THE EFFECT OF REPRESENTATIONAL COMPRESSION ON FLEXIBILITY ACROSS LEARNING IN HUMANS AND ARTIFICIAL NEURAL NETWORKS

Mia Whitefield
Department of Experimental Psychology
University of Oxford
Oxford, UK
mia.whitefield@linacre.ox.ac.uk

Christopher Summerfield
Department of Experimental Psychology
University of Oxford
Oxford, UK
christopher.summerfield@psy.ox.ac.uk

ABSTRACT

Humans can generalise from past experiences to novel situations as well as revise prior knowledge to flexibly adapt to changing contexts and goals. The representational geometry framework formalises how information is structured in the brain and suggests that abstraction involves a trade-off between generalisation and flexibility. However, how the geometry of task representations evolves across learning and how it corresponds to behaviour remains unclear. Here, we tested the hypothesis that task representations become compressed throughout learning, trading flexibility for task efficiency. Using an extra-dimensional shifting task, we manipulated the pretraining length to control the degree of compression. In both humans and artificial neural networks, longer pretraining was associated with decreased flexibility. Analysis of network dynamics suggest that greater compression incurs a higher representational reorganisation cost, restricting flexibility. However, the introduction of an auxiliary reconstruction loss maintains higher dimensionality, mitigating the impairment of flexibility. Our findings point towards a representational geometry-based mechanism that explains how representational compression constrains flexibility, and how preserving representational richness enhances flexibility.

1 INTRODUCTION

Humans acquire abstract knowledge by integrating information across experiences, which is essential for generalisation. However, abstraction entails information loss, potentially limiting flexibility in adapting to environmental changes or shifting goals, suggesting a trade-off between generalisation and flexibility.

Humans can exhibit both high flexibility and strong generalisation depending on the context. In contrast, artificial neural networks (ANNs) typically excel at in-domain generalisation but display limited flexibility, tending to overfit to the domain of the training data (Weiss et al., 2016) and suffering from catastrophic forgetting (French, 1999; Parisi et al., 2019). Recent work has shown that certain network modifications can overcome these limitations (Benna & Fusi, 2016; Kirkpatrick et al., 2017; Flesch et al., 2023), offering insights into potential biological mechanisms that support flexibility. Studying ANNs provides an opportunity to explore the flexibility-generalisation trade-off with full access to internal representations. This not only refines our understanding of artificial systems, but also informs hypotheses about biological learning (Saxe et al., 2021; Lillicrap & Kording, 2019). By identifying computational learning principles shared across biological and artificial agents, we work toward a more unified understanding of learning in intelligent systems.

The representational geometry framework formalises how activity patterns encode information. For example, the response evoked by a stimulus in the brain can be recorded as a neural activity vector. Each stimulus response can be represented as a point in neural space, with axes corresponding to the activity of each neuron (Fusi et al., 2016). The geometry of a set of responses in neural-space can provide insights into the perceived relational structure between stimuli, and offers a way to

quantitatively characterise task representations. Crucially, it enables comparison of representations across individuals, species and with artificial systems.

A key property of representational geometry is dimensionality, which refers to the number of dimensions required to capture the variation of a set of responses in neural-space (Fusi et al., 2016). Low dimensional representations compress irrelevant information, enhancing robustness and generalisation to novel stimuli, but reducing input separability and the diversity of downstream readouts (Badre et al., 2010; Collins & Frank, 2013; Anselmi et al., 2015; Badre et al., 2021). By contrast, high dimensional representations retain richer detail, supporting input separability (Fusi et al., 2016) and behavioural diversity, but at the cost of increased sensitivity to input noise and poorer generalisation to novel inputs (Cohen et al., 2020; Rigotti et al., 2013). This suggests representational dimensionality mediates a trade-off between generalisation and flexibility (Badre et al., 2021).

Here, we focus on control representations, primarily associated with the prefrontal cortex (PFC) (Miller & Cohen, 2001), which integrate sensory inputs with memory, goals, and context to guide behaviour (Badre et al., 2021; Kikumoto et al., 2023). While experimental work has largely focused on the properties of static representational geometries, control representations are inherently dynamic, evolving over the course of learning (Mill & Cole, 2023; Wójcik et al., 2023), as the environment and goals vary (Kikumoto & Mayr, 2020), and within the timescale of a task trial (Bernardi et al., 2020; Shi et al., 2023; Kikumoto et al., 2023; Shi et al., 2023).

It has been theorised that early in learning, high dimensional representations may be optimal for exploration (Enel et al., 2016; Fusi et al., 2016). Later in learning, task irrelevant dimensions may become compressed to minimise energy demand (Musslick & Cohen, 2021; Barak et al., 2013; Wójcik et al., 2023), improving robustness and generalisation (Bernardi et al., 2020). Wójcik et al. (2023) showed that the dimensionality of macaque PFC neural representations decreased over the course of learning a task. However, they did not directly test how the representational geometry relates to behaviour. How task representations are acquired over time, and how they constrain behaviour and future learning, remain open questions.

We hypothesise that representational compression reduces flexibility by limiting the ability to reintegrate previously compressed features. To test this, we designed an extra-dimensional shifting task and we manipulated the pretraining length to control the degree of representational compression. First we simulated the task in artificial neural networks and analysed the hidden layer representations to gain mechanistic insight into how geometry influences flexibility in adapting to the shift. Next we collected behavioural data from humans and compared their performance to ANN simulations.

2 RESULTS

2.1 TASK DESIGN

To test the hypothesis that representational compression impairs flexibility, we designed an extra-dimensional shifting task for both humans and ANNs. The task was designed to investigate how prior learning affects the ability to adapt when the relevant feature shifts.

