neural representations of Euclidean distance relate to map-like learning, we then tested how hippocampal-mPFC Euclidean distance representations related to individuals' Euclidean distance predictors of choice accuracy. We ran an inter-subject RSA (IS-RSA; 42,43), for which we computed an intersubject RDM (nxn, n=26) comparing beta coefficients of Euclidean distance to the closest boundary across participants for both the hippocampus and mPFC. We correlated these RDMs with a behavioral intersubject RDM, computed by comparing beta coefficients of Euclidean distance to the nearest boundary from the behavioral GLM(see Supplementary Figure 2). Notably, individual mPFC RSA 2D distance effects related to individual 2D distance-related improvements in choice accuracy(rho=0.180, p=.0011), which didn't significantly relate to individual hippocampal RSA Euclidean distance effects(rho=-0.087, p=.117). Consequently, these results link mPFC Euclidean distance representations to map-like learning of abstract knowledge.

Finally, to better link participants' 2D map-like drag-and-rate placement of the goods to the fMRI decision making task, we compared participants' overall choice accuracy and participants' ability to proportionally rescale the two dimensions of the boundary goods in the economic space. We now add these behavioral results on the Results on p.7-8, lines 184-196.

To further link the precision of participants' 2D map-like drag-and-rate placement of the goods to the fMRI task, participants' overall choice accuracy during the fMRI task was compared to participants' ability to proportionally rescale the two dimensions of the boundary goods in the economic space. We calculated the difference between the diagonals on the x and y axis reconstructed by participants and related it to participants' fMRI task performance, which yielded a significant correlation between participants proportionally rescaling the two drag-and-rate dimensions and individual fMRI task performance (rho=-0.47, p=.017; see Supplementary Figure 1B). In other words, this indicates that participants who performed the memory-guided choice task better exhibited a smaller difference between the two diagonals in their drag-and-rate reconstruction(e.g., better rescaling across the two axes). In parallel, we analyzed participants' placement of the landmark in relation to the centroid of their reconstructed shape. The distance of participants' drag-and-rate landmark placement from the reconstructed centroid significantly correlated with individual fMRI task performance(rho=-0.55, p=.003; see Supplementary Figure 1C), where participants who performed the memory-guided choice task placed the landmark closer to their reconstructed shape's centroid.

We believe these results highlight the importance of participants representing the boundary goods' respective 2D locations at both behavioral and neural levels.

Behavior: Relatedly, in figure 2D, the authors present mean arrangements in space of the sixteen goods, and assert via a monte carlo simulation that the cued goods fell within the predefined shapes above chance. However, I can't find any information within the manuscript on the actual statistics of cued stimuli falling within the boundary on a per-subject level. I certainly hope the authors aren't comparing the statistics of the monte carlo test to the mean cued good locations (that is, the locations plotted in 2D), which seems somewhat fraught with issues of regression to the mean. The actual methods for the monte carlo simulation are somewhat sparse, with no indication of how the sixteen points are sampled for each iteration (moreover, if they're being sampled uniformly, there's no need for a simulation at all here – the distribution should be possible to calculate empirically, no?)

2. We thank the reviewer for their feedback and pointing out we can calculate the distribution empirically. Consequently, we now remove the Monte Carlo simulation and present the empirical probabilities in the behavioral results. We also add further descriptive statistics to emphasize this point. We paste the revised Results from p.6-7, lines 136-154 section for this analysis, along with the revised Methods from p. 20-21 lines 507-529 below

Results

Further confirming that learning implicit 2D map-like representations of the different boundary-defined contexts informed participants' choice behavior, we used a surprise post-fMRI drag-and-rate task to investigate if participants retained a detailed map-like representation of an abstract boundary-defined context. We computed the probability of placing the four boundaries in the specific sequences provided in the task(see Methods). Participants consistently organized boundary goods in the correct order(each boundary good shared a side with its neighboring boundary good) above chance (p<.001; Fig. 2C). We then examined whether the arrangement of all observed goods during the task accurately reflected the underlying 2D structure of the abstract spaces, thereby ensuring that all goods were positioned within the designated boundaries. We calculated the area occupied by the square and distorted shape in the space(45.1%) and computed the empirical probability of sixteen points falling within the boundaries of the space, which was p<.001. The overall percentage of cued goods placed within the boundaries across all participants was 68%, with a 20% variance (see Fig. 2D for mean placement). We then tested whether participants reconstructed the boundary goods' respective freshness and price coordinates in a proportional manner during the drag-and-rate task. Notably, participants' reconstruction of the boundary goods was significantly wider along the freshness diagonal of the space(t(28)=6.7, p<.001), which is consistent with the freshness variable's longer range(1-33 days compared to the price variable's range of €1-€5.80). Confirming that the distortion of the freshness diagonal wasn't due to decreased placement precision along that dimension, participants exhibited no significant difference in boundary placement errors along the price versus freshness axes (t(28)=-1.50, p=0.13).

Methods

We quantified participants' reconstruction performance in the post-scan drag-and-rate task. We computed the area of the screen occupied by the square and the distorted shape using the shoelace formula. Subsequently, we evaluated the empirical probability of 16 uniformly

distributed points falling within these areas, which yielded a p< .001. To rule out a bias in clustering the goods' placement around the center of the screen, we calculated the distance of every cued good from each of the four replaced boundaries and the center of the screen. We then selected the minimum distance from the boundary of each cued good. This resulted in two distance vectors for each participant, one reflecting distances from the boundary and one reflecting distances from the center of the space. We compared the two distance vectors with a paired t-test, measuring the magnitude of the difference between those distances for each participant. We tested the t-statistic from the paired t-tests against zero to calculate group level statistics. To investigate the different cued good placement from the boundaries and centroid of the reconstructed figure, we computed the centroid of the polygon resulting from individual boundary placement and then followed the same procedure.

Asking whether participants' placement precision was related to individual differences in choice accuracy, we investigated participants' ability to equally rescale the price and freshness ranges during the drag-and-rate task. Consequently, we computed the length of reconstructed diagonals by calculating the Euclidean distance between boundaries of each axis, which resulted in a "Price" and a "Freshness" diagonal. We ran a paired t-test to check for differences in the length of the two diagonals. Testing for individual differences in diagonal rescaling to equal dimensions, we correlated the difference in the length of diagonals for each participant with their fMRI task accuracy.

