Parallel Decision-Making yields Disentangled World Models: Impact and Implications

Pantelis Vafidis
Computation & Neural Systems
Caltech
pvafeidi@caltech.edu

Aman Bhargava
Computation & Neural Systems
Caltech
abhargav@caltech.edu

Antonio Rangel
Humanities and Social Sciences
Caltech
arangel@caltech.edu

Abstract

Abstract, or disentangled, representations are a promising mathematical framework for efficient and effective out-of-distribution (OOD) generalization. Reflecting the topology of the real world in their representational geometry, they offer a compelling reason for why both biological and artificial systems might converge to such interpretable representations. We here highlight recent results demonstrating that disentanglement comes about naturally when (biological or artificial) agents solve canonical decision-making tasks that require evidence aggregation, in parallel. The tasks tie closely to Bayesian filtering theory, and should be solved by any agent that deals with a noisy world. Intriguingly, theory and experiments together suggest that solving such day-to-day decisions involving latent variables present in the real world directly leads to representations that (1) preserve the topology of the world, (2) isolate such factors of variation, and (3) are consistent across individuals. Furthermore, the highlighted work provides exact mathematical conditions for the emergence of such representations, and demonstrate the importance of noise in facilitating disentangling. The findings are consistent across tasks types and architectures, and we find that transformers are particularly suited for disentangling representations, which might explain their unique world understanding abilities. The universality of the tasks makes us believe that they present a prime candidate for OOD generalization in the brain. Hence, it should be no surprise that such disentangled, topology-preserving representations are widely found in the brain, in examples as disparate as navigation, decision-making and memory. We here expand upon and discuss in detail about potential implications of these findings, for machine learning and neuroscience alike.

1 Introduction

Humans and animals can generalize to new settings effortlessly, leveraging a combination of past experiences and world models [Lake et al., 2015, 2016]. Modern foundation models also display emergent out-of-distribution (OOD) generalization abilities, in the form of zero- or few-shot learning [Brown et al., 2020, Pham et al., 2021, Oquab et al., 2023].

One mechanism for generalization is through abstract, or *disentangled*, representations [Higgins et al., 2017, Kim and Mnih, 2018, Johnston and Fusi, 2023]. These two concepts are interrelated yet somewhat distinct [Ostojic and Fusi, 2024]. An abstract representation of $x_1, \ldots, x_n$ represents each

1st NeuroAI workshop @ 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

x_i linearly and approximately mutually orthogonally. Disentangled representations encode each x_i orthogonally, without the necessity of linearity. When a representation is abstract, a linear decoder (i.e. downstream neuron) trained to discriminate between two categories can readily generalize to stimuli not observed in training, due to the structure of the representation. Furthermore, the more disentangled the representation is, the lower the interference from other variables and hence the better the performance. This corresponds to decomposing a novel stimulus into its familiar features, and performing feature-based generalization. For instance, imagine you are at a grocery store, deciding whether a fruit is ripe or not. If the brain's internal representation of food attributes (ripeness, caloric content, etc.) is disentangled, then learning to perform this task for bananas would lead to zero-shot generalization to other fruit (e.g. mangos, Figure 1a). Crucially, the visual representation of a mango is high-dimensional, non-linear and noisy, making it particularly challenging to extract a low dimensional latent like "ripeness".

Several brain areas including the amygdala, prefrontal cortex and hippocampus have been found to encode variables of interest in an abstract format [Saez et al., 2015, Bernardi et al., 2020, Boyle et al., 2022, Nogueira et al., 2023, Courellis et al., 2024]. This raises the question of under which conditions do such representations emerge in biological and artificial agents alike. Here we argue that multi-task learning is crucial to get the kind of topology-preserving representations that yield generalization in biological systems, and that a parallel processing view of the brain, in line with the cortical architecture, is naturally conducive to that framework. To do so, we first summarize findings from Vafidis et al. [2024] which proves mathematical conditions for disentanglement and experimentally confirms them in autoregressive architectures (RNNs, LSTMs, transformers) that can deal with noisy sequential real-world data, and then discuss the implications of the work for machine learning and neuroscience alike.