In the human version of the task, participants learned to predict the locations (either North or South) of symbols on a map. In each trial, they were shown a symbol and freely clicked on the map to make their prediction (Figure 1A). A symbol set was defined by two features (e.g. colour and shape). During the pretraining phase, feature A determined the location of the symbol (Figure 1B). However, partway through the task, the relevant feature switched to feature B. During the post-shift phase, participants would have to shift their attention to feature B in order to learn the new rule. Continued use of the pretraining rule yields 50% accuracy as the locations for half the symbols are "flipped" after the rule change (Figure 1B).

Participants learned the locations of 2 sets of symbols, M1 and M2 (Figure 1C), each was defined by a distinct pair of visual features (abstract shape and pattern, or pictorial image and colour) and was associated with a unique map. M1 received longer pretraining than M2 before the shift-point, where the rule changed (Figure 1D).

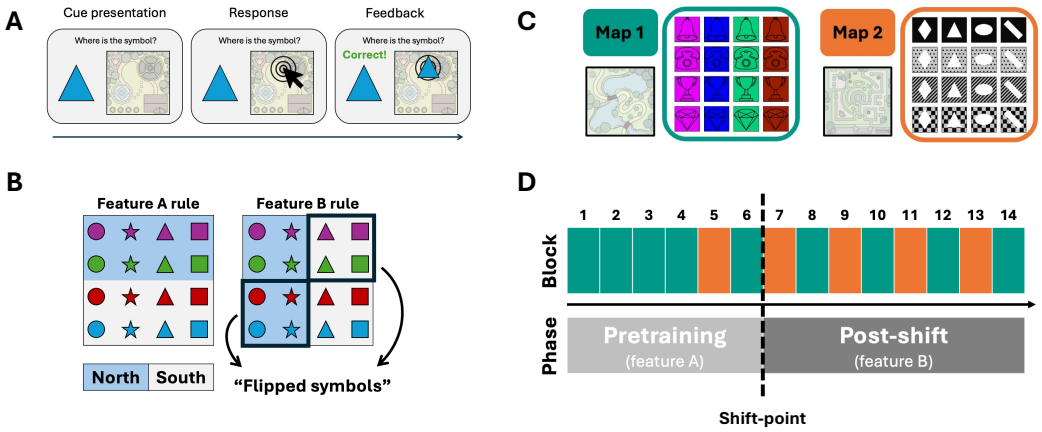

Figure 1: **(A)** Schematic of an example trial of the human task. **(B)** Symbol sets defined by two features, here shape and colour. Task A locations determined by feature A (left) and task B based on feature B (right), with "flipped" symbols changing location between the tasks. **(C)** M1 symbols defined by picture and colour, and M2 defined by pattern and geometric shape. **(D)** Schematic of curriculum. During pretraining, participants learned task A (5 M1 and 1 M2 block). After the shift-point (dashed line), they learned task B with alternating M1 and M2 blocks.

This design allowed us to compare the effect of pretraining length on post-shift flexibility within-subject. We hypothesised that the irrelevant feature (feature B) would be compressed over the course of pretraining. Therefore, we would expect M2 representations to be higher dimensional at the shift-point, facilitating better adaptation to the rule shift compared to M1.

3 Artificial Neural Networks

We first simulated a binary classification version of the task in multilayer perceptrons (MLPs) to formalise predictions about the impact of pretraining length on flexibility. Crucially, full access to the networks' hidden layer representations enabled us to directly measure representational dimensionality across learning. By examining how the representational geometry evolved across the extra-dimensional shift, we gained mechanistic insight into how dimensionality influences flexibility.

3.1 Setup

Previous work has shown that weight initialisations influence the dimensionality of representations in neural networks. Networks initialised with low-variance weights (rich regime) tend to discover task relevant features, forming low dimensional representations. In contrast, networks initialised with high-variance weights (lazy regime) tend to undergo minimal weight changes during learning to form high dimensional representations (Chizat et al., 2018; Flesch et al., 2021). We simulated networks initialised in both regimes to compare how representational geometry evolved at these two extremes of dimensionality. We expected the Rich networks to correspond most closely to human learning as we reasoned humans would solve the task by extracting the relevant feature. The Lazy networks served as a control, illustrating learning in the absence of significant representational change.

Additionally, we reasoned that humans would only partially compress representations, as their strong prior knowledge of the distinct stimulus features would prevent them from completely discarding information about the irrelevant feature within the time frame of the task. In order to instil similar prior knowledge in the MLPs, we first trained them to reconstruct the inputs before the classification task. A classification loss was applied to one output unit responsible for North/South classification, and a reconstruction loss was applied to the remaining 16 output units responsible for input reconstruction (Figure 2B). The classification and reconstruction losses were scaled by the constants α_{class} and