We next investigated how participants' placement precision of the landmark good. Since the landmark was originally close to the center of both square and distorted shapes, we asked how closely participants' placed the landmark to the center of their reconstructed shape. Consequently, we computed the Euclidean distance between the centroid of each participants' reconstruction and their replaced landmark, which was then correlated with their fMRI task performance.

RSA Analysis: This seems like a useful place to perhaps look at questions of Euclidean vs. independent axes of distances – the authors could potentially compare fits between such metrics. In addition, I wonder if it would be worth testing for the influence of other boundary stimuli that are in theory defining the abstract space here – however, I realize this may be somewhat difficult, as the closest stimulus (and thus the one that always co-occurred) is the only unique one of the four, and there may be some difficulty overcoming the resulting symmetry between the other three for establishing an RDM. Nonetheless, it seems to me that a representation that depended on the anchoring of all four points might not be so surprising.

3. Beyond the new RSA mPFC 2D map-like learning result we present in response to the first comment, we are unfortunately unable to implement the reviewer's suggested analysis here. As the reviewer hints in their comment, there is a significant collinearity(t(15)= 2.12; p= 0.03) in the Euclidean distances between the boundary goods with a respective cued good due to the symmetry of the shapes. Still, we believe the reviewer's comment is worth mentioning in the manuscript and add a caveat to our Discussion about how we only measure 2D Euclidean distance between a cued good and one of the four boundary goods on p.15 lines 320-324.

Furthermore, it's important to note with our experimental design that we can only test how neural representational similarity and behavioral performance relates to the Euclidean distance between a cued good and its' closest boundary good, not all four boundary goods.

Consequently, it's unclear how representing the Euclidean distance between the cued good and each boundary good influences choice behavior during the fMRI task.

MVPA Analysis: This seems mildly suggestive that the classifier accuracy is really just picking up on something performance-related rather than anything about geometry. The authors test earlier for an explicit behavioral effect of block type, but I only see statistical results mentioned, and can't find the actual methodology used here. The fact that classification for under-performing subjects doesn't seem distributed around chance or 50% but rather below-chance is somewhat confusing otherwise.

We apologize for the lack of information on our MVPA approach in the Results. We now add the following information to the Methods on p.24-25 lines 665-67. emphasizing that the below chance probabilities were due to the 5 possible discrete levels of classification accuracy in this between-subjects classifier analysis.

This approach yielded a decoding accuracy value for each participant. For each run of each specific shape, the decoding accuracy values were constrained to discrete levels (0%, 25%, 50%, 75%, or 100%).

We'd also like to point out there is no significant difference in performance between shapes(Fig. 2) to suggest that this is primarily driving our classifier accuracy. Furthermore, we now add that there is no correlation between classifier accuracy for a particular shape and individual performance to the Results on p.13 lines 298-301.

Notably, this correlation wasn't driven by performance effects related to an individual shape. There was no correlation between distorted context decision accuracy and hippocampal distorted context classifier performance (rho=0.24;p=.261), nor between square context decision accuracy and hippocampal square context classifier accuracy(rho=0.15;p=0.502).

Minor comments:

A schematic in figure 1 of the alternating square/distorted runs + the distinction between context-independent and context-dependent goods would be quite helpful.

We thank the reviewer for the suggestion and have added further illustration to Fig. 1B-C to highlight the difference between context-dependent v -independent goods and the alternating square distorted run structure of the task respectively.

Figure 2, C&E are swapped, and may want more informative labels than "x axis" and "y axis"

We have fixed this error.

The statistical methods are missing quite a bit of information, and much of the crucial information about task structure only appears distributed throughout the methods.

We apologize for the lack of clarity and revise the statistical methods with further detail. We now consolidate the information about the experiment task structure in more concentrated sections in the Methods.

Summary:

I believe this study could be a very nice addition to a growing body of work showing hippocampal representations of abstract spaces. In particular, existing studies have tended to focus on relative relationships between exchangeable stimuli, whereas here the authors have designed an experiment where subjects may plausibly learn certain stimuli as defining a geometric boundary and providing structure to others. However, at the moment, I have a number of reservations in drawing the same conclusions as the authors – first and foremost, I really would want to rule out that the current results might not depend on a 2D space at all, but only the univariate price and date relations, later translated through comparison to a 2D grid. Second, many of the methods are not fully elaborated on, and as such, I don't feel that I can comment much on the approach of the authors. Finally, I do worry (given the task design) that factors besides geometry may be underlying some of the MVPA results. If the authors can suitably address these factors, however, I think study as such would be quite reasonable.

Reviewer #2 (Remarks to the Author):

This study investigates how the hippocampus and medial prefrontal cortex (mPFC) represent abstract spatial boundaries in decision-making contexts. Using a novel memory-guided choice task involving produce items with price and freshness dimensions, the authors found that participants maintained 2D map-like representations of abstract spaces after the task. fMRI analyses revealed that Euclidean distances between cued and boundary goods modulated pattern similarity in both the hippocampus and mPFC. Notably, hippocampal activity patterns could distinguish between different abstract space configurations (square vs. distorted), with classification accuracy relating to task performance. These findings suggest that the hippocampus flexibly represents abstract boundary-defined contexts, potentially supporting adaptive decision-making in complex environments.

This is an innovative study with beautiful results. I have a few points to raise about this work, which could potentially improve it.

1. A potential confound in the RSA: The issue here stems from the design of the decision trials. In each trial, participants saw the cued good along with the landmark and the closest boundary good. This means that goods closer to each other in the abstract space would more often be presented with the same boundary good during decision trials. While the authors conducted a control analysis for the visual similarity of the cued goods themselves, they did not account for this task-induced visual similarity. This is problematic because their RSA analysis was based upon the decision trials. In these trials, the neural similarity between two cued goods could be influenced not just by their abstract spatial relationship, but also by whether they shared the same boundary good in their respective trials. This confound could explain why the authors found a significant effect of Euclidean distance to the closest boundary, but not to the landmark (which was present for all decision trials), on neural pattern similarity in the hippocampus and mPFC. To address this point, it might be beneficial to conduct a control analysis where neural similarity is compared between pairs of goods that share a boundary good versus those that don't, while controlling for abstract spatial distance (if there are enough trials to perform such an analysis). Alternatively, the authors can look for ways to explicitly model this boundary-item-based visual similarity as a separate model in the RSA.

