Zero-shot spatial planning in humans and deep reinforcement learning agents

Denis C. L. Lan (denis.lan@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford
Oxford, United Kingdom

Laurence T. Hunt (laurence.hunt@psy.ox.ac.uk) *
Department of Experimental Psychology, University of Oxford
Oxford, United Kingdom

Christopher Summerfield (christopher.summerfield@psy.ox.ac.uk) *
Department of Experimental Psychology, University of Oxford
Oxford, United Kingdom

(*** Contributed equally as senior authors)**

Abstract:

Humans are particularly good at planning 'zero-shot' (i.e. without prior experience of the environment), a skill that is especially apparent in spatial domains (e.g., navigating a new city). Zero-shot spatial planning likely depends on both 'transition-based' strategies that focus on connectivity between states and 'vector-based' strategies that focus on their relative spatial locations. We developed a novel behavioral paradigm to dissociate the use of the two strategies and show that human participants successfully arbitrate between them for zero-shot planning by using vector-based strategies to head in the general goal direction and transition-based strategies to fine-tune navigation near landmarks. Deep reinforcement learning models trained on the same task learn behavioral policies that are strikingly similar to that of humans. Analysis of the models' learnt representations reveal the emergence of functional 'modules' that implement these strategies, each with distinct informational content, representational geometries, and activation patterns.

Keywords: spatial, planning, navigation, cognitive maps

Introduction

Zero-shot planning likely depends on our ability to exploit both environment-specific state-transition structures and generalizable abstract structures. This roughly maps onto two classes of models of spatial navigation: the former is reminiscent of transition-based strategies, such as successor representation or tree search, that learn from experienced state transitions (de Cothi et al., 2022), while the latter is reminiscent of vector-based strategies that focus on the relative spatial locations of states in the environment (Banino et al., 2018). Both strategies are likely necessary for zero-shot planning, yet it is unclear how humans arbitrate between them. Here, we ask whether and how zero-shot planning depends on combining these two strategies. We hypothesize that humans rely predominantly on vector-based strategies to head in the general goal direction, and transition-based strategies to fine-tune navigation in areas where the transition structure is more familiar (e.g., near landmarks; Lan et al., 2023). We then ask whether human behavior resembles that of deep reinforcement learning (RL) models meta-trained for zero-shot spatial planning, and probe the models' representations to ask how these strategies might be differently implemented.

Method

We developed a behavioural paradigm designed to dissociate vector-based and transition-based strategies. Participants navigated through an 8x8 grid full of 'objects'. In the *learning* phase of each trial, participants saw a top-down view of the grid (Fig. 1A). They clicked on a sequence of squares, highlighted blue, successively revealing the 'landmark' objects that were at the corresponding locations. Participants learnt the locations of only a subset (2 to 16) of all objects. After clicking on the blue squares, participants clicked on a yellow square to reveal the 'goal' object location for the upcoming trial. Every trial involved a completely new grid of new objects, hence requiring participants to plan 'zero-shot' without prior navigational experience and based solely on the knowledge of a few landmarks. In the *test* phase, participants started in a random, previously unlearnt location and were required to navigate to the 'goal' object (Fig. 1B). The object associated with the current state was always displayed singly and centrally. On every step, participants could navigate the grid in one of two ways. They could click on arrows located on one side of the screen, which took them one step in the corresponding direction ('vector-based' strategy). Alternatively, they could click one of the adjacent objects (displayed in random order), which moved them to the state corresponding to that object ('transition-based' strategy). Crucially, both response methods allowed participants to move to the *same* adjacent states, but their choice revealed whether they were focusing on goal direction or state transitions. Every step cost 50 points and reaching the goal earned participants 1000 points.

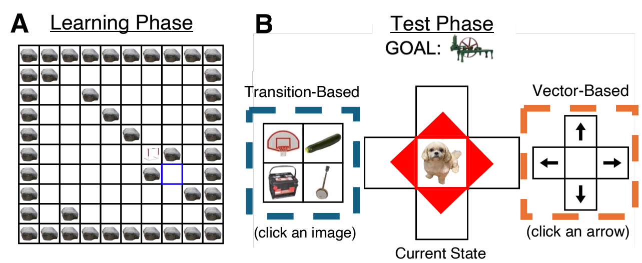

Figure 1: Task schematic for the learning (A) and test phases (B) of each trial

200 participants participated in Experiment 1 and 100 participants participated in Experiment 2, which was a pre-registered replication of results from Experiment 1. In Experiment 1, we manipulated across blocks which response strategies were available. There were four conditions: one where both strategies were always available ('Both'), where only one or the other strategy was available ('Vectors Only' or 'Transitions Only'), and where the type of strategy available randomly alternated on each step ('Random Alternation'). Half of the participants navigated in an open-field environment and half navigated in a cluttered environment with intervening obstacles. Moreover, we trained deep RL agents on all conditions in Experiment 1 with Proximal Policy Optimisation (PPO). The agent consisted of a shared LSTM with 100 units and separate policy and value heads (MLPs with 2 layers of 64 units each). We trained 10 models, each initialised with a different seed.