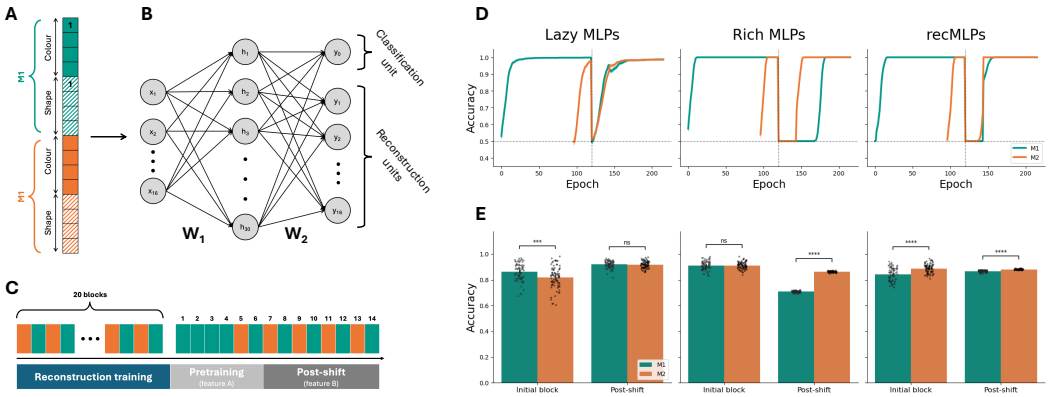

Figure 2: **(A)** Input two-hot vectors representing the feature values. **(B)** MLP architecture schematic. **(C)** MLP curriculum for recMLPs. Orange and green blocks indicate training on M1 and M2 inputs respectively. **(D)** Accuracy across epochs for rich (middle) and lazy (left) MLPs, and recMLPs (right) across the task. Vertical and horizontal dashed lines indicate the shift-point and chance accuracy respectively. **(E)** Mean accuracy for initial pretraining block (left) and across post-shift blocks (right). Black dots represent individual mean values. Stars denote Wilcoxon signed rank test significance (0.05 *, 0.01 ** , 0.001 ***, <0.000 ****).

α_{rec} respectively. During the N_{rec} blocks of reconstruction training, we applied L2 regularisation to the weights to keep them small, ensuring that the networks began the task in the rich regime.

We simulated the task under three conditions:

1. **Lazy MLPs** were initialised in the lazy regime, and received no reconstruction training ($N_{rec} = 0$; $\alpha_{rec} = 0$).
2. **Rich MLPs** were initialised in the rich regime, and received no reconstruction training ($N_{rec} = 0$; $\alpha_{rec} = 0$).
3. **recMLPs** were initialised in the rich regime, and trained on 20 blocks of the reconstruction task ($N_{rec} = 20$; $\alpha_{rec} = 1$).

3.2 Performance

In each condition, a set of MLPs (n = 80 to match the human cohort size) were trained on a binary classification version of the symbol task. The inputs were two-hot encoded vectors representing the feature values of the stimulus (Figure 2A).

We compared M1 vs. M2 performance during the initial block of pretraining to assess whether the MLPs generalised the task structure across maps. Only recMLPs exhibited evidence for generalisation, with greater initial block accuracy for M2 vs. M1 ($BF_{10} = 1.961 \times 10^{6}$; Wilcoxon signed rank test: V = 511, p-value = 1.771×10^{-7}). A Bayesian paired t-test provided moderate evidence for no difference in initial learning of M1 vs M2 in Rich MLPs ($BF_{10} = 0.1245$; Wilcoxon signed rank test: V = 1584, p-value = 0.8648), indicating there was no generalisation across maps. The Lazy MLPs displayed slower initial learning of M2 vs M1 ($BF_{10} = 145.9$; Wilcoxon signed rank test: V = 2405, p-value = 1.681×10^{-4}), implying that learning M1 interfered with learning of M2.

Our primary interest was how pretraining length affected flexibility, therefore, we compared M1 and M2 performance across the post-shift phase (Figure 2E). Rich MLPs exhibited greater mean accuracy across the post-shift phase for M2 vs. M1 ($BF_{10} = 2.072 \times 10^{101}$; Wilcoxon signed rank test: V = 0, p-value = 7.983×10^{-15}), indicating that pretraining length impaired flexibility. Greater M2 vs. M1 post-shift accuracy was also observed in recMLPs ($BF_{10} = 1.344 \times 10^{19}$; Wilcoxon signed rank test: V = 0, p-value = 1.169×10^{-14}), although, this performance difference was more subtle. For the Lazy MLPs, a Bayesian paired t-test provided moderate evidence for no difference in M1 vs M2 post-shift performance ($BF_{10} = 0.160$; Wilcoxon signed rank test: V = 1759, p-value = 0.3844), suggesting that pretraining length had no effect on flexibility.

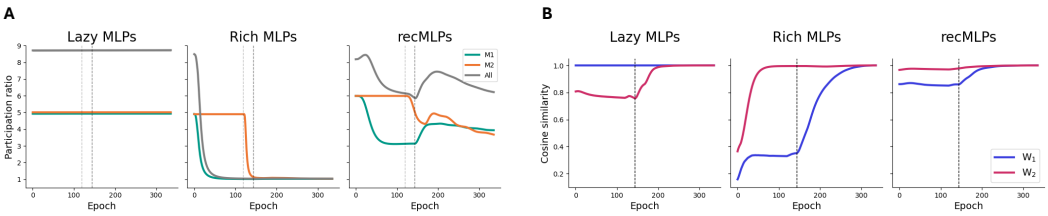

Figure 3: **(A)** Cosine similarity between weights and each time point and the final weights for layer 1 (blue) and layer 2 (pink). **(B)** Participation ratio of MLP hidden layer activations across the task for M1 (green) and M2 (orange) inputs, and all 32 inputs (grey). Black and grey vertical dashed lines indicate the shift-point and start of M2 learning respectively.

In summary, longer pretraining resulted in greater switch costs in Rich MLPs, whereas there was no such difference for Lazy networks. Additionally, during pretraining, Rich MLPs exhibited no difference in initial learning of M1 vs M2, while Lazy MLPs displayed interference between them.

3.3 Representational Geometry

The use of MLPs enabled us to analyse the geometry of the hidden layer activations and examine how task representations evolved across learning and rule shifting. We visualised these representations at various stages of the task using multidimensional scaling (MDS). After each epoch, we extracted the hidden layer activations for the input set and applied MDS to activations across all epochs to track how representational distances between inputs evolved over time. We also calculated the participation ratio, PR, (see Methods) as a measure of the effective dimensionality of the hidden layer representations.