We thank the reviewer for suggesting a task-induced visual similarity control analysis. To address this potential confound, we model the boundary-item-based visual similarity as a separate model in the RSA. Crucially, we observe no effect for the visual similarity control analysis in the hippocampus (t(25)=.323;p=0.750), or mPFC (t(25)=-0.502;p=0.620). We now present this control analysis in the Results on p.12 lines 248-252

Furthermore, since cued goods that shared the same boundary in the 2D space were also presented with the same boundary during decision trials, we ran a control analysis to account for task-induced visual similarity. We computed trial-visual-similarity RDMs and correlated them with RDMs of hippocampus and mPFC, where neither region showed a significant effect(hippocampus: (t(25)=1.07, p=.29; mPFC: t(25)=0.79, p=.44; see Methods for further information).

and describe the analysis in the Methods on p.24 lines 631-639.

Furthermore, we ran a task-induced-visual-similarity control analysis on decision trials, which featured generated vector embeddings (https://github.com/minimaxir/imgbeddings) from snapshots of decision trials of each cued good and computed a visual-similarity RDM (vsRDM). First, we checked whether trials in which cued goods shared the same boundary showed greater visual similarity. We correlated the vsRDM with a model RDM that encoded cued goods that shared the same boundary, which yielded a significant result (rho=0.61, p<.001). Next, we ran a partial correlation between the vsRDM and the neural RDM of both hippocampus and mPFC of each participant, while also controlling for the effect of spatial distances in the 2D space. Group-level statistics were conducted by comparing correlation coefficients against zero using a one-sample t-test.

2. The study used two features for the abstract space: price (ranging from 1-5.8 euros) and freshness (1-33 days). Looking at Figure 2D, there appears to be a wider range of responses on one axis (presumably the freshness axis?) compared to the other. This asymmetry in feature distributions could have significant implications for the results. Namly, I wonder whether the dominance of the distorted shape in the hippocampal findings and its stronger correlation with behavioral performance might be driven by this feature asymmetry rather than the shape distortion itself. If one feature (e.g., freshness) had a wider range or was more salient to participants, it could make the distorted shape more distinctive and easier to represent. If this systematic bias in responses does exist – is there a way to quantify it and perhaps perform an individual differences analysis testing whether the bias predicts behavioral performance and neural activity in the hippocampus/mPFC?

We appreciate the reviewer's intriguing suggestion. We investigated whether there is a systematic bias in responses along the different dimensions that make it easier to represent the distorted shape. To summarize, we observed a bias in the drag-and-rate task, where participants' reconstruction of the boundary goods resulted in a significantly wider freshness diagonal(t(28)=6.7; p<.001), which is consistent with its longer range(1-33 days compared to €1-€5.80). Moreover, this wider freshness diagonal wasn't a result of less precise memory(worse performance) for the freshness dimension during the drag-and-rate task since participants exhibited no significant difference in boundary placement errors along the price versus freshness axes(t(28)=-1.50, p= 0.13). Investigating further, we did find that participants who equivalently rescaled goods along the diagonals of the shape(i.e., had less of a bias) performed

better on the fMRI task. In other words, participants who performed the fMRI task well could more flexibly rescale the products' price and freshness coordinates to the drag-and-rate dimensions. In what follows, we comment more in depth below on what we found and the respective edits to the manuscript based on these observations.

First, It's important to mention that since we counterbalance the x and y axes in the drag-and-rate task, the y-axis shown in Fig. 2D actually is normalized across price and freshness, which we now clarify in the caption for Fig. 2D.

Next, we add mention of the significantly wider placement along the freshness dimension in the drag-and-rate and clarify what that means in the Results on p. 6 lines 149-154.

Notably, participants' reconstruction of the boundary goods was significantly wider along the freshness diagonal of the space(t(28)=6.7, p< .001), which is consistent with the freshness variable's longer range(1-33 days compared to the price variable's range of €1-€5.80). Confirming that the distortion of the freshness diagonal wasn't due to decreased placement precision along that dimension, participants exhibited no significant difference in boundary placement errors along the price versus freshness axes(t(28)=-1.50, p=.13).

The bias in freshness versus price motivated us to investigate whether participants' ability to equally resize the two boundary good dimensions of the shape's diagonals correlated with their fMRI task performance. We indeed observe a significant correlation(rho= -0.47, p= .017), which we now add to the behavioral results on p.7-8 lines 184-196 and Supplementary Figure 1.

To further link the precision of participants' 2D map-like drag-and-rate placement of the goods to the fMRI task, participants' overall choice accuracy during the fMRI task was compared to participants' ability to proportionally rescale the two dimensions of the boundary goods in the economic space. We calculated the difference between the diagonals on the x and y axis reconstructed by participants and related it to participants' fMRI task performance, which yielded a significant correlation between participants proportionally rescaling the two drag-and-rate dimensions and individual fMRI task performance (rho=-0.47, p=.017; see Supplementary Figure 1B). In other words, this indicates that participants who performed the memory-guided choice task better exhibited a smaller difference between the two diagonals in their drag-and-rate reconstruction(e.g., better rescaling across the two axes). In parallel, we analyzed participants' placement of the landmark in relation to the centroid of their reconstructed shape. The distance of participants' drag-and-rate landmark placement from the reconstructed centroid significantly correlated with individual fMRI task performance(rho=-0.55, p=.003; see Supplementary Figure 1C), where participants who performed the memory-guided choice task placed the landmark closer to their reconstructed shape's centroid.

These findings mesh well with recent findings about how cognitive control helps adjust map-like representations of abstract knowledge to reflect current task demands(Park et al., 2023 bioRxiv), which we now mention in the Discussion on p. 16lines 355-59.