Results

Human and Model Behaviour. Both participants and deep RL models performed best in the ‘Both’ condition, suggesting that zero-shot planning depends on freely arbitrating between vector- and transition-based strategies (Fig. 2A/B). Compared to the ‘Both’ condition, both humans and models took more steps to get to the goal in the ‘Vectors Only’, (linear mixed effects model; humans: $t(278) = 3.47$, $p < .001$, models: $t(22) = 14.26$, $p < .001$), ‘Transitions Only’, (humans: $t(207) = 19.08$, $p < .001$; models: $t(17) = 76.09$, $p < .001$) and ‘Random Alternation’ conditions (humans: $t(194) = 9.27$, $p < .001$; models: $t(36) = 23.54$, $p < .001$).

When agents could freely choose between strategies, they relied predominantly on ‘vector’ responses, but used ‘transition’ responses to fine-tune their navigation adjacent to goals (mixed effects logistic regression; humans: $z = 20.30$, $p < .001$, models: $z = 34.87$, $p < .001$) and landmarks (humans: $z = 14.67$, $p < .001$, models: $z = 29.75$, $p < .001$; Fig. 2C/D). While agents knew only a few landmarks, they could also learn about the environment’s transition structure during navigation itself: indeed, both humans and deep RL agents use ‘transition-based’ responses more at states that had been previously encountered during navigation (humans: $z = 4.90$, $p < .001$, models: $z = 11.09$, $p < .001$). These human behavioural effects replicated in pre-registered Experiment 2, where participants only experienced the ‘Both’ condition in a cluttered environment.

Model Representations. The results reported here represent those from the best-performing model, but findings replicate across models. We identified the LSTM units responsible for implementing ‘vector’ vs ‘transition-based’ strategies by taking the 10 units whose cell state responses correlated most strongly with the output logits of either the ‘vector’ or the ‘transition’ actions in the policy network (Fig. 2E). Lesioning these units led to a double dissociation in performance on the ‘Vectors Only’ and ‘Transitions Only’ conditions (Fig. 2F). Decoding analyses suggested that these units encoded different task variables: on held-out trials, ‘vector’ units had lower decoding error for spatial variables like x/y-coordinates ($t(150998) = -14.63$, $p < .001$), while ‘transition’ units had lower decoding error on landmark adjacency ($z = 56.10$, $p < .001$). PCA on cell state responses revealed that the representational geometry of ‘vector’ units respected spatial structure, especially after a landmark had been encountered (Fig. 2G), while ‘transition’ units represented landmarks and non-landmarks differently without spatial structure (Fig. 2H). Lastly, we looked at the activation patterns of the units during navigation. ‘Vector’ units responded strongly near borders of the grid across environments, reminiscent of ‘boundary’ cells in the entorhinal cortex (Solstad et al., 2008; Fig. 2I). ‘Transition’ units remapped their peak responses to locations of landmarks and goals, reminiscent of hippocampal firing fields shifting to landmark and goal locations (Gauthier & Tank, 2018; Gothard et al., 1996; Muhle-Karbe et al., 2023; Fig. 2J).

Summary. Overall, our results suggest that humans successfully combine vector- and transition-based strategies for zero-shot planning. Analysis of deep RL models’ learnt representations reveal different computational implementations of each strategy, making predictions for future neural experiments.