2 Problem formulation

In Vafidis et al. [2024] we consider canonical cognitive neuroscience tasks that involve evidence aggregation over time, mirroring decision-making under uncertainty. Tasks have a trial structure. In each trial, a ground truth vector $\mathbf{x}^* \in \mathbb{R}^D$ ($x_i^* \sim$ *Uniform*$(-0.5, 0.5)$) is sampled (Figure 1b). Each element x_i^* of $\mathbf{x}^*$ corresponds to different options a decision-maker might have, or to different attributes of the same item. The target output for the trial $\mathbf{y}(\mathbf{x}^*) \in \{-1, +1\}^{N_{\text{task}}}$ is a vector of N_{task} +1s and -1s, depending on whether $\mathbf{x}^*$ is above or below each of N_{task} classification boundaries (Figure 1b). The boundaries are fixed, and reflect criteria based on which decisions will be made. Imagine for example that x_1 corresponds to food and x_2 to water reward. Depending on the agent's internal state, one could take precedence over the other, and the degree of preference is reflected in the slope of the line.

We train RNNs and GPT-2 transformers to output the target labels $\mathbf{y}(\mathbf{x}^*)$ (Figure 1c). The networks do not have access to the ground truth $\mathbf{x}^*$ but rather a noised-up, non-linearly transformed version of it. Specifically, the input is $\mathbf{X}(t) \in \mathbb{R}^D$ where $\mathbf{X}(t) = \mathbf{x}^* + \sigma \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$, σ being the input noise standard deviation. The network should integrate noisy samples $\mathbf{X}(t)$ over time, viewed through an static, injective observation map (encoder) f, to estimate $\hat{\mathbf{Y}}_i(t) = \Pr\{y_i(\mathbf{x}^*) = 1 | f(\mathbf{X}(1)), \ldots, f(\mathbf{X}(t))\}$.

3 Contributions

We here summarize the main contributions of Vafidis et al. [2024]:

- We prove that any optimal multi-task classifier is guaranteed to learn an abstract representation of the ground truth contained in the noisy measurements in its latent state, if the classification boundary normal vectors span the input space. Furthermore, the representations are guaranteed to be disentangled if $N_{\text{task}} \gg D$. Intriguingly, noise in the observations is necessary to guarantee the latent state would compute an optimal, disentangled representation of the ground truth (for detailed proofs, see Appendix B of unpublished paper).

- We confirm that RNNs trained to multitask develop abstract representations that generalize OOD as quantified by regression generalization [Johnston and Fusi, 2023] when $N_{\text{task}} \geq D$ (Figure 1f), and that the latent factors are approximately orthogonal when $N_{\text{task}} \gg D$