We speculate that participants' internal map-like contextual representations are partially distorted for both task contexts since participants don't need to rescale the two differently ranged axis values to the same dimensions until after the fMRI task during the drag-and-rate. The idea of flexibly rescaling abstract map-like to current task demands meshes well with the

idea that cognitive control processes can reshape the format of cognitive maps of abstract knowledge(Park et al., 2023).

3. Whether the dominance of the distorted shape across multiple analyses was a consequence of the features asymmetry or not, it could have important implications for understanding how abstract spaces are represented in the brain and how they influence behavior. It might reflect fundamental aspects of how irregular vs. regular spatial configurations is processed and remembered. It is worth explicitly acknowledging and discussing the pattern of stronger effects for the distorted shape across different analyses and propose potential explanations for this pattern in the discussion section.

Following the previous response, we believe this new finding of relating individual 2AFC choice accuracy to their ability to rescale the two dimensions in the drag-and-rate reflects flexible adaptation of map-like representations of abstract knowledge to current task findings. We now add the following Discussion on p. 16 lines 354-61.

The question remains of why does the hippocampal classifier work significantly better on the distorted versus the square shape context? We speculate that participants' internal map-like contextual representations are partially distorted for both task contexts since participants don't need to rescale the two differently ranged axis values to the same dimensions until after the fMRI task during the drag-and-rate. The idea of flexibly rescaling abstract map-like to current task demands meshes well with the idea that cognitive control processes can reshape the format of cognitive maps of abstract knowledge(Park et al., 2023). Developing new paradigms that use two different dimensions with the same range values without the need for rescaling, or more systematically change the range of dimensions, could be better suited to answer this question.

4. In a similar way, I've noticed that in Figure 2C, which displays the post-fMRI mean boundary drag-and-rate placement, participants' reconstructions (pink dots) form more of a square shape than the original distorted shape (grey dots). However, the reported behavioral correlations and hippocampal findings show stronger effects for the distorted shape. How do the authors explain this discrepancy?

We believe this discrepancy in representation is due to the forced rescaling provoked by the drag-and-rate experimental format(that wasn't present during choices in the fMRI task). Hence, why participants that rescale the diagonal dimensions of their map-like representation on the drag-and-rate perform better on the fMRI task. Similar to our response to your comment #2, we agree that this is an important point to mention in the manuscript and have added mention of it to the new Discussion section on p.16 line 355-57.

We speculate that participants' internal map-like contextual representations are partially distorted for both task contexts since participants don't need to rescale the two differently ranged axis values to the same dimensions until after the fMRI task during the drag-and-rate.

5. The drag-and-rate task, as shown in Figure 1C, included the landmark item (the apple) along with the boundary and cued goods. However, the results reported in Figure 2C and the associated text do not mention the positioning of this landmark item. Did people place the landmark? If so, it would be interesting to assess the fidelity of landmark representation and its relationship with performance and neural activity.

Yes, we observed that participants' fMRI task performance negatively correlated with the distance between the centroid of their reconstructed shape and drag-and-rate place of the landmark good(rho=-0.55; p=.003). In other words, the closer they placed the landmark to the center of the shape they constructed, the better they performed on decisions in the fMRI task. Ruling out that participants weren't just placing the landmark in the center of the space/screen, there was no correlation between the participants' task performance and the distance between the landmark placement and center of the space(rho = -0.22, p=.30). We have added this new finding to the Results on p.7 lines 192-6 and Supplementary Figure 1C.

In parallel, we analyzed participants' placement of the landmark in relation to the centroid of their reconstructed shape. The distance of participants' drag-and-rate landmark placement from the reconstructed centroid significantly correlated with individual fMRI task performance(rho=-0.55, p=.003; see Supplementary Figure 1C), where participants who performed the memory-guided choice task placed the landmark closer to their reconstructed shape's centroid.

and description of the experimental procedure on p.19 line 489

including the landmark good

and analysis in the Methods on p.20-21, lines 525-9.

We next investigated how participants' placement precision of the landmark good. Since the landmark was originally close to the center of both square and distorted shapes, we asked how closely participants' placed the landmark to the center of their reconstructed shape. Consequently, we computed the Euclidean distance between the centroid of each participants' reconstruction and their replaced landmark, which was then correlated with their fMRI task performance.

5. Reading the Methods, it occurred to me that the practice session was quite extensive – including 5 blocks with 12 trials each. I am wondering whether it could have cued participants to the task design and the understanding that they should learn an abstract shape during encoding to be able to perform the decisions task. It would be beneficial to explain more why the authors decided to opt for this extensive practice session and discuss its potential effects on the results.

The practice session for our task only lasted approx. 15 minutes to ensure participants were capable of performing the task, which is extremely brief compared to the typical multi-day training session for studies investigating human cognitive maps of abstract knowledge(please see Constantinescu et al., 2016; Park et al., 2020, 2021, 2023; Mark et al., 2020 for some examples). Still, we agree with the reviewer that our training approach is worth further mention in the Discussion. We now add further discussion of our training approach in the Discussion on p.15-16, lines 339-42

In contrast, we explicitly cued participants with both price and freshness decision variables during memory encoding, where they only made decisions on one dimension at a time. This is further reinforced by our pre-fMRI training procedure, which involved participants encoding landmark and boundary good values immediately prior to making 12 forced-choices in five separate blocks.

We also further justify our training procedure in the Methods to ensure they could do it well on p.18, lines 441-444

The training session lasted approximately fifteen minutes. We chose a training session of this duration to ensure that participants could effectively perform the task and efficiently make choices. We created a novel 2D triangle-shaped space during training to avoid priming participants towards one of the shapes in the fMRI experiment.

6. It's worth clarifying some aspects of the methods –

* what was the structure of each run? (Encoding followed by decisions? How many trials each?)

We apologize for the lack of clarity. The decision phase of a block directly follows the encoding phase and is repeated four times over the course of a run. For example, participants alternate between encoding and decision phases a total of four times during the first presentation of the square context. We edit the text to better highlight this key experimental design detail in the Introduction, caption for Figure 1, and Methods.

