The Hebrew University of Jerusalem
Edmond and Lily Safra Center for Brain Sciences

Modeling Fluid Intelligence via Real-Time Adaptation

By Tomer Barak

Thesis for the degree of “Doctor of Philosophy”

Submitted to the Senate of the Hebrew University of Jerusalem
September 2025

This work was carried out under the supervision of
Professor Yonatan Loewenstein

Acknowledgements

I am deeply grateful to my supervisor, Professor Yonatan Loewenstein, for guiding me with focus and clarity. While my natural tendency was to branch out in many directions, Yonatan consistently brought me back to the heart of the story, teaching me how to build arguments that are both compelling and logically rigorous. One of his key insights—reducing our model's latent space to a single neuron—not only simplified the mathematical analysis but sharpened and focused the entire framework. I would also like to thank my accompanying committee, Omri Barak and Amir Globerson, for their constructive feedback throughout this research.

To my parents, who shaped me in complementary ways: my mother by nurturing curiosity and encouraging freedom, and my father by showing what it means to act with integrity and do things right. To my brother, sister, and extended family, thank you for being the supportive backdrop that made this possible.

Daniel and Yonatan, you may be too young to understand what a PhD is, but you gave this project meaning. Your laughter and presence reminded me, even in the hardest stretches, that my efforts were for something larger than myself and for the world you will inherit.

And at the very center of it all stands Moran, my wife. You filled my days with small delights, sending me to the office with your home-roasted oats, and you showed me what true work ethic looks like — giving me a model for how to treat my PhD. Beyond that, you made life joyful with our projects at home, from experimenting with coffee brewing to baking Neapolitan pizza. These shared joys made the long road lighter. More than anything, you turned this academic journey into a life we built together, and for that I am endlessly grateful.

Abstract

Fluid intelligence—the capacity to solve novel problems without relying on prior knowledge—is a hallmark of human cognition, yet its mechanisms remain elusive. Consider modern artificial intelligence (AI) systems as models of intelligence. While these systems exhibit remarkable performance across a wide range of tasks, it is still unclear whether they possess genuine fluid intelligence.

Skepticism arises from two main limitations. First, these systems rely on massive training datasets: state-of-the-art models require vastly more examples than human children need to acquire comparable skills. Second, their brittleness: they often fail when faced with problems that differ even slightly from the format of their training data. Taken together, these shortcomings suggest that current AI models function primarily as sophisticated pattern matchers, akin to humans' crystallized intelligence, rather than demonstrating the adaptive flexibility characteristic of human fluid intelligence.

This thesis explores the idea that fluid intelligence may arise from a system's capacity to adapt its internal structure while reasoning about novel problems. I formalize this idea through a computational framework in which an artificial neural network's parameters are optimized in real time during problem-solving. I consider both the extreme case, where adaptation is confined to a single problem instance (inference-time adaptation), and its extension across streams of inputs, which naturally situates the model in the online learning paradigm.

This thesis unfolds across three studies. The first demonstrates that a neural network, initialized with random parameters and therefore lacking any prior training, can solve abstract reasoning tasks analogous to human in-

telligence tests. In this extreme case, the network adapts its parameters using only the information within a single problem instance. This shows that abstract reasoning can, in some cases, succeed without memorization, challenging the prevailing view that such capabilities require vast stored knowledge and situating inference-time adaptation as a candidate mechanism for fluid intelligence.

The second study applies this framework to a stream of inputs—a reversal learning task—to test its explanatory power against a paradoxical finding in human learning: extreme violations of expectations can inhibit, rather than promote, belief updating. I attribute this phenomenon to a competition inherent in the model's architecture, which structurally decouples the parameters encoding sensory inputs from those encoding expectations. When an observation contradicts expectation, the architecture presents a choice: adapt the expectation or adapt the input representation. The results show that the model's real-time optimization dynamics naturally arbitrate this choice: moderate violations drive updates to relational expectations, whereas extreme violations favor adapting the input representation, thereby preserving the original expectation.

The final study grounds my computational framework by empirically testing its central prediction in humans. Specifically, I studied how people adapt to relational reversal. Based on my model, I hypothesized that the magnitude of the violation would shape the adaptive response, biasing participants toward one of two strategies: updating the relational expectation or reinterpreting the input. To test this, I designed a novel psychophysical experiment. Human behavior was qualitatively consistent with the model's dynamics: The effect was modest, yet larger violations significantly increased the likelihood that participants would reinterpret the input, as predicted. These findings provide important, albeit modest, empirical support for the model's core prediction, while underscoring the challenges of directly mapping computational mechanisms onto human cognition.

In summary, this thesis proposes real-time adaptation as a key principle for abstract reasoning and belief dynamics. This perspective has dual implica-

tions: for cognitive science, it offers a mechanistic model of fluid intelligence, and for artificial intelligence, it supports approaches that overcome brittleness by augmenting pre-trained models with adaptive capabilities at inference time. Together, the results suggest that real-time adaptation is a core aspect of intelligence, both natural and artificial.

A Letter of Contribution

This letter details the contributions of the student, Tomer Barak, and other collaborators to the research presented in this thesis.

- **Paper 1 (Published)** Tomer Barak conducted the simulations. Tomer Barak and Yonatan Loewenstein were responsible for the analysis of the results, and writing the manuscript.
- **Paper 2 (Published)** Tomer Barak conducted the simulations. Tomer Barak and Yonatan Loewenstein collaboratively designed the study, analyzed the results, and wrote the manuscript.
- **Paper 3 (Unpublished)** The first experiment was conducted by Tomer Barak and Ron Hafzadi for his final B.sc. project. The rest of the experiments were conducted by Tomer Barak. Tomer Barak led the study design and data analysis in collaboration with Ron Hafzadi and Yonatan Loewenstein. The manuscript was written by Tomer Barak and Yonatan Loewenstein.

I, Tomer Barak, confirm that the above accurately reflects my contribution to the research presented in this thesis.

Student: 

Tomer Barak

Supervisor: 

Prof. Yonatan Loewenstein

Contents

Introduction

When encountering an unfamiliar puzzle, navigating a new city, or reasoning through an unexpected social situation, we cannot simply retrieve prestored answers. Instead, our mind appears to actively *adapt* its approach, drawing connections, testing hypotheses, and refining its understanding as the problem unfolds. This ability to solve genuinely novel problems in real-time without relying solely on previously memorized solutions is termed **fluid intelligence** [1–3]. Despite decades of research in cognitive science, we lack a clear computational account of how this real-time adaptation works. What mechanisms allow the mind to flexibly reconfigure its processing in response to novel challenges?

Modern artificial intelligence (AI) systems, particularly large language models, can appear to demonstrate similar flexibility, processing novel prompts and generating sophisticated responses in real-time [4]. However, this apparent similarity masks a fundamental difference in underlying mechanism. During inference, these systems operate with completely fixed parameters, applying a vast but static body of pre-learned knowledge to new inputs. The network itself does not learn, update, or adapt based on the unique structure of the problem it is actively solving.

This static architecture becomes apparent in the phenomenon of **brittleness**: AI systems that perform well on training tasks often fail catastrophically on slight variations of the same problems [5–7]. Crucially, brittleness reveals something deeper than mere performance limitations – it suggests that these systems are engaging in sophisticated pattern matching rather than the kind of flexible, adaptive reasoning that characterizes human fluid intelligence.

One approach to overcome brittleness is to simply scale up the training

data to cover all possible variations a model might encounter [8]. However, this ”brute force” approach faces fundamental limitations. Research into the scaling laws of large language models (LLMs) suggests this path is intractable, arguing that the performance gains from increasing data or model size are so marginal that achieving the reliability needed for scientific inquiry is practically impossible. This problem is compounded by the fact that as datasets grow, they become overwhelmingly dominated by spurious correlations, making it even harder for models to learn true underlying principles [9].

Another approach to mitigate brittleness involves prompting models to generate a ”Chain-of-Thought” (CoT), a series of intermediate steps performed in real-time that mimic a reasoning process before providing a final answer. This technique is not merely a superficial trick; theoretical work has shown that it can fundamentally enhance the computational power of these models. For example, it has been demonstrated that by generating intermediate steps, autoregressive models can solve complex mathematical and logical problems that are impossible for them to solve directly [10]. Moreover, auto-regressive next-token prediction mechanism, when combined with CoT-style data, was shown to be powerful enough to approximate any function computable by a Turing machine, making these models universal learners [11]. However, this apparent universality may be illusory. A competing line of inquiry suggests that CoT reasoning is itself a brittle form of pattern matching, rather than genuine, flexible inference, framing CoT as a ”mirage” that reflects an inductive bias learned from the training data’s distribution. Consequently, its effectiveness is fundamentally bounded by the similarity between a test query and the data the model was trained on, and it fails when pushed beyond these distributional boundaries [12].

If static-parameter inference cannot capture human-like fluid intelligence, then what computational architecture could? This thesis argues that the answer lies in abandoning the rigid separation between learning and inference that characterizes current AI. Instead, I propose that fluid intelligence emerges from processes where inference and learning occur simultaneously

– where the very act of confronting a novel problem drives real-time adaptation of the cognitive system itself.

My framework draws inspiration from two machine learning paradigms. *Test-time adaptation* (TTA) has emerged as a technique to combat distribution shifts by temporarily optimizing a pre-trained model's parameters during inference [13–15]. *Online learning* updates model parameters incrementally after processing each data point, making it conceptually powerful for modeling biological knowledge acquisition [16]. My framework synthesizes these concepts: from TTA, I adopt the principle of optimizing parameters during problem-solving to handle novel inputs; from online learning, I incorporate the incremental accumulation of these adaptations, modeling how problem-solving leads to increasing knowledge. This hybrid perspective enables me to model both fluid intelligence (momentary adaptation) and its gradual conversion into crystallized intelligence (knowledge accumulation).

The research presented in this thesis unfolds across three studies, each addressing a fundamental question about modeling intelligence through real-time adaptation.

Study 1: Establishing Computational Sufficiency

Can inference-time adaptation alone generate abstract reasoning capabilities? My first paper [17] demonstrates that completely naive networks with randomly initialized weights can solve intelligence test-like problems by optimizing their parameters using only information within each specific problem. This challenges the prevailing view that such capabilities require memorization of vast datasets, showing instead that abstract reasoning can emerge from adaptive problem-specific learning. I then investigate how momentary adaptations accumulate into lasting knowledge. Specifically, I replicated the interleaving advantage from cognitive science: the finding that alternating between tasks during training improves learning compared to blocked practice [18, 19].

Study 2: Demonstrating Explanatory Power

Having established the model's computational viability, I extended it to provide a mechanistic explanation for a surprising cognitive puzzle: why extreme violations of an expectation can paradoxically lead to a *reduction* in belief updating, contrary to the predictions of standard learning theories [20–22]. My second paper [23] demonstrates that this phenomenon emerges naturally from the adaptation dynamics of my network. I examined the network in a relational reversal task and showed that the system's adaptation to violations is determined by a "race" between two competing pathways: updating relational expectations or re-interpreting the representation of the input stimuli. The model demonstrates that the magnitude of the violation determines the outcome of this race, providing a mechanistic explanation for resistance to belief change without needing to posit separate belief-protecting mechanisms.

Study 3: Testing Empirical Plausibility

Having established my model's computational power (Paper 1) and its ability to explain a known cognitive puzzle (Paper 2), the final step was to test its empirical plausibility. The dual-pathway model developed in my second paper predicts that the magnitude of an expectation violation systematically modulates which adaptive pathway an individual chooses in relational tasks. In my third paper (unpublished), I designed and conducted a behavioral experiment involving a relational task to test this directly. The results provide modest but statistically significant support for the model's core hypothesis, showing that a larger violation magnitude increases the likelihood that participants will adapt their representation of a stimulus rather than their relational expectation. This result provides preliminary empirical support for my theory and completes the progression from a novel computational idea to a behaviorally testable model of cognition.

Untrained neural networks can demonstrate memorization-independent abstract reasoning

Tomer Barak, Yonatan Loewenstein

Status: Published

Scientific Reports (2024).
https://doi.org/10.1038/s41598-024-78530-z

scientific reports

Check for updates

OPEN

Untrained neural networks can demonstrate memorization-independent abstract reasoning

Tomer Barak[1]✉ & Yonatan Loewenstein[1,2]

The nature of abstract reasoning is a matter of debate. Modern artificial neural network (ANN) models, like large language models, demonstrate impressive success when tested on abstract reasoning problems. However, it has been argued that their success reflects some form of memorization of similar problems (data contamination) rather than a general-purpose abstract reasoning capability. This concern is supported by evidence of brittleness, and the requirement of extensive training. In our study, we explored whether abstract reasoning can be achieved using the toolbox of ANNs, without prior training. Specifically, we studied an ANN model in which the weights of a naive network are optimized during the solution of the problem, using the problem data itself, rather than any prior knowledge. We tested this modeling approach on visual reasoning problems and found that it performs relatively well. Crucially, this success does not rely on memorization of similar problems. We further suggest an explanation of how it works. Finally, as problem solving is performed by changing the ANN weights, we explored the connection between problem solving and the accumulation of knowledge in the ANNs.

The topic of this paper is abstract reasoning, sometimes referred to as "fluid intelligence"[1]. Abstract reasoning is, broadly speaking, the ability to solve complex problems by identifying regularities and relations in the problem being solved and utilizing them for deducing the solution[2,3]. It is often studied using intelligence tests that comprise word analogy tests (e.g., infer that the relationship between "cow" and "milk" is the same as between "chicken" and "egg") and visual reasoning tests (e.g., Raven Progression Matrices)[4,5]. As artificial intelligence continues to advance, understanding the nature of abstract reasoning in both humans and machines is becoming a central question in cognitive science and AI research[6].

Abstract reasoning in artificial neural networks (ANNs) appears to be closely tied to training. While deep ANNs have shown impressive performance on various intelligence tests[7–12], their success relied heavily on extensive prior training. Additionally, questions have been raised about the nature of this performance. There are indications that ANNs' success may stem more from "contamination" – exposure to similar questions in their training data – rather than from genuine abstract reasoning[13,14]. This dependency on specific training data is further emphasized by findings that minor changes in problem phrasing, which do not affect human performance, can render problems unsolvable for ANNs[6,15]. Thus, while ANNs may exhibit some analogical reasoning capabilities, it is disputed that these are based on pattern matching or memorization rather than on general intelligence comparable to that of humans.

In this work, we investigated whether ANN tools commonly used in machine learning are capable of demonstrating general abstract reasoning. Specifically, we asked if these networks could solve intelligence test problems with novel inputs, relying only on the information provided by the specific problem at hand, without drawing on prior memorization.

Certain intelligence tests, by their nature, require some level of prior knowledge. For instance, a human unfamiliar with English or the relationship between "cow" and "milk" would struggle to relate "chicken" to "egg" in an analogy test. Consequently, general intelligence in humans is often assessed using *visual* reasoning tests, utilizing abstract shapes like squares and triangles to minimize the influence of language or cultural knowledge. Thus, to evaluate the abstract reasoning of ANN models, we employed visual abstract reasoning tests. These visual reasoning tests require identifying relations in a sequence of stimuli, a skill common to many intelligence tests[1–3,16].

For our network models, we used Relation Networks (RNs)[17], as members of this class of models were shown to be capable of identifying abstract relations and solving intelligence tests after extensive training[9,18]. Notably,

[1]The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Israel. [2]Department of Cognitive Sciences, The Federmann Center for the Study of Rationality, The Alexander Silberman Institute of Life Sciences, The Hebrew University, Jerusalem, Israel. ✉email: tomer.barak@mail.huji.ac.il

RNs were shown to successfully solve word-analogy problems without specific training on those problems, but with specific training on relevant relationships[19,20]. In contrast to these previous studies, our focus was on the ability of "naive" RNs, who were not exposed to *any* pre-training, to identify relations in visual reasoning tests and use them for solving the tests.

The structure of this paper is as follows: We begin by introducing the visual reasoning tests and the network models employed in our study. Next, we present the model's performance, showcasing its ability to solve non-trivial problems. We then analyze the mechanisms underlying this performance. Finally, as the model's problem-solving involves an optimization process that modifies network parameters in a manner similar to learning, we examine the relationship between the model's problem-solving capabilities and the networks' accumulation of knowledge.

Results

Sequential visual reasoning tests

We constructed a set of artificial problems in which the task is to evaluate the consistency of an image with a sequence of its preceding images (Fig. 1). Each problem comprises 5 gray-scale images and 4 optional-choice images. The images, 224×224 pixels each, are composed of identical abstract objects and differ along several dimensions: the shape of the objects, their size, their color, their number, and their arrangement. By construction, one of these features changes predictably over the 5 images. Formally, an image is characterized by a low-dimensional vector of features, $\mathbf{f}_j$, where f_j^i denotes the value of feature *i* in image *j*. An image in pixel space, $\mathbf{x}_j$, is constructed according to its characterizing features by a generative function $\mathbf{x}_j = \mathbf{G}(\mathbf{f}_j)$. One of the features f^p changes predictably along the sequence according to a simple deterministic rule $f_{j+1}^p = U(f_j^p)$ while the other features are either constant over the images or change randomly (values are i.i.d). Considering the optional-choice images, the predictable rule is followed in only one of them, and the task is to select this image. The other features are either constant in all 9 images (5 of sequence and 4 of optional choices) or change randomly (see Methods). We refer to a problem's predictably changing feature as the problem's Predictive Feature (PF) and to the randomly changing features as distracting features or *distractors*. Intuitively, the number of distractors is a measure of a problem's difficulty.

The computational task

Each image is characterized by a small number of features, of which one changes predictably. The challenge is to simultaneously identify the features and the rule that relates the features of the different images. Relation Networks[17] do exactly that. Taking a set of stimuli, they learn two functions: an encoder function $Z_\phi(\mathbf{x})$ that extracts relevant feature(s) from the stimulus, that is, a low-dimensional representation of the stimuli $\mathbf{x}$ and a relation module $R_\theta(Z_\phi(\mathbf{x}_i), Z_\phi(\mathbf{x}_j))$ that characterizes the relationship between the features of pairs of stimuli $\mathbf{x_i}$ and $\mathbf{x_j}$. In practice, the encoder and the relation modules are functions (typically networks) whose parameters (ϕ and θ, respectively) are learned from examples. It should be noted that, to some extent, the complexities of the encoder and the relation module are interchangeable. The reason is that a sufficiently-complex relation module

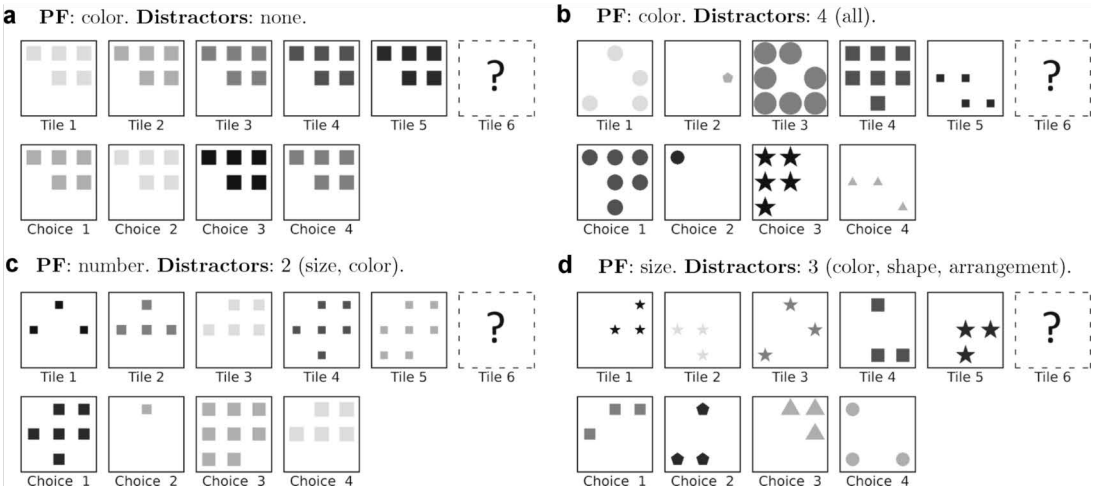

Fig. 1. Visual reasoning problems. The problems are characterized by the *Predictive Features (PF)* that can be the color (**a, b**), number (**c**), or size (**d**) of the abstract shapes. The values of the predictive features linearly increase along the sequence. The rest of the features (non-predictive) are either constant or random. We refer to the random features as *Distractors*, and their number determines the problem difficulty. *Note*: the shapes' type and arrangement are always non-predictive, and can either be constant or distracting. The correct choices in this figure are all 3.

can incorporate the feature extraction. Similarly, a sufficiently complex encoder can operate on the extracted features as to simplify the relation between them. For example, any monotonous relation between the features is also a linear relation between a (nonlinear) transformation of the features.

Previous studies have shown that with sufficient examples, relational networks can learn to extract the relevant features and their relations at a level sufficient for solving intelligence tests[9,18,20]. The challenge here is to perform a similar task without any pre-training. To do so, we defined the following loss function on a sequences of 5 images:

$$\mathscr{L}(\theta,\phi)=\frac{1}{4}\sum_{i=1}^{4}\left[R_\theta\left(Z_\phi\left(\mathbf{x}_i\right),Z_\phi\left(\mathbf{x}_{i+1}\right)\right)\right] \tag{1}$$

$\mathbf{x}_i$ is the i^{th} image in the sequence, the encoder $Z_\phi:\mathbb{R}^{224\times224}\to\mathbb{R}^n$ is a function that takes 224×224 pixel images to an n dimensional latent space and the relation module $R_\theta:\mathbb{R}^{2n}\to\mathbb{R}^+$ takes two consecutive latent variables, each of dimension n, and outputs a positive 1D relation score.

This loss is minimized for a relation function $R_\theta\left(Z_\phi\left(\mathbf{x}_i\right),Z_\phi\left(\mathbf{x}_j\right)\right)$ that outputs a minimal relation score for consecutive sequence images ($j=i+1$), requiring the identification of the regularity that characterizes these consecutive images. We updated the networks' weights θ and ϕ with 10 optimization steps over the loss $\mathscr{L}(\theta,\phi)$ using the RMSprop optimizer[21] (learning rate of 10^{-5}, the rest of the parameters are set to PyTorch[22] default). Eventually, after optimization, we evaluated the consistency of each choice image with the sequence based on their relation value R when they were placed as the sixth sequence image $R_\theta\left(Z_\phi\left(\mathbf{x}_5\right),Z_\phi\left(\,\cdot\,\right)\right)$ and selected the choice image with the lowest relation value as the answer.

To clarify, in these settings, the model does not need to learn the features and their relations in the generative sense to solve a test successfully. Instead, it is enough to find image representations and rules that are *sufficiently* correlated with a problem's predictive feature for selecting the most consistent image out of four options.

Vanilla model performance

The success of the model would depend on the specific choice of R and Z (their network structure), as they can be inductively biased towards certain types of features and rules. In our vanilla model, the encoder Z was a small CNN from input space to a 1D latent neuron, composed of 3 convolutional layers followed by 5 fully-connected (FC) layers with a single output neuron (see Methods and Supplementary Information Fig. S1). For the relation module, we used a simple function that asserts a linearly changing relation between the latent variables,

$$R_\theta\left(Z_\phi\left(\mathbf{x}_i\right),Z_\phi\left(\mathbf{x}_j\right)\right)=\left(Z_\phi\left(\mathbf{x}_i\right)-Z_\phi\left(\mathbf{x}_j\right)+\theta\right)^2 \tag{2}$$

where θ is a trainable constant that does not depend on Z.

We evaluated the performance of the vanilla model on the different tests, in which the predictive feature's values increased linearly, and found that it performed substantially better than chance (0.25) in almost all tasks and all levels of difficulty (Fig. 2). Without distractors, its performance on some tasks was close to perfect. We also found that performance decreased with the number of distractors, verifying that the number of distractors is a good measure of the task's difficulty. All these results were obtained using networks that were randomly initialized before each problem, thus demonstrating that ANNs can perform abstract reasoning that does not depend on memorization. Averaged over all conditions, the model's performance was $\mathbf{0.58\pm0.01}$. From this point in the paper, we use this global performance measure for comparisons (see Methods; complete performance results are in the Supplementary Information).

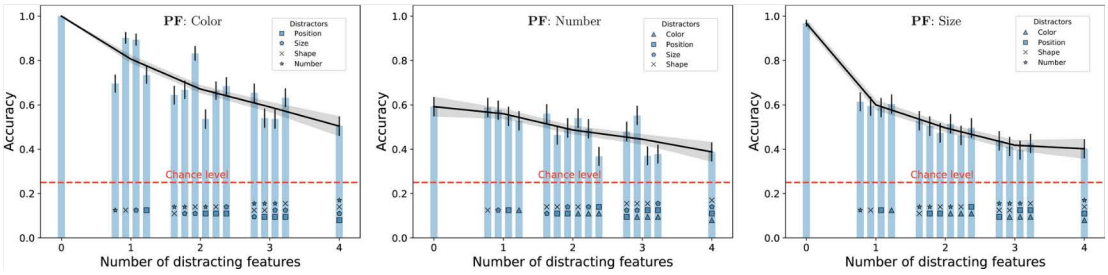

Fig. 2. Vanilla model performance. The performance of naive ANNs on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each predictive feature, we tested the networks over 16 test conditions where the predictive feature was linearly changing along the sequence, and the non-predictive features were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% Confidence Intervals (CI). The black line and its shade are the average accuracy per difficulty and the corresponding 95% CI. The dashed line denotes the chance level of problems with four choice images (0.25).

Determinants for success

Our model consists of two main components: the encoder Z and the relation module R. The parameters of both were changed in the direction of minimizing the loss function on the images of each problem, a process that we will refer to as optimization. To study the relative contribution of these components to problem-solving, we studied the model's performance when the parameters of only one of these components, either Z or R, were optimized. We found that optimizing the encoder was essential: when the parameters of the encoder Z remained unchanged, the model's performance, averaged over all conditions, was close to the chance level, 0.30 ± 0.01 (Fig. S2). By contrast, using random parameters for the relation module R and not optimizing it had no significant effect on performance, resulting in an average performance of 0.58 ± 0.01 (Fig. S3), which is not significantly different from that of the vanilla model. These results motivated us further to study the role of the encoder in the task.

The encoder

The encoder is an 8-layer network with 3 convolutional layers followed by 5 Fully-Connected (FC) layers. Removing the convolutional layers and connecting the FC layers directly to the inputs impaired the average performance of the model, reducing its performance to 0.48 ± 0.01 (Fig. S4), indicating that the convolutional layers are important for performance. In the vanilla model, the parameters of both the convolutional layers and the FC layers are optimized in the direction of minimizing the loss function. However, it turns out that the optimization of the parameters of the convolutional layers does not contribute to the performance. The average performance when the weights of the convolutional layers remained random, 0.57 ± 0.01, was not significantly different than that of the vanilla model (Fig. S5). By contrast, keeping the FC network weights fixed at their randomly-initialized values during problem-solving was detrimental to the performance (0.34 ± 0.01, Fig. S6).

So far, we saw that freezing either the weights of the convolutional layers or the relation module at their initial random values does not impair performance. This insensitivity does not change when *both* are frozen (0.58 ± 0.01, Fig. S7).

We conclude that the convolutional layers effectively operate as frozen feature extractors (features in the more general sense – not necessarily the features used for constructing the images) while the parameters of the FC layers are optimized to solve the task.

To test how the FC layers contribute to this task, we note that the task could be perfectly solved if the encoder could learn to identify the inverse generative function of the problem images $G^{-1}(\mathbf{x})$ and use it to extract the underlying predictive feature f^p and its rule $U(f^p)$. If this is done, we expect the optimization steps to increase the correlation of the encoder's output neuron with the predictive feature (but not with the distracting features). We tested this hypothesis in the vanilla model for all the predictive features. Indeed, the absolute Pearson correlation of the output neuron with the predictive feature (see Methods) increases with optimization steps, as depicted in Fig. 3a (black).

To better understand how such correlations emerge in the FC network, we also computed the absolute Pearson correlations of these features with the activities of all other neurons in the FC network (see Methods. Comprehensive results in Supplementary Fig. S8). These correlations, averaged over all neurons in a layer, are depicted in Fig. 3a. We found that the correlations with the predictive feature increase with the layer depth.

The higher the correlation of the output neuron with the predictive feature, the easier it is for the relation module to identify the regularity in the sequence of images. Along the same lines, we also expected the optimization process to decrease the correlation of the encoder output neuron with the other irrelevant features. This, however, is not the case. Considering the same features in problems in which they are not predictive features (either constant or distracting), we found that the correlation of the output neuron with these features also increases on average in the optimization process, albeit to a lesser extent (Fig. 3b). Considering the correlations of these features with neurons in the hidden layers of the encoder, we found that the correlations with these irrelevant features also increased with the layer depth.

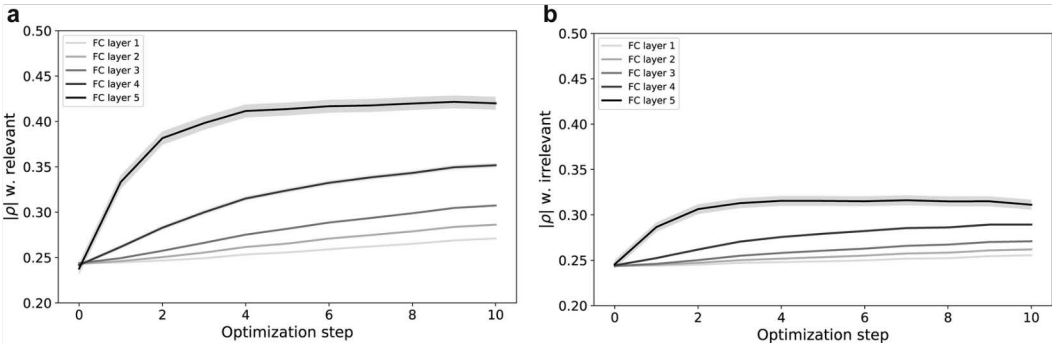

Fig. 3. The encoder's FC layers feature correlations. The average absolute correlations of encoders' FC layers with (**a**) the specific predictive feature of the problems they solved (either Color, Number, or Size), and (**b**) the other two non-predictive features (from either Color, Number, or Size). Error shades represent the 95% CI, based on the standard error of the means. The calculation of the correlations is detailed in the Methods section.

At the end of the optimization procedure, the output neuron of the encoder network is correlated with both the predictive feature and the irrelevant features (both distractors and constant). The stronger the correlation of the output neuron with the predictive feature, relative to its correlation with the distracting features, the better we expected the performance to be. To test this, we focused on the six problems in which the relevant feature was either color, number, or size, and there was one distracting feature, again: color, number, or size. For each of these problems, we computed the absolute Pearson correlations of the output neuron with the predictive and distracting features (taken from Fig. S8). We expected that performance in each of these problems would increase with the correlation with the relevant feature and decrease with the correlation with the irrelevant feature. Indeed, as depicted in Fig. 4, the logit of the performance ($\log \frac{p}{1-p}$ where the accuracies p are taken from Fig. 2) is correlated with the ratio of the absolute Pearson correlation of the output neuron with the predictive feature and the distracting feature (Wald t-test, p-value = 0.028).