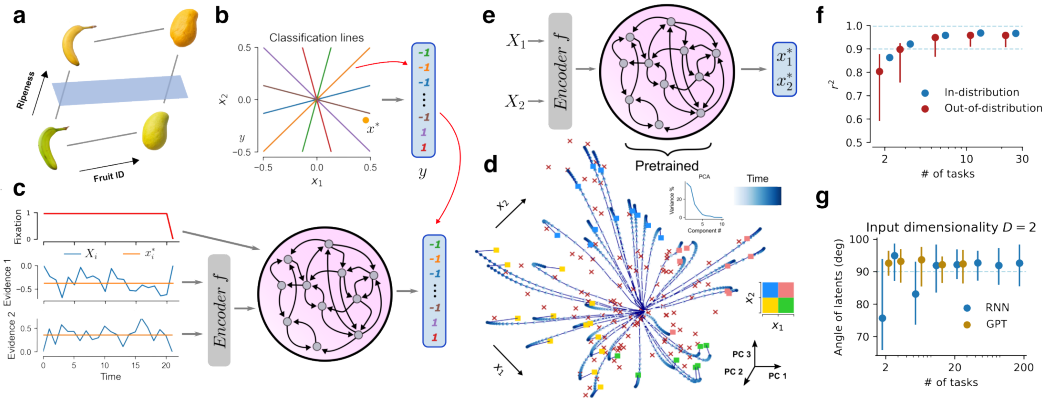

Figure 1: **Learning disentangled representations.** **(a)** A disentangled representation directly lends itself to OOD generalization: a downstream linear decoder that can differentiate ripe from unripe bananas can readily generalize to mangos, even though it has never been trained on mangos. **(b)** The task is to simultaneously report whether the ground truth $\mathbf{x}^*$ lies above $(+1)$ or below (-1) a number of classification lines. **(c)** RNNs are trained to report the outcome of all the binary classifications in **b** at the end of the trial (indicated by the fixation input turning 0). **(d)** Top 3 PCs of RNN activity. Each line is a trial, while color saturation indicates time. All trials start from the center and move outwards, towards the location of $\mathbf{x}^*$ in state space. The last timepoint in each trial (squares) is colored according to the quadrant this trial was drawn from. Red x's correspond to attractors. Input noise here is removed so that trajectories can be visualized easier. The network learns a two-dimensional continuous attractor that seems to provide a disentangled representation of the state space. **(e)** To evaluate OOD generalization, a linear decoder (see **a**) is trained to output the ground truth $\mathbf{x}^*$ at the end of the trial, while keeping network weights frozen. The decoder is trained in 3 out of 4 quadrants and tested OOD in the 4th quadrant. **(f)** ID and OOD generalization performance for networks trained in different number of tasks N_{task}. The 25, 50 and 75 percentiles of r^2 for each network size are reported. ID and OOD performance increase with number of tasks, and the generalization gap decreases, indicating that the networks have indeed learned abstract representations. **(g)** Angles between latent factor decoders. The angles approach 90 degrees as $N_{\text{task}} \gg D$ for RNNs, but they are already close to 90 degrees for $N_{\text{task}} \geq D$ for GPT-style models. The errors that remain for $N_{\text{task}} \geq 24$ for RNNs and for $N_{\text{task}} \geq 2$ for GPT can be attributed to variability in the linear decoder fits. Therefore, we conclude that the representations become disentangled for both models. RNNs are disentangled as $N_{\text{task}} \gg D$ as our theory predicts, but GPT style models disentangle as long as $N_{\text{task}} \geq D$, showcasing their unique ability in disentangling latent factors.

(disentanglement, Figure 1g). The computational substrate of these representations are continuous attractors [Amari, 1977] storing a ground truth estimate in a product space of the latent factors (Figure 1d). Furthermore, these representations preserve the topology of the real world in their structure, where a latent factor (e.g. x_1, x_2) corresponds to a direction in PC space of RNN hidden layer activity, and nearby trials get mapped to nearby trajectories in PC space, without the network being explicitly trained to do so (Figure 1d).

- We show that the setting is robust to a number of manipulations, including interleaved learning of linear and non-linear tasks and free reaction time decisions.
- We reproduce these findings in GPT-2 transformers, which generalize better due to them learning orthogonal representations for lower N_{task}, confirming their appropriateness for constructing world models.
- Finally, we demonstrate the strong advantage of multi-task learning, which scales linearly with D and leads to representations that can be used for any task that involves the same latent variables, over previously proposed mechanisms of representation learning in the brain ("context-dependent computation") [Mante et al., 2013, Yang et al., 2019], which scale linearly with N_{task} and exponentially with D.

Despite being framed in the context of canonical decision-making neuroscience tasks, these results are general; they apply to any system aggregating noisy evidence over time.