* If I understand correctly, each shape included two runs – did the coordinates of the cued goods in the same shape repeat? If so, are there performance differences between the first and second run of the same shape? Supp figure and/or results showing that

Yes, that is correct. Yes, we observe a significant increase in choice accuracy when a context is presented a second time, which we now include in the results and Supplementary Fig.1A. In fact in response to reviewer 1's first comment, when we run a GLM to test whether performance changes over the course of the experiment are best predicted by the relative 1D choice distance, Euclidean distance to the landmark good, or the Euclidean distance to the closest boundary(see Table 1). We find that 2D Euclidean distance to the good boundary for a given trial is the only variable that both predicts choice accuracy in general and significantly improves its predictability for accuracy over the course of the experiment. This effect also correlates with the strength of participants' mPFC RSA Euclidean distance to boundary effect, but not the hippocampal RSA effect. We now add these exciting results to the Abstract, Results, Figures 2 & 3, Discussion, and Methods.

* What trials are included when analyzing task performance? for example, in fig 3e -- was accuracy analyzed across the two shapes? or only in the distorted condition? do you find differences in correlation if context is included in the analysis? (same question about context regarding figure 4c).

All trials are included when analyzing task performance, which we now clarify in the captions for Fig. 2E and Fig. 4C. Yes, we analyze accuracy in both shapes. We don't observe any relationship between Procrustes distance from the square and square-shaped context task performance(rho=-0.09; p=.656), or Procrustes distance from the distorted shape and distorted context task performance(rho=-0.27, p = 0.17). For Fig. 4C, we observed no correlation between distorted context decision accuracy and hippocampal distorted context classifier performance (rho=-0.24;p=.261), nor between

square context decision accuracy and hippocampal square context classifier accuracy(rho=0.15;p=0.502).

7. In Figure 2 there appears to be a mismatch between the labels on the figure itself and their descriptions in the figure legend (C and E swapped?).

We have fixed this error in the revised version of the manuscript.

Reviewer #3 (Remarks to the Author):

In this paper, the authors investigated cognitive maps of abstract knowledge spaces and whether the hippocampus and mPFC represent changes to the boundaries of 2-D knowledge spaces. Participants were scanned while they completed a decision-making task in which they learned about grocery store goods along two dimensions (price and freshness). There were two different shapes of knowledge spaces that were implicitly learned by participants (square and distorted). Participants then completed a surprise drag-and-drop task outside of the scanner where they placed the goods on a grid according to their price and freshness. The authors report that the precision of placement on the drag-and-drop task was related to their choice accuracy during the decision-making task. Using multivariate fMRI analyses, the authors also found that the hippocampus and mPFC represented Euclidean distance between decisions cues and the closest boundary item during decision-making. Based on a multivariate classification analysis, the authors found that the hippocampus flexibly represents changes in abstract boundaries.

While the study is certainly interesting to the field of learning and cognitive maps, the paper was very difficult to read. The introduction lacked a clear research question and clear motivation for the study based on the previous literature. It was not clear what the gap in the literature is and how the current study will address the gap. The description of the methods and paradigm were confusing, and potentially lacking important details, which makes it difficult to determine whether the methods are well-suited for the research question and makes it difficult to interpret the results. Our detailed comments and concerned are outlined below:

We thank the reviewers for their interest and respond to their comments below.

Introduction:

In the introduction the authors set up the hypothesis, however the specific question and predictions are unclear. For example, the authors' hypothesis states that "map-like representations of knowledge in the human hippocampus and mPFC are sensitive to abstract boundary changes." It is unclear what the specific predictions would be for behavior and neural activity if the authors were to test this hypothesis. Is map-like knowledge defined by the neural representation or is there a behavioral construct (e.g., errors/error rate) that is a signature of map-like learning? How is the paradigm designed to specifically test this prediction?

1.To summarize, we thought of map-like knowledge as being defined by our Euclidean distance RSA and boundary-defined context classifier analyses, where the post-fMRI drag-and-rate served as a behavioral confirmatory control that participants perceived

boundary goods as 2D abstract boundaries. To better emphasize the importance of learning 2D Euclidean distances in our task, we now include a behavioral GLM to show how important retaining 2D representations of the task were to decision accuracy. We also relate the behavioral GLM to individual differences in RSA effects. Additionally, we now better highlight the specific questions and predictions for our behavioral, RSA, and classifier analyses in the excerpts listed below

in the Introduction about specific neural hypotheses on p.2, lines 23-26.

In particular, we were interested whether 2D Euclidean distances to abstract boundaries were represented in the hippocampal-prefrontal circuit and if different abstract boundary-defined decision contexts could be distinguished by hippocampal fMRI signals.

and behavioral GLM on p.4, lines 58-63.

This design aspect allowed us to confirm with a general linear model(GLM) if the 1D distance and 2D Euclidean distance of a choice were both significant predictors of choice accuracy and test whether this predictability changed over the course of the experiment. If map-like learning of abstract knowledge spaces related to behavioral performance, the predictability of the Euclidean distance on choice accuracy would increase over the course of the task, while the predictability of the 1D relative distance would stay the same.

This newly added information better ties into what we wrote at the end of the Methods about on p.4 lines 70-74.

Following the fMRI task, participants performed a surprise drag-and-rate task(Kriegeskorte & Mur, 2012) on a laptop outside the scanner, where they were asked to place all the supermarket goods they previously saw on a blank map according to their 2D decision variable coordinates they most strongly associated with that particular good(Fig. 1D). We relied on this post-scan session to determine if participants retained a 2D representation of the decision spaces after the task.

and the fMRI analyses on p.4-5 lines 75-89.

Our analyses first focused on univariate fMRI signal differences related to the contextual relevance of a cued good versus the context type(type of shape) of its' run during 2AFC trials. Ensuring that participants were representing 2D abstract knowledge in a similar way as previous studies(Park et al., 2021), we tested whether the Euclidean distance between the cued goods to the landmark and closest boundary modulated fMRI representational similarity in the hippocampus and mPFC during the 2AFC. Testing whether hippocampus and mPFC signals were sensitive to geometrically changing abstract boundaries, we then used fMRI classifiers on boundary good encoding trials when decision variables weren't visible to determine if there were distinct hippocampal-prefrontal representations for each boundary-defined abstract space. We could then relate each of these analyses back to choice and drag-and-rate performance. We expected hippocampal involvement in detecting contextual relevance and hypothesized that both mPFC and hippocampus would represent the Euclidean distance from the cued good to the nearest boundary good. Given the role of hippocampal cognitive maps in flexibly remapping their internal representation of the environment(O'Keefe & Nadel, 1978; Julian et al., 2021; Keinath et al., 2021; Doeller et al., 2008; Eichenbaum et al., 2014), we predicted that hippocampal fMRI signals would be sensitive to abstract boundary-defined contextual changes. To link any neural representations to behavioral performance,

multivariate fMRI analyses were related to choice behavior, including the GLM choice accuracy analysis.