Next, we studied how the correlation with the features increases during optimization. The loss function "seeks" *some* 1D predictable representation of the sequence of inputs. Considering the individual neurons at the output layer of the convolutional network part of the encoder, some co-vary with the sequence while others do not. Examples of two such neurons are depicted in Fig. 5a (left). As shown in Fig. 5a-b for a single problem, the optimization process makes larger changes to the synaptic weights from those neurons that co-vary strongly with the sequence order (e.g., blue in Fig. 5a-b) compared with low co-variance neurons (e.g., orange in Fig. 5a-b). This is the case across all problems (Fig. 5c). Consequently, the neurons in the encoder's FC layer become strongly correlated with the sequence order (Fig. 5d). As a result, the encoder amplifies the representation of those features that co-vary with the sequence order (independently of whether they are the predictive or irrelevant features).

Together, our results indicate that the ANN's ability to execute abstract reasoning without prior learning stems from two important properties: (1) The random convolutional layers extract features that are correlated with the relevant features. (2) The optimization process amplifies the response to those features that monotonically vary along the sequence.

The relation module

By construction, the vanilla model's relation module is simple, implicitly assuming that the features change linearly. Therefore, one may naively expect that identifying a non-linear change in the predictive feature will be more challenging. However, any monotonically changing rule can be mapped into a linearly changing rule with a sufficiently complex encoder. Therefore, we tested our vanilla model in problems in which the change in the feature was *non-linear* (Fig. 6). We found that when the relevant feature was size, the performance for an exponential increase or a square root increase of this feature was comparable to that of a linear increase (Linear: 0.53 ± 0.01; Exp: 0.54 ± 0.01; Sqrt: 0.52 ± 0.01. Fig. S9-10 right). Similarly, when the relevant feature was color, the model achieved comparable performance to the linear case, although with higher variability: performance was better for an exponential increase and worse for a square root increase (Linear: 0.70 ± 0.01 ; Exp: 0.75 ± 0.01; Sqrt: 0.64 ± 0.01. Fig. S9-10 left). These results suggest that a relation module that assumes linear relationships can capture general monotonic relationships, substantially downsizing the hypothesis space of possible relationships. It would, however, be more difficult for the model to deal with non-monotonic rules. Indeed, when tested in problems where the predictive feature alternated between two of its values, the vanilla model performance was at a chance level (0.24 ± 0.01. Fig. S11).

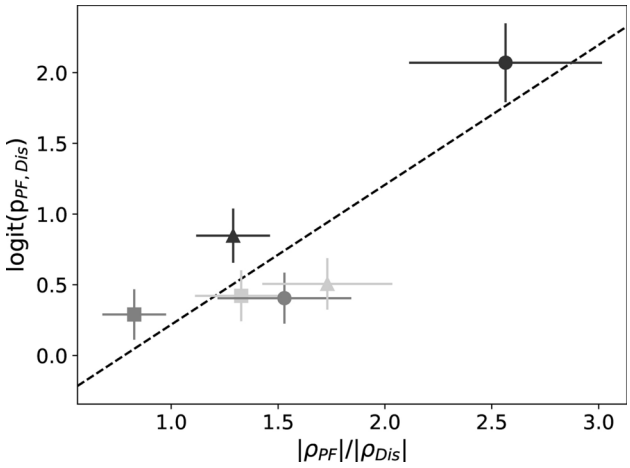

Fig. 4. The effect of distractors on accuracy. The figure depicts the relationship between the absolute correlation ratio with the relevant Predictive Feature ($|\rho_{PF}|$) and the Distracting feature $|\rho_{Dis}|$, and its consequential effect on networks' accuracy in problems of that predictive feature with the corresponding distracting feature ($p_{PF,Dis}$). The predictive features were either Color (Dark Gray), Number (Medium Gray), or Size (Light Gray). The distracting features were either Color (Square), Number (Triangle), or Size (Circle). Error bars represent the 95% CI. The black dashed line depicts a linear regression analysis.

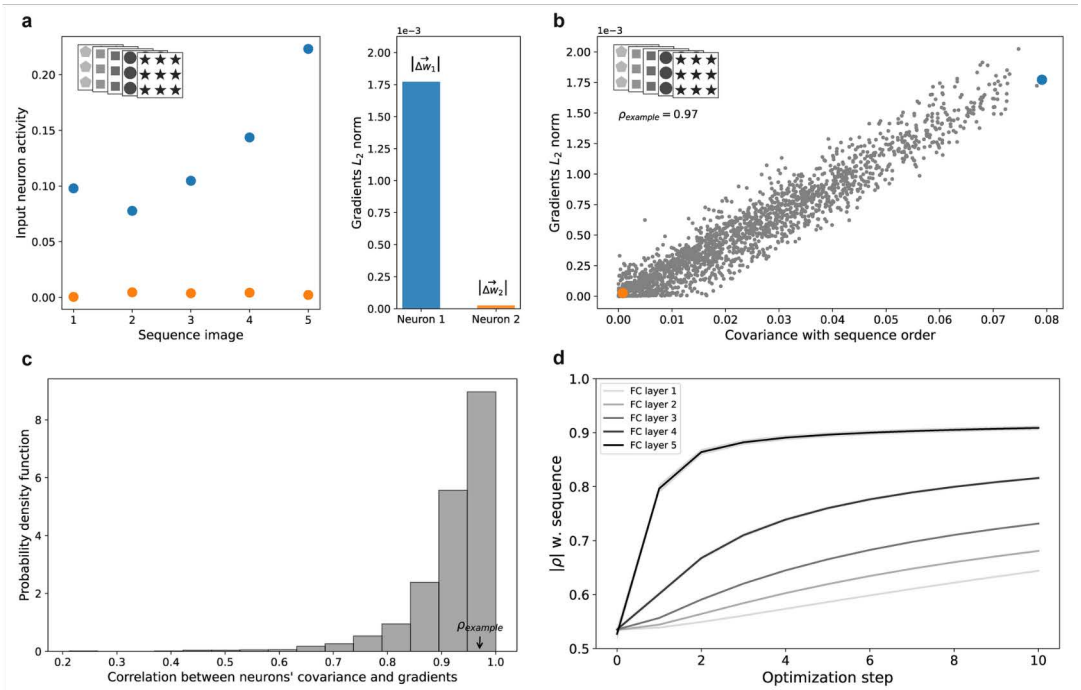

Fig. 5. Problem-solving mechanism. (**a**) Two example neurons' activity from the convolutional layers' output of a network (before optimization) when presented with the example problem of the inset. The blue neuron has a large covariance with the problem's image order, and the orange neuron has a small covariance with the order. The neurons' L2 gradient norms correlate with their respective image order covariances. **(b)** In this example network, the L2 gradient norms of the convolutional layers' output neurons are strongly correlated with their image-order covariances ($\rho_{\text{example}} = 0.97$). The two example neurons presented in (a) are highlighted. **(c)** Distribution of the correlations between L2 gradient norms and images' sequence-order across all problems. (**d**) The absolute correlation of the encoder's FC layers with the sequence order during the optimization process. Error shades represent 95% CI. Neurons' covariance and correlation calculations are explained in Methods.

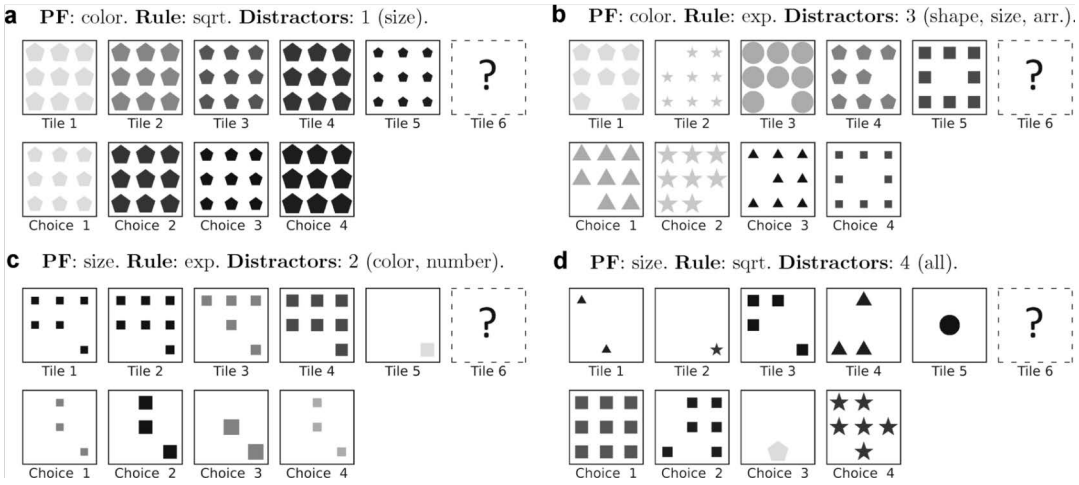

Fig. 6. Non linear rules. The predictive features' values in these tests increased as a square root (**a** and **d**) or exponentially (**b** and **c**). The rest of the features (non-predictive) are either constant or random. *The correct choices are all 3.*

In the vanilla model, the relation module is simple and general, and the encoder that finds appropriate image representations carries most of the "computational load". However, we expect the complexity of the encoder and the complexity of the relation module to be interchangeable, to some extent. Thus, we can move some of the computational load from the encoder to the relation module without changing the performance. To test this, we simplified the encoder by removing the fully-connected layers, leaving only the convolutional layers, and complicated the relation module, by making it a more complex and expressive,

$$R_\theta\left(Z_\phi\left(\mathbf{x}_i\right), Z_\phi\left(\mathbf{x}_j\right)\right) = H_\theta\left(Z_{\text{conv}}\left(\mathbf{x}_i\right) \oplus Z_{\text{conv}}\left(\mathbf{x}_j\right)\right) \tag{3}$$

where the relation module H_θ takes a concatenation of the convolutional layers' outputs to a single output neuron and has a network architecture similar to the vanilla model's encoder's FC layers (with twice the input dimension). Rather than optimizing both the encoder and the relation module, as in the vanilla model, we optimized only the relation module. This version of the model achieved an average accuracy of 0.59 ± 0.01 (Fig. S12) comparable to that of the vanilla model, demonstrating that it is possible to move the computational load from the encoder to the relation module without paying in performance.

To conclude this section, we demonstrated two ways for carrying the computational load. Either the encoder carries most of the load by extracting the relevant feature in a manner that a simple linear relation module is sufficient for capturing the rule. Alternatively, the relation module can carry the computational load. In that case, the relation module finds a *specific relation* between high-dimensional input representations, keeping the input representations fixed during problem-solving.

Knowledge crystallization

Our focus so far was the ability of the networks to solve problems without any training, that is, without any accumulation of information between problems. Embedded in our model, however, is the ability to accumulate knowledge. This is because problem-solving in our model is achieved through changes in synaptic weights. This motivated us to study how solving multiple problems affects performance. In humans, the improvement of performance due to the accumulation of knowledge by training is referred to as knowledge crystallization[23].

We first studied the extent to which the model can improve its performance on one predictive feature by practicing on that feature. Notably, in these practice sessions there was no feedback about the correct answer (in fact, the networks were exposed only to the sequences of 5 images and not to the possible answers). We found that networks that solved 1,000 easy problems with a specific predictive feature (without resetting the weights between problems) improved their accuracy on problems with that same predictive feature to 0.74 ± 0.01 (averaged over the three predictive features, Fig. 7), a substantial improvement from the average accuracy without prior training (0.58 ± 0.01). Notably, the improvement was not uniform across features. While performance on Number and Size substantially improved (Size: from 0.53 ± 0.01 to 0.69 ± 0.01, Number: from 0.50 ± 0.01 to 0.84 ± 0.01), training on Color did not affect performance in Color problems (0.70 ± 0.01 in both conditions).

Interestingly, freezing the weights of the relation module resulted in an even better performance (0.80 ± 0.01 , Fig. S13). On the other hand, the improvement was only modest when the convolutional layers' weights were frozen (0.65 ± 0.01, Fig. S14).

This improvement in performance is analogous to knowledge crystallization. However, will training on one predictive feature improve performance when other predictive features are used? In humans, training on one abstract reasoning task does not necessarily generalize to other tasks[24]. Similarly, we found that while training on one predictive feature improved performance in problems with that same predictive feature, it was detrimental when the networks were tested on problems with a different predictive feature (Fig. S15).

Will training on several predictive features improve network performance on those trained features? To address this question, we focused on the two predictive features that exhibited improvement with training, Number and Size. We used block training and tested performance on the most difficult problems of both types.

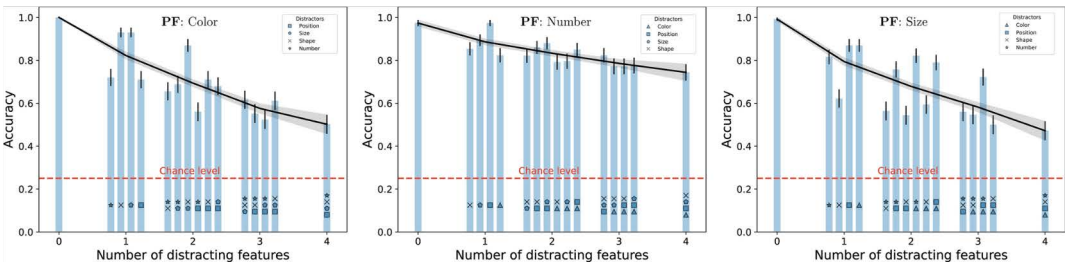

Fig. 7. Knowledge crystallization. The performance of networks that trained on 1,000 easy problems (without distracting features) of a certain predictive feature and tested on the different test conditions of that same predictive feature. Error bars correspond to 95% CI. The black line and its shade are the average accuracy per difficulty and 95% CI corresponding to the mean. The dashed line denotes the chance level given four choice images (0.25).

Considering the first block of training, extensively training the network with one predictive feature improves performance on that feature but not on the other feature (Fig. S16). Considering the second block, when this network is trained on the other predictive feature, the network quickly improves on that feature, but improvement on the first predictive feature quickly diminishes when the network trains on the other feature (Fig. S16a, b). Trying to resolve this by interleaving these two predictive-feature problems in short blocks of 5 problems does not change the result and the network seems unable to simultaneously improve on two predictive features (Fig. S16c, d).

To minimize conflict between the two features, we trained and tested the network in problems in which the competing non-predictive feature (Size or Number) was set at *the same constant value* (see Methods). We found that when training was done in two long blocks, the network only improved on the trained feature (Fig. 8a, b). By contrast, when training was done by interleaving many short blocks of 5 problems, the network improved in both features (Fig. 8c, d).

The fact that the network forgets one feature when training on the other is known in the machine learning literature as *catastrophic forgetting*[25], and indeed, interleaving has been shown to address this problem effectively[26]. Similarly, the fact that interleaving is more effective than block training for learning is also well known in the cognitive literature as the interleaving effect[27,28].

Discussion

We found that naive randomly-initialized ANNs can perform abstract reasoning that does not rely on memorization when they are optimized at test time. This result has implications both the cognitive sciences and for machine learning.

Traditionally, abstract reasoning in humans has been considered a symbolic computation – a type of digital processing distinctly different from the analog nature of computation in ANNs[29,30]. Recently, however, studies have shown that complex computations once attributed solely to symbolic processing can be accomplished by extensively trained ANNs[31,32]. This is especially evident with large language models, which appear capable of performing certain forms of abstract reasoning[12]. Nevertheless, critics of abstract reasoning in ANNs argue that this success may be due more to sophisticated memory retrieval than to genuine abstract reasoning[13,14]. Our contribution is that we show that the tools used for training ANNs can also be used for exhibiting what resembles symbolic abstract reasoning without any training, hence without relying on memory recall.

A key element in our network's ability to perform abstract reasoning tasks is the convolutional part of the encoder. We found that these random convolutional layers are instrumental in extracting features correlated with relevant latent features. Interestingly, optimization of the convolutional layers was not necessary for achieving the performance, but did support knowledge accumulation. These results resonate with human-brain studies. In humans, early visual cortex regions act as general-purpose feature extractors, sensitive to basic features like

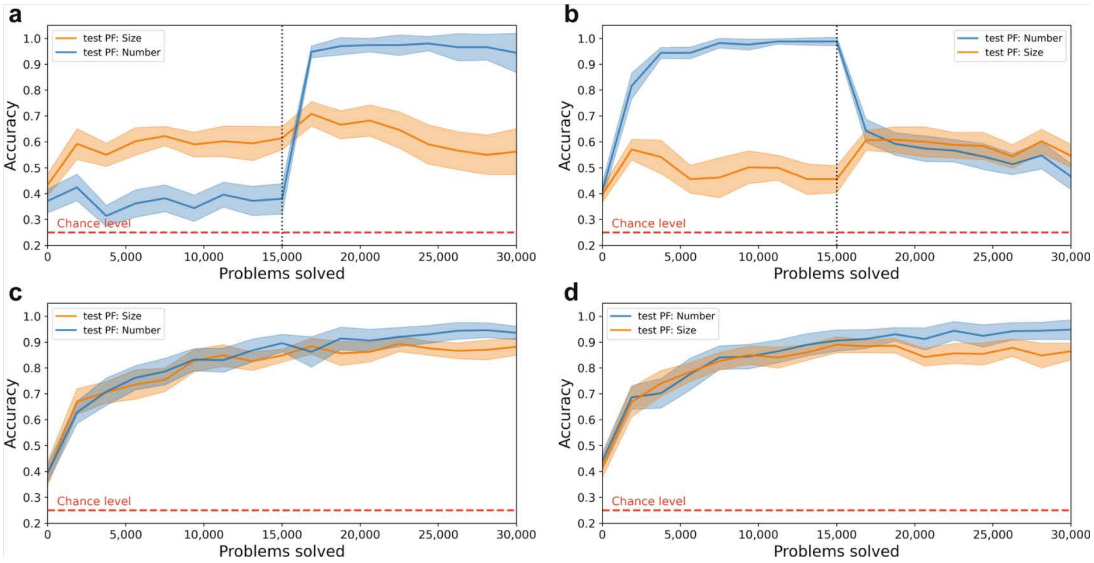

Fig. 8. Interleaving effect. Networks were trained on 15,000 problems in which the predictive feature was Size and 15,000 problems with predictive feature Number. The training problems were either presented in two large consecutive blocks (Size and then Number (**a**); Number and then Size (**b**)) or interleaved at a rate of five problems per predictive feature (Size and then Number (**c**); Number and then Size (**d**)). Errors correspond to 95% CI (see Methods).

orientation and direction. It seems unlikely that this low-level feature extraction changes with every problem presented to a human participant. They may, however, change with extensive training[33].

Highlighting *the* features that are relevant for *the* particular sequence of images of a particular problem was done in the deep layers of the encoder (fully-connected layers), together with the relation module. In humans, imaging studies suggest that higher cortical regions, such as the lateral prefrontal cortex, play an important role in abstract reasoning[34] and rule learning[35]. We hypothesize that these higher cortical regions perform the analog of the optimization-based computation of highlighting the relevant feature (fully-connected layers of the encoder) and identifying its regularity (the relation module).

Notably, the computations performed by the fully-connected layers of the encoder and the relation module are somewhat interchangeable. This is because either of those networks can carry the main computational load. This suggests an interesting approach for finding relations by implementing a few very simple and general (applicable to different problems) relation modules, transferring a significant computational load of finding appropriate input representations to the encoder. Furthermore, given the interchangeability of complexity in the fully-connected layers of the encoder and the relational module, the separation between the encoder and relational module in the brain analog of this computation may not exist.

Our framework naturally generalizes to explaining knowledge acquisition through problem-solving (the practice in Fig. 7 was unsupervised, with no feedback). We found that training with many short interleaved blocks was substantially more effective than training with two long blocks. This resembles a similar observation in the cognitive sciences known as the interleaving effect[27,28]. In the cognitive sciences, two competing theories have been used to explain this effect. In one, the interleaving effect is due to the enhanced problem-identification and feature-distinction required when solving two types of problems in close proximity[36]. The second theory explains the interleaving effect by proposing that with interleaved training, the brain is continually engaged at retrieving the responses from memory – a process that enhances the consolidation of those memories[37,38]. In contrast to these theories, our model has no explicit problem-identification or memory-consolidation mechanisms implemented. Rather, the interleaving effect is a manifestation of the well-known catastrophic forgetting phenomenon in machine learning[25,26].

We focused in this paper on the abstract reasoning of ANNs optimized by gradient descent. These models have shown to achieve performance levels that sometimes rival or even exceed human capabilities, especially in areas like language and visual processing, key aspects of human cognition[10,39]. However, these successes have been achieved by scaling up model size and training data, with models trained on datasets vastly larger than what human children require to learn comparable skills[40,41]. Thus, the sustainability of simply increasing network size and data volume as a path to further improvements has been doubted[42,43]. Our findings suggest a potential alternative approach, in which ANNs, by optimizing their weights at test time, exhibit computational capacity in the absence of massive datasets. This approach offers a promising direction for AI development that prioritizes efficiency over scale.

Abstract reasoning consists of several computational facets. In this work, we focused on only one of them: the identification of relationships between images in order to infer the sequence completion, also termed inductive reasoning. Inductive reasoning is needed for solving many types of intelligence tests[16]. While modeling this facet, our model does not encapsulate other facets of abstract reasoning observed in humans. Specifically, our model does not incorporate working memory, limiting the regularities it can identify. It also does not explicitly perform the mapping computation required for analogical reasoning[20]. Additionally, the model cannot solve a problem by breaking it into its sub-components[44]. For example, to solve a Raven Progression Matrix, humans use the strategy of identifying common regularities in the rows and the columns. Our model was constructed only to find a regularity in a sequence. As with most ANN models, the model cannot interpret its choices. Finally, it lacks the ability to generate new images that follow the regularity it identifies. These limitations present opportunities for future research and suggest areas for improvement.

In humans, evidence suggests that abstract reasoning operates as a general computational process, analogous to a general-purpose computer that can handle any input. For example, an individual's performance on various cognitively demanding tests tends to correlate[45]. As the tests require different prior knowledge, these correlations are taken as support for the hypothesis that a general ability, often termed general intelligence[46], underlies these diverse cognitive skills. Additionally, training on a specific cognitive task usually does not improve performance on unrelated tasks[24]. This lack of transfer suggests that human abstract reasoning is indeed general, relying on general cognitive processes rather than specific learned patterns or memorized solutions. This somewhat resembles our model.

In conclusion, our work demonstrates that ANNs can exhibit abstract reasoning abilities without reliance on memory recall, opening pathways for further exploration of abstract reasoning mechanisms in both artificial systems and humans.

Methods

Code availability

The code for this paper was written using PyTorch[22]. The code that generates test problems and applies the model to solve them is available at https://github.com/Tomer-Barak/learning-independent_abstract_reasoning.

Network architectures

The encoder ($Z_\phi(\mathbf{x})$) consisted of two main components (see Supplementary Fig. S1): three convolutional layers (kernel sizes: 2, 2, and 3; strides: all 1; padding: all 1) and five Fully-Connected (FC) layers (number of neurons: 200, 100, 50, 10, 1). Three ReLU activation functions were applied after each convolutional layer, and two Max-Pool layers (kernels: 4 and 6, strides: all 1) were applied after the second and third convolutional (+ReLU) layers.

Four tanh activation functions were applied after each FC layer, except the last one, which had no activation function and remained a linear transformation.

The vanilla model's relation module consisted of a single parameter as written in equation (2). The more complex relation module written in equation (3) was implemented by a five-layer fully-connected network (number of neurons: 200, 100, 50, 10, 1). Four tanh activation functions were applied after each of this relation module's layers, except the last one, which had no activation function and remained a linear transformation.

Sequential visual reasoning tests

Each image of the tests was constructed using the following five features: the number of objects in an image (possible values: 1 to 9), their shade (6 linearly distributed grayscale values), their shape (circle, triangle, square, star, hexagon), their size (6 linearly distributed values for the shapes' enclosing circle circumference), and arrangement (a vector of grid positions that was used to place the shapes in order).

As written in the paper, the choice images' non-predictive features followed the same rules they abide by in the sequence (constant or randomly changing). The predictive feature followed the sequence rule only in the correct choice and was randomly chosen from the remaining feature values in the incorrect choices. We restricted the possibility of having a repeated choice image in the same problem. If a repeated image was generated by chance, we generated another one to replace it.

Average accuracies

In the paper, we report networks' average accuracies/performance in different experiments. For example, the vanilla model's average accuracy was 0.58 ± 0.01. These numbers were obtained (except in the knowledge crystallization section, discussed below) in the following way. For each predictive feature relevant to the experiment, we considered all its test conditions of different difficulties. There were five features, one predictive and the other four either constant or distracting, amounting to $2^4 = 16$ test conditions per predictive feature. We tested randomly initialized networks in 500 problems in each test condition (each problem with a different initialized network) and obtained their success rate in that test condition. To estimate the errors, we calculated the standard error of the mean of a sample of Binomial random variables based on the success rate and the number of samples (500). To obtain the average accuracy of that predictive feature, we averaged the success rates over all test conditions and propagated the errors accordingly. For the total average accuracy, we averaged the accuracies of the experiment's relevant predictive features and propagated the errors.

In the knowledge crystallization section, the average accuracies (e.g., Fig. 7) were obtained by training 50 networks in each predictive feature on 1000 easy problems (without distractors) of that predictive feature. After training, the networks solved 10 test problems in each of the 16 test conditions of a given predictive feature (results for networks that trained on one predictive feature and tested on another are shown in Fig. S15). We then calculated the average success rate of the networks in each test condition, using the standard error of the mean of Binomial random variables as errors. For the total average accuracy, we averaged across the different test conditions and all the relevant predictive features of the experiment, propagating the errors.

In the blocks versus interleaving experiments (Fig. 8, S16-S18), we trained 20 networks (in each of the figure panels) on 30,000 easy problems of two training predictive features. The training was either in two big blocks or interleaved into small five-problem blocks. After every 1875 training problems, we tested the networks on 25 difficult problems of the two predictive features. In Fig. S16, the easy and difficult problems were such that there were no distractors or all of the distractors. In Fig. 8, both the easy and difficult problems of Size had a fixed value of Number (5 shapes). Accordingly, the easy and difficult problems of Number had a fixed value of Size (the 5^{th} size value).

Correlations

To calculate an encoder neurons' correlations with a particular feature (color, number, or size; Fig. 3 and S9), we generated artificial testing examples corresponding to that feature: 20 images for each of the feature's six possible values (120 examples overall) where the rest of the features' values were drawn randomly. We applied these examples to the network and recorded its neurons' activity. Based on the neurons' activity, we calculated each of the neurons' correlation with the feature values. Finally, we averaged the correlations across the layers.

To generate Fig. 3a, we average the correlations of networks that solved the three possible predictive features with the predictive features they solved. For Fig. 3b, we calculated the correlations with the other (non-predictive) features. In both figures, we averaged over the 16 test conditions of a given predictive feature, 50 problems per test condition, each problem solved by a different naive network. Thus, overall, the results are average over $3 \times 50 \times 6 = 900$ networks. The complete results of these simulations, before averaging over networks and test conditions, are shown in Fig. S8. To calculate the errors of the correlations, we estimated the standard error of the mean of the average correlations of the different networks. We propagated these errors when we averaged the correlations across test conditions and different predictive features.

To calculate the correlations (or covariance) of neurons with the sequence (Fig. 5b-d), we applied the sequence images to the network and recorded its neural activity. Then, we calculated for each neuron the correlation (or covariance) between its activity and the sequence order indices of the images. In Fig. 5b-c, we calculated the covariance with the sequence of neurons taken from the output of the encoder's 3^{rd} convolutional layer, while in Fig. 5d, we calculated correlations with the sequence of the FC layers' neurons. The errors in the latter case were calculated like those of the feature correlations.

Figure 4

To obtain the values of this plot, we considered test conditions with one distracting feature that was either Color, Number, or Size (as those are the features for which we were able to calculate networks' correlations). In total,

there were 6 such test conditions (2 for each predictive feature). For each of the test conditions, the y-axis value is the logit function ($\text{logit}(p) = \ln \frac{p}{1-p}$) of the success rate of 500 problems of that test condition, each solved by a different naive network, obtained from Fig. 2. For the errors of these values, we propagated the success rate errors through the logit function. For the x-axis values, we obtained networks' average absolute correlations with the predictive feature of the problems they solved (after solving them) and compared that with the absolute correlation of the distracting features. The values and errors were obtained from Fig. S8, and the errors were propagated through the ratio. The linear regression analysis was conducted using SciPy's[47] linear regression function.

Figure 5c

The histogram in Fig. 5c was obtained by considering the three predictive features (color, number, and size) and each of their 16 test conditions, 50 problems in each test condition. For each problem, we calculated the correlations between the convolutional layers' output neurons' co-variance with the sequence order and the L2 gradient norm of those neurons and averaged over the neurons. Finally, we plotted the histogram of those averages.

Data availability

No datasets were generated or analyzed during the current study. The visual reasoning tests were generated in real-time by an algorithm (included in the Supplementary materials).