4 Implications for representation learning

Topology-preserving representation learning These results have implications for the learning of representations that inherit the topological structure of the world. It suggests that this naturally happens, as long as there are enough tasks to uniquely identify the location of the ground truth $\mathbf{x}^*$ when solving these classifications (see Appendix B of the unpublished paper). Crucially, the constraints from different tasks need to be placed simultaneously on the representation, which explains why representations emerging from context-dependent computation are typically not disentangled. A prime example of such multiple imposed constraints from neurobiology is the fly head-direction system [Vafidis et al., 2022, Wilson, 2023], where a ring-like topology-preserving representation of head direction might be enforced exactly because of it's functional role in driving many downstream circuits for navigation.

Consistency across individuals Potentially even more far reaching, this work implies guarantees about representational alignment across individuals or neural networks. It suggests that as long as we are faced and solve similar problems in the day-to-day world, we are bound to arrive at similar, disentangled representations of latent factors governing these decisions. This is reminiscent of the Platonic representation hypothesis [Huh et al., 2024], which suggests that the convergence in deep neural network representations is driven by a shared statistical model of reality, like Plato's concept of an ideal reality. This could explain why for example modern LLMs come to encode high-level, human-interpretable concepts [Templeton et al., 2024].

Interplay between number of tasks and fine-grainness of representations Intriguingly, this work reveals a fundamental interplay between richness of tasks performed and complexity/detail of the representation learned. If only a small number of tasks are performed, the resulting representations will be fundamentally limited to lie within the space spanned by these tasks. However, as more tasks are added, finer details could be discerned. Therefore, the theorem and experimental results provided are not a one-way-street from dimensionality D of the latent factors to how many tasks N_{task} are required to uncover such latents. Rather, in a complicated and high-dimensional world, the richness of the tasks at hand directly affects the dimensionality D of the latents that can be extracted, allowing for "ground truths" $\mathbf{x}^*$ at different levels of granularity to be explored. The richer the label information available, the more fine-grained the resulting world model will be.

Disentanglement and axis-alignment Axis-alignment is the property by which individual neurons encode distinct latent factors, or equivalently factors are encoded across standard axis of the representation. Computer science [Higgins et al., 2017, Kim and Mnih, 2018, Chen et al., 2018, Hsu et al., 2023, Eastwood et al., 2022] and some recent neuroscience [Whittington et al., 2022] work has incorporated axis-alignment in the definition of disentanglement. However, under our definition above axis-alignment is not a requirement for disentanglement (also see Higgins et al. [2018]). Instead, we suggest that the computer science and computational neuroscience communities should adopt this broader definition of disentanglement, because otherwise we might be missing cases where the factors are not axis-aligned, but they are still orthogonal and can still be isolated by a linear decoder. Our argument is that there is nothing special about individual factors being encoded by individual neurons. Rather, we think that allowing for mixed representations within the definition of disentanglement leads to a more holistic view of disentanglement. A contribution of our work, along with others [Johnston and Fusi, 2023], is to bring this argument to the forefront.

5 Connections to neuroscience and machine learning

5.1 Correspondence to brain processes

The brain encodes variables of interest in a disentangled format, in processes as disparate as memory [Boyle et al., 2022], emotion [Saez et al., 2015], and decision making [Bongioanni et al., 2021]. Furthermore, performance in tasks has been shown to degrade once said neural representations collapse [Saez et al., 2015], supporting the role of abstract representations in guiding generalizable

behavior. Crucially, the cortical architecture lends itself to parallel processing, which readily yields such representations; the cortical column has long been posited as a fundamental unit acting independently and processing complementary parts of sensory information, the results of such parallel processing combined later [Hawkins et al., 2019]. Another candidate brain area for such processing is the thalamus. It has been posited that thalamocortical loops operate in parallel, and combined with internal state-dependent mechanisms lead to state-dependent action selection (e.g. prioritizing water when thirsty over food), while evidence integration occurs in corticostriatal circuits [Rubin et al., 2020]. The algorithmic efficiency of multi-task learning compared to alternatives ("context-dependent computation", Mante et al. [2013], Yang et al. [2019]), makes us think that it is no coincidence that the cortex is built for parallel processing; all the pieces are there, and we feel that the brain has to leverage this feature to construct faithful models of the world, as it does.