Also, the authors cite previous papers that have demonstrated that the hippocampus does represent cognitive maps of abstract knowledge space, but I think the novelty of the current study and how it diverges from past work to add something new to the literature got a little bit lost.

2. We place additional emphasis on the importance of boundary representations in cognitive maps of abstract knowledge space and the dearth of literature addressing this question in the Introduction on p.2 lines 18-23.

Yet, whether hippocampus or mPFC representation are sensitive to geometric changes to the 'boundaries' or extreme coordinates of abstract knowledge spaces is unclear. Still, the ability of cognitive maps to flexibly code abstract knowledge raises the possibility that abstract boundaries–continuous(e.g., price range) or categorical(e.g., types of cars) limits–could help guide map-like learning of abstract knowledge. Here, we test the hypothesis that map-like representations of knowledge in the human hippocampus and mPFC are sensitive to abstract boundary changes.

Paradigm:

The design and analyses are not well described or motivated by previous literature. Specifically, more work is needed to explicitly explain why participants completed alternating blocks of distorted and square knowledge spaces. The surprise drag-and-drop task at the end is conducted outside of the scanner, after participants had exposure to both the distorted and square knowledge spaces. How did the authors predict the participants would respond on this drag-and-drop task? Should they be responding based on one of the shape spaces that they had learned? Or demonstrate some morphed/in-between version of both square and distorted?

3. The experimental design follows a typical human spatial memory experiment, where participants learn the location of stimuli and then are cued to remember where that stimuli is located in an environment. The key differences here with traditional spatial memory paradigms are: 1) To test memory, we include forced-choice proximity judgments instead of virtual navigation to a remembered location. 2) Participants need to remember the boundary and landmark, as opposed to the cued goods located between them. 3) In spatial navigation, environmental boundary changes are immediately noticeable by participants, so trials can more noticeably alternate between different contexts. In contrast, due to the subtlety of the abstract boundaries and corresponding changes in our paradigm, we couldn't switch between boundary-defined contexts within the same run and expect participants to implicitly learn these differences. We predicted that participants maintaining an implicit 2D representation of the abstract contexts would facilitate participants' fMRI task performance, despite being unnecessary to make the 1D forced-choice. Given the reviewers' comment, we now make a greater effort to highlight this rationale in the Introduction on p.2 lines 34-39 and on p.4, lines 42-51.

Inspired by human spatial memory experiments that have participants learn the location of stimuli and then cue them to remember where that stimuli is located in an environment, the

fMRI task consisted of blocks containing encoding and decision components in the abstract space (Fig. 1A). Differing from the typical spatial memory experiment that has participants first remember cued goods and relate their location to visible boundaries and landmarks, we had participants remember boundary and landmark abstract good coordinates and then relate those values to a visible cued good.

Immediately following their encoding of the landmark and boundary goods, participants were presented with the price or freshness of a cued good and chose whether it was more similar along that dimension to either the landmark good or the most proximal boundary good. The most proximal boundary good in each trial was selected based on its 2D Euclidean distance to the presented cued good. During 2AFC trials, the cued good was the only stimulus with a visible price or freshness variable. Crucially, the boundary produce goods had two distinct sets of price and freshness variable coordinates depending on which run they were featured. Given the increased cognitive demand of changing invisible abstract contexts during the middle of an fMRI run compared to spatial navigation tasks, each run was dedicated to one of two distinct boundary-defined abstract spaces. In contrast, the landmark coordinates were consistent in the two spaces(Fig. 1B).

4. Additionally, what were participants instructed to do during encoding? Were they watching the screen and trying to learn the price/freshness of each item? **Yes, they needed to remember the price and freshness for each presented good during encoding(please see p.4 lines 41-42; p.4 lines 53-54; Fig.1 caption and Methods).** During the decision-making task it wasn't clear whether participants were given feedback on their responses (correct/incorrect) to reinforce the learning of the abstract spaces. Were participants continuing to learn during the decision-making trials? **They didn't receive any feedback(please see edits to p.2 lines 29-30 in Intro; Fig.1 caption), but likely continued to learn during the decision-making trials.** Or did they do all the learning of food + price + freshness during encoding? **An fMRI block consisted of an encoding phase immediately followed by a decision phase, which was repeated four times per run. We now more clearly highlight this design aspect on p.4 lines 42-43 and p. 19 lines 480-81.** How was this evaluated? **This was evaluated by the accuracy of their choices during the decision trials(please see p.4 lines 53-63 in the Intro and Methods on p.19 line 474-82).** Were trials in which participants were early in learning excluded from the fMRI analyses? Or were all trials included in the analyses? **All trials were included in analyses, which we now mention in the Methods on p.20 line 492.** Also after each block of trials did the participants know they were switching to a new "space" with different foods/characteristics? **No, they weren't explicitly aware that they alternated between boundary-defined contexts beyond encoding slightly different price and freshness values for the boundary goods, as well as viewing different price and freshness values for the context-dependent cued goods.**

I'm not sure why the contextual relevance manipulation was included. What does this tell us? Why was this manipulation needed? **No, the identity of cued, boundary, and landmark goods never changed between the contexts, just the price and freshness values of boundary and context-independent cued goods. We apologize for the lack of detail here. We have some sentences in the Introduction to explain that this helps us differentiate boundary good location changes versus cued good changes similar to how we differentiate item v. environmental novelty during spatial memory tasks on p.4, lines 64-74.**

The same cued goods were presented in both spaces. Following previous work on object and contextual novelty(Kaplan et al., 2014, Mizrak et al., 2021), the experimental design distinguished context-related boundary changes (based on the manipulation of the boundaries in the space) from the associative structure related to the actual spaces(the contextual relevance of cued goods). Consequently for the cued goods, half of the cued goods inside the space maintained the same coordinates across abstract spaces (context-independent goods), while the other eight cued goods changed both their price and freshness variables depending on the space (context-dependent goods).