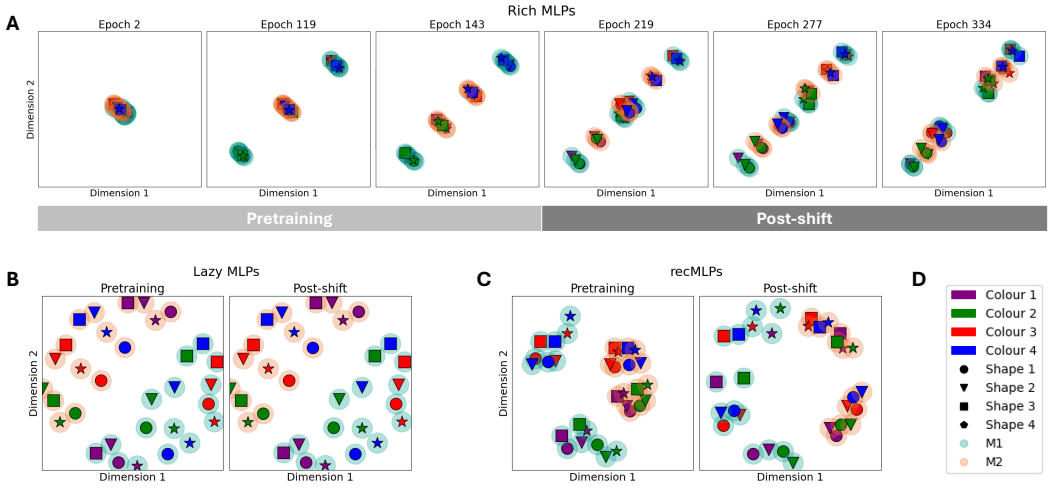

Figure 4: Multidimensional scaling (MDS) visualisations for M1 (orange background circle) and M2 (green background circle) symbol hidden layer representations. **(A)** Rich MLP MDS plots at stages during pretraining and the post-shift phase. **(B)** Rich recMLPs and **(C)** lazy MLPs MDS plots at the end of pretraining (left) and the end of the post-shift phase (right).

3.3.1 Rich MLPs

Figure 4A displays the hidden representations in Rich MLPs across the task. During pretraining, the symbol representations divided into 2 clusters based on the class label (North/South). After the rule shift, the flipped symbols migrated across the decision boundary to join the alternate cluster. Figure 3A confirms that the dimensionality decreased across the task, approaching 1, as the input representations clustered into two groups along the relevant feature dimension. Furthermore, the

total dimensionality of both M1 and M2 inputs also approached 1, indicating the representations of both maps became aligned in a common subspace.

To quantify how much the weights changed across the task, we calculated the cosine similarity between the weights after each epoch and the final weights. Figure 3B shows that the readout weights (W_2) learned during pretraining remained stable across the post-shift phase, while the embedding weights (W_1) changed significantly after the shift point. Together, these results suggest that reorganisation of the embedding weights drove post-shift adaptation, consistent with the hidden layer visualisations.

This analysis provides a geometric explanation for the flexibility difference between M1 and M2 in Rich MLPs: the symbol representations were more expanded along the relevant dimension, requiring a greater degree of reorganisation to move across the decision boundary. This suggests that compression imposes a representational reorganisation cost.

3.3.2 RECMLPS

In line with our predictions, the recMLPs maintained higher dimensional representations than Rich MLPs, reaching a PR of approximately 3 compared to 1 (Figure 3A). After the shift-point, M1 dimensionality expanded indicating that representational reorganisation was not constrained to the compressed subspace. MDS visualisations (Figure 4C) show that the embedding geometry changed across the post-shift phase. Additionally, the weight similarity dynamics (Figure 3B) indicate that both the readout and embedding weights changed after the shift-point, although the readout weights exhibited only minor adjustments. Additionally, the degree of embedding weight change was smaller in the recMLPs compared to the Rich MLPs. Together, these findings suggest that reconstruction training resulted in higher dimensional embeddings, which maintained access to both relevant and irrelevant features. Therefore, less representational reorganisation was required to adapt to the extra-dimensional shift, and so a smaller flexibility impairment resulted from longer pretraining. This suggests that the structured priors acquired through reconstruction training enabled the networks to retain task-general information, which supported both generalisation and flexibility.

3.3.3 LAZY MLPS

In contrast to the Rich MLPs, Lazy MLPs maintained high dimensional representations throughout the task (Figure 3B). MDS visualisations show minimal changes in embedding geometry after the rule shift. Figure 3A shows that the embedding weights remained stable throughout the task and post-shift adaptation was driven by readout weight changes. As hidden layer representations remained static, no representational reorganisation cost was incurred. This supports the hypothesis that compression impairs flexibility, as in the absence of compression differences, no difference in flexibility was observed.

3.4 HUMANS

3.4.1 PERFORMANCE

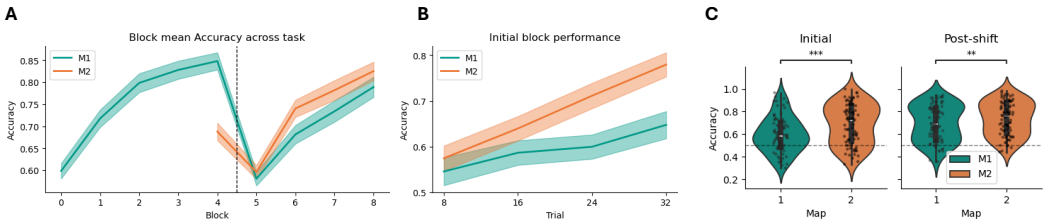

Figure 5: **(A)** Human accuracy for M1 (green) and M2 (orange) trials. Shaded ribbon represents the standard error. Vertical and horizontal dashed lines indicated the shift-point and chance accuracy respectively. **(B, C)** Mean accuracy for the initial block of pretraining and across reversal blocks. Black dots show individual means connected by lines. Stars denote Wilcoxon signed rank test significance (0.01 ** , 0.001 ***).