5.2 Multitasking vs. Multi-task learning

While our theory stems from parallel processing, i.e. multi-task learning, it is not contingent upon the parallel *execution* of multiple tasks, i.e. multitasking. Behaviorally, the agent need only perform one action, the one most appropriate to it's current internal state (e.g. thirst vs. hunger in the example above). What we posit is that tasks that have been performed by the agent before and rely on the same input are still resolved somewhere in the brain, by the brain circuits (e.g. cortical columns Hawkins et al. [2019]) previously responsible for them, instead of the entire decision-making brain area focusing only on the current task [Mante et al., 2013]. We feel that this is a more natural way of thinking about how the brain manages different tasks, with older tasks still leaving traces somewhere in the brain [Losey et al., 2024], and this theory is closely related to the widely observed phenomenon of memory replay [Foster and Wilson, 2006].

5.3 Relation to path-integration and value-based decision-making

Our findings directly link to two important neuroscientific findings: spatial cognition and value-based decision-making. First, the tasks here bear close resemblance to path-integration, i.e. the ability of animals to navigate space only relying on their proprioceptive sense of linear and angular velocity [Mittelstaedt and Mittelstaedt, 1980, Burak and Fiete, 2009, Vafidis et al., 2022]. In path-integration animals integrate velocity signals to get location, while here we integrate noisy evidence to get rid of the noise. In path-integration, networks have to explicitly report distances, while in our setting distances are estimated implicitly, as a by-product of estimating $\Pr\{y_i(\mathbf{x}^*) = 1\}$ (see Lemma B.3 in proof in Vafidis et al. [2024]). We learn abstract representations in the form of a 2D "sheet" continuous attractor, while the computational substrate for path integration is a 2D toroidal attractor [Gardner et al., 2022, Sorscher et al., 2023] – not an abstract representation. The conditions under which a 2D sheet vs. toroidal continuous attractor is learned is a potential area of future research.

Second, decision making experiments in monkeys result in a 2D abstract representation in the medial frontal cortex, which supports novel inferential decisions [Bongioanni et al., 2021]. Likewise, context-dependent decision-making experiments in humans also resulted in orthogonal, abstract representations [Flesch et al., 2022].

5.4 Relation to machine learning paradigms

The experiments in Vafidis et al. [2024] are inspired by canonical cognitive neuroscience tasks, rather than state-of-the-art ML paradigms. Yet, the conclusions drawn are informative concerning the fundamental nature of generalization. For instance, why do foundation models generalize well in various domains? We suggest that parallel processing forces learning of generalizable world models, and our setting directly applies to settings where neural networks predict a rich representation of the world from partial observations. Some examples are predictive coding where high-dimensional next states have to be predicted [Gornet and Thomson, 2024], which is equivalent to the classification objective of predicting which objects are going to be in the field of view (and where), and self-supervised learning, where multiple missing image patches have to filled up at once [Dosovitskiy et al., 2020].

Finally, an alternative to multi-task learning that we explore is slow, interleaved learning. This allows the weights of a neural network to be effectively conditioned to solve all the tasks simultaneously. The relation between multi-task and interleaved learning is a promising topic for future research.

6 Acknowledgements

PV would like to thank the Onassis Foundation and AR the NOMIS Foundation for funding. AB thanks the NIH PTQN program for funding. No competing interests to declare. We would like to thank Yisong Yue for early discussions and Stefano Fusi for early feedback.

References

S Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2):77–87, 1977. doi: 10.1007/bf00337259. URL https://doi.org/10.1007/bf00337259.

Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4): 954–967.e21, November 2020. doi: 10.1016/j.cell.2020.09.031. URL https://doi.org/10.1016/j.cell.2020.09.031.