Received: 28 November 2023; Accepted: 31 October 2024
Published online: 08 November 2024

References

1. Lohman, D. F. Complex Information Processing and Intelligence. In Sternberg, R. J. (ed.) *Handbook of Intelligence*, 285–340, https://doi.org/10.1017/CBO9780511807947.015 (Cambridge University Press, 2000).
2. Sternberg, R. J. Component processes in analogical reasoning. *Psychol. Rev.* **84**, 353–378. https://doi.org/10.1037/0033-295X.84.4.353 (1977).
3. Sternberg, R. J. Components of human intelligence. *Cognition* **15**, 1–48. https://doi.org/10.1016/0010-0277(83)90032-X (1983).
4. Kaplan, R. M. & Saccuzzo, D. P. *Psychological Testing: Principles, Applications, and Issues*. (Wadsworth/Thomson Learning, 2009).
5. Raven, J., Raven, J. C. & Court, J. H. *Manual for Raven's Progressive Matrices and Vocabulary Scales* (Pearson, 1998) (**OCLC: 697438611**).
6. Mitchell, M. How do we know how smart AI systems are?. *Science* **381**, eadj5957. https://doi.org/10.1126/science.adj5957 (2023) https://www.science.org/doi/pdf/10.1126/science.adj5957.
7. Hersche, M., Zeqiri, M., Benini, L., Sebastian, A. & Rahimi, A. A neuro-vector-symbolic architecture for solving Raven's progressive matrices. *Nat. Mach. Intell.* **5**, 363–375. https://doi.org/10.1038/s42256-023-00630-8 (2023).
8. Santoro, A. *et al.* A simple neural network module for relational reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4974–4983 (Curran Associates Inc., Red Hook, NY, USA, 2017).
9. Barrett, D., Hill, F., Santoro, A., Morcos, A. & Lillicrap, T. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, 511–520 (PMLR, 2018).
10. OpenAI. GPT-4 Technical Report, https://doi.org/10.48550/arXiv.2303.08774 (2023). ArXiv:2303.08774 [cs].
11. Kim, Y., Shin, J., Yang, E. & Hwang, S. J. Few-shot visual reasoning with meta-analogical contrastive learning. *Adv. Neural. Inf. Process. Syst.* **33**, 16846–16856 (2020).
12. Webb, T., Holyoak, K. J. & Lu, H. Emergent analogical reasoning in large language models. *Nature Human Behaviour* 1–16 (2023). Publisher: Nature Publishing Group UK London.
13. McCoy, R. T., Yao, S., Friedman, D., Hardy, M. & Griffiths, T. L. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve, https://doi.org/10.48550/arXiv.2309.13638 (2023). ArXiv:2309.13638 [cs].
14. Biever, C. ChatGPT broke the Turing test - the race is on for new ways to assess AI. *Nature* **619**, 686–689. https://doi.org/10.1038/d41586-023-02361-7 (2023).
15. Azulay, A. & Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations?. *J. Mach. Learn. Res.* **20**, 1–25 (2019).
16. Siebers, M., Dowe, D. L., Schmid, U., Hernández-Orallo, J. & Martínez-Plumed, F. Computer models solving intelligence test problems: Progress and implications. *Artif. Intell.* **230**, 74–107. https://doi.org/10.1016/j.artint.2015.09.011 (2015) (**ISBN: 9780999241103 Publisher: Elsevier B.V.**).
17. Sung, F. *et al.* Learning to compare: Relation network for few-shot learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
18. Hill, F., Santoro, A., Barrett, D., Morcos, A. & Lillicrap, T. Learning to Make Analogies by Contrasting Abstract Relational Structure (2018).
19. Lu, H., Wu, Y. N. & Holyoak, K. J. Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences* **116**, 4176–4181, https://doi.org/10.1073/pnas.1814779116 (2019).
20. Lu, H., Ichien, N. & Holyoak, K. J. Probabilistic analogical mapping with semantic relation networks. *Psychol. Rev.* **129**, 1078–1103. https://doi.org/10.1037/rev0000358 (2022).
21. Dauphin, Y. N., de Vries, H., Chung, J. & Bengio, Y. RMSProp and equilibrated adaptive learning rates for non-convex optimization. *CoRR* arxiv:1502.04390 (2015)
22. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).
23. Horn, J. L. Intelligence-why it grows, Why it declines. In *Human Intelligence*, 22 (Routledge, 1972).
24. Jaeggi, S. M., Buschkuehl, M., Jonides, J. & Perrig, W. J. Improving fluid intelligence with training on working memory. *Proc. Natl. Acad. Sci.* **105**, 6829–6833 (2008).
25. French, R. M. Catastrophic forgetting in connectionist networks. *Trends Cogn. Sci.* **3**, 128–135. https://doi.org/10.1016/S1364-6613(99)01294-2 (1999).
26. Robins, A. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect. Sci.* **7**, 123–146. https://doi.org/10.1080/09540099550039318 (1995).
27. Kornell, N. & Bjork, R. A. Learning concepts and categories: Is spacing the enemy of induction?. *Psychol. Sci.* **19**, 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x (2008).
28. Brunmair, M. & Richter, T. Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychol. Bull.* **145**, 1029–1052. https://doi.org/10.1037/bul0000209 (2019).

29. Turing, A. M. & others. On computable numbers, with an application to the Entscheidungsproblem. *J. Math.***58**, 5 (1936).
30. Turing, A. M. Computing machinery and intelligence. *Mind* LIX, 433–460. https://doi.org/10.1093/mind/LIX.236.433 (1950) https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf.
31. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, vol. 25 (Curran Associates, Inc., 2012).
32. Campbell, D., Kumar, S., Giallanza, T., Cohen, J. D. & Griffiths, T. L. Relational constraints on neural networks reproduce human biases towards abstract geometric regularity, https://doi.org/10.48550/arXiv.2309.17363 (2023). ArXiv:2309.17363 [q-bio].
33. Ahissar, M. & Hochstein, S. The reverse hierarchy theory of visual perceptual learning. *Trends Cogn. Sci.* **8**, 457–464. https://doi.org/10.1016/j.tics.2004.08.011 (2004).
34. Gray, J. R., Chabris, C. F. & Braver, T. S. Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* **6**, 316–322. https://doi.org/10.1038/nn1014 (2003).
35. Mansouri, F. A., Freedman, D. J. & Buckley, M. J. Emergence of abstract rules in the primate brain. *Nat. Rev. Neurosci.* **21**, 595–610. https://doi.org/10.1038/s41583-020-0364-5 (2020).
36. Rohrer, D. Interleaving helps students distinguish among similar concepts. *Educ. Psychol. Rev.* **24**, 355–367. https://doi.org/10.1007/s10648-012-9201-3 (2012).
37. Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J. & Willingham, D. T. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychol. Sci. Public Interest* **14**, 4–58 (2013).
38. Krug, D., Davis, T. B. & Glover, J. A. Massed versus distributed repeated reading: A case of forgetting helping recall?. *J. Educ. Psychol.* **82**, 366 (1990).
39. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models, https://doi.org/10.48550/arXiv.2006.11239 (2020). ArXiv:2006.11239.
40. Laurence, S. & Margolis, E. The poverty of the stimulus argument. *Br. J. Philos. Sci.* **52**, 217–276. https://doi.org/10.1093/bjps/52.2.217 (2001).
41. Frank, M. C. Bridging the data gap between children and large language models. *Trends Cogn. Sci.*[SPACE]https://doi.org/10.1016/j.tics.2023.08.007 (2023).
42. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era*, 843–852 (2017).
43. Hestness, J. *et al.* Deep learning scaling is predictable, empirically, https://doi.org/10.48550/arXiv.1712.00409 (2017). ArXiv:1712.00409 [cs, stat].
44. Carpenter, P. A., Just, M. A. & Shell, P. What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychol. Rev.* **97**, 404–431. https://doi.org/10.1037/0033-295X.97.3.404 (American Psychological Association, 1990).
45. Gottfredson, L. S. *The General Intelligence Factor* (1998).
46. Spearman, C. General intelligence, objectively determined and measured. *Am. J. Psychol.* **15**, 201–293. https://doi.org/10.2307/1412107 (1904).
47. Virtanen, P. et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272. https://doi.org/10.1038/s41592-019-0686-2 (2020).

Acknowledgements

This work was supported by the Gatsby Charitable Foundation. Y.L. holds the David and Inez Myres Chair in Neural Computation.

Author contributions

T.B. conducted the experiments, T.B. and Y.L. analyzed the results, wrote, and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-78530-z.

Correspondence and requests for materials should be addressed to T.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Untrained neural networks can demonstrate abstract reasoning without memorization (supplementary information)

Tomer Barak[1,*] Yonatan Loewenstein[1,2]

[1]The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Israel

[2]Department of Cognitive Sciences, The Federmann Center for the Study of Rationality, The Alexander Silberman Institute of Life Sciences, The Hebrew University, Jerusalem, Israel

[*]tomer.barak@mail.huji.ac.il

Encoder architecture

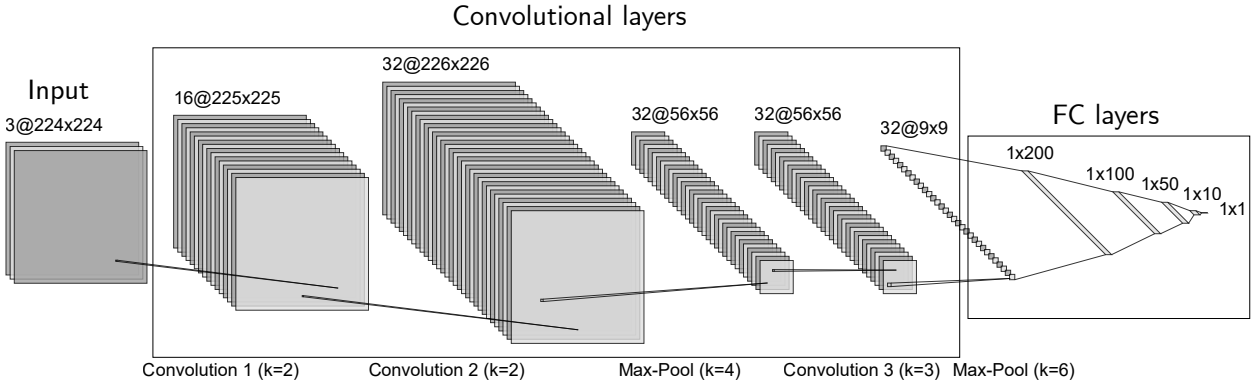

Figure S1: **Encoder's architecture** The encoder consisted of two main parts: three convolutional layers (kernel sizes (k): 2, 2, and 3) and five Fully-Connected (FC) layers. Three ReLU activation functions were applied after each convolutional layer, and two Max-Pool layers were applied after the second and third convolutional layers. Four Tanh activation functions were applied after each FC layer, except the last one, which had no activation function and remained a linear transformation. This figure was generated by NN-SVG [1].

Performance figures

Determinant of success

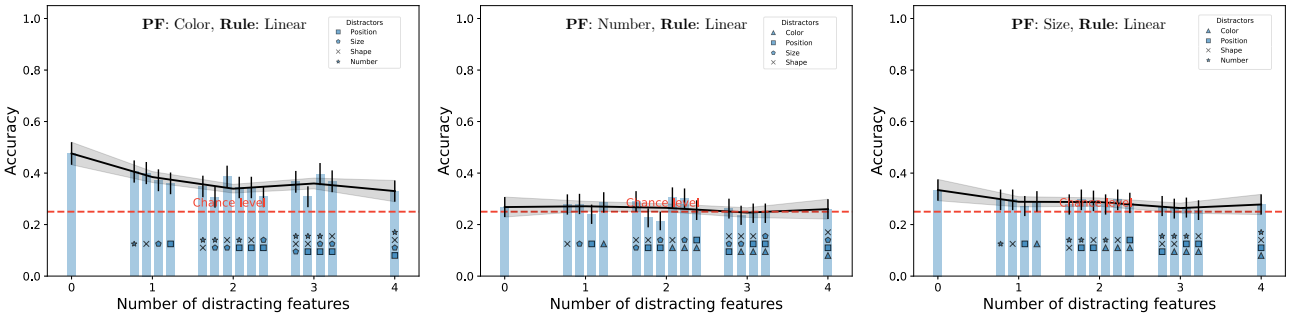

Figure S2: **Frozen encoder** The performance of naive ANNs with frozen encoder weights on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

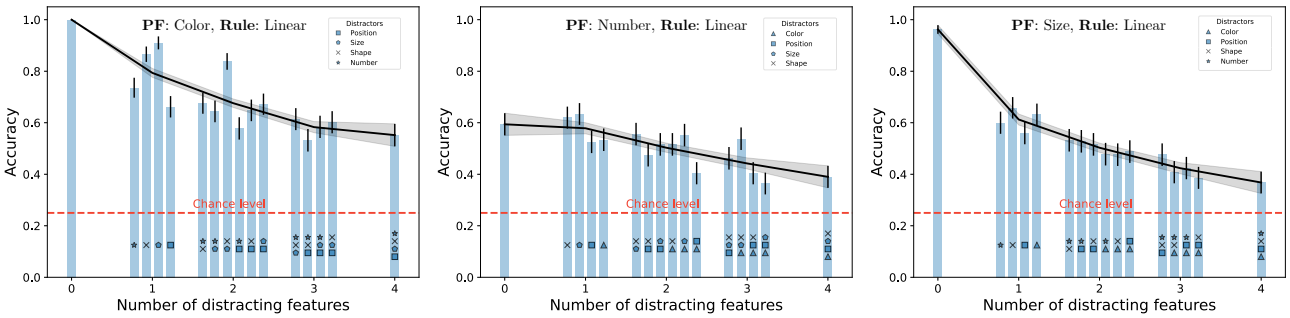

Figure S3: **Frozen relational module** The performance of naive ANNs with a frozen relational module on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

Encoder

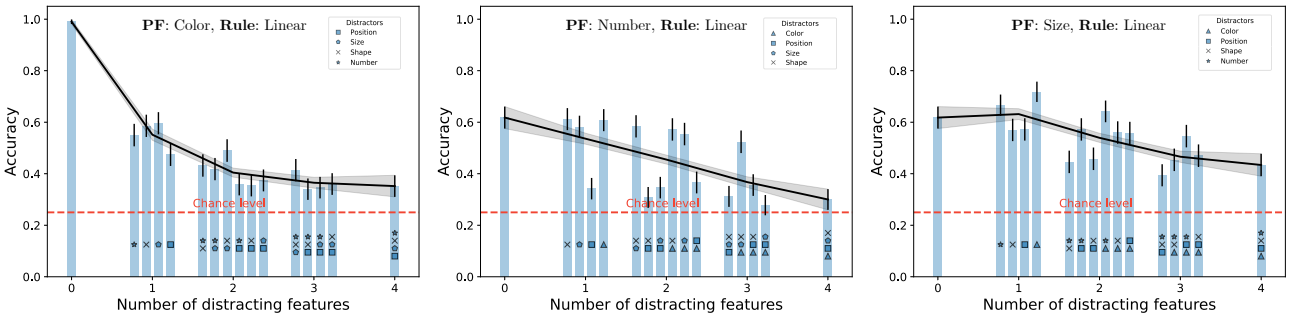

Figure S4: **Without convolutional layers** The performance of naive ANNs without convolutional layers, where the FC layers are directly connected to the images, on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

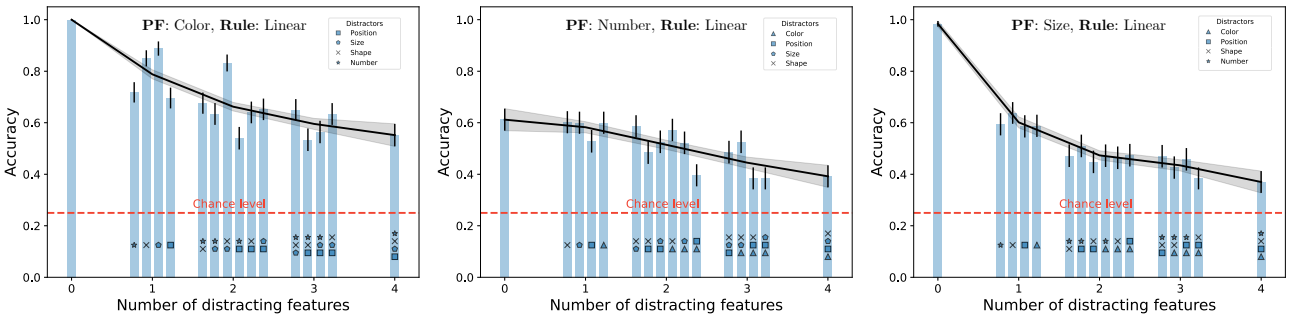

Figure S5: **Frozen convolutional layers** The performance of naive ANNs with frozen convolutional layers on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

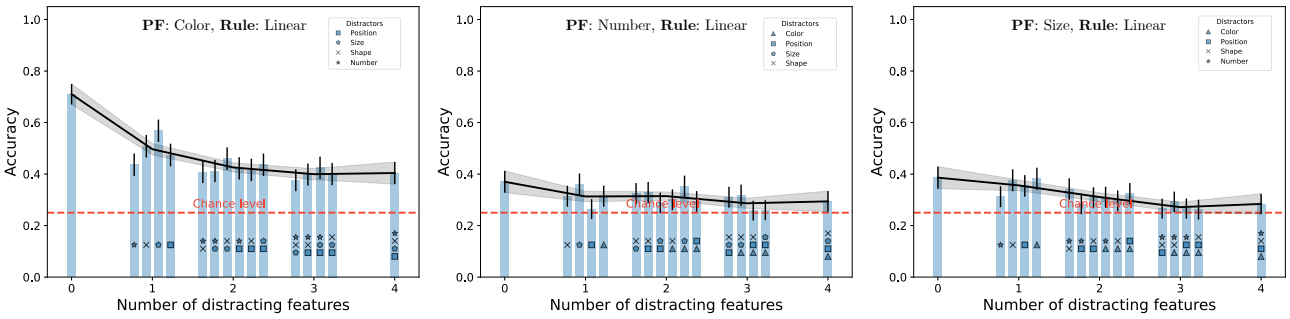

Figure S6: **Frozen FC layers** The performance of naive ANNs with frozen FC layers on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

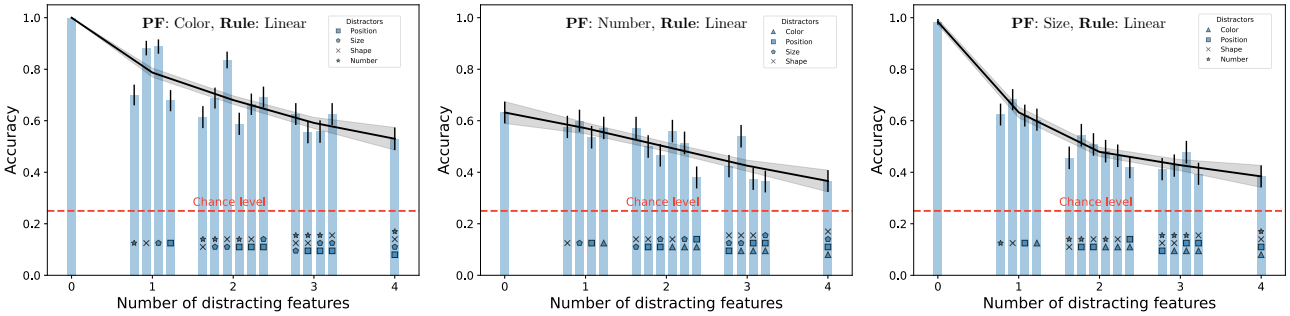

Figure S7: **Frozen convolutional and FC layers** The performance of naive ANNs with frozen convolutional and FC layers on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

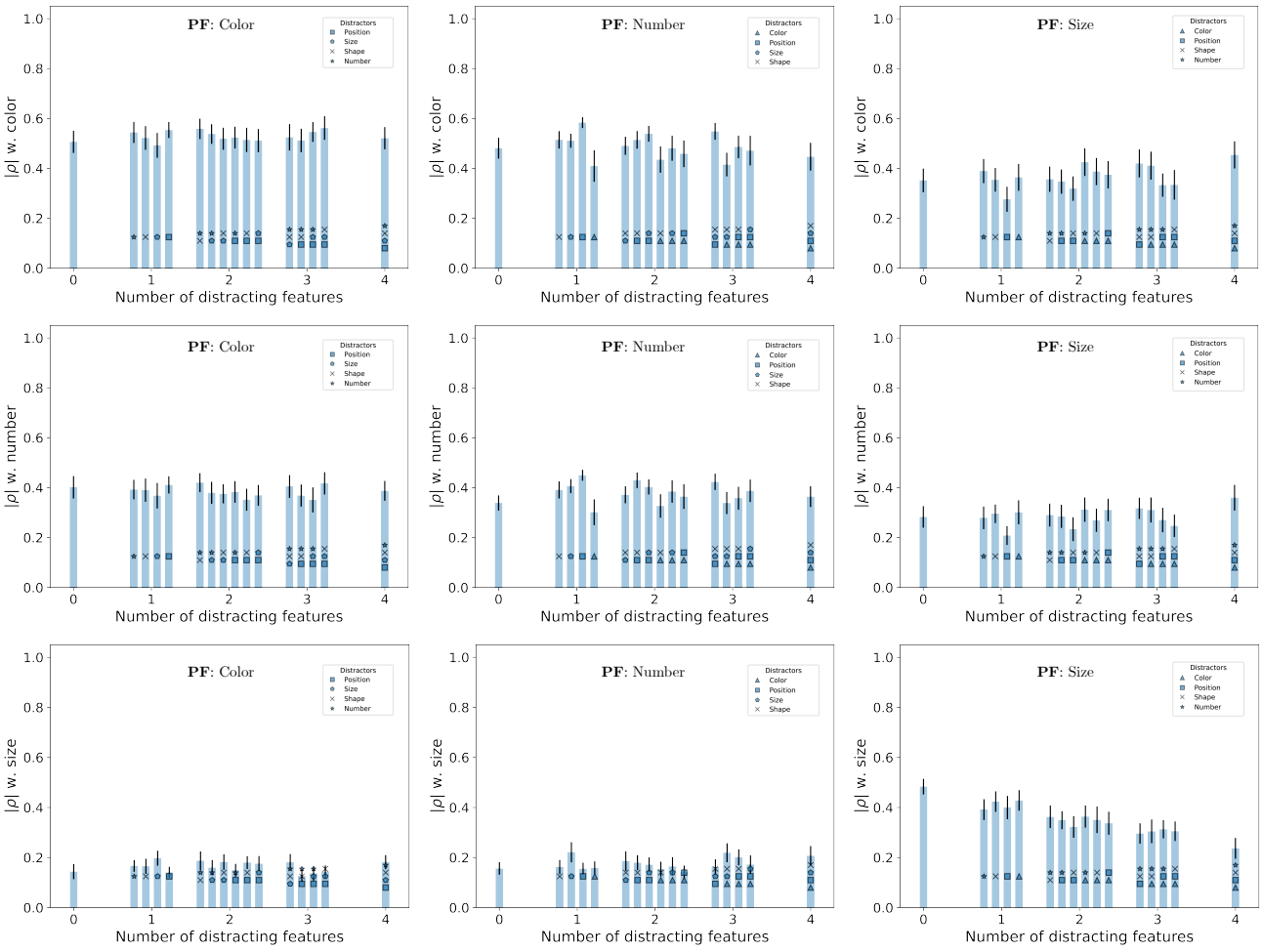

Figure S8: **Feature correlations** For each predictive feature (rows) and test conditions (number of distractors), we measured the average correlation of the output neurons of 50 networks after optimization with the color, number, and size features (columns). Error bars correspond to a 95% confidence interval.

Relation module

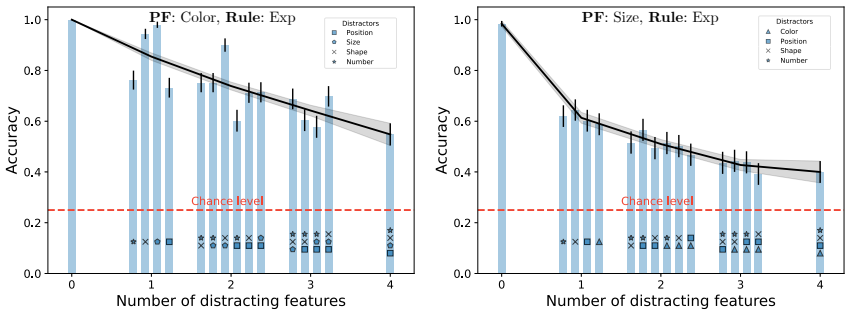

Figure S9: **Exponential relations** The performance of naive ANNs on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was changing exponentially along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

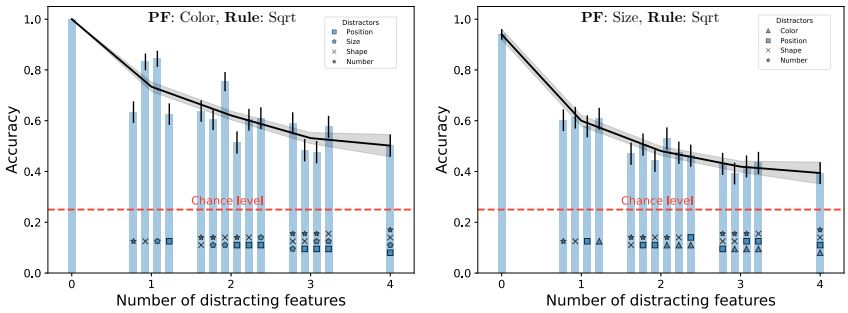

Figure S10: **Square root relations** The performance of naive ANNs on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was changing as a square root along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

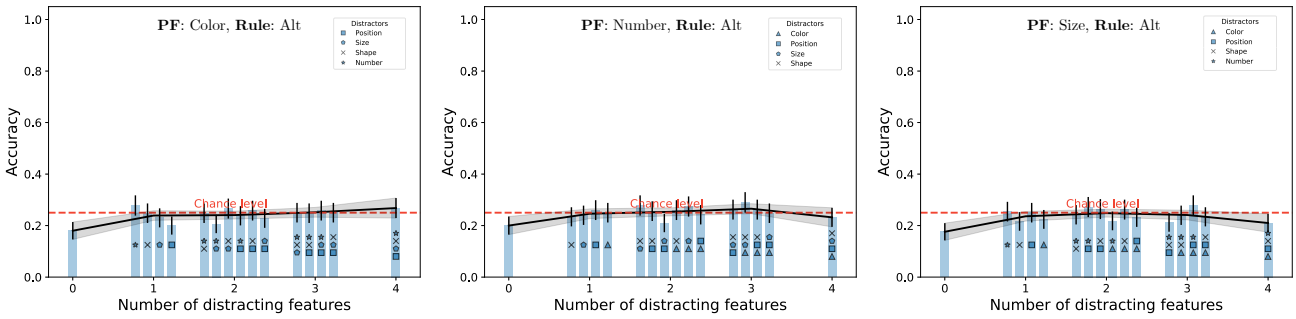

Figure S11: **Alternating relations** The performance of naive ANNs on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was alternating between two values along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

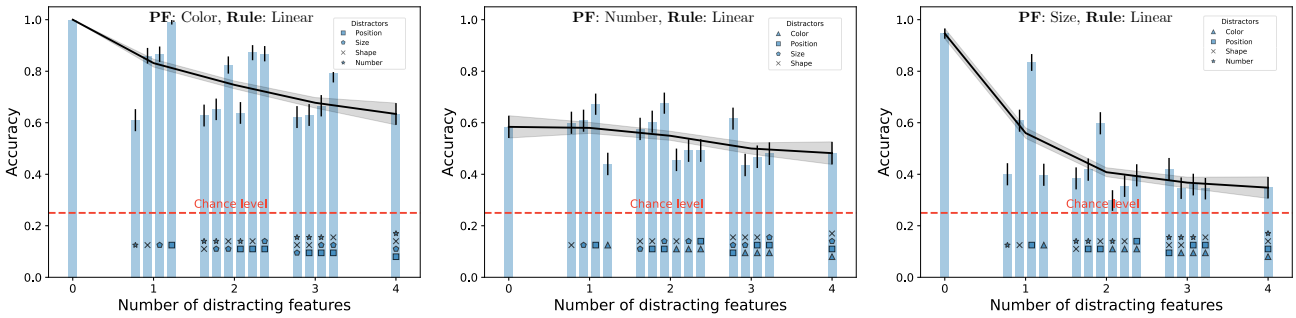

Figure S12: **Complex Relation Module** The performance of naive ANNs with a complex relational module (Eq. (3)) on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested the networks over 16 test conditions where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Each test condition included 500 randomly generated problems. Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

Knowledge crystallization

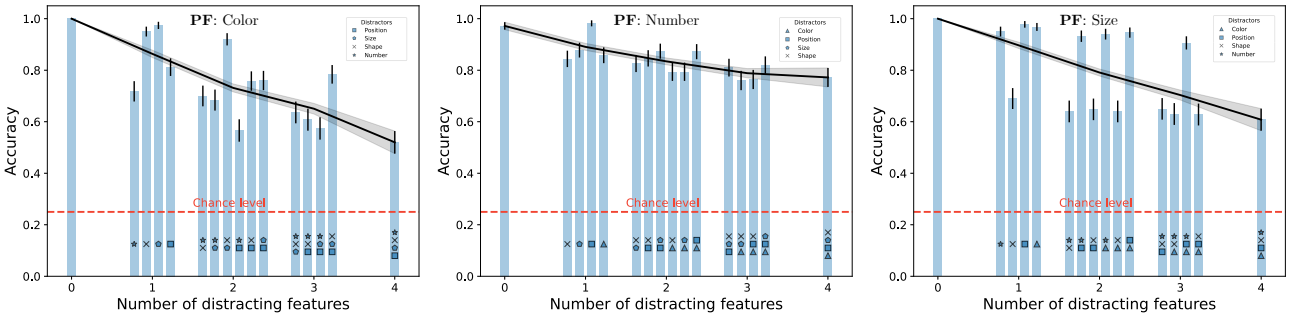

Figure S13: **Extensive training with frozen relation module** The performance of extensively trained networks with frozen relational module on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested 50 networks over 16 test conditions, 10 problems in each test condition, where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

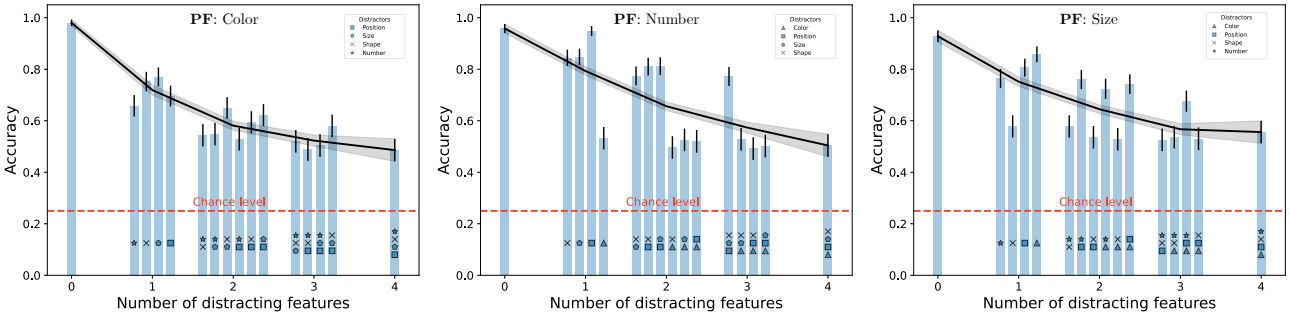

Figure S14: **Extensive training with frozen convolutional layers** The performance of extensively trained networks with frozen convolutional layers on the three Predictive Features (PFs): Color (left), Number (center), and Size (right). For each PF, we tested 50 networks over 16 test conditions, 10 problems in each test condition, where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

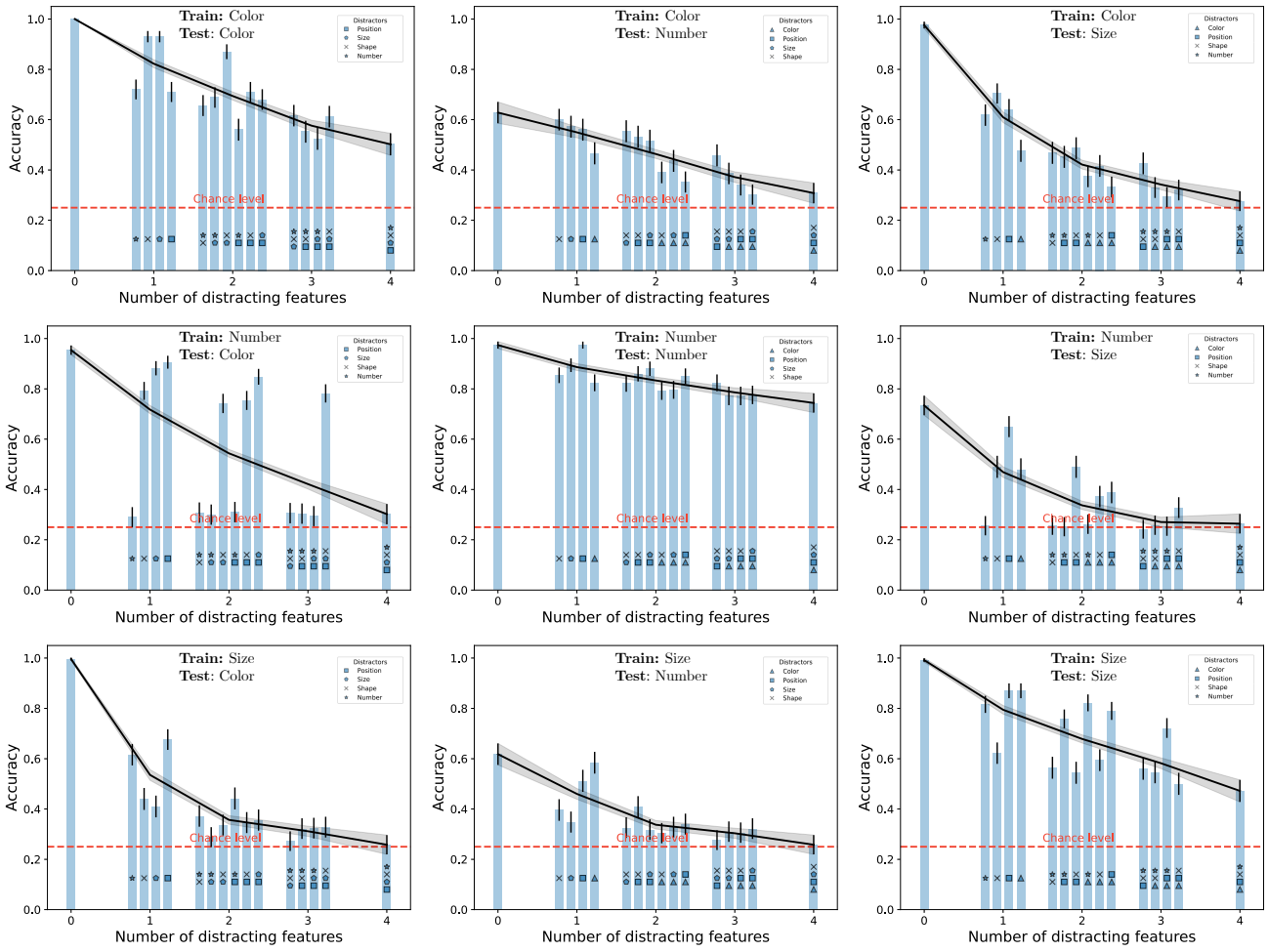

Figure S15: **Full extensive training results with conflicting features.** The performance of extensively trained networks that trained easy tests of either of the three Predictive Features: Color (top row), Number (middle row), Size (bottom row) and were tested on these features: Color (left), Number (center), and Size (right). For each test PF, we tested 20 networks over 16 test conditions, 25 problems in each test condition, where the PF was changing linearly along the sequence, and features that were not predictive were either distractors (marked according to the legend) or constant (not marked). Error bars are 95% confidence intervals. The black line and its shade are the average accuracy per difficulty and the standard deviation. The dashed line denotes the chance level given 4 choice images (0.25).