Eighty adult human participants completed the online task (Figure 5A). During the initial block of pretraining, participants learned M2 faster than M1, achieving greater mean accuracy (Wilcoxon signed-rank test: Z = 924, p = 0.0008) (Figure 5B, C). This suggests that participants generalised the task structure from M1 to M2, mirroring the behaviour of the recMLPs.

To assess the effect of pretraining length on flexibility, we compared M1 and M2 performance during the post-shift phase. Participants showed higher mean accuracy for M2 vs. M1 (Wilcoxon signed rank test: Z = 992, p = 0.0026) (Figure 5C). This pattern was observed in both Rich MLPs and recMLPs, though the small difference in humans more closely resembles the recMLPs.

Together, these human behavioural signatures align most closely with the recMLP simulations, which similarly exhibit positive transfer from M1 to M2 during pretraining as well as the subtle difference in switch cost with longer pretraining.

3.4.2 Evidence for compression in humans

While we could not directly measure representational geometry in humans from the behavioural data, reaction times can provide indirect behavioural evidence for representational compression. Prior work suggests that selective attention distorts representational geometries (Chapman & Störmer, 2024), which we expected would contribute to representational compression in this task (Mack et al., 2020). Therefore, we investigated whether participants displayed selective attention to the relevant feature, as this would imply compression of the irrelevant feature.

Task-switching psychology research has shown that RTs are longer on task-switching trials compared to task-repetition trials (Monsell, 2003; Kiesel et al., 2010), as there is a local switch cost incurred when changing strategy. In this task, a repetition (match) trial was defined as a trial where the attended symbol feature value matched that of the previous trial, while switch trials were defined as trials where the value differed (Figure 6A).

We calculated the difference between **S**witch and **M**atch trial mean **RT**s (SMRT) for a given feature, across each block. We expected the magnitude of the SMRT to reflect the degree of attention to that feature, i.e. the SMRT would be positive if the participant was attending to the feature, and zero if they were not. Next we subtracted the feature B SMRT from that of feature A to obtain the SMRT difference (ΔSMRT), a measure of selective attention to feature A over B ($\Delta\text{SMRT} = \text{SMRT}_A - \text{SMRT}_B$).

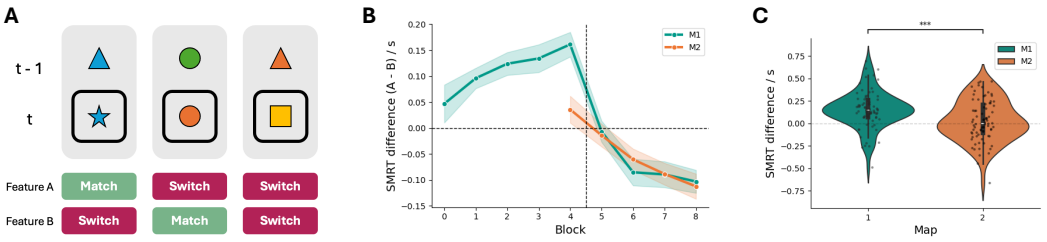

Figure 6: Schematic of switch and match trials. For a given feature, match trials refer to trials where the symbol feature value matches that of the previous trial. Switch trials refer to those where the feature value differs to that of the previous trial. For example, a blue star following a blue triangle is a feature A (colour) match, and a feature B (shape) switch. **(B)** Switch - Match Reaction Time (SMRT) difference ($\text{SMRT}_A - \text{SMRT}_B$) for each block across the task, as a measure of selective attention to Feature A over B. Horizontal line indicates zero difference, and the vertical line represents the shift-point. **(C)** M1 vs M2 mean SMRT difference during the block prior to the shift-point. Black dots represent individual mean values. The stars indicate the Wilcoxon signed rank test significance. Significance codes: 0.05 *, 0.01 ** , 0.001 ***, <0.000 ****.

Figure 6B shows that ΔSMRT increased across pretraining for M1, indicating that participants learned to selectively attend to the relevant feature (A) across this phase of the task. After the shift-point, ΔSMRT for both M1 and M2 began to fall and became increasingly negative, indicating that participants shifted their attention from feature A to feature B. This implies the compression of the irrelevant feature across each phase of the task. Importantly, in the block preceding the shift-point,

ΔSMRT for M1 was greater than M2 (Wilcoxon signed rank test: Z = 848, p = 0.0003), suggesting that the more extensive pretraining on M1 resulted in greater selective attention to the relevant feature relative to M2 (Figure 6C). This analysis provides indirect evidence for greater compression of M1 vs M2 representations prior to the rule-shift, which is consistent with our hypothesis that greater compression of M1 representations contributed to the greater switch cost.

In summary, the recMLP behavioural signatures of human performance align most closely with the recMLP simulations: the positive transfer from M1 to M2 during training as well as the subtle difference in switch cost with longer pretraining.

4 Discussion

We set out to test the hypothesis that representational compression across learning impairs flexibility by comparing extra-dimensional shifting in two stimulus sets (M1 and M2) that differed in pretraining duration.