A. Bongioanni, D. Folloni, L. Verhagen, J. Sallet, M. C. Klein-Flügge, and M. F. S. Rushworth. Activation and disruption of a neural mechanism for novel choice in monkeys. *Nature*, 591(7849): 270–274, January 2021. doi: 10.1038/s41586-020-03115-5. URL https://doi.org/10.1038/s41586-020-03115-5.

Lara M. Boyle, Lorenzo Posani, Sarah Irfan, Steven A. Siegelbaum, and Stefano Fusi. Tuned geometries of hippocampal representations meet the demands of social memory, January 2022. URL https://doi.org/10.1101/2022.01.24.477361.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Yoram Burak and Ila R. Fiete. Accurate path integration in continuous attractor network models of grid cells. *PLoS Computational Biology*, 5(2):e1000291, February 2009. doi: 10.1371/journal.pcbi.1000291. URL https://doi.org/10.1371/journal.pcbi.1000291.

Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *CoRR*, abs/1802.04942, 2018. URL http://arxiv.org/abs/1802.04942.

Hristos S. Courellis, Juri Minxha, Araceli R. Cardenas, Daniel L. Kimmel, Chrystal M. Reed, Taufik A. Valiante, C. Daniel Salzman, Adam N. Mamelak, Stefano Fusi, and Ueli Rutishauser. Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632 (8026):841–849, August 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07799-x. URL http://dx.doi.org/10.1038/s41586-024-07799-x.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

Cian Eastwood, Andrei Liviu Nicolicioiu, Julius von Kügelgen, Armin Kekić, Frederik Träuble, Andrea Dittadi, and Bernhard Schölkopf. Dci-es: An extended disentanglement framework with connections to identifiability, 2022. URL https://arxiv.org/abs/2210.00364.

Timo Flesch, David G. Nagy, Andrew Saxe, and Christopher Summerfield. Modelling continual learning in humans with hebbian context gating and exponentially decaying task signals, 2022. URL https://arxiv.org/abs/2203.11560.

David J. Foster and Matthew A. Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680–683, February 2006. ISSN 1476-4687. doi: 10.1038/nature04587. URL http://dx.doi.org/10.1038/nature04587.

Richard J. Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A. Baas, Benjamin A. Dunn, May-Britt Moser, and Edvard I. Moser. Toroidal topology of population activity in grid cells. *Nature*, 602(7895):123–128, January 2022. doi: 10.1038/s41586-021-04268-7. URL https://doi.org/10.1038/s41586-021-04268-7.

James Gornet and Matt Thomson. Automated construction of cognitive maps with visual predictive coding. *Nature Machine Intelligence*, 6(7):820–833, July 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00863-1. URL http://dx.doi.org/10.1038/s42256-024-00863-1.

Jeff Hawkins, Marcus Lewis, Mirko Klukas, Scott Purdy, and Subutai Ahmad. A framework for intelligence and cortical function based on grid cells in the neocortex. *Frontiers in Neural Circuits*, 12, January 2019. ISSN 1662-5110. doi: 10.3389/fncir.2018.00121. URL http://dx.doi.org/10.3389/fncir.2018.00121.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations, 2018. URL https://arxiv.org/abs/1812.02230.

Kyle Hsu, Will Dorrell, James C. R. Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via latent quantization, 2023. URL https://arxiv.org/abs/2305.18378.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

W. Jeffrey Johnston and Stefano Fusi. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, 14(1), February 2023. doi: 10.1038/s41467-023-36583-0. URL https://doi.org/10.1038/s41467-023-36583-0.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, December 2015. doi: 10.1126/science.aab3050. URL https://doi.org/10.1126/science.aab3050.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, November 2016. doi: 10.1017/s0140525x16001837. URL https://doi.org/10.1017/s0140525x16001837.