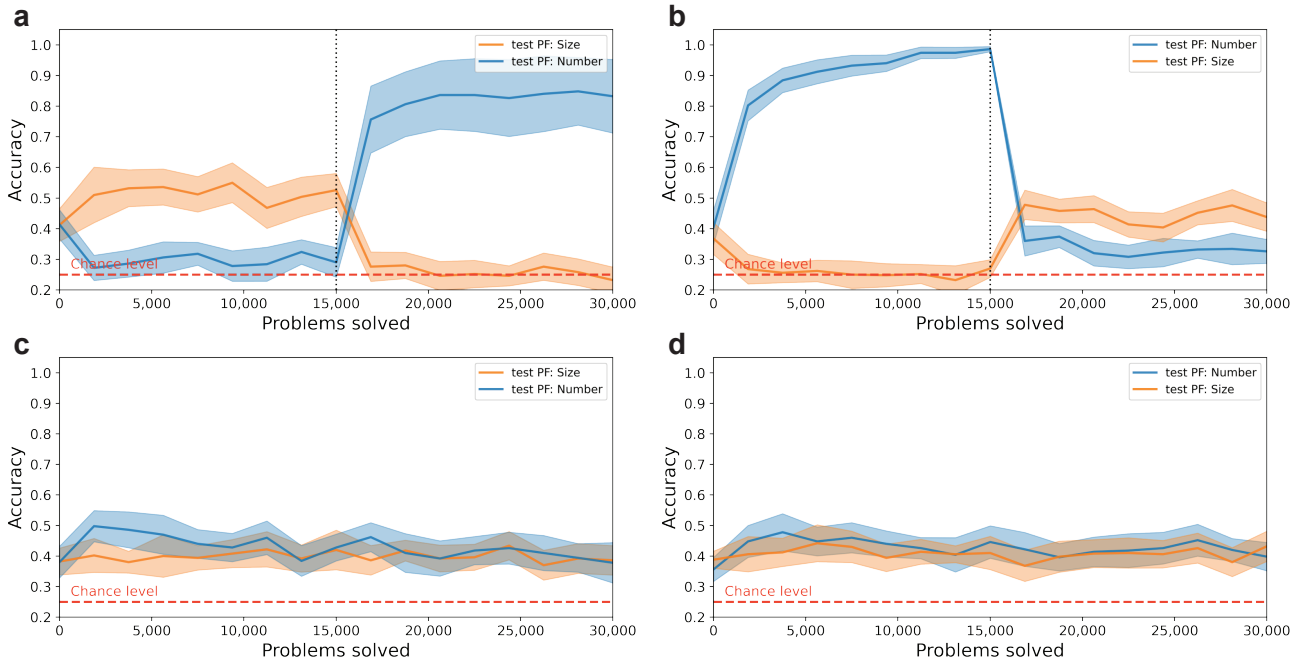

Figure S16: **Solving two inconsistent problem types.** Networks were trained on 15,000 problems in which the predictive feature was Size and 15,000 problems with predictive feature Number. Training problems of a certain PF were easy in the sense that all the non-PFs were constant along the sequences. However, these features' constant values, including those of the other PF, changed between the problems. Furthermore, the testing problems of a certain PF included the other PF as distractors. The training problems were either presented in two large consecutive blocks (Size and then Number (**a**); Number and then Size (**b**)) or interleaved at a rate of five problems per predictive feature (Size and then Number (**c**); Number and then Size (**d**)). Errors correspond to 95% CI (see Methods).

References

[1] LeNail, (2019). NN-SVG: publication-Ready Neural Network Architecture Schematics. Journal of Open Source Software, 4(33), 747, https://doi.org/10.21105/joss.00747

Two pathways to resolve relational inconsistencies

Tomer Barak, Yonatan Loewenstein

Status: Published

Scientific Reports (2025).
https://doi.org/10.1038/s41598-025-16135-w

scientific reports

Check for updates

OPEN

Two pathways to resolve relational inconsistencies

Tomer Barak[1]✉ & Yonatan Loewenstein[1,2]

When individuals encounter observations that violate their expectations, when will they adjust their expectations and when will they maintain them despite these observations? For example, when individuals expect objects of type A to be smaller than objects B, but observe the opposite, when will they adjust their expectation about the relationship between the two objects (to A being larger than B)? Naively, one would predict that the larger the violation, the greater the adaptation. However, experiments reveal that when violations are extreme, individuals are more likely to hold on to their prior expectations rather than adjust them. To address this puzzle, we tested the adaptation of artificial neural networks (ANNs) capable of relational learning and found a similar phenomenon: Standard learning dynamics dictates that small violations would lead to adjustments of expected relations while larger ones would be resolved using a different mechanism—a change in object representation that bypasses the need for adaptation of the relational expectations. These results suggest that the experimentally-observed stability of prior expectations when facing large expectation violations is a natural consequence of learning dynamics and does not require any additional mechanisms. We conclude by discussing the effect of intermediate adaptation steps on this stability.

Imagine strolling through an art museum, expecting awe-inspiring masterpieces. Suddenly, disrupting your expectations, you encounter, displayed in the vitrine, ... a banana. This can be framed as a violation of a relational expectation: On the one hand, the artistic value of museum displays is expected to be *greater* than the artistic value of mundane objects. On the other hand, a banana seems to have a limited artistic value. There are two ways to resolve this inconsistency: recalibrate the expected relationship between the museum displays and mundane objects, or maintain this expectation and find an alternative explanation for the observation (e.g., find a deeper appreciation for the artistic value of bananas).

Previous studies suggested the existence of distinct cognitive modules associated with the generation of representations and the encoding of relations both in humans and other species[1–7]. In the banana example, the *representational* module, which extracts task-relevant features from inputs, determines the artistic value of objects, whether a Rodin sculpture or a banana. Adaptation of this module would correspond to finding merit in bananas. The *relational* module encodes the expected relationship between representations, or between these representations and a predefined anchor. For instance, the relational module in the banana scenario encodes the expectation that objects displayed in an art museum surpass a certain threshold of artistic value, and its adaptation would result in decreasing this threshold. While in the banana example both modules can simultaneously adapt—slightly changing the representation of bananas and the expectation from museums—this is not the case in all violations of relational expectations.

Consider a scientist who has consistently observed that particles of type B are larger than particles of type A. With new experimental techniques, experiments suggest the opposite: particle A is actually larger than particle B. This unexpected finding forces the scientist to choose between two alternative adaptation pathways: maintain the view that B is larger than A by, for example, questioning the validity of the experimental results (adapt the representational module), or alternatively, update their view and conclude that A is, indeed, larger than B (adapt the relational module). This scenario sets two distinct possible adaptation pathways.

Our study was motivated by recent experiments that found that the choice between resolution pathways exhibits an inverted U-shaped dependence on the size of the violation[8–11]: When the violation is minute, its effect on the expectations is small. Larger violations have a larger effect on the expectations. Even larger violations (*extreme* violations), however, fail to alter expectations. Somewhat similar findings were also observed in non-human animals in the framework of associative learning. When moderate, stronger unconditional stimuli elicit stronger conditional response. However, when extreme, the magnitude of the resultant conditional response *decreases* with the magnitude of unconditional stimulus[12–14]. These results are surprising because naively, operant learning, predictive coding and Bayesian models assert that the larger the prediction error, the greater

[1]The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Israel. [2]Department of Cognitive and Brain Sciences, The Federmann Center for the Study of Rationality, The Alexander Silberman Institute of Life Sciences, The Hebrew University, Jerusalem, Israel. ✉email: tomer.barak@mail.huji.ac.il

the adaptation[15–19]. Therefore, gating mechanisms that modulate the magnitude of the adaptation[10], for example, "immunization" mechanisms that prevent adaptation in extreme violation settings have been proposed[20–23].

In this paper we show that the inverted U-shaped dependence of adaptation on the size of the violation naturally emerges in artificial neural networks (ANNs) that are trained to identify relations using standard (gradient) learning. We first present the results from a behavioral perspective using deep networks. Then, we explain these results by mathematically analyzing a simplified model.

Results

The order discrimination task

We study adaptation using the order discrimination task: Agents are presented with image pairs and are instructed to determine, for each pair, if it was presented in the correct or reverse order. Each image depicts shapes arranged on a 3×3 grid and is characterized by five features: the grayscale color of the shapes, their number, size, grid arrangement, and shape type (Fig. 1). The first three features—color, number, and size—are described by a scalar number, establishing natural possible order relations between images. The "correct" order in the task depends on the identity of the relevant feature (color, size, or number), termed the *predictive feature*, and whether this feature increases or decreases from left to right.

To teach the underlying rule, the agents are presented with a series of "correctly" ordered image pairs (Fig. 1a), such that the predictive feature changes according to the underlying rule, while the remaining features, irrelevant to deciding the correct order, are randomly chosen for each pair but remain constant within the pair (as in the test images). In the main text of this paper we present the results when the predictive feature was the size. We demonstrate the generality of our findings by presenting similar results in the Supplementary Information, when the predictive features were color or number (Figures S1–S4).

A key parameter in this task is the difference in the predictive feature values between the two images, a quantity that we denote by α. A positive α implies that the predictive feature increases from left to right, whereas a negative α implies that it decreases. The absolute value of α determines the magnitude of change: $\alpha = 0$ implies that the feature does not change between the two images, whereas $\alpha = \pm 1$ signifies maximal difference. In the real world, α would correspond to the (scaled) objective difference between objects. In the example of the scientist measuring particle sizes, α is the (scaled) objective difference between the sizes of particle A and particle B.

The ANN

As agents, we used ANNs that were designed to emulate relational learning[24–27]. Specifically, our networks were comprised of two modules. The first is a representational module, an encoder which we denote by Z. Its goal is to extract a relevant feature from inputs. For each of the two images of a pair, x and $\mathbf{x}'$ (left image and right image, respectively), it maps the $n \times n$ image to one-dimensional variables $Z_{\mathrm{w}}(\mathrm{x})$ and $Z_{\mathrm{w}}(\mathbf{x}')$, where w are trainable parameters. We implemented this mapping using a multilayered convolutional network. A similar architecture was shown in a previous study to be able to extract the relevant features in an intelligence test[28]. The difference between the representations of the two images is given by $\Delta Z = Z_{\mathrm{w}}(\mathbf{x}') - Z_{\mathrm{w}}(\mathrm{x})$.

The relational module, which we denote as R_θ, characterizes the expected relation between the representations of the two images of the pair. In general, a relational module could be any function $R_\theta\left(Z_{\mathrm{w}}\left(\mathrm{x}'\right), Z_{\mathrm{w}}\left(\mathrm{x}\right)\right)$. We used a constant, single-parameter function that encodes the expected difference between representations, $R_\theta = \theta$.

Formally, we define a loss function for a pair of images as:

$$\mathscr{L}(\mathrm{w}, \theta) = \left(\left(Z_{\mathrm{w}}\left(\mathrm{x}'\right) - Z_{\mathrm{w}}\left(\mathrm{x}\right)\right) - R_\theta\right)^2 = \left(\Delta Z - \theta\right)^2 . \tag{1}$$

If $\Delta Z = \theta$, that is, if the difference between the two representations ΔZ is equal to the expected relationship θ then the loss function is minimized. Formulated this way, however, if $Z_{\mathrm{w}}(\mathrm{x}) = 0$ for all x then $\Delta Z = 0$ for all x and $\theta = 0$ will trivially minimize the loss. To avoid the model collapsing into this trivial solution, we defined a regularized loss function, in which ΔZ and θ are driven to reside on a ring,

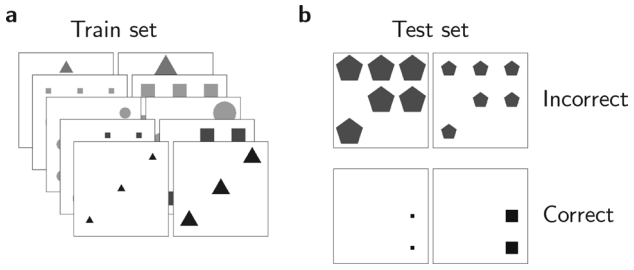

Fig. 1. Order discrimination task. (**a**) Training set that demonstrates an underlying rule between images. The shapes in the right images are larger than those in the left images. The difference is characterized by α. In these examples, $\alpha = 0.5$, corresponding to half of the maximal size difference possible in our simulations. (**b**) The task is to determine whether two images are in the correct order according to the rule that characterizes the training set.

$$\tilde{\mathscr{L}}(\mathrm{w},\theta) = \mathscr{L} + \lambda\left(\Delta Z^2 + \theta^2 - r^2\right)^2. \quad (2)$$

$\lambda > 0$ and r are hyper-parameters. With this additional regularization term, solutions such that $\tilde{\mathscr{L}}(\mathrm{w},\theta) = 0$, if attainable, would reside in the points where the line $\Delta Z = \theta$ intersects with the ring $\Delta Z^2 = \theta^2 = r^2/2$, $\Delta Z = \theta = r/\sqrt{2}$ and $\Delta Z = \theta = -r/\sqrt{2}$.

The model parameters w and θ are trainable and we used Stochastic Gradient Descent (SGD) on $\tilde{\mathscr{L}}(\mathrm{w},\theta)$ to learn them (see Methods). To evaluate the performance of the ANN, we measured its ability to determine the order of novel test images. Specifically, we presented the ANN with two images, measured the values of the loss function associated with the two possible orders of these images (left-right or right-left), and chose the order that minimized the loss function. Figure 2a, depicts the average performance of 100 ANNs as a function of the size of the training set when $\alpha = 0.5$, showing that the networks successfully learned to solve the task after less than 40 image pairs. The high performance holds for other values of α. We tested α values ranging from 0.1 to 1, training them on 160 image pairs, and found that throughout this range, the ANNs performance exceeded 90% accuracy (Fig. 2b). Larger differences in the predicted feature between the two images (α) were associated with higher performance, signifying that the magnitude of α is a measure of the difficulty of the task. This robust learning performance, in a model with distinct representational and relational modules, reflects the ability shown in humans and non-human animals to learn relationships regardless of absolute attributes. Moreover, the higher performance with larger α values align with observations from those studies, where clear distinctions between stimuli support more robust relational behavior[3,4,29,30].

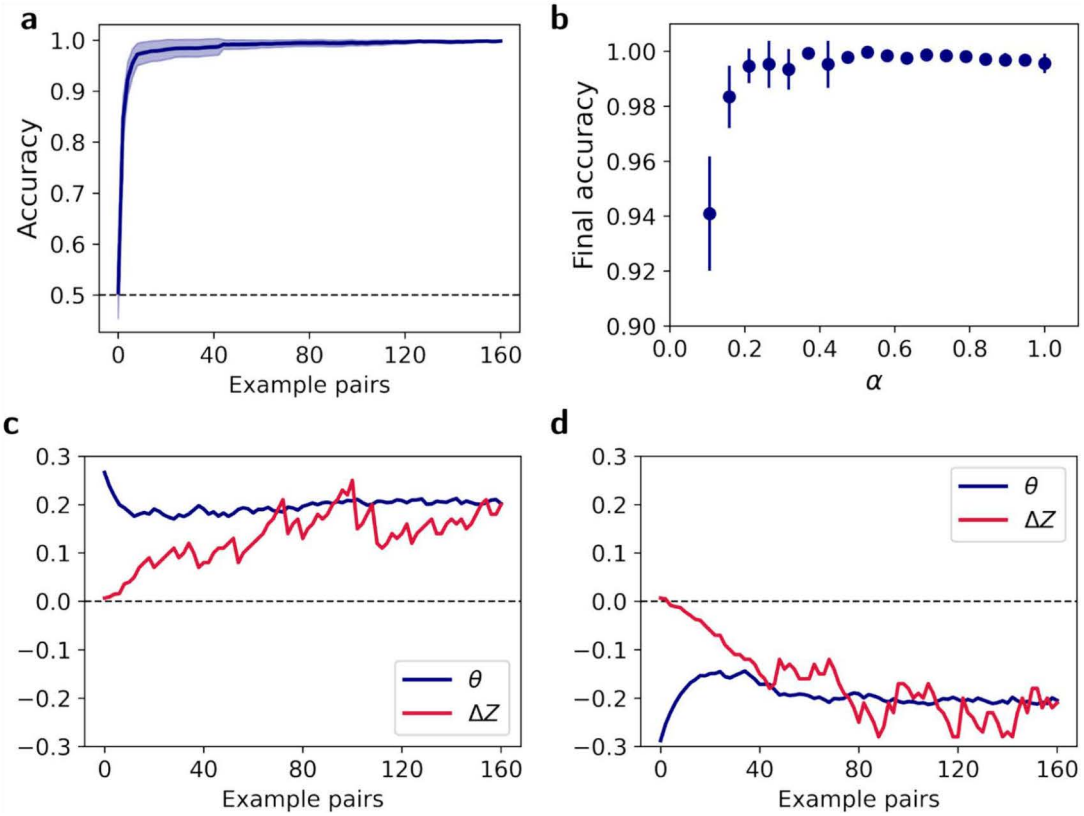

Fig. 2. Task performance and solutions. (**a**) The average test accuracy of 100 networks trained on a task where the predictive feature, size, changed by $\alpha = 0.5$. (**b**) The final test accuracies for various values of α, averaged over 100 networks per α. Error shades and bars correspond to 95% CI. (**c**, **d**) The model can solve the task using two different internal strategies. In one strategy, (**c**) a representative network learned to measure "largeness," where it perceived the shapes getting larger (positive ΔZ) and expected them to get larger (positive θ). In an equally effective strategy, (**d**) another network learned to measure "smallness," where it perceived the shapes becoming less small (negative ΔZ) and expected "smallness" to decrease (negative θ). Both solutions are equally valid for solving the task.

It is instructive to separately consider how each of the two modules adapt in the process of learning. This is depicted in Fig. 2c for one representative network, where we plot the values of ΔZ and θ as a function of example pairs. According to Eq. (2), the regularized loss is minimized when $\Delta Z^2 = \theta^2 = r^2/2$. In this simulation, $r^2 = 0.1$. While the values of ΔZ and θ vary between example pairs, they approach $\Delta Z \approx \theta \approx r/\sqrt{2} \approx 0.2$. Figure 2d depicts another example network. In this simulation, which differed only in the initial parameters, ΔZ and θ converged to a negative solution, where $\Delta Z \approx \theta \approx -r/\sqrt{2} \approx -0.2$. From a point of view of task performance, the positive and negative solutions ($\pm r/\sqrt{2}$) are identical. In the positive solution, the representational module learns to extract the size of the shapes, that is, how large they are, and the relational module learns that this size increases between the two images (from left to right). In the negative solution, the representational module learns to extract the "smallness" of the shapes (the negative of the size) and the relational module learns that the "smallness" decreases between the two images.

Dual adaptation pathways in ANNs

To study the violation of relational expectations in the ANNs, we trained them with sequences of image pairs, characterized by a specific predictive feature that changes by $\alpha_1 > 0$ between the two images of the pair. After learning, we changed the training set's relation rule to $-\alpha_2$ where $\alpha_2 > 0$. That is, the shapes' sizes decreased rather than increased from left to right, violating the relational expectation. We continued training the ANNs with the new rule and measured the performance of the ANNs as a function of examples.

The rule sizes before and after reversal, α_1 and α_2, determine the magnitude of the violation. Specifically, we expect that large α_1 and α_2 would correspond to a large violation while small α_1 and α_2 correspond to a small violation. To illustrate the core phenomenon, we begin our analysis by considering the specific case of a symmetric rule reversal: $\alpha_1 = \alpha_2$. To simplify notations, we write $\alpha_1 = \alpha$ and $\alpha_2 = \alpha$, therefore analyzing the case of rule reversal, $\alpha \to -\alpha$. Later, we study how α_1 and α_2 independently affect the adaptation pathway.

Crucially, there are two ways of resolving a rule reversal violation. Recall that in Fig. 2 we saw two different solutions that networks trained on the task identified. In one, both ΔZ and θ were positive, while in the other, both were negative. There, we discussed the fact that the solution that minimizes the loss function is not unique: $\Delta Z = \theta = r/\sqrt{2}$ and $\Delta Z = \theta = -r/\sqrt{2}$ both minimize the loss function. These two solutions are depicted schematically in Fig. 3a. Following rule reversal, ΔZ necessarily changes its sign, violating the expectation set by θ. As illustrated in Fig. 3b, one possibility for resolving the expectation violation is by changing the weights of the representational module, Z_w, so that ΔZ would return to its pre-reversal sign keeping the relational module unchanged. The other possibility is that the representational module retains its post-reversal sign of

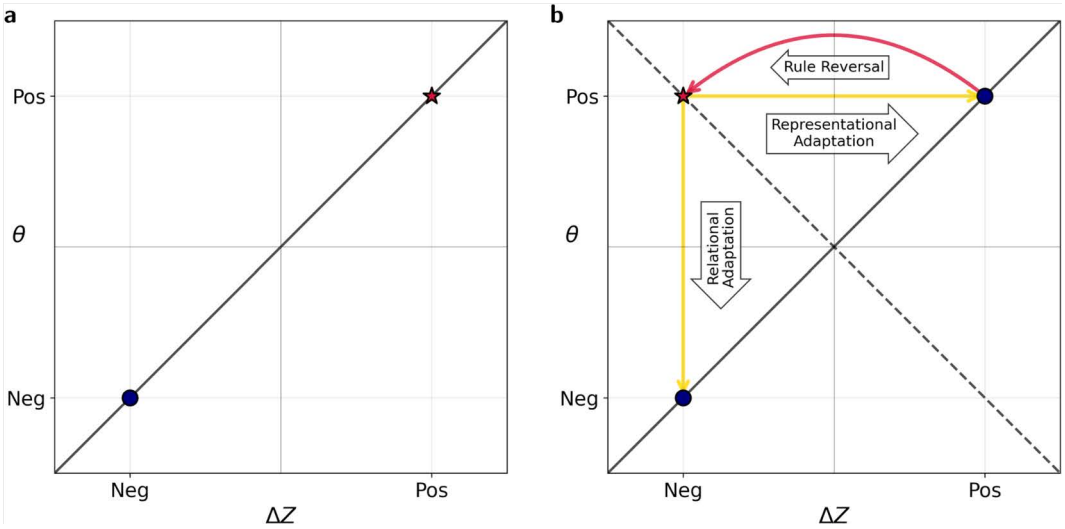

Fig. 3. Illustration of the two adaptation pathways. (**a**) The two equivalent solutions for the relational task. One solution is to encode the size of the objects and expect that the size increases (positive ΔZ and θ, marked by a star). In the other solution, the smallness decreases (negative ΔZ and θ). The diagonal line all solutions in which $\Delta Z = \theta$. (**b**) Adaptation pathways following rule reversal. Without loss of generality, we consider an agent that has learned the positive solution. After learning the initial rule, the rule is reversed, flipping the sign of ΔZ, making it inconsistent with positive θ (marked by a star). There are two adaptation pathways: *relational adaptation*—maintaining the sign of ΔZ but reversing the sign of the expectation θ; or *representational adaptation* - changing the sign of ΔZ to encode the smallness of objects, while maintaining the expectation that the relevant feature increases ($\theta > 0$). The dashed diagonal line represent expectations that are opposite to the observations ($\theta = -\Delta Z$).

ΔZ while the relational module θ changes its sign. Both can lead to exactly the same level of performance. We hypothesized that magnitude of α would determine which of the two adaptation pathways would be taken by the ANNs.

Going back to the ANNs, we first considered two violation magnitudes: a larger violation ($\alpha = 0.8$, Fig. 4a left) and a smaller violation ($\alpha = 0.2$, Fig. 4a right). We measured the ANNs performance as a function of examples before and after reversing the rule (Fig. 4b). In both the larger and the smaller violation conditions, performance levels just before the sign reversal (image pair 160) were almost perfect. The first image pairs immediately following the reversal were almost always incorrectly ordered by the ANNs (performance level close to 0). With examples, however, the networks adapted to the new rule, achieving almost perfect performance after additional 160 image pairs. Notably, learning was slower for the more difficult task associated with the smaller value of α. Also, adaptation to the reversal was slower than initial learning, a phenomenon that has been previously reported in human learning[31].

To dissect the roles of the two modules in this reversal adaptation, Fig. 4c depicts the values of ΔZ (red) and θ (blue) in representative networks adapting to the large and small violations. Before reversal, both networks converged to comparable values of ΔZ and θ. Immediately after the rule reversal, ΔZ flipped. This is because the image order was reversed – the sizes of the shapes decreased, rather than increased between the two images – and the representation module reflected it. The value of θ, however, remained unchanged. This is because the network "expected" the sizes of the shapes to increase rather than decrease. With training, the system resolves this violation.

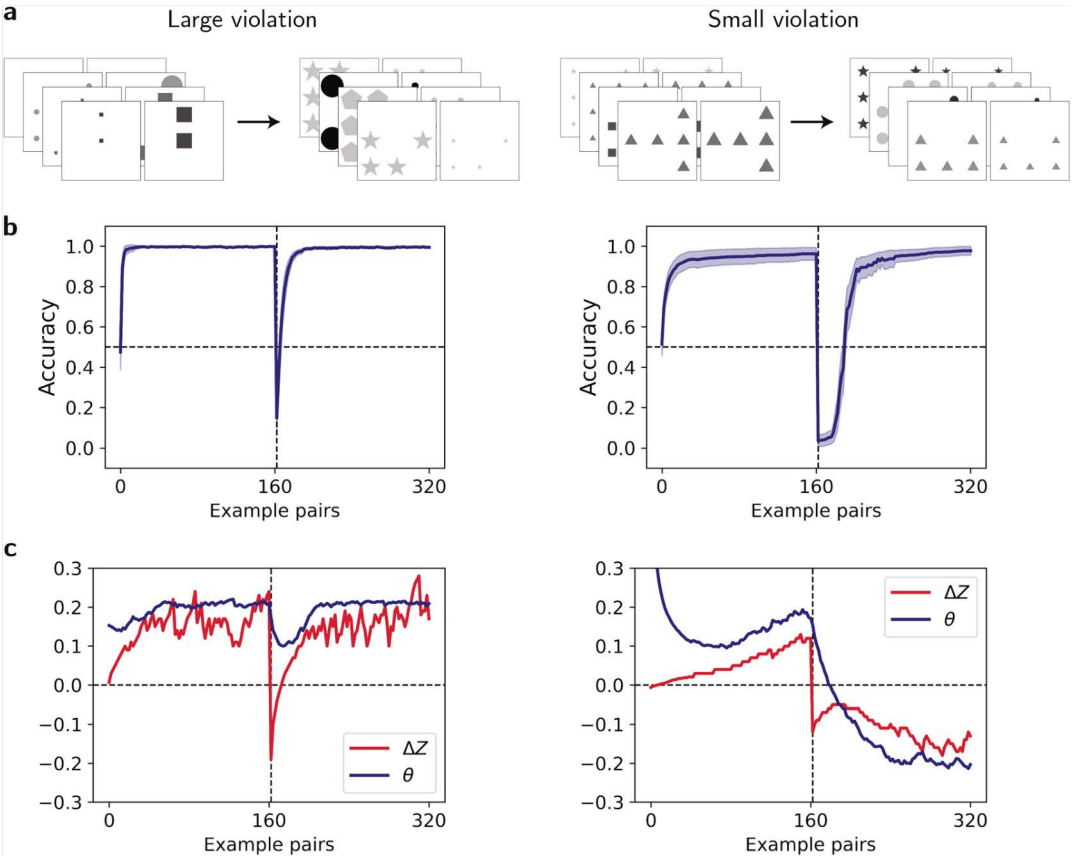

Fig. 4. Dual adaptation pathways following rule reversal. Demonstration of the two adaptation pathways, depending on the violation magnitude, when the rule changes from α to $-\alpha$. (**a**) Example image pairs before and after the rule reversal for a large violation ($\alpha = 0.8$, left) and a small violation ($\alpha = 0.2$, right). (**b**) Performance across 100 networks as a function of examples (shaded regions: 95% CI). (**c**) Representative demonstration of the evolution of ΔZ (red) and θ (blue) for each violation magnitude. Initially, both networks converge to the positive solution ($\theta, \Delta Z > 0$). After reversal, ΔZ flips immediately due to the change in the image order. Optimization drives ΔZ and θ towards each other, approaching $\Delta Z = 0$ and $\theta = 0$. For large violations (left), the representation module "wins" and adaptation restores ΔZ while keeping θ unchanged. For small violations (right), adaptation eventually occurs via the relational module, flipping the sign of θ to match the negative ΔZ.