Statistical/fMRI methods:

I have a few questions as well about the specific methods used.

First, the use of the Monte Carlo simulation is slightly opaque and needs a bit more unpacking. For example, the defined geometric shape seems to comprise almost all of the screen (based on Figure 2C). Providing the proportion of the screen that is within the shape would be useful. Additionally, when reading that Monte Carlo was used, we thought that the authors were going to define a baseline, accounting for the participant bias for placing items close together, or account for some other noise. That is, the Monte Carlo approach assumes that all points are independently sampled, but we know that there are biases in participants fixating towards the center of the screen as well as placing items clustered around each other at random. For instance, if one were to drag and drop all of the items in the center, and add some noise, you would still get a very low p-value on the Monte Carlo comparison, but it would not mean much for your data. These sorts of biases could be integrated into your Monte Carlo sampling in order to create a meaningful baseline to compare to the alternative null. In the case of centroid bias + noise, the distances that comprise the behavioral RDMs would show some distribution around the null as opposed to being consistently positive or negative. Showing the distribution of average participant distances that informed each behavioral RDM would therefore be useful in addition to the other behavioral metrics shown in Fig 2. I think the work could be strengthened by thinking through different null scenarios in the placement other than simple independent random points assumption.

5. Following this comment and reviewer 1's comment #2 suggesting we should present the empirical probability without needing Monte Carlo simulations, we remove the Monte Carlo simulation results and focus on presenting the empirical probability, along with the relevant descriptive statistics that informed each behavioral RDM. We also provide the proportion of the screen that is within the shape(45% of the screen) on p. lines. Further addressing reviewer 3 & 4's comment, we ran a control analysis to ensure that participants' good placement within the boundaries was significantly above chance (p< 0.001)compared to a screen-aligned centroid bias (T(28) = -2.36, p = .025), or shaped-aligned centroid bias (t(28)= -3.63;p= .002). In a paragraph in the revised Results section on p.6-7, lines 136-147, we list the empirical probability results, descriptive stats, and centroid bias control analyses

Further confirming that learning implicit 2D map-like representations of the different boundary-defined contexts informed participants' choice behavior, we used a surprise post-fMRI drag-and-rate task to investigate if participants retained a detailed map-like representation of an abstract boundary-defined context. We computed the probability of placing the four boundaries in the specific sequences provided in the task(see Methods). Participants consistently organized boundary goods in the correct order(each boundary good shared a

side with its neighboring boundary good) above chance (p<.001; Fig. 2C). We then examined whether the arrangement of all observed goods during the task accurately reflected the underlying 2D structure of the abstract spaces, thereby ensuring that all goods were positioned within the designated boundaries. We calculated the area occupied by the square and distorted shape in the space (45.1%) and computed the empirical probability of sixteen points falling within the boundaries of the space, which was p<.001. The overall percentage of cued goods placed within the boundaries across all participants was 68%, with a 20% variance (see Fig. 2D for mean placement).

We also list here the updated Methods on p.20, lines 507-518.

We quantified participants' reconstruction performance in the post-scan drag-and-rate task. We computed the area of the screen occupied by the square and the distorted shape using the shoelace formula. Subsequently, we evaluated the empirical probability of 16 uniformly distributed points falling within these areas, which yielded a p<.001. To rule out a bias in clustering the goods' placement around the center of the screen, we calculated the distance of every cued good from each of the four replaced boundaries and the center of the screen. We then selected the minimum distance from the boundary of each cued good. This resulted in two distance vectors for each participant, one reflecting distances from the boundary and one reflecting distances from the center of the space. We compared the two distance vectors with a paired t-test, measuring the magnitude of the difference between those distances for each participant. We tested the t-statistic from the paired t-tests against zero to calculate group level statistics. To investigate the different cued good placement from the boundaries and centroid of the reconstructed figure, we computed the centroid of the polygon resulting from individual boundary placement and then followed the same procedure.

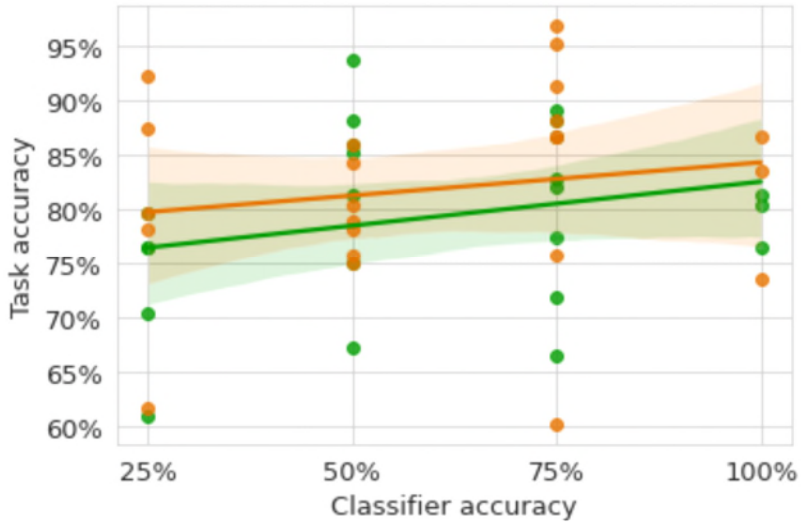
Second, we are a bit confused by the classifier results. The distorted shape has points that are confounded with the extreme ends of each dimension (e.g., the most expensive food and freshest food + the cheapest foods and least fresh food). Instead of getting a representation of the shape of the space per se, the classifier might instead be picking up on extremity of options. This leads to why we are concerned with Figure 4C. Our original expectation was that the magnitude of performance difference would be positively correlated with classification accuracy, regardless of which task participants performed better on. It is hard to estimate visually, but we are not certain that the absolute value of performance difference would correlate with the classification accuracy. Instead, we are wondering why worse performance on the distorted shape leads to worse classifier accuracy. If the participants were not tracking the extremum of the distorted shape properly and the classifier was simply picking up on extremum (instead of abstract knowledge shape), then you would expect this sort of correlation.