First we simulated the task in multilayer perceptrons (MLPs) and analysed how representational geometry evolved under different learning conditions to investigate how representational compression impaired flexibility. In Rich MLPs, dimensionality decreased across pretraining, requiring greater representational reorganisation for rule-shift adaptation, leading to impaired flexibility. In contrast, MLPs first trained to reconstruct the inputs (recMLPs) maintained higher dimensional representations, requiring a smaller degree of representational reorganisation after the rule shift, reducing the flexibility cost. These networks displayed only modest flexibility differences based on pretraining length. Lazy networks served as a control; their representations remained uncompressed, and they showed no flexibility impairment. Together, these results suggest that high dimensional representations promote flexibility, supporting the idea that compression drives the flexibility impairment associated with pretraining length.

Both rich and lazy networks overfit to the initial task, yet rich networks adapted more slowly to the rule shift, suggesting that compressed representations were more rigid. Higher dimensionality may enhance flexibility by allowing for a greater diversity of downstream readouts (Badre et al., 2021; Kaufman et al., 2022) and making the representational space more accessible by providing more avenues for representational change. This is consistent with biological studies demonstrating that high dimensional representations support faster learning (Tang et al., 2019), efficient task switching (Ritz et al., 2024), and flexible action selection (Kikumoto et al., 2023).

Our findings indicate that representational compression incurs a cost on flexibility, as adaptation to an extra-dimensional shift requires the reconfiguration of representations to incorporate previously compressed features. According to this account, post-shift adaptation is primarily driven by changes in the representation embedding, while the downstream readout remains largely stable. This mechanism is consistent with work by Jahn et al. (2024), which showed that primate attentional template representations were incrementally updated as the target changed in a visual search task. They suggest that smooth changes in neural population activity may underlie the adaptation to new tasks. Moreover, Sadtler et al. (2014) demonstrated, using a brain-computer interface paradigm, that in-manifold perturbations are more easily learned than out-of-manifold perturbations in macaques. These findings highlight that existing neural representations constrain the space for reconfiguration, indicating that flexibility is influenced by the capacity for reorganisation within the manifold. Together, these studies lend support to the idea that representational compression impairs flexibility by limiting the space accessible for representational reorganisation.

We collected human behavioural data for the task and found that longer pretraining (M1) was associated with lower post-shift performance, mirroring the patterns observed in Rich MLPs and recMLPs, and supporting the idea that flexibility decreases across learning. Human behavioural signatures aligned most closely with the recMLPs, showing both positive transfer from M1 to M2 during pretraining as well as the subtle difference in switch cost with longer pretraining. This suggests that the structured priors acquired through reconstruction training enabled the networks to retain task-general information, thereby promoting both generalisation and flexibility. The parallels between human and MLP behaviour support the idea that prior knowledge plays a critical role in enabling flexible behaviour in biological systems (Behrens et al., 2018; Tenenbaum et al., 2011), and high-

light structured priors as a promising avenue for improving flexibility in artificial systems (Lake et al., 2016).

While we could not directly examine representational geometry in humans from the behavioural data, we leveraged reaction times to infer patterns of selective attention to the stimulus features. We reasoned that increasing selective attention to the relevant feature over time would correspond to representational compression. Consistent with this prediction, our analyses suggest that participants showed increasing selective attention to the relevant feature across each phase of the task. Critically, we found that participants' selectivity for feature A was greater for M1 than M2 prior to the shift-point, suggesting that M1 representations were more compressed. These findings provide indirect evidence that learning induces representational compression, with longer training resulting in attentional narrowing and the impairment of flexibility.

As we did not directly measure neural activity in this study, future research is needed to determine whether the representational reorganisation cost observed here aligns with biological neural dynamics. While our study focused on a simple, extra-dimensional shifting task, further work could explore the trade-off between compression and flexibility in more complex, noisy, and naturalistic settings where the separation of dimensions is less clear. Furthermore, here we used simple, linear, feedforward networks to maximise interpretability, however, it remains to be established whether the relationship between compression and flexibility extends to more complex architectures.

In summary, this experiment demonstrated that pretraining length impaired flexibility in an extra-dimensional shift task. MLP simulations indicated that the compression of representations across learning hindered adaptation to the rule shift, as it increased the degree of representational reorganisation required. However, networks with prior knowledge of the inputs retained higher dimensional representations, which mitigated the impact of pretraining on flexibility. This suggests that maintaining representational richness enhances flexibility. These findings contribute to the development of a representational account of cognitive flexibility across biological and artificial systems.

5 METHODS

5.1 HUMAN TASK

We recruited participants using the online platform Prolific, restricting selection to adults fluent in English, aged between 18-40 years, and without colour-blindness. Participants completed a pre-screening 2-back test and were invited to take part in the experiment if they achieved at least 80% true positive accuracy and 80 % overall accuracy. 80 adults (32 female and 48 male) aged 30.9 $\pm$ 10 years completed the task. All participants provided informed consent and were paid £9/hour.

On each trial, a symbol was presented on screen for a maximum duration of 1.5 s. On train trials, 0.6 s elapsed between prediction and feedback. The feedback ("Correct" or "Incorrect" alongside the true location) was presented for 1s, with 0.1 s interval between trials. Each block consisted of 32 trials (3 repetitions of the 8 train symbols and 1 repetition of the 8 test symbols, with the order randomised). The pretraining phase consisted of 5 M1 blocks and 1 M2 block. The post-shift phase consisted of alternating M1 and M2 blocks (4 each). Responses were considered correct if the participant clicked within the correct half of the map (North/South or East/West - randomised per participant).