In our simulations, we found that when the violation was large (Fig. 4c left), the representational module Z adapted so that ΔZ returned to its pre-reversal values. By contrast, when the violation was small (Fig. 4c right), ΔZ remained negative and the violation was resolved by a change in the sign of the relational module θ.

Are these results representative? We simulated reversal adaptation for different values of α, each time simulating 100 randomly-initialized networks, and measuring the fraction of times in which the adaptation involved a change in the sign of θ, which indicates that the relational module dominates the adaptation. In line with the examples of Fig. 4c, the smaller α, the larger was the fraction of networks in which the relational module flipped its sign in response to the rule reversal (Fig. 5a). The transition between the two adaptation pathways, an inflection point marked by $\bar{\alpha}$, was at 0.34 ± 0.02. These results demonstrate that large violations ($\alpha > \bar{\alpha}$) can inhibit adjustments to relational expectations, leading to adaptation in object representations instead.

The results of this section are reminiscent of the surprising part of the experimentally-observed inverted U-shaped dependence of relational adaptation to the size of the violation discussed in the Introduction. When the violation is modest (α is small), the relational expectation adapts: θ changes its sign when the order of images is reversed. By contrast, it remains unchanged when the violation is large. Instead, the violation is resolved by the network changing its representation. This behavior does not require any explicit immunization mechanism. Rather, it naturally emerges from the dynamics of learning.

General rule reversals

In the analysis above, we considered the special case of symmetric rule reversal, in which $\alpha_1 = \alpha_2$. Now, we study the specific contribution of these two parameters to the choice of an adaptation pathway. To that end, we studied the adaptation of ANNs to a different pairs of α_1 and α_2, as depicted in Fig. 5b. For each pair (α_1, α_2) we measured the fraction of networks in which adaptation was associated with a change in the sign of θ (color coded). To better visualize the transition point (which was denoted by $\bar{\alpha}$ in the case of $\alpha_1 = \alpha_2$), the symbol (square vs. triangle) denotes whether this fraction was smaller or larger than 50%. These simulations show that the adaptation pathway depends both on α_1 and α_2. The larger α_1 and the larger α_2, the more likely it is that the representational module will change its sign. However, a large α_1 can compensate for by a small α_2 and vice versa. The boundary between the two adaptation pathways resembles a hyperbolic curve. This is not a coincidence. Below we prove, in a simplified model that in the limit of weak regularization, the boundary is, indeed, a hyperbolic function in the $\alpha_1 \times \alpha_2$ plane.

Next we study the implications regarding our ability to use shaping to facilitate or inhibit adaptation of relational expectations.

Shaping adaptation through an intermediate rule

Understanding how individuals adapt to violations of their expectations is important for clinical psychology, as expectation persistence and change are central to mental health interventions. Clinical research has shown that maladaptive expectations contribute to disorders such as anxiety and depression, where individuals often maintain dysfunctional expectations even in the face of disconfirming evidence[23]. Effective psychological treatments leverage expectation violations to induce cognitive and behavioral change. Yet, while moderate expectation violations are most effective in altering beliefs, extreme violations risk reinforcing rigid mental models, preventing adaptation[10,23]. In this section we show that adding an intermediate rule in the reversal task can alter the adaptation pathway, thereby steering the adaptation process toward a preferred adaptation strategy.

We compared adaptation in reversal task in which the rule is reversed in one step, as before ($\alpha \to -\alpha$) to adaptation when our agents also adapt to an intermediate step ($\alpha \to \beta \to -\alpha$). We hypothesized that the value of $\bar{\alpha}$ would inversely depend on the magnitude of the intermediate step: larger intermediate rules would lower $\bar{\alpha}$,

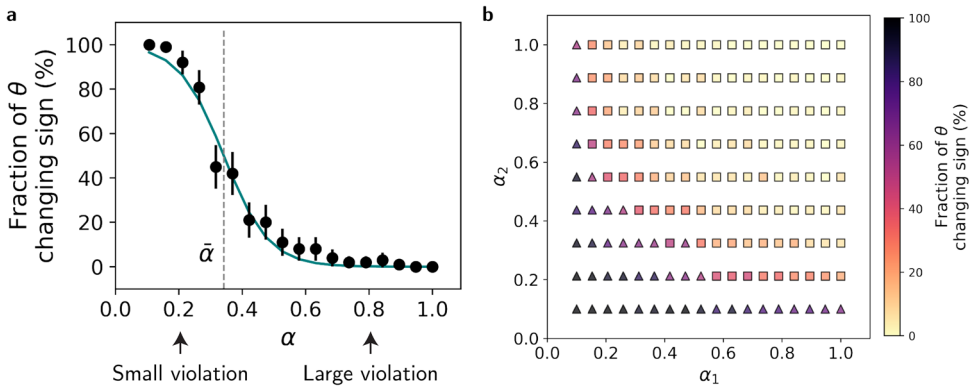

Fig. 5. Adaptation pathway dependence on violation magnitude. (**a**) Fraction of networks ($n = 100$) adapting via relational module (θ sign change) for rule reversal $\alpha \to -\alpha$, fitted with logistic function (green) to estimate inflection point $\bar{\alpha}$. Arrows mark the α values from Fig. 4. The logistic function fit has two parameters, the inflection point, $\bar{\alpha} \sim 0.34$, and the slope parameter, ~ 14.10 (see Methods). Error bars: 95% CI. (**b**) Adaptation pathways for general rule changes $\alpha_1 \to -\alpha_2$. Color indicates fraction (out of $n = 50$) adapting via relational module (θ). Squares: representational module (Z) dominant (ratio $< 50\%$); triangles: relational module dominant (ratio $\geq 50\%$).

making relational adaptation less likely, whereas smaller intermediate steps would increase $\bar{\alpha}$, favoring relational adaptation. The intuition behind this hypothesis is that the relational module can change its sign only when the rule reverses. Therefore, when $\beta > 0$, if $\beta > \alpha$ then the intermediate step enhances the violation (increases α_1 of Fig. 5b) and therefore decreases the probability of a relational adaptation. By contrast, $\beta < \alpha$ decreases the magnitude of the violation and therefore, increases the probability of a relational adaptation. The effect of a negative β is similar. This time, the focus is on the transition ($\alpha \rightarrow \beta$), where $-\beta$ takes the role of α_2 of Figure 5b.

To test this hypothesis, we trained ANNs using the $\alpha \rightarrow \beta \rightarrow -\alpha$ paradigm, using 160 image pairs for each rule (total $160 \times 3 = 480$ image pairs). For each pair (α, β) we trained 50 ANNs and computed the fraction of networks in which the sign of θ changed from image pair 160 (after the network learned the α rule) to image pair 480 (after the network learned the $-\alpha$ rule). The results are depicted in Fig. 6a. Then, for each value of α and β, we computed $\bar{\alpha}$ by fitting a logistic function to the computed fraction as a function of α, as in Fig. 5a. The values of $\bar{\alpha}$ as a function of β are presented in Fig. 6b. Indeed, for large values of $|\beta|$, $\bar{\alpha}$ was smaller than that computed in the $\alpha \rightarrow -\alpha$ paradigm (dashed horizontal line, computed in Fig. 5a), while small values of $|\beta|$ increased $\bar{\alpha}$. These findings demonstrate that adaptation pathways can be influenced by the sequence of changes between them. By strategically introducing an intermediate step, we can shift the boundary between relational and representational adaptation, effectively shaping the learning process. Specifically, small values of β increase the probability that the relational network would adapt.

Simplified model analysis

Setting up the model and loss function

To gain an analytical understanding as to why small and large violations lead to qualitatively different adaptation mechanisms, and the extent to which these results depend on the particularities of the model that we studied, we considered adaptation to violation in a simplified model, in which the pair of images was replaced by a pair of scalars, denoting the predictive features in the pair of images. That is, $\mathrm{x} = x$ and $\mathrm{x}' = x'$ and their relation is their difference, $x' - x = \alpha$.

Now, that the feature is explicitly provided to the agent, we model the representational module using a single-weight linear encoder, $Z_{\mathrm{w}}(x) = wx$. Under this formulation, the difference in the representations of the pair of stimuli is simply $\Delta Z = wx' - wx = w\alpha$. The corresponding loss function then becomes

$$\tilde{\mathscr{L}}(w, \theta) = (w\alpha - \theta)^2 + \lambda \left((w\alpha)^2 + \theta^2 - r^2\right)^2 . \tag{3}$$

This loss function captures two key components: (1) the squared error between the representational difference $w\alpha$ and the relational expectation θ, and (2) a regularization term ensuring that solutions remain on a ring.

SGD dynamics and differential equations

The advantage of this simplified formulation is that we can now use mathematical techniques, borrowed from the field of non-linear dynamics, to analytically characterize the adaptation dynamics[32]. In the limit of an infinitesimally small learning rate, the SGD dynamics that acts to minimize the loss function can be expressed as a set of differential equations governing the evolution in time (a proxy of example pairs) of w and θ[33]:

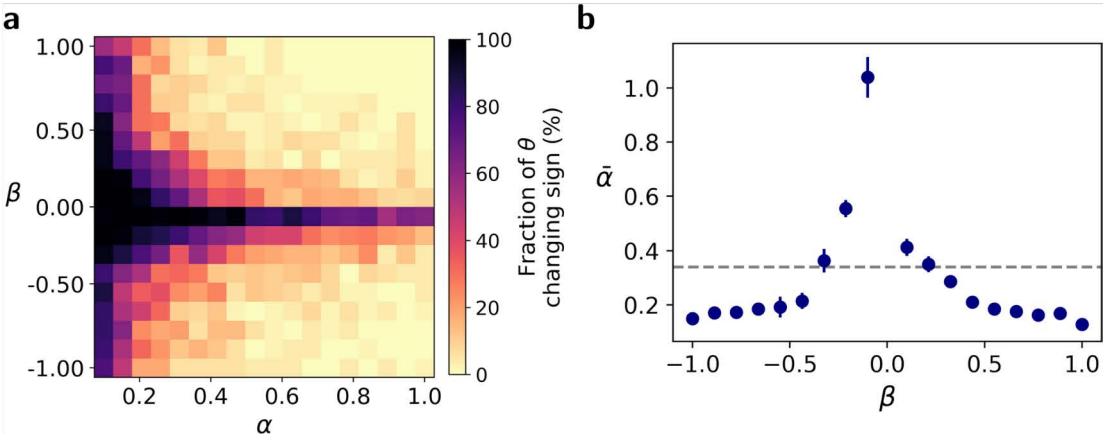

Fig. 6. Intermediate learning step: influence on adaptation pathways. The threshold between relational and representational adaptation ($\bar{\alpha}$) is modulated by the magnitude of an intermediate step β. (**a**) The fraction of networks that adapt θ for training with an intermediate step $\alpha \rightarrow \beta \rightarrow -\alpha$ for various pairs of (α, β). (**b**) Large values of $|\beta|$ lower $\bar{\alpha}$, promoting representational adaptation, while small $|\beta|$ increases $\bar{\alpha}$, favoring relational adaptation. The gray dashed line represents $\bar{\alpha}$ in the absence of an intermediate step, highlighting how structured transitions can shape the adaptation process.

$$\begin{aligned}\dot{w} &= -\alpha\left(w\alpha-\theta\right)-2\lambda\alpha^2\left(\left(w\alpha\right)^2+\theta^2-r^2\right)w\\ \dot{\theta} &= \left(w\alpha-\theta\right)-2\lambda\left(\left(w\alpha\right)^2+\theta^2-r^2\right)\theta.\end{aligned} \tag{4}$$

The evolution of w and θ over time represents the adaptation of these parameters through learning with successive example pairs. Thus, by analyzing the temporal dynamics of w and θ, we can gain insight into the adaptation pathway of the system. Instead of directly examining the dynamics of w, it is more instructive to focus on the dynamics of the output of the representational module, $\Delta Z = w\alpha$. Substituting this into the differential equations, we obtain the following system of equations:

$$\begin{aligned}\frac{1}{\alpha^2}\dot{\Delta Z} &= \left(\theta-\Delta Z\right)-2\lambda\left(\Delta Z^2+\theta^2-r^2\right)\Delta Z\\ \dot{\theta} &= \left(\Delta Z-\theta\right)-2\lambda\left(\Delta Z^2+\theta^2-r^2\right)\theta.\end{aligned} \tag{5}$$

Recall that when studying the adaptation of the ANN to the image pairs, we identified two solutions that minimize the loss: $\Delta Z = \theta = \pm r/\sqrt{2}$. The two solutions are associated with the same level of performance, they differ in how the system encodes the relationship: when $\Delta Z = \theta = r/\sqrt{2}$, the predictive feature increases, whereas when $\Delta Z = \theta = -r/\sqrt{2}$, it decreases, as discussed in Fig. 2.

In gradient-based systems, learning dynamics converge to a stable fixed point where $\dot{\Delta Z} = \dot{\theta} = 0$. In the Methods section we prove that $\Delta Z = \theta = \pm r/\sqrt{2}$ are the only stable fixed points of the dynamics, Eq. (5). Thus in general, the dynamics would converge to either the positive or the negative fixed point.

Adaptation pathways following rule reversal

To investigate the dynamics following rule reversal, we consider a system that has converged to the positive fixed point $\Delta Z = \theta = r/\sqrt{2}$. Following the reversal of the rule, the sign of the representational difference ΔZ flips to $-r/\sqrt{2}$. As a result, the value of $\left(\Delta Z-\theta\right)^2$ in the loss becomes non-zero, reflecting the violation of the expected relationship.

At this point, the gradient dynamics would resolve the violation by either driving the system back to the positive fixed point $\Delta Z = \theta = r/\sqrt{2}$, adapting the representational module, or to the negative fixed point $\Delta Z = \theta = -r/\sqrt{2}$, which adapts the relational expectation (as illustrated in Fig. 3).

To understand how the magnitude of α affects this adaptation pathway, we first considered the dynamics of a weakly regularized system, where $\lambda \ll 1$. In this case, the dynamics first minimize the unregularized part of the loss, $\left(\Delta Z-\theta\right)^2$, driving the system to $\Delta Z = \theta$, and then the regularization kicks in to set the system on the ring $\Delta Z^2 = \theta^2 = r^2/2$. We show below that this sequential adaptation pattern – first aligning ΔZ and θ, then enforcing the regularization constraint – provides a key insight into the behavior of the more complex ANN.

Without regularization, the dynamical equations simplify to

$$\begin{aligned}\dot{\Delta Z} &= -\alpha^2\left(\Delta Z-\theta\right)\\ \dot{\theta} &= \left(\Delta Z-\theta\right).\end{aligned} \tag{6}$$

When $\alpha > 1$, the dynamics of ΔZ is faster than that of θ (because of the α^2 prefactor). Consequently, adaptation is expected to be dominated by a change in the sign of ΔZ. By contrast, when $\alpha < 1$, adaptation is expected to rely on a change in θ. This can be shown more formally. In the Methods section we show that the dynamics converge to

$$\Delta Z = \theta = \frac{\alpha^2\theta(0)+\Delta Z(0)}{\alpha^2+1}, \tag{7}$$

where $\Delta Z(0)$ and $\theta(0)$ denote the values of ΔZ and θ before the reversal.

Substituting the initial conditions $\Delta Z = -r/\sqrt{2}$ and $\theta = r/\sqrt{2}$, we find that after reversal, θ will change its sign if and only if $\alpha < 1$.

Considering now the full model (Eq. (5)), we identify the following symmetry: Immediately after reversal $-\Delta Z\left(0\right) = \theta\left(0\right)$. Considering the equations for $-\dot{\Delta Z}$ and $\dot{\theta}$, they are also symmetric to swapping $-\Delta Z$ and θ, as long as α^2 is replaced by $1/\alpha^2$. Consequently, if for a particular value of α', ΔZ would change its sign in the reversal protocol, it would be θ which changes its sign if $1/\alpha'$ is used, indicating that also in the full model, $\alpha = 1$ is the transition point between the two, qualitatively-different modes of adaptation.

Figure 7 depicts these dynamics in the reversal paradigm using a phase portrait for two values of α. When α is large (left) adaptation is dominated by a change in ΔZ. The opposite, a dynamics that is dominated by a change of θ manifests when α is small (right). Notably, in simulating the model, we used a relatively weak regularization, $\lambda = 0.1$. As a result, initially, the dynamics drive the system to the line $\Delta Z = \theta$, minimizing the unregularized term by either changing the sign of ΔZ or θ, depending on the size of the violation. Then, when $\Delta Z \approx \theta$, the regularization term pushes the system towards one of the two fixed points, where $\Delta Z^2 = \theta^2 = r^2/2$.

The more general case of $\alpha_1 \to -\alpha_2$

We also studied the system's dynamics in the more general adaptation of $\alpha_1 \to -\alpha_2$. In this case, the initial state $\Delta Z(0)$ would generally not be $-r/\sqrt{2}$ like in the symmetric rule reversal. Instead, it would be $\Delta Z(0) = -\frac{r}{\sqrt{2}}\frac{\alpha_2}{\alpha_1}$ (see Methods). This is while $\theta(0)$ remains the same. Therefore, the ratio α_2/α_1 adjusts the distance of the system from the ordinates (y-axis), whose crossing corresponds to changing the sign of ΔZ. The

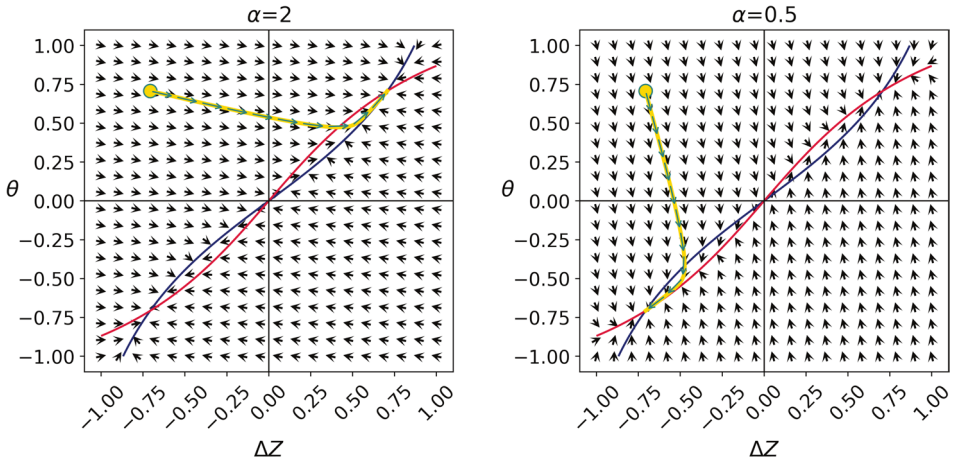

Fig. 7. The adaptation dynamics in the $\Delta Z \times \theta$ plane. Visualization of the two adaptation pathways as two trajectories in the simplified model's internal state represented by ΔZ and θ. The arrows represent the "downhill" direction for learning: the opposite direction of the loss gradient with respect to ΔZ and θ. The blue and red nullclines represent the loci where $\dot{\Delta Z} = 0$ and $\dot{\theta} = 0$, respectively. Their intersections correspond to fixed points in which learning halts. They intersect at the stable fixed points at $\Delta Z = \theta = \pm r/\sqrt{2}$, and at the unstable fixed point $\Delta Z = \theta = 0$. Before reversing the rule, the network start with positive ΔZ and θ. Following the reversal of the rule, ΔZ flips its sign (yellow circles). Adaptation when α is large (left panel) leads to reversing the sign of ΔZ, returning to the positive fixed point, whereas when α is small (right panel) it is θ that changes its sign to match the negative ΔZ.

larger α_2/α_1, the longer is the path required for a change in the sign of ΔZ (Fig. 8a). Another modification to the dynamics is that α in Eq. (6), which sets the gradient of ΔZ compared with those of θ, is replaced by α_2. Therefore, the value of α_2 determines the direction of the gradient, or the relative speed of adaptation of the two modules: The larger α_2, the more the gradient is aligned with ΔZ compared to θ (more horizontal) (Fig. 8b).

Together, we can view the dynamics as a "race" between the two adaptation pathways. The ratio $\frac{\alpha_2}{\alpha_1}$ determines the starting points of the two adaptation pathways, and hence the relative distance to the finish line. The larger this ratio, the smaller is the relative distance required for adaptation by the relation module. The larger α_2, the faster is the dynamics of ΔZ, compared to that of θ. Together, the relation module is more likely to adapt when the ratio $\frac{\alpha_2}{\alpha_1}$ is large and when α_2 is small (Fig. 8c).

As discussed above, when regularization is weak, we can analytically predict the "winner" from the ratio between the relative distances of ΔZ and θ to the finish line, α_2/α_1, and the ratio between their speeds, α_2^2. If $\alpha_1\alpha_2 > 1$, then the representational module would adapt; when $\alpha_1\alpha_2 < 1$ the relational module adapts (see Methods). This prediction is a good fit for the regularized simplified model of Fig. 8c, in which $\lambda = 0.1$. This prediction also fits the adaptation pattern of the ANN in the more complex images task, when adjusted such that $\alpha_1\alpha_2 > \bar{\alpha}^2$ is the condition for adapting the representational module (Fig. 5b and S5). Overall, our results suggest that the multiplication of $\alpha_1\alpha_2$, which precisely determines the adaptation pathway in the limit of weak regularization, is a good fit to to how α_1 and α_2 affect the adaptation pathway in this model.

Discussion

In this study, we explored how artificial neural networks (ANNs) resolve relational inconsistencies when faced with violations of expected relationships. Our findings reveal two distinct adaptation pathways that naturally emerge from gradient-based learning dynamics: when violations are small, networks primarily adjust their relational expectations, whereas extreme violations lead to modifications in object representations, preserving the initial relational expectation. This dichotomy can account for the experimentally observed inverted U-shaped dependence of expectation adaptation on violation magnitude, where moderate expectation violations lead to learning, but extreme violations often result in resistance to change.

The key component of our theoretical finding is that a violation of expectation can be resolved in more than just one way. Therefore, in case of relational inconsistency, whether or not the relational component will eventually adapt depends on whether adaptation of this component is "sufficiently fast", compared with the alternatives route for adaptation—the adaptations of the representational module. In response to the violation of expectation, a "race" between the two adaptation pathways begins. The ratio between the new rule and the original rule determines the starting points of the two adaptation pathways, and hence the distance to the finish line. The larger this ratio, the smaller is the distance required for adaptation by the relation module, relative to the representational module. Moreover, despite the fact that we used the same learning rule for the representational and relation modules, SGD, which is characterized by a single learning rate, the "speed" of adaptation of the two modules is not equal. The larger the magnitude of the "new" rule (α_2) the slower the relation module adapts relative to the representational module.

Our results indicate that the product of the initial and final rules $\alpha_1\alpha_2$ is a crucial parameter in determining the adaptation pathway. When this product is large, the representational module adapts and when it is small the

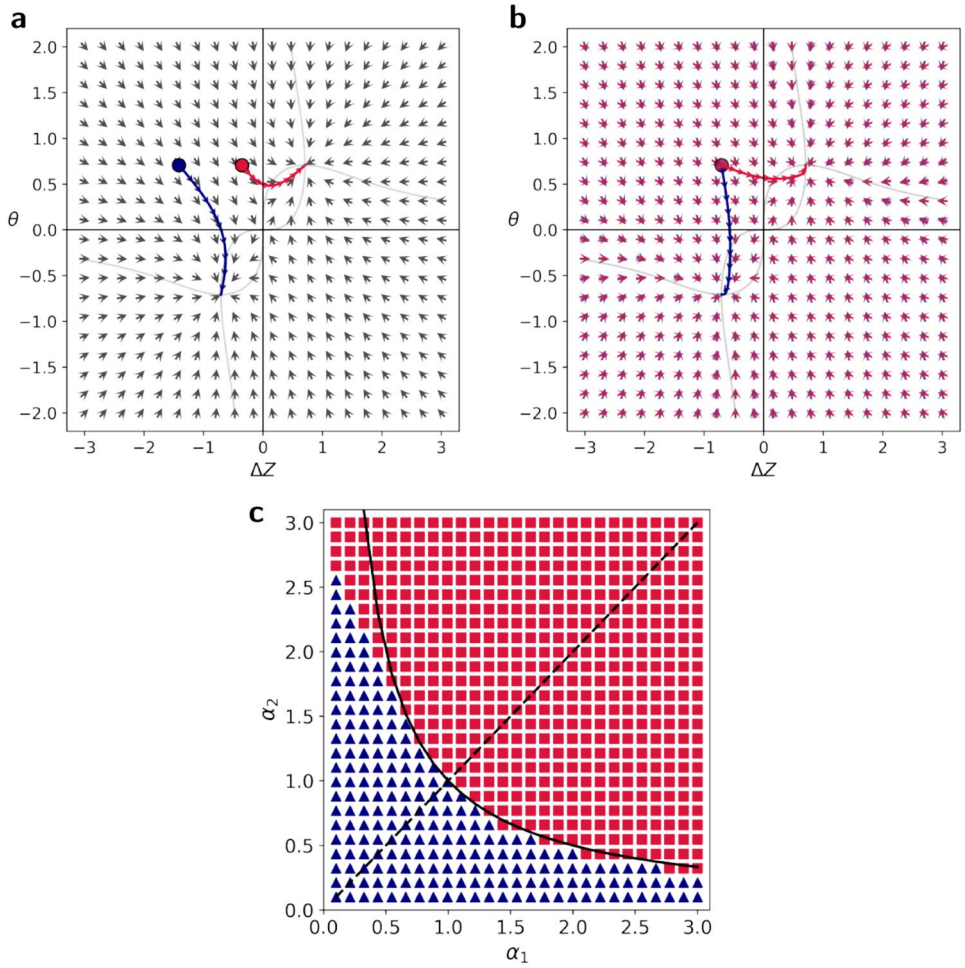

Fig. 8. How initial and new rule strengths determine the adaptation pathway. This figure explores how the network adapts when the rule changes from an initial strength (α_1) to an opposite rule of a different strength ($-\alpha_2$). The choice of adaptation pathway can be understood as a "race" between the representational module (ΔZ) and the relational module (θ). Two factors determine the winner. First, as illustrated in panel (**a**), the ratio of the new rule's strength to the old one (α_2/α_1) determines the "starting position" for the race (here, α_2 is fixed at 1, while α_1 is 2 for the blue trajectory and 0.5 for the red). Second, as shown in panel (**b**), the strength of the new rule (α_2) influences the relative "speed" of adaptation, changing the direction of the learning process (we used $\alpha_2 = 0.5$ for the blue trajectory and $\alpha_2 = 2$ for the red, while fixing the starting points by setting $\alpha_1 = \alpha_2$). (**c**) The final race outcome for various combinations of α_1 and α_2. The boundary between the two pathways is determined by the combined strength of the rules. Specifically, the line $\alpha_1\alpha_2 = 1$ (black line) is a good fit for this boundary.

relational module adapts. In the real world, the rule sizes correspond to an objective difference between objects. Extrapolating our results to the scientist example, if particles A are expected to be larger than particles B by α_1, and the novel observation is that they are smaller by α_2, then the observation would be dismissed if $\alpha_1\alpha_2$ would be larger than some threshold that depends on the scale α.

In our study, the core relationships we investigated were defined by changes in specific predictive features: shapes size in the main manuscript, and grayscale color and number of shapes in the Supplementary Materials. All other features varied randomly between image pairs and were categorized as irrelevant features. Our prior research has demonstrated that the quantity of these irrelevant features directly correlates with the difficulty of similar tasks[34]. Consequently, we integrated these irrelevant features into our experimental design to slightly increase task complexity, aiming for a more realistic scenario. We hypothesize that as the networks achieve high performance, they learn to effectively disregard these irrelevant features. Crucially, when the rule is reversed in our experiments, only the predictive feature's rule is altered. Therefore, we do not anticipate that the number or identity of these irrelevant features will influence our reported results.

Reconciling prior beliefs with conflicting evidence is commonly framed within a Bayesian framework[17,18]. To connect our findings to this approach, consider a scenario with two competing hypotheses: one suggesting the predictive feature (e.g., size) increases, and the other suggesting it decreases. When an agent begins with a stronger belief in the "increase" hypothesis and then encounters evidence challenging this belief, Bayesian updating provides a formal way to revise these beliefs. This revision depends on two key factors: the initial

confidence in each hypothesis (prior beliefs) and how well each hypothesis explains the new observation (likelihood of the evidence). The Bayesian framework provides a clear decision threshold – the point at which belief should shift from one hypothesis to the other. For the Bayesian model to account for the inverted U-shaped dependence of relational adaptation to the size of the violation, observed experimentally as well as in our ANN model, we need a model in which large α_1 and α_2 support belief consistency. If we interpret the magnitudes of α_1 and α_2 as measures of the strengths of evidence they provide, it is easy to see why a strong initial evidence, in the form of a large α_1 would do that. However, it is challenging to interpret belief consistency when α_2 is large. This is because stronger contradictory evidence (larger α_2) is expected to increase the likelihood of hypothesis revision rather than decrease it. It is possible to account for the inverted U-shaped dependence in a Bayesian framework if we add a measure of confidence in the observations, and posit that the confidence in the first set of observations (associated with α_1) is larger than that of the second set of observations (associated with α_2).

Traditional psychological perspectives often interpret belief persistence in the face of contradictory evidence as a cognitive bias[35,36]. Our findings, however, suggest an alternative view: rather than mere bias, this persistence may reflect an adaptive learning strategy that stabilizes relational expectations by integrating violations through representational adjustments. A more recent theory, *ViolEx*[20–23], posits that when violations are extreme, immunization mechanisms act to devalue or reframe the new information. Our findings are consistent with ViolEx, but suggest that the immunization mechanism naturally emerges from the same learning process that drives expectation updates, with the outcome determined by which adaptation pathway "wins the race."