Darby M. Losey, Jay A. Hennig, Emily R. Oby, Matthew D. Golub, Patrick T. Sadtler, Kristin M. Quick, Stephen I. Ryu, Elizabeth C. Tyler-Kabara, Aaron P. Batista, Byron M. Yu, and Steven M. Chase. Learning leaves a memory trace in motor cortex. *Current Biology*, 34(7):1519–1531.e4, April 2024. ISSN 0960-9822. doi: 10.1016/j.cub.2024.03.003. URL http://dx.doi.org/10.1016/j.cub.2024.03.003.

V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, November 2013. doi: 10.1038/nature12742. URL https://doi.org/10.1038/nature12742.

M. L. Mittelstaedt and H. Mittelstaedt. Homing by path integration in a mammal. *Naturwissenschaften*, 67(11):566–567, November 1980. doi: 10.1007/bf00450672. URL https://doi.org/10.1007/bf00450672.

Ramon Nogueira, Chris C. Rodgers, Randy M. Bruno, and Stefano Fusi. The geometry of cortical representations of touch in rodents. *Nature Neuroscience*, 26(2):239–250, January 2023. doi: 10.1038/s41593-022-01237-9. URL https://doi.org/10.1038/s41593-022-01237-9.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Srdjan Ostojic and Stefano Fusi. Computational role of structure in neural activity and connectivity. *Trends in Cognitive Sciences*, 28(7):677–690, July 2024. ISSN 1364-6613. doi: 10.1016/j.tics.2024.03.003. URL http://dx.doi.org/10.1016/j.tics.2024.03.003.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning, 2021. URL https://arxiv.org/abs/2111.10050.

Jonathan E. Rubin, Catalina Vich, Matthew Clapp, Kendra Noneman, and Timothy Verstynen. The credit assignment problem in cortico-basal ganglia-thalamic networks: A review, a problem and a possible solution. *European Journal of Neuroscience*, 53(7):2234–2253, May 2020. ISSN 1460-9568. doi: 10.1111/ejn.14745. URL http://dx.doi.org/10.1111/ejn.14745.

A. Saez, M. Rigotti, S. Ostojic, S. Fusi, and C.D. Salzman. Abstract context representations in primate amygdala and prefrontal cortex. *Neuron*, 87(4):869–881, August 2015. doi: 10.1016/j.neuron.2015.07.024. URL https://doi.org/10.1016/j.neuron.2015.07.024.

Ben Sorscher, Gabriel C. Mel, Samuel A. Ocko, Lisa M. Giocomo, and Surya Ganguli. A unified theory for the computational and mechanistic origins of grid cells. *Neuron*, 111(1):121–137.e13, January 2023. doi: 10.1016/j.neuron.2022.10.003. URL https://doi.org/10.1016/j.neuron.2022.10.003.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Pantelis Vafidis, David Owald, Tiziano D'Albis, and Richard Kempter. Learning accurate path integration in ring attractor models of the head direction system. *eLife*, 11:e69841, jun 2022. ISSN 2050-084X. doi: 10.7554/eLife.69841. URL https://doi.org/10.7554/eLife.69841.

Pantelis Vafidis, Aman Bhargava, and Antonio Rangel. Disentangling representations through multi-task learning, 2024. URL https://arxiv.org/abs/2407.11249.

James C. R. Whittington, Will Dorrell, Surya Ganguli, and Timothy E. J. Behrens. Disentanglement with biological constraints: A theory of functional cell types, 2022. URL https://arxiv.org/abs/2210.01768.

Rachel I. Wilson. Neural networks for navigation: From connections to computations. *Annual Review of Neuroscience*, 46(1):403–423, July 2023. ISSN 1545-4126. doi: 10.1146/annurev-neuro-110920-032645. URL http://dx.doi.org/10.1146/annurev-neuro-110920-032645.

Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, Feb 2019. ISSN 1546-1726. doi: 10.1038/s41593-018-0310-2. URL https://doi.org/10.1038/s41593-018-0310-2.