6. We apologize for any confusion. To clarify, the square shape has more extreme values on the price dimension, while the distorted shape only has more extreme values on the freshness dimension. We now clarify this point of confusion in the Methods on p.18, lines 417-418.

The square shape had more extreme values on the price dimension, while the distorted shape had more extreme values on the freshness dimension.

Therefore, one shape isn't more extreme in magnitude than the other, so we can rule out that our hippocampal classifier is just measuring the most expensive and freshest food, etc. Additionally, we confirmed that the MVPA distorted v. square context

performance effect wasn't simply due to worse performance on the distorted shape, rather it was related to the difference in performance between the shapes regardless of how well participants performed overall. To this end, we observed no correlation between distorted context decision accuracy and hippocampal distorted context classifier performance(rho=0.24;p=.261), nor between square context decision accuracy and hippocampal square context classifier accuracy(rho=0.15;p=0.502). We include a figure showing both correlations below with the distorted correlation in orange and square correlation in green.



And now report this null effect in the Results on p.13, lines 298-301.

Notably, this correlation wasn't driven by performance effects related to an individual shape. There was no correlation between distorted context decision accuracy and hippocampal distorted context classifier performance (rho=0.24;p=.261), nor between square context decision accuracy and hippocampal square context classifier accuracy(rho=0.15;p=0.502).

So if it isn't due to the aforementioned confound, why does the hippocampal classifier work significantly better on the distorted versus the square shape context? We speculate that participants' internal map-like contextual representations are partially distorted for both task contexts since participants don't need to rescale the two differently ranged axis values to the same dimensions until after the fMRI task with the drag-and-rate. Still, we agree with the reviewers that this aspect of the classifier warrants further consideration in the manuscript and add the following section in the Discussion on p.16, lines 354-61.

The question remains of why does the hippocampal classifier work significantly better on the distorted versus the square shape context? We speculate that participants' internal map-like

contextual representations are partially distorted for both task contexts since participants don't need to rescale the two differently ranged axis values to the same dimensions until after the fMRI task during the drag-and-rate. The idea of flexibly rescaling abstract map-like to current task demands meshes well with the idea that cognitive control processes can reshape the format of cognitive maps of abstract knowledge(Park et al., 2023). Developing new paradigms that use two different dimensions with the same range values without the need for rescaling, or more systematically change the range of dimensions, could be better suited to answer this question.

There are a couple important papers that are missing from the paper. We have copied them below.

In relating event boundaries and hippocampal cognitive maps, a relevant paper is S

Sherrill, K. R., Molitor, R. J., Karagoz, A. B., Atyam, M., Mack, M. L., & Preston, A. R. (2023). Generalization of cognitive maps across space and time. Cerebral Cortex, bhad092. https://doi.org/10.1093/cercor/bhad092

A very relevant preprint in our minds is this one by Park et al. following up on work you have cited.

Park, S. A., Zolfaghar, M., Russin, J., Miller, D. S., O'Reilly, R. C., & Boorman, E. D. (2023). The representational geometry of cognitive maps under dynamic cognitive control [Preprint]. Neuroscience. https://doi.org/10.1101/2023.02.04.527142

7. We thank the reviewers for their reference suggestions. In the revised manuscript, we cite the Sherrill et al., 2023 paper on p.17 lines 385-6.
However, there is evidence that expectations derived from spatial experience generalize to support temporal predictions(Sherrill et al., 2023)
and the Park et al., 2023 preprint on p.16, lines 358-9.
The idea of flexibly rescaling abstract map-like to current task demands meshes well with the idea that cognitive control processes can reshape the format of cognitive maps of abstract knowledge(Park et al., 2023).

Response to Reviewers

We thank the reviewers for their constructive feedback and enthusiasm about our manuscript.

Reviewer #1 (Remarks to the Author):

I appreciate the authors' thorough efforts addressing the robustness of the 2D spatial encoding. The fact that this metric (as opposed to the one-dimensional distances) shows a learning effect is quite interesting, and perhaps speaks to different timeframes of integrating simpler scalar distances into an abstract metric space. Likewise, exploring this in the space of mPFC RSA is a solid addition, and significantly reinforces the notion that participants are relying on boundary effects to form a cognitive map of this space versus cognitive approaches that treat the two axes as entirely independent. I believe this to be a much improved manuscript as a result.

Overall, I am quite satisfied with the changes here.

Reviewer #2 (Remarks to the Author):

The authors have addressed all of my concerns—thank you.
I would, however, suggest including a detailed description of all behavioral analyses and the coefficients used in these models (including t-tests and ANOVAs) in the Methods section. This addition would enhance the clarity of the results.

We thank the reviewer for the helpful suggestion and now add the following paragraph to the Choice Behaviour section of the methods.

We investigated whether participants responded faster when giving correct answers by computing the difference between RTs for correct and incorrect answers. We then tested whether these group differences were significant using a one sample t-test. Additionally, participants' accuracy and RTs relationship to relative choice distance (RD) for each trial was calculated with a Spearman's rank correlation. Next, asking whether the type (contextual relevance) of cued good and boundary shape affected task performance, we ran a 2x2 ANOVA for both accuracy and RTs, with each factor included as a separate variable. Evidence of performance changes as the experiment progressed were tested using a 2x2 ANOVA that included boundary shape and repetition as factors(i.e., whether it was the first or second time participants made choices within a shape context).

Reviewer #2 (Remarks on code availability):

code was not provided

We now add a link to the relevant code in our final draft of the manuscript.

Reviewer #3 (Remarks to the Author):

The authors have done a good job of addressing our concerns. In the Introduction, the motivation and predictions/hypotheses are now clearly outlined. The authors also have added additional rationale for the paradigm and have made clear how the present study is

distinct from past work. The authors have clarified all our questions and confusions about the details of the paradigm and have updated the manuscript with more detail about the task and methods (including an updated task figure). Notably, the authors have now added a new analysis of the choice accuracy data that is a great addition to the paper and significantly strengthens the results.