In the paper, we found that intermediate adaptation steps can be used to reshape the adaptation pathway. Specifically, a small intermediate step promoted the adaptation of the relational expectation. This result aligns with findings in cognitive-behavioral therapy and education, where gradual exposure to conflicting information enhances adaptive outcomes. For example, therapies designed to modify dysfunctional expectations, such as exposure therapy for anxiety disorders, benefit from structured interventions that introduce moderate violations instead of extreme ones[37,38]. Similarly, in education, conceptual change is more effective when scaffolded gradually rather than introduced through abrupt contradiction[19,39,40].

The two adaptation pathways discussed in this work are categorical (representational vs. relational). But notably, there are multiple ways to adapt, even within each category. Specifically, the representational module is characterized by a large number of parameters, more parameters than examples. Therefore, there can be multiple combinations of parameters that, for the same set of examples, yield the same representational adaptation. The relational module in our model is rather simple, but in a more general model we expect a similar multiplicity of possible adaptation pathways within the relational pathway. Indeed, different adaptation pathways within a module have been a subject of research in the cognitive sciences. For example, an inconsistency between a person's unhealthy habit of smoking and the warning against the harmful effects of smoking presented on a tobacco package can be resolved in several ways (that do not change the habit): The smoker can posit that benefits associated with appetite suppression outweigh the cancer-related health risks, or alternatively, question the research that links smoking to increased mortality[41]. Both solutions are consistent with a change to the representation of smoking. A major limitation of our model is that it is not informative about the determinants of *within-category* adaptation pathways.

Another limitation of our model is that it modeled passive agents who get a sequence of pairs of inputs and are required to learn their relationships. However, humans actively engage in the learning process by performing various comparison and manipulation patterns of stimuli[29,30]. This active participation is useful for establishing relational behavior. The active manipulation of stimuli can be interpreted as a form of representational adaptation, where the learner actively re-frames or re-processes the sensory input by physically interacting with it, allowing the learner to adjust their internal representations of objects to resolve inconsistencies and make new information compatible with existing or emerging relational rules.

By examining the intrinsic learning dynamics of neural systems, our work provides insights into the mechanisms that govern adaptation to inconsistencies. The competition between representational and relational adaptation pathways naturally produces the non-monotonic patterns of belief updating observed in human cognition. These findings have implications for understanding learning processes across cognitive, educational, and therapeutic contexts. Further work is needed to explore the nuances of adaptation within the representational and relational categories, as well as to experimentally test the model's predictions.

Methods

Code availability

A PyTorch[42] code that generates the results and figures of this paper is available at:https://github.com/Tomer-Barak/relational_expectation_violations.

Order discrimination task

The order discrimination task was designed to assess the ability of ANNs to determine the correct order of image pairs based on a specific feature. As written in the main text, each image in the pair depicted shapes arranged on a 3×3 grid and was characterized by five features: grayscale color, number of shapes, size, grid arrangement, and shape type. The images were 224×224 pixels in size and grayscale (they consisted of 3 channels with identical values). The size of the shapes was defined as the diameter of the circle enclosing them.

The "correct" order in the task was determined by the identity of the relevant feature (color, size, or number), termed the *predictive feature*, and whether this feature increased or decreased from left to right. To construct the training set, the predictive feature values were randomly selected from a uniform distribution over possible values for the left image. The corresponding values for the right image were then calculated by applying the rule parameter α to the left image's values. Non-predictive features were randomly selected from a uniform distribution for each image pair, remaining constant within the pair.

For each α and tested network, we constructed a training set that consisted of 160 image pairs that demonstrated that rule. To evaluate the performance of an ANN in this task, we tested its ability to classify the correct order of 32 novel image pairs. In Figs. 2a-b and 4b, we averaged the classification accuracy of 100 ANNs and estimated the confidence interval based on the standard error.

In the main text of this paper, we presented the results when the predictive feature is the size. Similar results were obtained when the predictive features were color or number, and these are presented in the Supplementary Information (Figs. S1-S4). For full implementation of the task, see Images.py in the paper's GitHub site:https://github.com/Tomer-Barak/relational_expectation_violations

The ANN

The representational module $Z_{\mathrm{w}}(\mathrm{x})$ consisted of three convolutional layers (number of filters: 16, 32, 32; kernel sizes: 2, 2, 3; strides: all 1; padding: all 1) followed by one fully-connected linear layer (taking a 2592-dimensional vector to a one-dimensional output). Three ReLU activation functions were applied after each convolutional layer, and two Max-Pool layers (kernels: 4 and 6, strides: all 1) were applied after the second and third convolutional (+ReLU) layers. The parameters of Z_{w} were randomly initialized using PyTorch's[42] default initialization (uniform distribution scaled by $1/\sqrt{N}$ where N is the number of the layer's input neurons).

Given a training set, we optimized the randomly initialized ANN's parameters to minimize the regularized loss function (2) with the vanilla SGD optimizer ($lr = 0.004$). For the regularization term, we used the hyperparameters $\lambda = 4$ and $r^2 = 0.1$. We used a batch size of 2 image pairs and applied 20 optimization steps per batch.

To assess the adaptation pathway of an ANN, we measured its parameter θ during training. To complement this measure, we also measured the average ΔZ of the ANN over 32 test image pairs from the same training set distribution (with the same α). A network that changed its sign of θ and kept the sign of ΔZ after rule reversal was classified as adapting its relational module. A network that kept the sign of θ while changing the sign of ΔZ has adapted its representational module. We excluded networks that kept or changed the signs of both θ and ΔZ together. These networks necessarily failed the task. The fraction of excluded networks was less than 1%: Fig. 5a: 3/1800. The fraction of networks that adapted their relational module (e.g., in Fig. 5a) was obtained by $\#\theta/\left(\#Z+\#\theta\right)$ where $\#Z$ is the number of networks that adapted ΔZ, and $\#\theta$ is the number of networks that adapted θ.

Calculating the inflection point in Fig. 5a

To calculate the inflection point $\bar{\alpha}$, we fitted a logistic function to the results of how many networks adapted their relational module as a function of α. Specifically, we fitted the two parameters c and d of the logistic function $\frac{1}{1+e^{c(\alpha-d)}}$. The inflection point was defined as $\bar{\alpha} = d$. The 95% CIs of $\bar{\alpha}$ correspond to $1.96 \cdot SE(d)$ where $SE(d)$ is the standard deviation error of the estimation of d using SciPy's[43] curve fitting function.

Simplified model: two attractive fixed points

Because this is a gradient system, the dynamics will necessarily converge to the (stable) fixed point(s) of the dynamics. To find the fixed point(s), we consider the two nullclines, $\dot{\Delta Z} = 0$ and $\dot{\theta} = 0$. From these equations we write,

$$\begin{aligned} |(\Delta Z - \theta)| &= 2\lambda \left|\left(\Delta Z^2 + \theta^2 - r^2\right)\right| |\Delta Z| \\ |(\Delta Z - \theta)| &= 2\lambda \left|\left(\Delta Z^2 + \theta^2 - r^2\right)\right| |\theta| \,. \end{aligned} \tag{8}$$

Subtracting the equations, we get that $\left|\left(\Delta Z^2 + \theta^2 - r^2\right)\right| \left(|\Delta Z| - |\theta|\right) = 0$. If $\left|\left(\Delta Z^2 + \theta^2 - r^2\right)\right| = 0$ then from Eq. (5), $\Delta Z = \theta$ at the fixed point. Therefore together, $|\Delta Z| = |\theta|$.

The nullclines are depicted in Fig. 7. The fixed points can be computed analytically by substituting $|\Delta Z| = |\theta|$ in the nullclines equations. We find that there is a trivial fixed point at $\Delta Z = r = 0$. A linear stability analysis reveals that this fixed point is unstable. Additionally, there are two additional fixed points $\Delta Z = \theta = \pm\frac{r}{\sqrt{2}}$. These fixed points satisfy both $\mathscr{L}$ and the regularization term. We will discuss their stability shortly. When the regularization term is large, $\lambda > \frac{1}{r^2}$, there are two additional fixed points, $\Delta Z = -\theta = \pm\sqrt{\frac{r^2 - \frac{1}{\lambda}}{2}}$, but a linear stability analysis reveals that they are unstable. Because the dynamics is driven by a gradient of a loss function, then it necessarily converges to a fixed point. Because the $\Delta Z = \theta = \pm\frac{r}{\sqrt{2}}$ are the only non-unstable fixed points, they are necessarily the only attractors of the dynamics.

Weakly regularized simplified model: exact solution

To understand how the magnitude of α affects this adaptation pathway, it is useful to consider the dynamics of a weakly regularized system, where $\lambda \ll 1$. In this case, the dynamics first minimize the unregularized part of the loss, $(\Delta Z - \theta)^2$, driving the system to $\Delta Z = \theta$, and then the regularization kicks in to set the system on the ring $\Delta Z^2 = \theta^2 = r^2/2$.

Without regularization, the dynamical equations simplify to

$$\begin{aligned} \dot{\Delta Z} &= -\alpha^2 \left(\Delta Z - \theta\right) \\ \dot{\theta} &= \left(\Delta Z - \theta\right) . \end{aligned} \tag{9}$$

These two equations are linearly dependent, implying that the unregularized system converges to a point on a line attractor that depends on its initial state. Specifically, due to the equations being linearly dependent, the value $\Delta Z + \alpha^2\theta$ is conserved during optimization and its value depends on the initial state $\Delta Z(0) + \alpha^2\theta(0)$. This is true also at the fixed points, where $\Delta Z^* = \theta^*$. Plugging the fixed point solution to the conservation law provides the exact point the system would reach on the line attractor $\Delta Z = \theta$:

$$\Delta Z = \theta = \frac{\alpha^2\theta(0) + \Delta Z(0)}{\alpha^2 + 1}, \tag{10}$$

In the rule reversal case, assuming that the system starts from the positive fixed point, the initial values are $\Delta Z(0) = -r/\sqrt{2}$ and $\theta = r/\sqrt{2}$. Substituting this initial state, we find that the unregularized system is driven to the following point on the line attractor

$$\Delta Z^* = \theta^* = \frac{r}{\sqrt{2}}\frac{\alpha^2 - 1}{\alpha^2 + 1}. \tag{11}$$

When approaching the the line attractor, $(\Delta Z - \theta)^2$ becomes small, comparable to the regularization term in Eq. (5). Therefore, the regularization would then become more dominant and drive the system towards $\Delta Z^{*2} = \theta^2 = r^2/2$. The result we arrived at, Eq. (11), shows that whether α^2 is smaller or larger than 1 determines the sign of the fixed point. $\alpha^2 > 1$ corresponds to a fixed point where both ΔZ and θ are positive, keeping the original sign of θ, whereas $\alpha^2 < 1$ leads to a negative fixed point, changing the sign of θ. Therefore, the value of α^2 distinguished between the two adaptation pathways, and the inflection point is at $\bar{\alpha} = 1$. We verified this analysis by simulating the simplified model with weak regularization. For example, Fig. 7 demonstrates the dynamics when $\lambda = 0.1$ for a strong violation $\alpha > \bar{\alpha}$ and a weak violation $\alpha < \bar{\alpha}$. Initially, the dynamics drive the system to the line $\Delta Z = \theta$, minimizing the unregularized term by either changing the sign of ΔZ or θ, depending on the size of the violation. Then, when $\Delta Z \approx \theta$, the regularization term pushes the system towards one of the two fixed points, where $\Delta Z^{*2} = \theta^2 = r^2/2$.

In the more general case, where the rule changes from α_1 to $-\alpha_2$, the initial state of ΔZ before the adaption changes. To see this, remember that $\Delta Z = w\alpha$ where α is the current rule. At the first learning phase, assuming that the system converged to the positive fixed point, the value of the representational module's weight at the fixed point, w^*, is given by $\Delta Z^* = w^*\alpha_1 = r/\sqrt{2}$. When flipping the rule, w^* remains as it is, while ΔZ is now defined with α_2. Therefore, $\Delta Z(0) = -w^*\alpha_2 = -\frac{r}{\sqrt{2}}\frac{\alpha_2}{\alpha_1}$. The point on the line $\Delta Z = \theta$ where the system approaches depends on this initial state (Eq. (7)):

$$\Delta Z^* = \theta^* = \frac{r}{\sqrt{2}}\frac{\alpha_2^2 - \frac{\alpha_2}{\alpha_1}}{\alpha_2^2 + 1}. \tag{12}$$

This equation shows that whenever $\alpha_1\alpha_2 > 1$, the system adapts its representational module, while for $\alpha_1\alpha_2 < 1$ it would adapt its relational module. We verified this prediction in the weakly regularized simplified model in Figures 5b and S5.

Data availability

The data that was used for the order discrimination tasks was generated in real-time by an algorithm. The generating code is available on this project's GitHub page: https://github.com/Tomer-Barak/relational_expectation_violations.

Received: 25 March 2025; Accepted: 13 August 2025
Published online: 21 August 2025

References

1. Gentner, D. Structure-mapping: A theoretical framework for analogy. *Cognit. Sci.* **7**, 155–170. https://doi.org/10.1016/S0364-0213(83)80009-3 (1983).
2. Doumas, L. A. A., Hummel, J. E. & Sandhofer, C. M. A theory of the discovery and predication of relational concepts. *Psychol. Rev.* **115**, 1–43. https://doi.org/10.1037/0033-295X.115.1.1 (2008).
3. Lazareva, O. F., Miner, M., Wasserman, E. A. & Young, M. E. Multiple-pair training enhances transposition in pigeons. *Learning Behav.* **36**, 174–187. https://doi.org/10.3758/LB.36.3.174 (2008).
4. Lazareva, O. F. Relational learning in a context of transposition: A review. *J. Exp. Anal. Behav.* **97**, 231–248. https://doi.org/10.1901/jeab.2012.97-231 (2012).
5. Mansouri, F. A., Freedman, D. J. & Buckley, M. J. Emergence of abstract rules in the primate brain. *Nat. Rev. Neurosci.* **21**, 595–610. https://doi.org/10.1038/s41583-020-0364-5 (2020).
6. Holyoak, K. J. & Monti, M. M. Relational integration in the human brain: A review and synthesis. *J. Cognit. Neurosci.* **33**, 341–356. https://doi.org/10.1162/jocn_a_01619 (2021).
7. Miconi, T. & Kay, K. Neural mechanisms of relational learning and fast knowledge reassembly in plastic neural networks. *Nat. Neurosci.* **28**, 406–414. https://doi.org/10.1038/s41593-024-01852-8 (2025).
8. Filipowicz, A., Valadao, D., Anderson, B. & Danckert, J. Rejecting outliers: Surprising changes do not always improve belief updating. *Decision* **5**, 165–176. https://doi.org/10.1037/dec0000073 (2018).
9. Hird, E. J., Charalambous, C., El-Deredy, W., Jones, A. K. P. & Talmi, D. Boundary effects of expectation in human pain perception. *Sci. Rep.* **9**, 9443. https://doi.org/10.1038/s41598-019-45811-x (2019).
10. Spicer, S. G., Mitchell, C. J., Wills, A. J. & Jones, P. M. Theory protection in associative learning: Humans maintain certain beliefs in a manner that violates prediction error. *J. Exp. Psychol. Animal Learning Cognit.* **46**, 151–161. https://doi.org/10.1037/xan0000225 (2020).

11. Kube, T., Kirchner, L., Lemmer, G. & Glombiewski, J. A. How the discrepancy between prior expectations and new information influences expectation updating in depression-the greater, the better?. *Clin. Psychol. Sci.* **10**, 430–449. https://doi.org/10.1177/21677026211024644 (2022).
12. Pavlov, I. P. *Lectures on conditioned reflexes. Vol. II. Conditioned reflexes and psychiatry*. Lectures on conditioned reflexes. Vol. II. Conditioned reflexes and psychiatry (International Publishers, New York, NY, US, 1941). Pages: 199.
13. Leaton, R. N. & Borszcz, G. S. Potentiated startle: Its relation to freezing and shock intensity in rats. *J. Exp. Psychol. Animal Behav. Process.* **11**, 421 (1985).
14. Davis, M. & Astrachan, D. I. Conditioned fear and startle magnitude: Effects of different footshock or backshock intensities used in training. *J. Exp. Psychol. Animal Behav. Process.* **4**, 95 (1978).
15. Rescorla, R. A theory of Pavlovian conditioning : Variations in the effectiveness of reinforcement and nonreinforcement (1972).
16. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87. https://doi.org/10.1038/4580 (1999).
17. Knill, D. C. & Pouget, A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719. https://doi.org/10.1016/j.tins.2004.10.007 (2004).
18. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *Science* **331**, 1279–1285. https://doi.org/10.1126/science.1192788 (2011).
19. Gopnik, A. & Wellman, H. M. Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychol. Bull.* **138**, 1085–1108. https://doi.org/10.1037/a0028044 (2012).
20. W, R. et al. Expectancies as core features of mental disorders. *Curr. Opin. Psychiatry*. https://doi.org/10.1097/YCO.0000000000000184 (2015).
21. Pinquart, M., Endres, D., Teige-Mocigemba, S., Panitz, C. & Schütz, A. C. Why expectations do or do not change after expectation violation: A comparison of seven models. *Consciousness Cognit.* **89**, 103086. https://doi.org/10.1016/j.concog.2021.103086 (2021).
22. Panitz, C. et al. A revised framework for the investigation of expectation update versus maintenance in the context of expectation violations: The ViolEx 2.0 model. *Front. Psychol.* https://doi.org/10.3389/fpsyg.2021.726432 (2021).
23. Rief, W. et al. Using expectation violation models to improve the outcome of psychological treatments. *Clin. Psychol. Rev.* **98**, 102212. https://doi.org/10.1016/j.cpr.2022.102212 (2022).
24. Santoro, A. et al. A simple neural network module for relational reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4974–4983 (Curran Associates Inc., Red Hook, NY, USA, 2017).
25. Barrett, D., Hill, F., Santoro, A., Morcos, A. & Lillicrap, T. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, 511–520 (PMLR, 2018).
26. Sung, F. et al. Learning to Compare: Relation Network for Few-Shot Learning. 1199–1208 (2018).
27. Hill, F., Santoro, A., Barrett, D., Morcos, A. & Lillicrap, T. Learning to Make Analogies by Contrasting Abstract Relational Structure (2018).
28. Barak, T. & Loewenstein, Y. Untrained neural networks can demonstrate memorization-independent abstract reasoning. *Sci. Rep.* **14**, 27249. https://doi.org/10.1038/s41598-024-78530-z (2024).
29. Ribes-Iñesta, E., León, A. & Andrade-González, D. E. Comparison patterns: An experimental study of transposition in children. *Behav. Process.* **171**, 104024. https://doi.org/10.1016/j.beproc.2019.104024 (2020).
30. León, A. et al. RBDT: A computerized task system based in transposition for the continuous analysis of relational behavior dynamics in humans. *J. Visualized Exp. (JoVE)* https://doi.org/10.3791/62285 (2021).
31. Monsell, S. Task switching. *Trends Cognit. Sci.* **7**, 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7 (2003).
32. Strogatz, S. H. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (CRC Press, 2018).
33. Benaïm, M. Dynamics of stochastic approximation algorithms. In Azéma, J., Émery, M., Ledoux, M. & Yor, M. (eds.) *Séminaire de Probabilités XXXIII*, 1–68, https://doi.org/10.1007/BFb0096509 (Springer, Berlin, Heidelberg, 1999).
34. Barak, T. & Loewenstein, Y. Naive Few-Shot Learning: Uncovering the fluid intelligence of machines, https://doi.org/10.48550/arXiv.2205.12013 (2023).
35. Lord, C. G., Ross, L. & Lepper, M. R. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Personality Social Psychol.* **37**, 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098 (1979).
36. Kahneman, D. *Thinking, Fast and Slow* (Macmillan, 2011).
37. Foa, E. B. & Kozak, M. J. Emotional processing of fear: Exposure to corrective information. *Psychol. Bull.* **99**, 20–35 (1986).
38. Barlow, D. H. *Anxiety and its disorders: The nature and treatment of anxiety and panic, 2nd ed.* Anxiety and its disorders: The nature and treatment of anxiety and panic, 2nd ed (The Guilford Press, New York, NY, US, 2002). Pages: xvi, 704.
39. Bruner, J. S. *Toward a Theory of Instruction* (Harvard University Press, 1974).
40. Vygotskij, L. S. F. A. & John-Steiner, V. *Mind in Society: The Development of Higher Psychological Processes* (Harvard University Press, 1979).
41. Festinger, L. *A Theory of Cognitive Dissonance*. (Stanford University Press, 1957). Pages: xi, 291.
42. Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).
43. Virtanen, P. et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272. https://doi.org/10.1038/s41592-019-0686-2 (2020).

Acknowledgements
This work was supported by the Gatsby Charitable Foundation. Y.L. holds the David and Inez Myres Chair in Neural Computation. We thank David Hansel and Ilya Nemenman for insightful discussions.

Author contributions
T.B. conducted the experiments, T.B. and Y.L. designed the study, analyzed the results and wrote the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-16135-w.

Correspondence and requests for materials should be addressed to T.B.

Two pathways to resolve relational inconsistencies: Supplementary information

The results for other predictive features: Color and Number

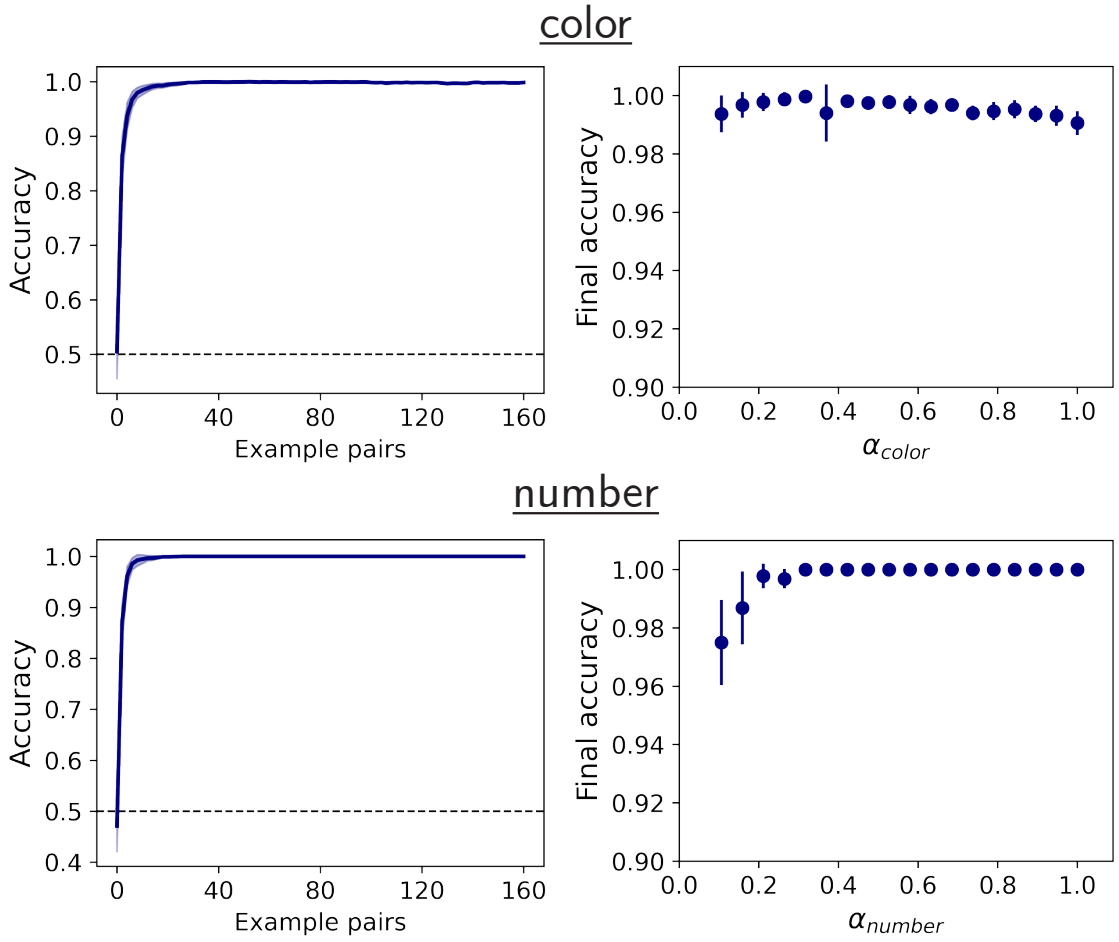

Figure S1: **Task performance.** Left: The average test accuracy of 100 networks trained on a task where the predictive features were color (top) or number (bottom). The change rule was $\alpha = 0.5$. Right: The final test accuracies for various values of α, averaged over 100 networks per α. Error shades and bars correspond to 95% CI.

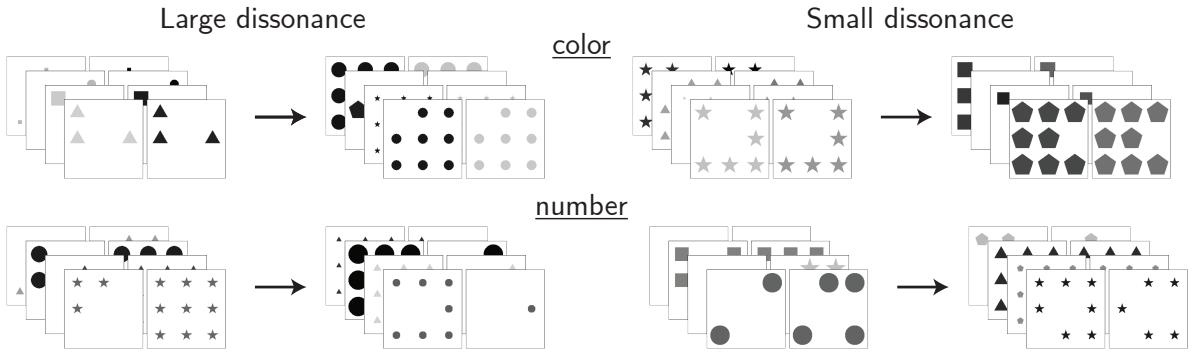

Figure S2: **Simulating a dissonance.** Initially, the predictive feature (top: color, bottom: number) increases by α. Then, the relationship is reversed to $-\alpha$. The dissonance magnitude is represented by α. Left: large dissonance $\alpha = 0.8$; Right: small dissonance $\alpha = 0.2$.

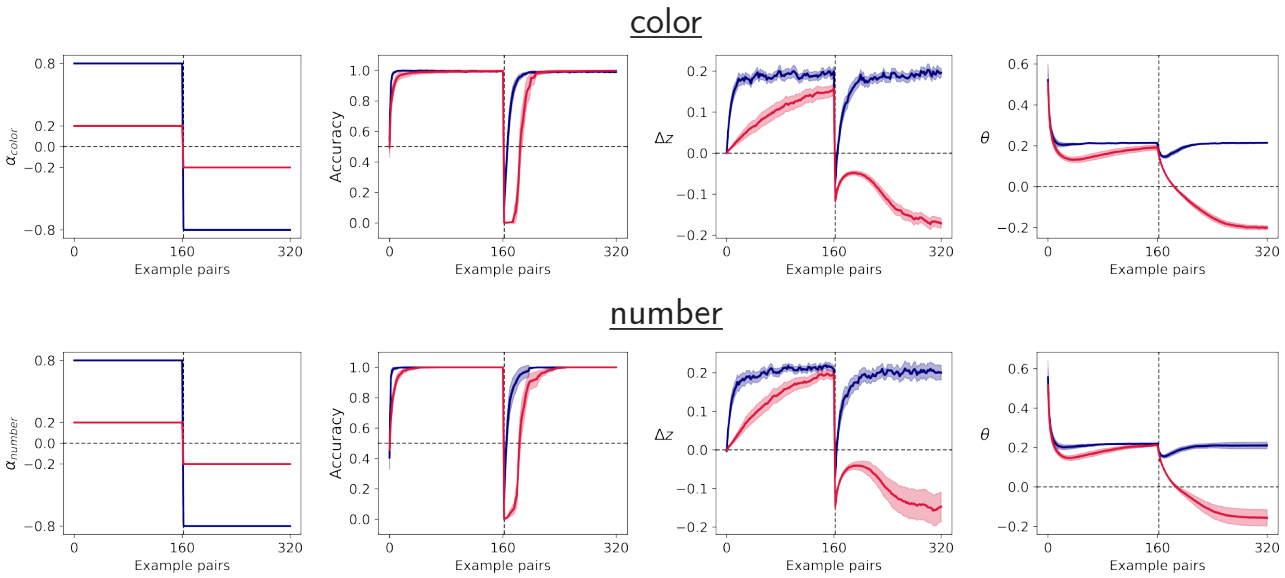

Figure S3: **Performance and adaptation following rule reversal.** 100 networks were presented with 160 examples in which $\alpha = 0.8$ (blue) or $\alpha = 0.2$ (red). Then, we reversed this rule, simulating a cognitive dissonance. From left to right, as a function of training examples: The rule α; Classification accuracy of the networks during the task; The ANNs' ΔZ; The ANNs' θ. The lines are the averages, and the shades correspond to 95% CI.

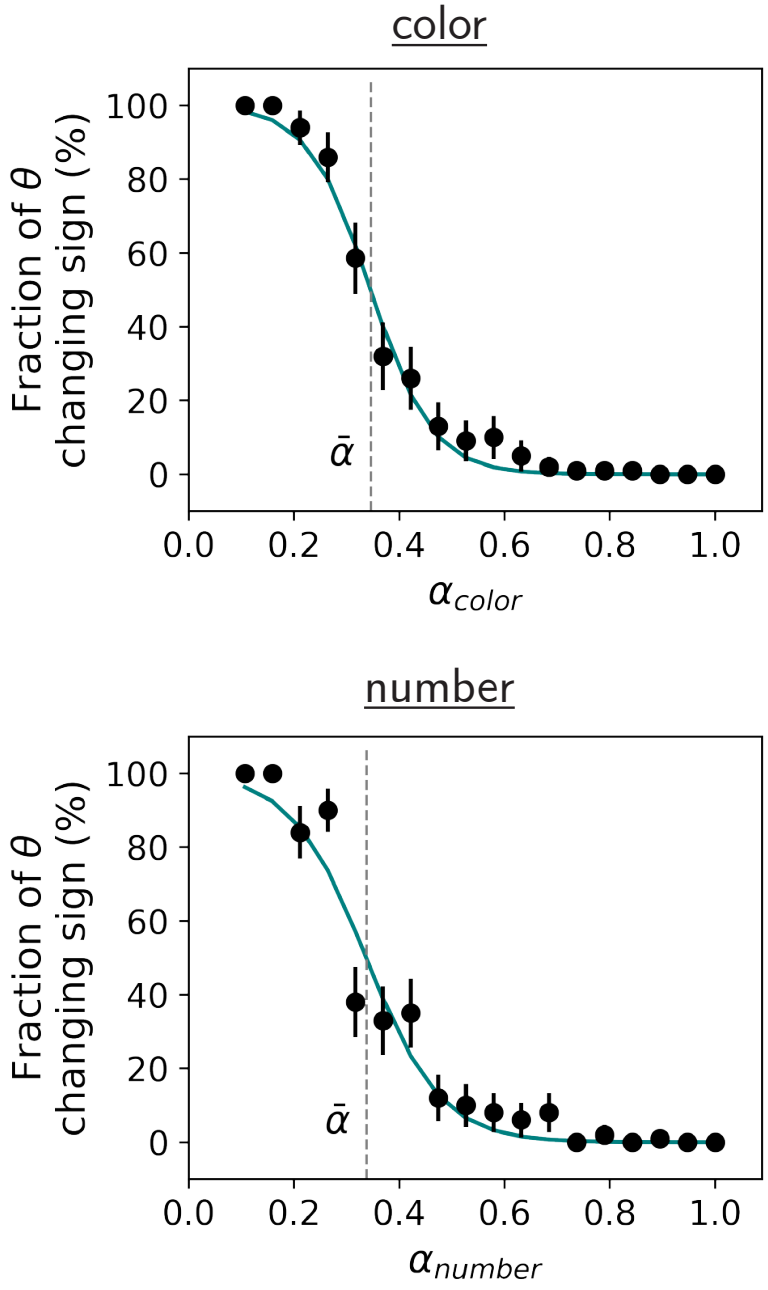

Figure S4: **Adaptation pathway versus dissonance magnitude.** For the color (left) and number (right) predictive feature, the percentage of networks that adapted their input representation ΔZ to match the expected θ increases a function of α. The inflection points between the adaptation pathways, $\bar{\alpha}$, were obtained by fitting the results (black) with a logistic function (green). Error bars correspond to 95% CI (Wilson estimation).