5.2 MLP SETUP

The MLP architecture consisted of an input layer with 16 units, a hidden layer with 30 units, and a 17 unit output layer (1 classification unit and 16 reconstruction inputs). A classification and reconstruction loss was applied to the corresponding units, and were each scaled a scaling factor. The networks were fully linear, with no activation functions or biases applied. All MLPs were trained using the Adam optimizer and the binary cross entropy loss function. The learning rate was set to 0.004. MLPs were initialised with weights drawn from a Xavier normal distribution with scale factor of 0.1 for rich MLPs and 50 for lazy MLPs (for the embedding weights only).

The human task was adapted into a binary classification task for the MLPs, with the input data was represented as two-hot vectors. The MLPs were trained on a curriculum similar to that used in the

human task, where each block consisting of 3 epochs of the training trials, with a batch size of 1. For recMLPs, the classification loss scaling factor was set to 0 for the 20 blocks of reconstruction training, then set to 1, and reconstruction scaling factor set to 1 throughout. For the rich and lazy MLPs, the reconstruction scaling factor was set to 0 throughout the task.

5.3 Multidimensional Scaling

We used the MDS function from the scikit-learn package to visualise the MLP representations. We extracted the input representations at different stages of the task, computed the pairwise Euclidean distances between all representation vectors across all time points, and applied MDS to the resulting distance matrix.

5.4 Participation ratio

The effective dimension of a set of inputs was estimated by computing the participation ratio:

$$\text{PR} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2} \tag{1}$$

Where λ_i correspond to the eigenvalues of the covariance matrix. We used the scikit-dimension python module to calculate PR.

References

Fabio Anselmi, Lorenzo Rosasco, and Tomaso Poggio. On Invariance and Selectivity in Representation Learning, 2015.

David Badre, Andrew S. Kayser, and Mark D'Esposito. Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron*, 66(2):315–326, April 2010. ISSN 08966273. doi: 10.1016/j.neuron.2010.03.025.

David Badre, Apoorva Bhandari, Haley Keglovits, and Atsushi Kikumoto. The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, 38:20–28, 2021.

Omri Barak, Mattia Rigotti, and Stefano Fusi. The Sparseness of Mixed Selectivity Neurons Controls the Generalization–Discrimination Trade-Off. *The Journal of Neuroscience*, 33(9):3844–3856, February 2013. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.2753-12.2013.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Timothy E.J. Behrens, Timothy H. Muller, James C.R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, 100(2):490–509, October 2018. ISSN 08966273. doi: 10.1016/j.neuron.2018.10.002.

Marcus K Benna and Stefano Fusi. Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19(12):1697–1706, December 2016. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.4401.

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. ISSN 00928674. doi: 10.1016/j.cell.2020.09.031.

Angus F. Chapman and Viola S. Störmer. Representational structures as a unifying framework for attention. *Trends in Cognitive Sciences*, 28(5):416–427, May 2024. ISSN 13646613. doi: 10.1016/j.tics.2024.01.002.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming, 2018.

Uri Cohen, SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1):746, February 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-14578-5.

Anne G. E. Collins and Michael J. Frank. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, 120(1):190–229, January 2013. ISSN 1939-1471, 0033-295X. doi: 10.1037/a0030852.

Pierre Enel, Emmanuel Procyk, René Quilodran, and Peter Ford Dominey. Reservoir Computing Properties of Neural Dynamics in Prefrontal Cortex. *PLOS Computational Biology*, 12(6): e1004967, June 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004967.

Timo Flesch, Keno Juechems, Tsvetomira Dumbalska, Andrew Saxe, and Christopher Summerfield. Rich and lazy learning of task representations in brains and neural networks, April 2021.

Timo Flesch, David G. Nagy, Andrew Saxe, and Christopher Summerfield. Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *PLOS Computational Biology*, 19(1):e1010808, January 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1010808.

R French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4): 128–135, April 1999. ISSN 13646613. doi: 10.1016/S1364-6613(99)01294-2.

Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74, April 2016. ISSN 09594388. doi: 10.1016/j.conb.2016.01.010.

Caroline I. Jahn, Nikola T. Markov, Britney Morea, Nathaniel D. Daw, R. Becket Ebitz, and Timothy J. Buschman. Learning attentional templates for value-based decision-making. *Cell*, 187(6): 1476–1489.e21, March 2024. ISSN 00928674. doi: 10.1016/j.cell.2024.01.041.

Matthew T. Kaufman, Marcus K. Benna, Mattia Rigotti, Fabio Stefanini, Stefano Fusi, and Anne K. Churchland. The implications of categorical and category-free mixed selectivity on representational geometries. *Current Opinion in Neurobiology*, 77:102644, December 2022. ISSN 09594388. doi: 10.1016/j.conb.2022.102644.

Andrea Kiesel, Marco Steinhauser, Mike Wendt, Michael Falkenstein, Kerstin Jost, Andrea M. Philipp, and Iring Koch. Control and interference in task switching—A review. *Psychological Bulletin*, 136(5):849–874, 2010.

Atsushi Kikumoto and Ulrich Mayr. Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences*, 117 (19):10603–10608, May 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1922166117.

Atsushi Kikumoto, Apoorva Bhandari, Kazuhisa Shibata, and David Badre. A Transient High-dimensional Geometry Affords Stable Conjunctive Subspaces for Efficient Action Selection. Preprint, Neuroscience, June 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1611835114.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building Machines That Learn and Think Like People, 2016.

Timothy P. Lillicrap and Konrad P. Kording. What does it mean to understand a neural network?, 2019.