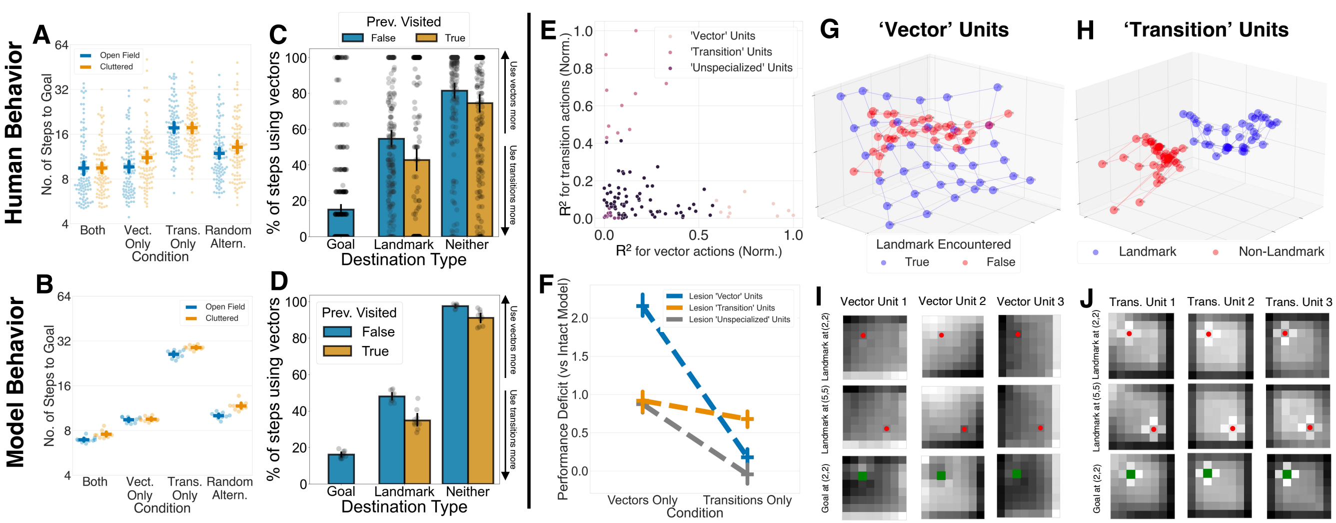

Figure 2: A/B: Human and model performance by condition. C/D: Human and model use of vectors by type of destination and whether the new state had been visited before. E: R^2 value for each unit’s cell-state activity predicting the output for the ‘vector’ or ‘transition’ actions. F: Model performance after lesions. G/H: Representational geometry of ‘vector’/‘transition’ units. I/J: Response patterns of ‘vector’/‘transition’ units.

Acknowledgments

D.L. is supported by the Clarendon Fund, the Christopher Welch Trust, and a Social Science Research Council (Singapore) Graduate Research Fellowship. L.H. is supported by the Wellcome Trust and Royal Society. C.S. is supported by an ERC Consolidator Award and a Wellcome Trust Discovery Award.

References

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Wayne, G., Soyer, H., Viola, F., Zhang, B., Goroshin, R., Rabinowitz, N., Pascanu, R., Beattie, C., Petersen, S., … Kumaran, D. (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), Article 7705. https://doi.org/10.1038/s41586-018-0102-6

de Cothi, W., Nyberg, N., Griesbauer, E.-M., Ghanamé, C., Zisch, F., Lefort, J. M., Fletcher, L., Newton, C., Renaudineau, S., Bendor, D., Grieves, R., Duvelle, É., Barry, C., & Spiers, H. J. (2022). Predictive maps in rats and humans for spatial navigation. *Current Biology*, *32*(17), 3676-3689.e5. https://doi.org/10.1016/j.cub.2022.06.090

Gauthier, J. L., & Tank, D. W. (2018). A Dedicated Population for Reward Coding in the Hippocampus. *Neuron*, *99*(1), 179-193.e7. https://doi.org/10.1016/j.neuron.2018.06.008

Gothard, K. M., Skaggs, W. E., & McNaughton, B. L. (1996). Dynamics of mismatch correction in the hippocampal ensemble code for space: Interaction between path integration and environmental cues. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *16*(24), 8027–8040. https://doi.org/10.1523/JNEUROSCI.16-24-08027.1996

Lan, D. C. L., Hunt, L. T., & Summerfield, C. (2023). Use of Vector- and Transition-based Strategies is Modulated by Knowledge of the Environment in Human Spatial Planning. *2023 Conference on Cognitive Computational Neuroscience*. 2023 Conference on Cognitive Computational Neuroscience, Oxford, UK. https://doi.org/10.32470/CCN.2023.1426-0

Muhle-Karbe, P. S., Sheahan, H., Pezzulo, G., Spiers, H. J., Chien, S., Schuck, N. W., & Summerfield, C. (2023). Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron*, *111*(23), 3885-3899.e6. https://doi.org/10.1016/j.neuron.2023.08.021

Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B., & Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science (New York, N.Y.)*, *322*(5909), 1865–1868. https://doi.org/10.1126/science.1166466