Fitting the adaptation pattern of the ANN

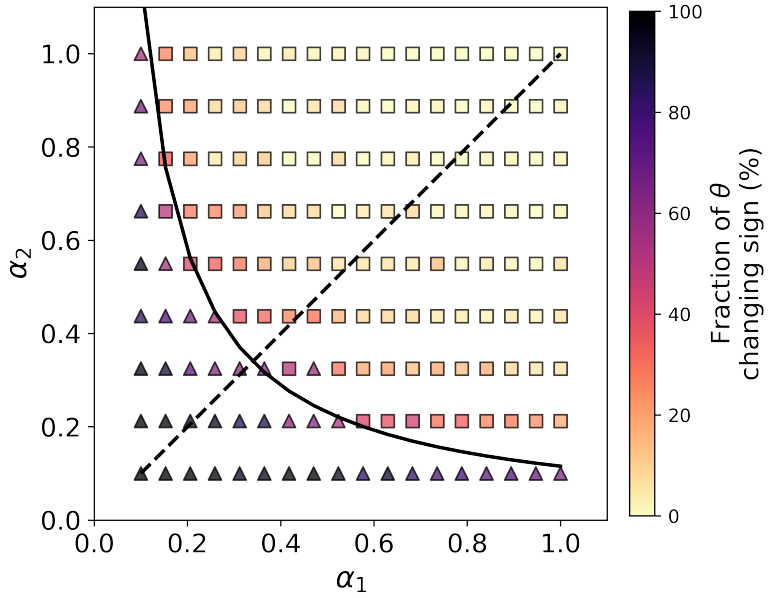

Figure S5: **Adaptation patterns for general rule reversals for the ANN.** The adaptation pattern, taken from (Fig. 5), with the predicted inflection line $\alpha_1\alpha_2 = \bar{\alpha}^2 = 0.34^2$ (black solid line). The value $\bar{\alpha}$ was obtained from the symmetric rule reversal case, represented by the dashed black line.

Addendum

In the Discussion of our second paper [23], we addressed how Bayesian models might account for the inverted U-shaped relationship between the magnitude of an expectation violation and the resulting adaptation. We hypothesized that the stability of expectations in the face of extreme violations could be modeled by adding a "confidence-in-observation" variable that prevents the updating of prior beliefs when data is deemed untrustworthy. A rigorous computational realization of this principle is found in the "latent cause" theory of memory modification proposed by Gershman et al. [24].

Gershman et al. [24] propose a normative framework consisting of two interacting sub-systems: an associative module that learns correlations, and a structure learning module that partitions experience into latent causes. In this framework, the agent's internal representation of a state is not merely the sensory object, but the object conjugated with its inferred latent cause. Consequently, the inference of a new latent cause constitutes a fundamental representational shift.

This structural mechanism effectively internalizes the "confidence" variable we hypothesized. The model continuously evaluates the likelihood that the current observation was generated by the active latent cause. Small prediction errors maintain a high likelihood (high confidence in the current structure), leading to parametric updating of the existing association. Conversely, large prediction errors—such as those occurring during extreme violations—drastically reduce the likelihood that the current cause is valid. Rather than forcing an update on a low-confidence representation, the structure learning module infers a new latent cause. This segmentation protects the original memory trace from interference and re-represents the object as belonging to a distinct generative context.

Testing human adaptation to relational violations

Tomer Barak, Ron Hafzadi, Yonatan Loewenstein

Status: Unpublished

Testing human adaptation to relational violations

Tomer Barak[1,*], **Ron Hafzadi**[1], **and Yonatan Loewenstein**[1,2]

[1]The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Israel
[2]Department of Cognitive and Brain Sciences, The Federmann Center for the Study of Rationality, The Alexander Silberman Institute of Life Sciences, The Hebrew University, Jerusalem, Israel
[*]tomer.barak@mail.huji.ac.il

ABSTRACT

Previous research suggests that when humans encounter observations that strongly contradict their expectations, they tend to avoid revising those expectations. Instead, they seek alternative explanations to preserve their original beliefs. In earlier work, we reconstructed this nonmonotonic adaptation pattern in a model capable of learning relationships between pairs of stimuli. The model had distinct representational and relational modules and we found that the learning dynamics constitutes a "race" between two adaptation pathways: adapting relational expectations or reinterpreting the stimuli. The model predicted that larger violations bias the outcome of that race toward stimuli reinterpretation. In this paper, we test whether this prediction holds for humans. Participants learned a relational rule between ambiguous visual stimuli and were then exposed to rule reversals of varying magnitudes. To evaluate the adaptation pathway of the participants, we primed them towards one input interpretation and evaluated whether this interpretation has changed after adaptation. A preliminary experiment yielded non-significant results, possibly because we used the same task for priming and for testing whether reinterpretation has occurred. A subsequent experiment, redesigned to circumvent this confound, found that a large violation magnitude significantly increased the likelihood that participants would reinterpret the stimuli rather than adjust their relational expectation, though the effect size was modest ($p=0.0427$, Cohen's $d=0.277$). This result is consistent with the model predictions. However, further tests of more nuanced model predictions, while consistent with the theory, but did not reach statistical significance. We conclude that while our findings provide modest support for the dual-pathway model's core prediction, our study primarily highlights the critical challenges in translating this theory to human behavior.

Introduction

Recent experimental studies indicate that when humans' expectations are violated, their adaptation is characterized by an inverted U-shape with the size of violation: They slightly update their expectations following small violations; larger violations lead to larger modification of their expectations; however, extreme violations diminish the updating of the expectations[1–4]. The latter transition, where increasing the violation reduces the adaptation, challenges traditional learning models such as predictive coding and Bayesian updating, which typically posit more adaptation following larger prediction errors[5–8].

In our previous work[9], we reconstructed this nonmonotonic adaptation pattern in a model capable of learning relationships between pairs of stimuli. The model had distinct representational and relational modules. The representational module encoded low-dimensional object-specific features, while the relational module encoded the expected relationships between those features. The model was trained with a specific relationship, forming an expectation of it, and was then exposed to stimuli with a reverse relationship, violating its expectation. When the violations were of moderate magnitude, the relational module was more likely to adapt. However, when faced with extreme violations, the likelihood of adaptation occurring within the representational module significantly increased, biasing the system to change stimuli representations and preserve the relational expectation.

Existing empirical studies in humans could, in principle, be reexamined through the lens of the relational framework to test the model's validity. However, such post hoc re-examinations would necessarily limit the ability of these studies to support or disprove the model[10]. Therefore, to test whether this dual-pathway model of adaptation applies to humans, we designed a novel experiment structured around the violation of relational expectations. Consider the scenario depicted in Fig. 1a-b. A participant is presented with two rectangles and is instructed to choose one of them. Unbeknownst to the participant, the "correct" response is to choose the rectangle associated with a higher ratio of blue (in the figure, the correct rectangles are marked with green V) and they learn this rule from a binary, "correct"/"wrong" feedback. This establishes an initial state where the stimulus is represented by its "blue ratio" and the relational expectation is "more is better" (Fig. 1c). Following this, the rule is reversed

without warning, creating a violation of their expectation. To adapt, the participant can take one of two pathways (Fig. 1d). They could engage in *relational adaptation*, maintaining their focus on the blue color but reversing their expectation (i.e., "*less* blue is now correct"). Alternatively, they could pursue *representational adaptation*, switching their representational focus to the complementary color and preserving their original rule (i.e., "more *orange* is now correct"). The central question of this study is: which pathway do humans prefer, and is this choice systematically influenced by the magnitude of the rule violation?

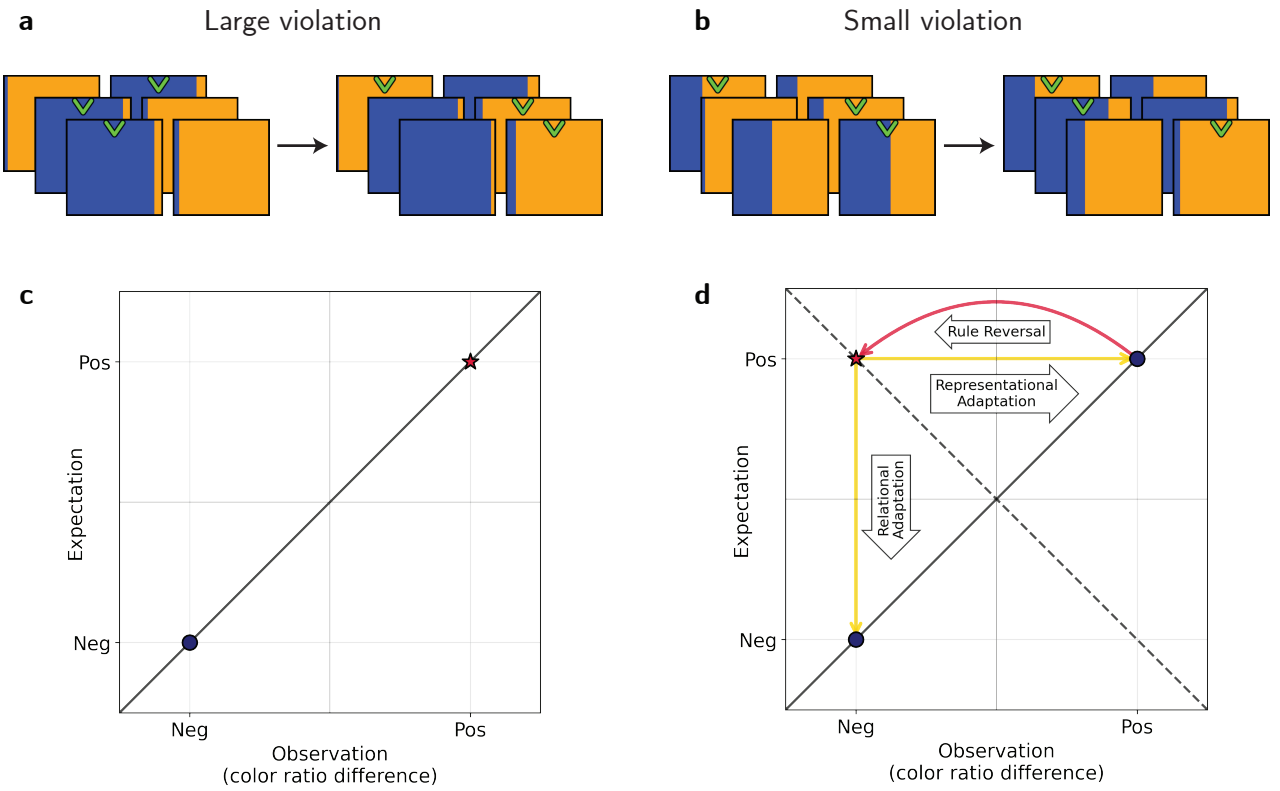

Figure 1. Experiment schema. (**a**) Example of a large violation, where the color ratio difference between "correct" and "wrong" rectangles is substantial. (**b**) Example of a small violation, where this difference is subtle. (**c**) Two equivalent solutions for learning the initial rule before reversal: Participants can focus on blue and learn that the correct rectangle has "more blue" (starred point), or equivalently focus on orange and learn that it has "less orange" (circled point). The diagonal line marks states where the learned rule matches observation. Our design primes participants toward the "more blue" solution. (**d**) Two adaptation pathways after a rule reversal. Upon "more blue" being violated, participants can perform *relational adaptation* (vertical arrow) by keeping focus on blue but updating the rule to "less blue is now correct," or *representational adaptation* (horizontal arrow) by preserving the "more is correct" rule while switching focus to orange.

Disentangling these two adaptation pathways in a behavioral experiment is challenging, as we cannot directly observe a participant's internal input representation or relational rule. To overcome this, our experimental design has two critical components. First, to establish a known starting point, we primed participants to adopt a specific representational focus – in this case, on the ratio of blue. Second, after the participants adapt to the rule reversal, we implemented a final evaluation to test whether this initial focus was maintained (signaling relational adaptation) or if it changed (signaling representational adaptation).

This paper details the experiments that implement this design. We will first describe the experimental flow chronologically, from the initial priming task, through the relational learning and reversal phases, to the final evaluation. We then report the findings, including an exploration of two main approaches for the final evaluation query, and compare them with the model's predictions. The results provide modest support for the core theoretical prediction while highlighting important methodological considerations for translating computational theories to human behavior.

Results

Experimental Procedure

The experiments designed to test our hypotheses followed a three-phase structure, as illustrated in Figure 2a: (1) an initial priming phase, (2) a relational learning and rule-reversal phase, and (3) a final evaluation phase to assess the adaptation pathway.

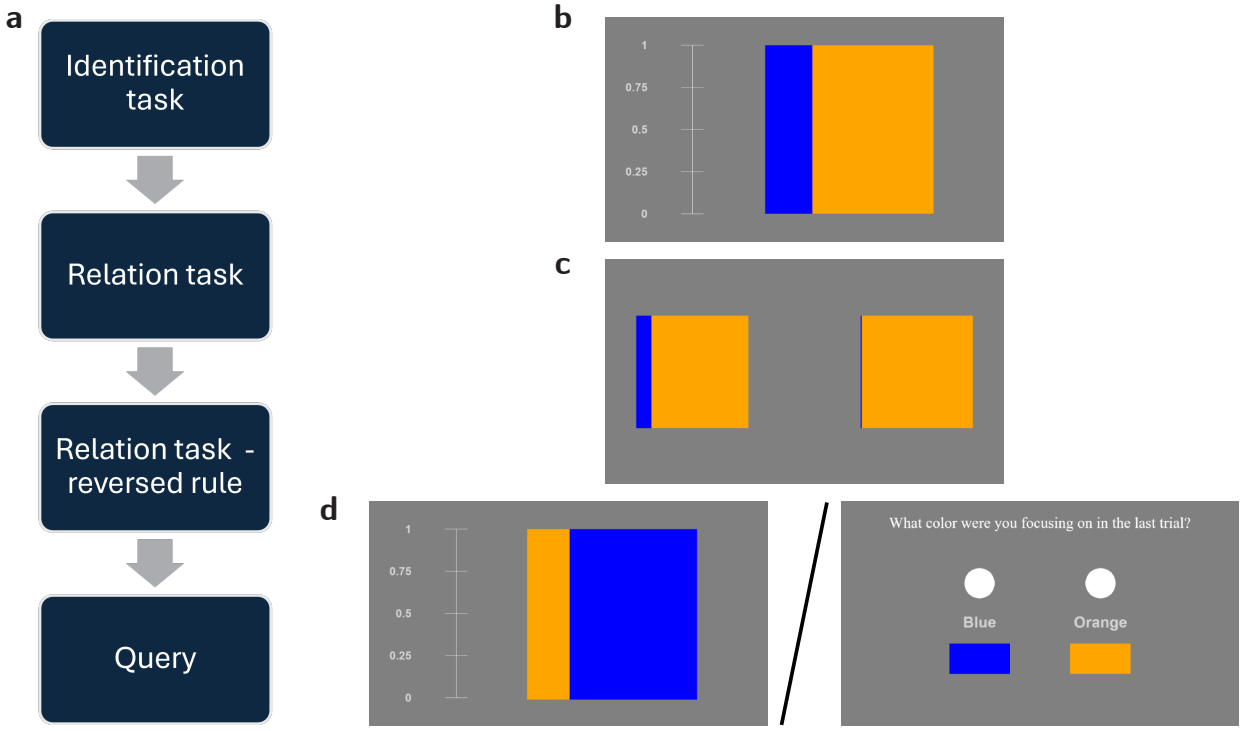

Figure 2. Experiment flow. (a) The three main stages of the experiment. **(b)** The initial priming task, an identification task which primed participants to focus on the ratio of a specific color. **(c)** The relation task, where participants first learned a rule (e.g., "more blue is correct") and then adapted to a rule reversal. (d) The final evaluation stage, which was designed to probe which of the colors the participant are finally focused on. The image depicts the two alternatives that we used: A repeated identification task in Experiment 1 (left), and a direct query method used in Experiment 2 (right).

The experiment began with an *initial priming phase* consisting of an identification task (Fig. 2b). In this task, participants were presented with single rectangles composed of two colors, and instructed to click on a slider. Through trial and error, using a binary feedback, they learned that the relevant feature is the fraction of one of the colors (e.g., blue) and that the task is to use the slider to indicate this fraction. For example, if one quarter of the area of the rectangle is blue, the correct answer is $0.25 \pm \varepsilon$, where ε is an error toleration term (see Methods). This procedure was designed to prime the participants to focus on the area of a specific color (e.g., blue) when considering the rectangles.

Next, participants proceeded to the *relational learning phase*. Here, they were presented with pairs of rectangles, of the same type as in the first part of the experiment, and were instructed to select one of them. Unbeknownst to them, the "correct" rectangle was the one associated with either a larger area or a smaller area of the primed color, e.g., more blue (Fig. 2c). Again, binary feedback was provided to indicate whether the participant answered correctly.

After a participant demonstrated they had learned this rule, the *rule-reversal phase* began without warning. The rule was inverted, such that the correct rectangle now had a smaller ratio of that color if it was larger before, or a larger ratio of that color if it was smaller before. In the example we use here, the rule changed so that the correct rectangle was associated with a smaller ratio of blue (or, equivalently, a higher ratio of orange). Based on the theory, the violation magnitude is determined by the magnitudes of the rules before and after the reversal. To test the effect of the violation magnitude on the adaptation pathway,

we divided participants into two groups based on the magnitude of this violation: In the "small violation" group, the color ratio difference between rectangles in each pair was 12%, both before and after reversal, whereas in the "large violation" group this difference was 87%.

We predicted that a small violation would more often lead to a change in the relational expectation whereas a large violation would more often lead to a change in the representation — changing the color focus of the participants. Thus, the crucial third phase of the experiment was the *final evaluation*, designed to determine whether participants had adapted via the relational or representational pathway. We explored two different methods for this evaluation across two experiments.

Experiment 1: Repeated identification task

In our preliminary experiment, the final evaluation consisted of repeating the identification task from the priming phase with just one item (Fig. 2d left). We hypothesized that participants who underwent representational adaptation (i.e., switched their focus from the color associated with the first identification task to the other color) would consequently evaluate the ratio of the other color when presented with the task again.

The single identification item used for this final evaluation consisted of a rectangle with a 75% ratio of blue. We classified participants clicking above 50% as focusing on blue and below 50% as focusing on orange.

Fig. 3a depicts the results of this experiment. Contrary to our prediction, there was no significant difference between the two groups (one-tailed independent samples t-test, $t(200) = -0.99$, $p = 0.84$). We did observe a bias in favor of responding in accordance with the strategy that was successful the first time they performed the task. Thus, we suspect that this "primacy" effect might have interfered with the current task. Hence, repeating the same task twice may be an unreliable measure of the representational state they had adopted during the intermediate relational task.

We, therefore, adopted a different experimental strategy in the second experiment.

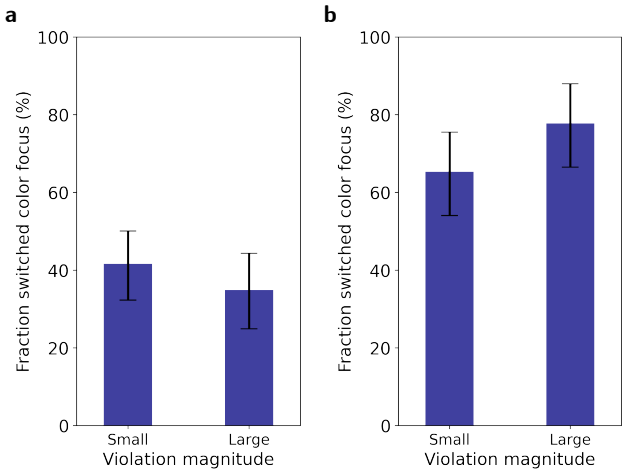

Figure 3. Adaptation pathways' dependence on violation magnitude. The fraction of participants who switched their color focus after rule reversal, categorized by violation magnitude (**a**) Experiment 1, evaluated with a repeated identification task. The analysis revealed no significant difference in the rate of representational switching between participants experiencing large violations ($N = 106$) and those experiencing small violations ($N = 96$). (**b**) Experiment 2, evaluated with a direct query. Participants experiencing large violations ($N = 81$) showed significant higher rates of representational adaptation compared to those experiencing small violation ($N = 75$), $t(154) = 1.73$, $p = 0.0427$, but the effect size was small (Cohen's $d = 0.277$). Error bars represent Wilson score interval estimates.

Experiment 2: Direct query of color focus

To address the potential confounds identified in Experiment 1, we designed a second experiment whose main modification was that it decoupled the final evaluation from the priming task. Specifically, after participants adapted to the reversed rule, we

directly queried their representational focus by asking, "What color were you focusing on in the last trial?" (See full changes in the Methods).

The results of this revised experiment confirmed our central hypothesis (Fig. 3b). Participants in the large violation group were significantly more likely to switch their color focus than those in the small violation group, providing modest but statistically significant support for the model's core prediction that the magnitude of a violation systematically modulates the chosen adaptation pathway. ($t(154) = 1.73$, $p = 0.0427$, Cohen's $d = 0.277$).

Caution, however, should be exercised when interpreting these results. We implicitly assumed that participants who chose to change the representation of the stimuli, did so by focusing on the other color. An alternative way could have been to focus on the *lack* of the original color, rather than its presence. For instance, participants who initially learned that "the correct rectangle has more blue" and then experienced a relational violation could adapt their input representation to either "more orange" or "more lack of blue". The latter would undermine the experiment's ability to separate the adaptation pathways, as their focus would remain on blue even though they had changed their input representation. However, this will not explain the difference between the two groups.

Experiment 3: Intermediate rules

One of the more nuanced predictions of the theory is that it is possible to influence the adaptation pathway by introducing intermediate steps before the rule is fully reversed. Based on the theory, we hypothesized that for participants whose initial and final rules constitute a small violation, introducing an intermediate step of a large magnitude (in either direction) would increase the propensity for representational adaptation (switching color focus). Conversely, for participants whose initial and final rules constitute a large violation, an intermediate step with a smaller rule magnitude would decrease this propensity.

We designed four conditions to test this prediction. To decrease the propensity for representational adaptation relative to a standard large violation condition, participants experienced intermediate steps of a smaller magnitude:

- **Contrast reducing:** The rule sequence was $87\% \rightarrow 12\% \rightarrow -87\%$.
- **Undershooting:** The rule sequence was $87\% \rightarrow -12\% \rightarrow -87\%$.

To increase the propensity for representational adaptation relative to the small violation condition, participants experienced intermediate steps of a larger magnitude:

- **Contrast enhancing:** The rule sequence was $12\% \rightarrow 87\% \rightarrow -12\%$.
- **Overshooting:** The rule sequence was $12\% \rightarrow -87\% \rightarrow -12\%$.

Our findings were directionally consistent with the hypothesis but did not reach statistical significance (Fig. 4). For instance, the 'Contrast enhancing' group showed a higher rate of switching than the 'Small violation' group, but the difference was not significant ($t(172) = 1.18, p = 0.12$). Similar results were true for the 'Undershooting' group ($t(134) = 1.06, p = 0.14$). The conditions that were supposed to reduce the proportion of representational adaptation did not reach significance as well. The 'Contrast reducing' group switched color focus less frequently than the 'Large violation' group, but this difference was also not significant ($t(176) = -0.55, p = 0.29$). A similar non-significant result was obtained for the 'Overshooting' group ($t(149) = -0.89, p = 0.19$).

When combining all four intermediate conditions ($N = 327$), the overall tendency to switch color focus (73.4%) was positioned between the small and large violation groups, as predicted. However, the combined group was not significantly different from either the small violation group ($t(400) = 1.333, p = 0.0928$) or the large violation group ($t(406) = -0.834, p = 0.2028$). Therefore, while the results align with the model's qualitative predictions, they do not provide statistically robust evidence that these intermediate steps effectively steer the adaptation pathway in human participants.

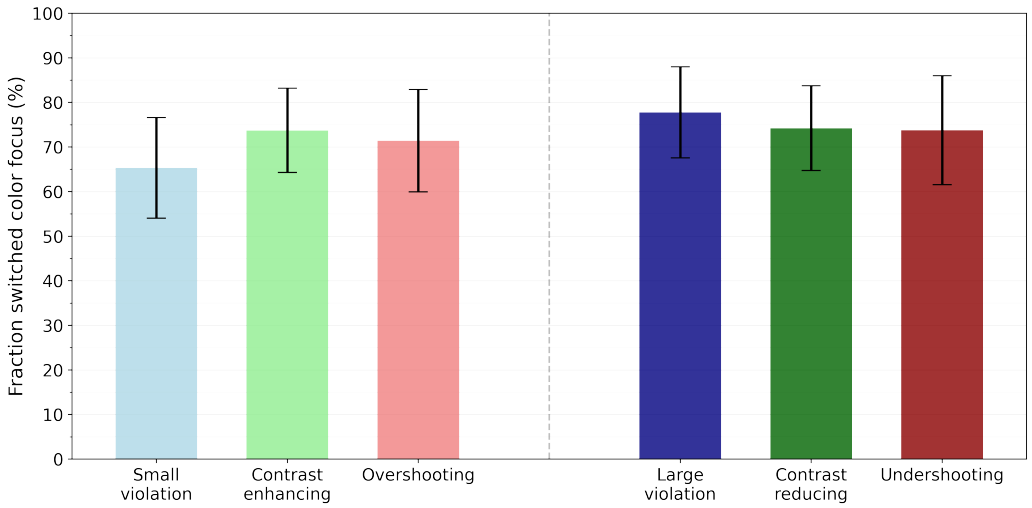

Figure 4. The effect of intermediate adaptation steps on the final adaptation pathway. The fraction of participants who switched their color focus is shown for the baseline small violation group (light blue; $N = 75$) and large violation group (dark blue; $N = 81$), alongside four intermediate conditions. Two conditions were designed to potentially increase color focus switching from the small violation baseline: Contrast enhancing (light green; $N = 99$) and Overshooting (light red; $N = 70$). Two other conditions were designed to potentially decrease switching from the large violation baseline: Contrast reducing (dark green; $N = 97$) and Undershooting (dark red; $N = 61$). While the trends for the intermediate groups were in the hypothesized directions relative to the baseline groups, the differences were not statistically significant. Error bars represent Wilson score interval estimates.

Discussion

This study provides the first direct empirical test of a dual-pathway model of adaptation to relational expectation violations. Our central hypothesis was that the magnitude of a violation would determine the adaptation pathway, with larger violations making it more likely for individuals to reinterpret stimuli (representational adaptation) rather than update their existing rule (relational adaptation). While the findings offer modest support for this core prediction, they also underscore the significant challenges in translating the computational theory into a behavioral experiment with human participants.

Our primary finding that participants experiencing large violations were more likely to switch their representational focus compared to those experiencing small violations provides statistically significant support for the model's central prediction. This result suggests that humans, like the model, exhibit different adaptation strategies depending on the severity of expectation violations. The finding aligns with emerging evidence from other domains showing that extreme violations can lead to qualitatively different cognitive responses than moderate ones. However, the small effect size (Cohen's $d = 0.277$) and p-value near the significance threshold ($p = 0.0427$) raise questions about the robustness and importance of this result. The magnitude of the difference between conditions, while statistically detectable, may reflect relatively subtle cognitive processes that are easily overwhelmed by individual differences or task-specific factors.

Our experiments demonstrate the methodological difficulties in measuring internal cognitive states. Our initial experiment (Experiment 1) failed to support the hypothesis and showed a non-significant trend in the opposite direction. We believe this was due to a methodological confound where participants' memory of the initial priming task interfered with the final evaluation. This led us to redesign the evaluation in Experiment 2 to be a direct query about the participant's focus. The contrast between the outcomes of Experiment 1 and 2 demonstrates how human-specific factors, particularly memory, can be a powerful confounding variable that must be carefully managed in experimental design.

A primary methodological contribution of this study is the application of a psychophysical framework to investigate the abstract cognitive process of relational learning. We attempted to distill the complexity of learning a relationship into a simple, quantifiable task: determining whether a single, continuous variable—the ratio of blue to orange in a rectangle—should be increased or decreased to achieve a correct outcome. This design moves beyond traditional paradigms by creating a tightly controlled environment where the "rule" is not a complex logical statement but a simple directional judgment on a perceptual

continuum ("more is better" vs. "less is better"). While the modest effect sizes and methodological challenges we encountered suggest that human cognition does not map onto this simplified model as neatly as an artificial neural network, this experimental paradigm lays the groundwork for future research. It provides a template for how psychophysical methods can be used to probe the internal representations and adaptation mechanisms that underlie how we learn and modify abstract relationships.

Beyond these methodological considerations, further experiments designed to test more nuanced predictions of the theory yielded results that were directionally consistent with the model but did not achieve statistical significance. Specifically, introducing intermediate adaptation steps (Experiment 3) produced trends in the hypothesized directions, but the differences were not significant. While these results do not contradict the model, they fail to provide the robust evidence needed to confirm its more detailed predictions, suggesting that human learning may be more flexible or variable than that of the model used to generate the theory.