Michael L. Mack, Alison R. Preston, and Bradley C. Love. Ventromedial prefrontal cortex compression during concept learning. *Nature Communications*, 11(1):46, January 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-13930-8.

Ravi D. Mill and Michael W. Cole. Neural representation dynamics reveal computational principles of cognitive task learning. Preprint, Neuroscience, June 2023.

Earl K. Miller and Jonathan D. Cohen. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, 24(1):167–202, March 2001. ISSN 0147-006X, 1545-4126. doi: 10.1146/annurev.neuro.24.1.167.

Stephen Monsell. Task switching. *Trends in Cognitive Sciences*, 7(3):134–140, March 2003. ISSN 13646613. doi: 10.1016/S1364-6613(03)00028-7.

Sebastian Musslick and Jonathan D. Cohen. Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, 25(9):757–775, September 2021. ISSN 13646613. doi: 10.1016/j.tics.2021.06.001.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, May 2019. ISSN 08936080. doi: 10.1016/j.neunet.2019.01.012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497 (7451):585–590, May 2013. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12160.

Harrison Ritz, Aditi Jha, Jonathan Pillow, Nathaniel D. Daw, and Jonathan D. Cohen. Humans actively reconfigure neural task states, September 2024.

Patrick T. Sadtler, Kristin M. Quick, Matthew D. Golub, Steven M. Chase, Stephen I. Ryu, Elizabeth C. Tyler-Kabara, Byron M. Yu, and Aaron P. Batista. Neural constraints on learning. *Nature*, 512(7515):423–426, August 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13665.

Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67, January 2021. ISSN 1471-003X, 1471-0048. doi: 10.1038/s41583-020-00395-8.

Yuelin Shi, Dasheng Bi, Janis K. Hesse, Frank F. Lanfranchi, Shi Chen, and Doris Y. Tsao. Rapid, concerted switching of the neural code in inferotemporal cortex. Preprint, Neuroscience, December 2023.

Evelyn Tang, Marcelo G. Mattar, Chad Giusti, David M. Lydon-Staley, Sharon L. Thompson-Schill, and Danielle S. Bassett. Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nature Neuroscience*, 22(6):1000–1009, June 2019. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-019-0400-9.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285, March 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1192788.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, December 2016. ISSN 2196-1115. doi: 10.1186/s40537-016-0043-6.

Michał J. Wójcik, Jake P. Stroud, Dante Wasmuht, Makoto Kusunoki, Mikiko Kadohisa, Nicholas E. Myers, Laurence T. Hunt, John Duncan, and Mark G. Stokes. Learning shapes neural geometry in the prefrontal cortex. Preprint, Neuroscience, April 2023.

A Appendix

A.1 Generalised Mixed-Effects Modelling

Generalised mixed-effects modelling (GLMM) analyses were conducted in R (Version 2024.04.2+764) (R Core Team, 2023) using the lme4 package (Bates et al., 2015). For each analysis,

we selected the data over the task period of interest and normalised the trial indices to a range of 0 to 1 for each map. We then fit the GLMM to the accuracy data, where accuracy was a binary outcome indicating whether a participant's response was correct. Models were specified with a binomial family and the optimiser was set to "bobyqa".

A.1.1 INITIAL PRETRAINING BLOCK GLMM

We fit the following GLMM to the trial-by-trial accuracy data:

$$\text{Accuracy} \sim \text{Trial} * \text{Map} + (1 + \text{Trial} + \text{Map} \mid \text{Participant}) \tag{2}$$

$Accuracy$ is a binary outcome (correct/incorrect), t represents trial number and Map indicates M1 or M2. The final term accounts for random effects over participants.

Table 1: GLMM results for human initial pretraining block accuracy data.

Fixed effect	Estimate	Std. err	z value	Pr($>\|z\|$)
β_0	0.131	0.124	1.052	0.293
t	0.619	0.205	3.027	0.002 *
Map_{M2}	-0.015	0.181	-0.083	0.933
$t * Map_{M2}$	0.988	0.259	3.815	< 0.001 ***

Table 1 shows that trial number (t) was a positive predictor of accuracy, indicating that accuracy increased across trials. Additionally, there was a positive, significant interaction effect between M2 and trial number ($t*Map_{M2}$), indicating faster learning for M2 than M1 during the initial pretraining block. This suggests that participants generalised the task structure from M1 to M2.

A.1.2 POST-SHIFT GLMM

We fit the following GLMM to the post-shift accuracy data:

$$\begin{aligned}\text{Accuracy} \sim \text{Trial} * \text{Map} + \text{Type} \\ + (1 + \text{Trial} * \text{Map} + \text{Type} \mid \text{Participant})\end{aligned} \tag{3}$$

$Accuracy$ is a binary outcome (correct/incorrect), t represents trial number, Map indicates M1 or M2, and $Type$ denotes flipped or unflipped trials after the rule shift. The final term accounts for random effects over participants.

Table 2: GLMM results for human post-shift accuracy data.

Fixed effect	Estimate	Std. err	z value	Pr($>\|z\|$)
β_0	-0.210	0.078	-2.707	0.008
t	2.314	0.263	8.808	< 0.000
Map_{M2}	-0.002	0.106	-0.015	0.988
$Type_{unflipped}$	0.587	0.062	9.532	< 0.000
$t * Map_{M2}$	0.718	0.266	2.700	0.007

The GLMM results (Table 2) show that trial number (t) and trial type ($Type_{unflipped}$) were significant positive predictors of accuracy, indicating that accuracy increased over trials and was higher for unflipped trials. A significant, positive interaction effect between trial number (t) and M2 (Map_{M2}) suggests faster learning of M2 than M1, supporting our hypothesis that longer pretraining impairs flexibility.