Several limitations of this study must be acknowledged. A key limitation is the high rate of representational switching observed in Experiment 2, where it was the dominant strategy in both the small (65.3%) and large (77.8%) violation groups. This general preference for switching is consistent with the principle that humans find it easier to represent positive relations (e.g., "more orange") than negative ones (e.g., "less blue")[11]. However, this strong baseline preference may have created a ceiling effect that masked a larger underlying difference between the conditions. Additionally, the direct query in Experiment 2, while avoiding the memory confound of Experiment 1, relies on the assumption that participants can accurately report their internal strategy. Finally, the artificial nature of the stimuli and task may not fully capture how expectation violations are handled in more complex, real-world scenarios.

Several avenues for future research emerge from our findings. Developing more sensitive measures and refined experimental designs could strengthen empirical tests of the dual-pathway model. To counteract the potential bias towards representational adaptation, future studies could aim to increase the cognitive cost of representational switching. For example, using more complex, non-complementary features instead of simple colors would make it harder to find an alternative representation, potentially creating a scenario where relational adaptation becomes the "easier" path, which might align results more closely with the theory's predictions for small violations. Methodologies like eye-tracking could provide real-time data on attentional focus, potentially resolving the ambiguity between the subjective query of Experiment 2 and the memory-confounded task of Experiment 1. Additionally, exploring individual differences in adaptation strategies could illuminate the sources of variability observed in our study. Factors such as cognitive flexibility, working memory capacity, or personality traits related to openness to experience might predict which adaptation pathway individuals prefer. Finally, extending the paradigm to more ecologically valid contexts could test whether the framework applies beyond artificial laboratory tasks. Studies using social expectations, causal relationships, or real-world learning scenarios would provide stronger tests of the model's generalizability.

In conclusion, this research represents a preliminary step in testing the dual-pathway model. We found modest, statistically significant evidence for the model's core prediction that violation magnitude influences the adaptation pathway. However, the small effect size and the null results for more nuanced predictions indicate that the direct translation of this computational model to human behavior is not straightforward. The findings highlight that human adaptation is complex, variable, and subject to cognitive factors like memory that are not always captured in simpler models. Despite these limitations, this work establishes a foundation and highlights clear methodological directions for future research into how humans adapt their beliefs in a changing world.

Methods

Participants

A total of 1027 individuals participated in the experiments and completed the entire study (including those participated in the preliminary study of Experiment 2). The participants were recruited online via Prolific (www.prolific.com) and received monetary compensation for their time. The sample had a mean age of 36.27 years ($SD = 10.94$, range: 18–60 years), with approximately equal representation across sex (Male: 50.8%, Female: 49.2%). Participants were primarily from the United Kingdom (71.2%) and the United States (28.8%), and English was the primary language for 91.5% of participants. All participants provided informed consent before beginning the study, and the experimental protocol was approved by the Hebrew University of Jerusalem Ethics Committee.

Apparatus and stimuli

The experiment was built using PsychoPy[12] and was presented to participants in their web browser.

The stimuli consisted of rectangles presented on a neutral gray background. Each rectangle was filled with a combination of blue and orange. The key feature of each stimulus was the ratio of the two colors.

In Experiment 1, the sides of the color fillings were random, i.e., the blue side could have been on either the left or right. In Experiments 2 and 3, the blue side was always on the left.

In the relational task, the rule strength corresponded to the difference in this ratio between pairs of rectangles. For example, in the "small violation" condition, the difference in the blue-to-orange ratio between the two rectangles was 12%. Similarly, in the "large violation" condition, the difference between the blue-to-orange ratio was 87%.

After each choice, participants received feedback in the form of "Correct :)" or "Wrong :(".

1 Experimental Procedures

1.1 Experiment 1

1.1.1 Overview

Participants initiated the experiment online through a procedure consisting of four main phases: an identification task, an initial relation-learning task, a relation task with a reversed rule, and a final query.

1.1.2 Instructions

Participants received the following sequential instructions:

1. **Welcome and Consent:** “Welcome. In this experiment you are requested to identify rules that are associated with blue and orange rectangles presented on the screen. You will need to select an answer based on the ratios of orange and blue in the rectangles. After each choice, you will get feedback. Try to identify the rules based on these feedbacks. Please follow the specific instructions of each part. If you consent, press SPACE to continue. Otherwise, press ESC anytime throughout the experiment.”
2. **Slider Task Instructions:** “In this part, you will need to click on the slider with your mouse. Press SPACE to continue.”
3. **Choice Task Instructions:** “In this part, you will need to choose left or right using the arrows of your keyboard. Press SPACE to continue.”
4. **Final Task Instructions:** “Now you are ready for the final task. It is the slider task with a single item. Press SPACE to continue.”

1.1.3 Experimental Phases

Priming Phase (Identification Task) Participants were first presented with a series of rectangles combining blue and orange colors. They were instructed to click on the slider based on a rule related to the rectangles. Participants received binary feedback and used it to learn that the relevant feature was the fraction of one of the colors (e.g., blue) and to indicate this fraction with the slider. For instance, if a quarter of the rectangle’s area was blue, the correct slider answer was $0.25 \pm \varepsilon$. This phase was designed to prime participants to focus on the area of a specific color and ended when participants succeeded on 6 consecutive trials.

Relational Learning and Rule-Reversal Phase Following the priming phase, participants were presented with pairs of rectangles and had to select the “correct” one. Initially, the rule specified that the correct rectangle had a larger area of one of the colors (e.g., blue). After the participant learned this rule, it was reversed without warning. The new rule specified that the correct rectangle had a smaller ratio of the original color. Both the color used for priming (blue or orange) and the initial relational rule were counterbalanced across participants.

Final Evaluation (Repeated Identification Task) To determine the adaptation pathway, the final evaluation repeated the identification task from the priming phase using a single item—a rectangle with a 75% blue ratio. Participants who clicked above the 50% mark on the slider were classified as focusing on blue, while those who clicked below 50% were classified as focusing on orange.

1.1.4 Parameters

Error tolerance for identification task (ε): ± 0.15

Trials to success for priming phase: 6 consecutive correct trials

1.1.5 Time Limits

Welcome and consent screen: Maximum 60 seconds (auto-advance or SPACE press)

Evaluation task instructions: Maximum 30 seconds (auto-advance or SPACE press)

Individual evaluation trials: Maximum 8 seconds per trial (response or timeout)

Feedback after evaluation trials: Fixed 2 seconds

Choice task instructions: Maximum 30 seconds (auto-advance or SPACE press)

Individual choice trials: Maximum 5 seconds per trial (key press or timeout)

Feedback after choice trials: Fixed 2 seconds

Final test instructions: Maximum 30 seconds (auto-advance or SPACE press)

Final test trial: Maximum 30 seconds (response or timeout)

Feedback after final test: Fixed 2 seconds

Concluding screen: Fixed 1 second

1.2 Experiment 2

1.2.1 Overview

The procedure for Experiment 2 was similar to Experiment 1, with key modifications designed to simplify the analysis and address a potential memory confound. To standardize the experimental pathway, priming was always conducted for the ratio of blue, and the relational task always began with "higher ratio of blue" as the correct rule before inversion. This created a single, consistent experimental pathway for all participants before the rule reversal.

1.2.2 Instructions

Participants received the following sequential instructions:

1. **Welcome and Consent:** "Welcome. In this experiment you are requested to identify rules associated with blue and orange rectangles (like below). Use the provided feedback to identify the rules. *Hint*: the answers correspond to the ratio of orange versus blue in the rectangles. If you consent, press SPACE to continue. Otherwise, press ESC anytime throughout the experiment."
2. **Slider Task Instructions:** "In this part, you will need to find a rule related to a single rectangle and click on the slider accordingly. Press SPACE to continue."
3. **Choice Task Instructions:** "In this part, you will need to choose left or right using the arrows of your keyboard. Press SPACE to continue."
4. **Final Query:** "What color were you focusing on in the last trial?"

1.2.3 Final Evaluation (Direct Query)

After participants adapted to the reversed rule, their representational focus was directly assessed by asking: "What color were you focusing on in the last trial?"

1.2.4 Parameters

Error tolerance for identification task (ε): ± 0.15

Trials to success for priming phase: 7 consecutive correct trials

1.2.5 Time Limits

An initial, exploratory version of this experiment revealed a trend in the predicted direction. Participants experiencing large violations ($N = 175$) showed a greater tendency to switch their representational focus than those experiencing small violations ($N = 167$), though the difference did not reach statistical significance ($t(340) = 1.5$, $p = 0.07$). A post-hoc analysis of these results indicated that the total time taken to complete the task was a significant source of variance; applying a time limit retroactively increased the significance of the result (reducing the groups to $N = 122$ in the large violation groups and $N = 96$ in the small violation group). Based on this finding, we conducted a subsequent experiment with 156 new participants that included an ad-hoc 200-second time limit for task completion. The results of this revised experiment are presented in Figure 3b.

Overall timeout: 200 seconds for revised version (no limit for initial version)[1]

Welcome and consent screen: Maximum 60 seconds (earlier with SPACE press)

Evaluation task instructions: Maximum 30 seconds (earlier with SPACE press)

Individual evaluation trials: Maximum 8 seconds per trial (response or timeout)

Feedback after evaluation trials: Fixed 2 seconds

Choice task instructions: Maximum 20 seconds (earlier with SPACE press)

Individual choice trials: Maximum 5 seconds per trial (arrow key response or timeout)

Feedback after choice trials: Fixed 2 seconds

Final question screen: Maximum 30 seconds (slider response)

Concluding screen: Fixed 1 second

1.3 Experiment 3

1.3.1 Overview

The procedure for Experiment 3 was identical to Experiment 2, with two main modifications: an additional intermediate step in the rule-reversal phase and an extended overall time limit of 250 seconds. The intermediate step was introduced to test more nuanced theoretical predictions across four experimental conditions:

Contrast reducing: Rule sequence $87\% \to 12\% \to -87\%$

Undershooting: Rule sequence $87\% \to -12\% \to -87\%$

Contrast enhancing: Rule sequence $12\% \to 87\% \to -12\%$

Overshooting: Rule sequence $12\% \to -87\% \to -12\%$

[1] Participants who did not complete the test within the time limit were excluded from analysis.

1.3.2 Time Limits

Global timeout: 250 seconds (automatic termination)

Welcome and consent screen: Maximum 60 seconds (SPACE press or timeout)

Evaluation task instructions: Maximum 30 seconds

Individual evaluation trials: Maximum 8 seconds per trial (slider response or timeout)

Feedback after evaluation trials: Fixed 2 seconds

Choice task instructions: Maximum 20 seconds

Individual choice trials: Maximum 5 seconds per trial (arrow key response or timeout)

Feedback after choice trials: Fixed 2 seconds

Final question screen: Maximum 30 seconds (slider response)

Concluding screen: Fixed 1 second

Data analysis

The primary dependent variable was the participant's final color focus, as determined by their response in the final evaluation phase (repeated identification task in Experiment 1, direct query in Experiments 2 and 3). A one-tailed independent samples t-test was used to compare the fraction of participants who switched their color focus between the small and large violation groups.

References

1. Filipowicz, A., Valadao, D., Anderson, B. & Danckert, J. Rejecting outliers: Surprising changes do not always improve belief updating. *Decision* **5**, 165–176, DOI: 10.1037/dec0000073 (2018).

2. Hird, E. J., Charalambous, C., El-Deredy, W., Jones, A. K. P. & Talmi, D. Boundary effects of expectation in human pain perception. *Sci. Reports* **9**, 9443, DOI: 10.1038/s41598-019-45811-x (2019).

3. Spicer, S. G., Mitchell, C. J., Wills, A. J. & Jones, P. M. Theory protection in associative learning: Humans maintain certain beliefs in a manner that violates prediction error. *J. Exp. Psychol. Animal Learn. Cogn.* **46**, 151–161, DOI: 10.1037/xan0000225 (2020).

4. Kube, T., Kirchner, L., Lemmer, G. & Glombiewski, J. A. How the Discrepancy Between Prior Expectations and New Information Influences Expectation Updating in Depression—The Greater, the Better? *Clin. Psychol. Sci.* **10**, 430–449, DOI: 10.1177/21677026211024644 (2022).

5. Rescorla, R. A theory of Pavlovian conditioning : Variations in the effectiveness of reinforcement and nonreinforcement (1972).

6. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87, DOI: 10.1038/4580 (1999).

7. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719, DOI: 10.1016/j.tins.2004.10.007 (2004).

8. Ahissar, M. & Hochstein, S. The reverse hierarchy theory of visual perceptual learning. *Trends Cogn. Sci.* **8**, 457–464, DOI: 10.1016/j.tics.2004.08.011 (2004).

9. Barak, T. & Loewenstein, Y. Two pathways to resolve relational inconsistencies. *Sci. Reports* **15**, 30738, DOI: 10.1038/s41598-025-16135-w (2025). Publisher: Nature Publishing Group.

10. Kerr, N. L. HARKing: Hypothesizing After the Results are Known. *Pers. Soc. Psychol. Rev.* **2**, 196–217, DOI: 10.1207/s15327957pspr0203_4 (1998). _eprint: https://doi.org/10.1207/s15327957pspr0203_4.

11. Deschamps, I., Agmon, G., Loewenstein, Y. & Grodzinsky, Y. The processing of polar quantifiers, and numerosity perception. *Cognition* **143**, 115–128, DOI: 10.1016/j.cognition.2015.06.006 (2015).

12. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203, DOI: 10.3758/s13428-018-01193-y (2019).

Acknowledgements

This work was supported by the Gatsby Charitable Foundation. Y.L. holds the David and Inez Myres Chair in Neural Computation.

Author contributions statement

T.B. and R.H. conducted the experiments, T.B., R.H. and Y.L. designed the study and analyzed the results. T.B. and Y.L. wrote the manuscript.

Additional information

Competing interests statement

The authors declare no competing interests.

Discussion and Conclusion

This thesis explored *real-time adaptation* as a computational framework for modeling fluid intelligence. The work demonstrates the potential of this paradigm across three studies: solving intelligence test-like abstract reasoning problems, providing a mechanistic account of paradoxical findings in human belief updating, and testing a novel prediction of the model in a behavioral experiment. While the empirical validation remains preliminary, these studies collectively highlight both the promise and the limitations of viewing fluid intelligence through the lens of real-time adaptation.

Abstract Reasoning as Real-Time Optimization

A key finding of this work is that artificial neural networks can perform abstract reasoning without relying on extensive pre-training. This finding challenges the dominant view that modern AI must rely on large-scale training for succeeding in reasoning tasks. Instead, I propose that abstract reasoning may be understood as a process of real-time optimization, where parameters adapt dynamically based on the information presented by the problem. This perspective offers an alternative to the prevailing trend of ever-larger models and datasets in AI, suggesting a path toward more efficient and adaptable systems that emphasize inference-time learning over exhaustive prior training.

The model architecture also draws parallels to human cognition. The use of fixed, random convolutional layers as general-purpose feature extractors, combined with adaptive higher layers, echoes the division between early sensory cortices and higher-order association regions, such as the prefrontal cortex, which support flexible reasoning [25–27]. While this

analogy should be treated with caution, it suggests that structural biases in networks, like biological priors in the brain, may scaffold adaptive reasoning.

A Mechanistic Account of Expectation Violation

Building on this foundation, the second study applied the model in an online learning framework to model belief updating. Specifically, it provided a process-level account of the inverted U-shaped adaptation pattern observed in humans, where moderate expectation violations drive belief change more effectively than extreme ones [20–22].

Within my framework, this effect arises naturally from competition between adaptation pathways. When contradictions occur, the system can either adjust its relational rule (e.g., “the rule has reversed”) or reinterpret the stimulus (e.g., “I was attending to the wrong feature”). My analysis revealed that the choice between these pathways is determined by a ”race” where the relative speeds are governed by the magnitudes of the original and new rules. This view aligns with computational perspectives in which apparent “non-updating” can reflect inference over structure (e.g., latent causes) rather than a failure of learning. Gershman and colleagues, for example, argue that memory modification depends on structured inference that determines when new information should update an existing memory versus be stored separately [24]. This explanation contrasts with psychological theories that invoke specialized “immunization” or bias mechanisms [28, 29]. Instead, resistance to updating emerges as a by-product of optimization dynamics.

From Theory to Empirical Test: Challenges and Insights

The third component of this work sought to empirically test the dual-pathway model in humans. I developed a novel psychophysical paradigm, translating relational rules into perceptual comparisons of "more" or "less." By parametrically varying the violation magnitude, I assessed how the strength of contradiction influenced adaptation behavior. The results—a modest but significant effect of violation magnitude on representational switching—provide initial support for the dual-pathway model.

The experimental process also revealed important limitations. An initial experiment design was confounded by carryover of prior task strategies, highlighting differences between isolated network models and human cognition. The second, more successful design, was more controlled but it introduced assumptions about metacognitive reporting and was subject to considerable variability across participants. Moreover, a general bias toward representational switching suggested the influence of additional factors, such as preferences for positive relations [30]. These findings emphasize that while the model might capture a core mechanism, its interaction with broader cognitive systems must be taken into account.

Strengths, Limitations, and Future Directions

A central contribution of this thesis is demonstrating that artificial neural networks can perform abstract reasoning through real-time parameter adaptation, without extensive pre-training. This challenges prevailing assumptions about the necessity of large-scale training for reasoning tasks. The work provides a specific computational mechanism – test-time adaptation – that can be implemented, tested, and compared with human behavior, moving beyond purely descriptive accounts of fluid intelligence.

The iterative progression across studies represents a methodical approach

to theory development: establishing the computational possibility in abstract reasoning tasks, extending the framework to explain a specific paradox in human belief updating, and designing behavioral experiments to test model predictions. While the empirical support remains limited, this progression demonstrates how computational frameworks can generate testable hypotheses about cognitive mechanisms.

The work also illustrates how tools from deep learning can be repurposed to model cognitive phenomena, potentially opening new avenues for understanding adaptive reasoning. However, the biological plausibility and broader applicability of these mechanisms require substantial further investigation.

Nonetheless, important limitations constrain the scope of these contributions. First, the framework isolates a single mechanism in abstraction from the broader architecture of human cognition. Future models should integrate test-time adaptation with systems for working memory, hierarchical planning, and goal-setting to approximate human reasoning more faithfully. Second, the empirical validation presented here is preliminary. The observed effect sizes were small, and several predictions were not confirmed. More sensitive behavioral measures (e.g., eye-tracking), task designs that prevent ceiling effects, and studies incorporating reasoning in more natural domains will be necessary to test the framework's generalizability.

Third, the biological plausibility of test-time adaptation remains limited. The implementation used here abstracts away from neural learning rules, resource constraints, and interactions across brain systems. While the computational approach provides useful insights, bridging to neuroscience will require addressing how such adaptive mechanisms could be implemented in biological networks. Finally, the tasks explored were deliberately simplified; it remains an open question whether inference-time optimization can scale to the richness and diversity of real-world reasoning.

The implications of this work speak directly to the dominant paradigm in modern artificial intelligence: large-scale, pre-trained foundation models.

These models, while powerful, are often static and brittle. Their performance can degrade unpredictably when they encounter data that differs from their vast training sets—a critical vulnerability for systems deployed in the real world, from an autonomous vehicle facing an unforeseen road obstacle to a medical diagnostic tool analyzing a rare disease presentation.

Embedding the principles of test-time adaptation, as explored here, offers a potential solution, promising to make AI systems more robust and context-aware. However, this introduces a profound and practical challenge: do we truly want our most critical systems to learn and change on the fly? For instance, an autonomous vehicle that continuously adapts its driving model could accumulate unsafe behaviors from idiosyncratic local driving habits, leading to unpredictable and potentially dangerous outcomes. The real engineering problem is not simply enabling adaptation but controlling it. Unconstrained adaptation risks catastrophic forgetting, where new learning overwrites essential pre-trained knowledge, or model destabilization, turning a reliable system into an unreliable one.

Therefore, the challenge for future research is to develop frameworks that strike a balance between adaptability and stability—systems that know when and how to update their internal models in a safe, verifiable manner. If such controlled adaptation could achieve competitive performance, it might steer the trajectory of AI development away from a sole reliance on ever-larger models and toward more dynamic, efficient, and ultimately more intelligent systems. While the work presented here is preliminary, it underscores that moving beyond static AI is a crucial next step, though demonstrating this potential safely and effectively will require substantial advances.

Conclusion

In sum, this thesis has argued that fluid intelligence can be understood as an active process of real-time adaptation. The studies presented here

provide initial evidence that adapting artificial neural networks in real-time offers a useful computational framework for understanding abstract reasoning and belief updating. While the empirical findings are preliminary, they demonstrate the value of process-level models that can generate testable predictions and connect mechanistic principles with human behavior. Taken together, this work suggests that adaptation in real-time is a central principle of intelligence – one that may ultimately help bridge the gap between artificial and human minds.

Bibliography

[1] C. Spearman. 'General intelligence,' objectively determined and measured. *The American Journal of Psychology*, 15(2):201–293, 1904.

[2] Robert J. Sternberg. Components of human intelligence. *Cognition*, 15(1-3):1–48, 1983.

[3] David F. Lohman. Complex Information Processing and Intelligence. In Robert J. Sternberg, editor, *Handbook of Intelligence*, pages 285–340. Cambridge University Press, Cambridge, 2000.

[4] OpenAI. GPT-4 Technical Report, March 2023.

[5] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019.

[6] R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve, September 2023.

[7] Celeste Biever. ChatGPT broke the Turing test — the race is on for new ways to assess AI. *Nature*, 619(7971):686–689, July 2023.

[8] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. arXiv:2001.08361 [cs].

[9] Peter V. Coveney and Sauro Succi. The wall confronting large language models, July 2025. arXiv:2507.19703 [cs] version: 2.

[10] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective, December 2023. arXiv:2305.15408 [cs].

[11] Eran Malach. Auto-Regressive Next-Token Predictors are Universal Learners, July 2024. arXiv:2309.06979 [cs].

[12] Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. Is Chain-of-Thought Reasoning of LLMs a Mirage? A Data Distribution Lens, August 2025. arXiv:2508.01191 [cs].

[13] Jian Liang, Ran He, and Tieniu Tan. A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts. *International Journal of Computer Vision*, 133(1):31–64, January 2025.

[14] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts, July 2020. arXiv:1909.13231 [cs].

[15] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-time Adaptation by Entropy Minimization, March 2021. arXiv:2006.10726 [cs].

[16] Léon Bottou. Online algorithms and stochastic approximations. *Online learning in neural networks*, 1998. Publisher: Cambridge University Press.

[17] Tomer Barak and Yonatan Loewenstein. Untrained neural networks can demonstrate memorization-independent abstract reasoning. *Scientific Reports*, 14(1):27249, November 2024. Publisher: Nature Publishing Group.

[18] Nate Kornell and Robert A. Bjork. Learning Concepts and Categories: Is Spacing the "Enemy of Induction"? *Psychological Science*, 19(6):585–592, June 2008.

[19] Matthias Brunmair and Tobias Richter. Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, 145(11):1029–1052, 2019.

[20] Stuart G. Spicer, Chris J. Mitchell, Andy J. Wills, and Peter M. Jones. Theory protection in associative learning: Humans maintain certain beliefs in a manner that violates prediction error. *Journal of Experimental Psychology: Animal Learning and Cognition*, 46(2):151–161, 2020.

[21] E. J. Hird, C. Charalambous, W. El-Deredy, A. K. P. Jones, and D. Talmi. Boundary effects of expectation in human pain perception. *Scientific Reports*, 9(1):9443, July 2019.

[22] Alex Filipowicz, Derick Valadao, Britt Anderson, and James Danckert. Rejecting outliers: Surprising changes do not always improve belief updating. *Decision*, 5(3):165–176, 2018.

[23] Tomer Barak and Yonatan Loewenstein. Two pathways to resolve relational inconsistencies. *Scientific Reports*, 15(1):30738, August 2025. Publisher: Nature Publishing Group.

[24] Samuel J. Gershman, Marie-H. Monfils, Kenneth A. Norman, and Yael Niv. The computational nature of memory modification. *eLife*, 6:e23763, March 2017.

[25] Jeremy R. Gray, Christopher F. Chabris, and Todd S. Braver. Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, 6(3):316–322, March 2003.

[26] Farshad Alizadeh Mansouri, David J. Freedman, and Mark J. Buckley. Emergence of abstract rules in the primate brain. *Nature Reviews Neuroscience*, 21(11):595–610, November 2020.

[27] E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202, 2001.

[28] Charles G. Lord, Lee Ross, and Mark R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109, 1979.

[29] Christian Panitz, Dominik Endres, Merle Buchholz, Zahra Khosrowtaj, Matthias F. J. Sperl, Erik M. Mueller, Anna Schubö, Alexander C. Schütz, Sarah Teige-Mocigemba, and Martin Pinquart. A Revised Framework for the Investigation of Expectation Update Versus Maintenance in the Context of Expectation Violations: The ViolEx 2.0 Model. *Frontiers in Psychology*, 12, November 2021.

[30] Isabelle Deschamps, Galit Agmon, Yonatan Loewenstein, and Yosef Grodzinsky. The processing of polar quantifiers, and numerosity perception. *Cognition*, 143:115–128, October 2015.

להצליח ללא הסתמכות על שינון ובכך מאתגר את התפיסה הרווחת לפיה יכולות כאלו דורשות ידע נרחב שנאגר מהזיכרון, ומבסס את הרעיון לפיו אדפטציה בזמן אמת מסוגלת למדל בהצלחה אינטליגנציה נוזלית.

המחקר השני מיישם מסגרת זו על זרם של קלטים כדי לבחון את כוחה ההסברי מול ממצא פרדוקסלי בלמידה אנושית: הפרות קיצוניות של ציפיות יכולות לעכב, ולא לקדם, עדכון של הציפיות. אני מייחס תופעה זו לתחרות הטבועה בארכיטקטורה של המודל, אשר מפרידה מבנית בין הפרמטרים המקודדים קלטים חושיים לבין אלה המקודדים ציפיות. כאשר תצפית אינה עולה בקנה אחד עם הציפייה, ארכיטקטורה זו מציבה בחירה: לשנות את הציפייה עצמה או לשנות את ייצוג הקלטים. התוצאות מראות שדינמיקת האופטימיזציה בזמן אמת של המודל מכריעה באופן טבעי בבחירה זו: הפרות מתונות מניעות עדכונים בפרמטרים של הציפייה, בעוד הפרות קיצוניות מעדיפות שינוי בפרמטרים של ייצוג הקלטים, ובכך משמרות את הציפייה הראשונית.

המחקר האחרון מבסס את המסגרת החישובית שלי על ידי בחינה אמפירית של הניבוי המרכזי שלה בבני אדם. בדקתי האם אנשים מראים הטיה שיטתית באופן שבו הם מסתגלים להפרה של ציפיות יחסיות. בהתבסס על המודל שלי, שיערתי שעוצמת ההפרה תעצב את התגובה האדפטיבית, תטה משתתפים לאחת משתי אסטרטגיות: עדכון הציפייה היחסית או פירוש מחדש של הקלט. כדי לבחון זאת, עיצבתי ניסוי פסיכופיזי. ההתנהגות האנושית הייתה עקבית באופן איכותני עם הדינמיקה של המודל: האפקט היה צנוע, אך הפרות גדולות יותר הגבירו באופן מובהק את הסבירות שמשתתפים יפרשו מחדש את הקלט, כפי שניבא המודל. ממצאים אלה מספקים תמיכה אמפירית ראשונית אך חשובה למודל שלי, ובה בעת מדגישים את האתגרים במיפוי של מנגנונים חישוביים לקוגניציה האנושית.

לסיכום, תזה זו מציעה שאדפטציה בזמן אמת היא עיקרון מפתח להסקה מופשטת ולדינמיקה של ציפיות. יש לכך שתי השלכות מרכזיות: עבור מדע הקוגניציה, היא מציעה מודל מכניסטי לאינטליגנציה נוזלית, ועבור בינה מלאכותית, היא תומכת בהתגברות על שבריריות המודלים על ידי הוספת היכולת לאדפטציה בזמן אמת במודלים שאומנו מראש. יחד, התוצאות מציעות שאדפטציה בזמן אמת היא היבט מרכזי של אינטליגנציה, טבעית ומלאכותית כאחד.

תקציר

אינטליגנציה נוזלית – היכולת לפתור בעיות חדשות ללא הסתמכות על ידע קודם – היא תכונה מובהקת של הקוגניציה האנושית, אך המנגנונים העומדים בבסיסה עדיין אינם מובנים במלואם. בחינה לצורך העניין מערכות בינה מלאכותית מודרניות כמודלים של אינטליגנציה מגלה שאף שמערכות אלו מפגינות ביצועים מרשימים במגוון משימות, ועשויות להיראות כאילו מחזיקות באינטליגנציה נוזלית, עדיין לא ברור אם הן מחזיקות באינטליגנציה נוזלית אמיתית.

ספקנות זו נובעת משתי מגבלות עיקריות של מערכות בינה מלאכותית: ראשית, הן תלויות במאגרי נתונים עצומים: ילדים אנושיים רוכשים יכולות דומות עם הרבה פחות דוגמאות. שנית, השבריריות שלהן: הן נכשלות לעיתים קרובות כאשר מופיעים קלטים שונים במעט מהפורמט עליו התאמנו. חסרונות אלו מרמזים שמודלים אלו פועלים בעיקר כמתאימי־תבניות מתוחכמים, המזכירים יותר אינטליגנציה גבישית ופחות את הגמישות של אינטליגנציה נוזלית אנושית.

תזה זו בוחנת את הרעיון שאינטליגנציה נוזלית עשויה לנבוע מיכולתה של מערכת להתאים את המבנה הפנימי שלה תוך כדי פתרון בעיות, אדפטציה בזמן אמת. אני מפרמל רעיון זה באמצעות מסגרת חישובית שבה הפרמטרים של רשת נוירונים מלאכותית עוברים אופטימיזציה בזמן אמת. אני בוחן הן את המקרה הקיצוני, שבו האופטימיזציה היא עבור בעיה בודדת)הסתגלות בזמן היסקו, והן את המקרה בו ישנו זרם של בעיות, מה שמציב את המודל בפרדיגמת למידת אונליין.

תזה זו נפרשת על פני שלושה מחקרים. הראשון מדגים כי רשת נוירונים, המאותחלת עם פרמטרים אקראיים ולכן חסרה כל אימון קודם, יכולה לפתור משימות הסקה הדומות למבחני אינטליגנציה אנושיים. במקרה קיצוני זה, הרשת מתאימה את הפרמטרים שלה תוך שימוש רק במידע שבתוך בעיה בודדת. מחקר זה מראה כי הסקה מופשטת יכולה במקרים מסוימים

עבודה זו נעשתה בהדרכתו
של פרופסור יונתן לוינשטיין

האוניברסיטה העברית בירושלים
מרכז אדמונד ולילי ספרא למדעי המוח

מידול אינטליגנציה נוזלית באמצעות אדפטציה בזמן אמת

מוגש על ידי תומר ברק

חיבור לשם קבלת תואר דוקטור לפילוסופיה

הוגש לסנט האוניברסיטה העברית בירושלים
ספטמבר 2025