PNAS RESEARCH ARTICLE | NEUROSCIENCE APPLIED MATHEMATICS

A mathematical theory of relational generalization in transitive inference

Samuel Lippl[a,b,c,1], Kenneth Kay[a,b,d], Greg Jensen[a,c,e], Vincent P. Ferrera[a,c,f], and L. F. Abbott[a,b,c]

Edited by James McClelland, Stanford University, Stanford, CA; received August 22, 2023; accepted May 30, 2024

Humans and animals routinely infer relations between different items or events and generalize these relations to novel combinations of items. This allows them to respond appropriately to radically novel circumstances and is fundamental to advanced cognition. However, how learning systems (including the brain) can implement the necessary inductive biases has been unclear. We investigated transitive inference (TI), a classic relational task paradigm in which subjects must learn a relation ($A > B$ and $B > C$) and generalize it to new combinations of items ($A > C$). Through mathematical analysis, we found that a broad range of biologically relevant learning models (e.g. gradient flow or ridge regression) perform TI successfully and recapitulate signature behavioral patterns long observed in living subjects. First, we found that models with item-wise additive representations automatically encode transitive relations. Second, for more general representations, a single scalar "conjunctivity factor" determines model behavior on TI and, further, the principle of norm minimization (a standard statistical inductive bias) enables models with fixed, partly conjunctive representations to generalize transitively. Finally, neural networks in the "rich regime," which enables representation learning and improves generalization on many tasks, unexpectedly show poor generalization and anomalous behavior on TI. We find that such networks implement a form of norm minimization (over hidden weights) that yields a local encoding mechanism lacking transitivity. Our findings show how minimal statistical learning principles give rise to a classical relational inductive bias (transitivity), explain empirically observed behaviors, and establish a formal approach to understanding the neural basis of relational abstraction.

relational learning | compositional generalization | transitive inference | relational representations

Significance

The ability to infer how elements are related is fundamental to our cognition: when we encounter new circumstances composed of familiar elements, grasping relationships helps us generalize. An important instance is transitive inference (TI): if we know that $A > B$ and $B > C$, we can infer that $A > C$. However, it has been unclear how the brain (and other learning systems) implement such relational generalizations. Here, we investigated artificial learning systems (such as neural networks) that do not have transitivity built in. Remarkably, we found that they perform TI and show behaviors seen in humans and animals. Our findings explain how simple learning models can implement the kind of relational generalization that is essential for successful behavior.

Humans and animals have a remarkable ability to generalize to circumstances that are radically different from their prior experience. They are able to do so, in part, by learning relationships between different events or items and extending these relationships to novel combinations of components (1). Such relational generalization is important across a broad range of domains: for example, subjects can reason about social relationships between individuals they have never seen interact (2), take a novel route between familiar locations (3), or apply familiar tools to novel problems (4). Accordingly, relational cognition has been implicated in a variety of cognitive abilities, including social cognition (5), spatial navigation (6), and logical and causal cognition (7).

It is unclear how living subjects, and learning systems more generally, can learn the kinds of abstractions needed for relational generalization. To generalize from limited experience, subjects (whether they are humans, animals, or learning models) need an "inductive bias": a disposition toward certain behaviors among the many that are consistent with past experience (8). Much attention has been devoted to clarifying suitable inductive biases on standard statistical tasks that require generalization to nearby data points ("near transfer" or "in-distribution generalization") (9, 10). An important instance of such a "statistical inductive bias" is norm minimization, which selects the model parameters with the smallest weight norm and is at the core of many successful learning models (11–13). Both theoretical and practical insights suggest that such a solution is likely to generalize well in distribution (14, 15). In contrast, our understanding of how learning models perform relational tasks, which require generalization to radically different circumstances ("far transfer" or "out-of-distribution generalization"), has been much more limited (16, 17). Addressing this gap is essential to understanding how living subjects perform these tasks.

To investigate this question, we studied transitive inference (TI; Fig. 1*A*; 18–20), a classical cognitive task that tests whether humans or animals can generalize transitively. In this task, subjects are presented with pairs of items (Fig. 1*B*) and must pick the "larger" item according to an implicit hierarchy ($A > B > \cdots > G$, Fig. 1*A*).

Author affiliations: [a]Mortimer B. Zuckerman Mind Brain Behavior Institute, Department of Neuroscience, Columbia University, New York, NY 10027; [b]Center for Theoretical Neuroscience, Department of Neuroscience, Columbia University, New York, NY 10027; [c]Department of Neuroscience, Columbia University Medical Center, New York, NY 10032; [d]Grossman Center for the Statistics of Mind, Columbia University, New York, NY 10027; [e]Department of Psychology, Reed College, Portland, OR 97202; and [f]Department of Psychiatry, Columbia University Medical Center, New York, NY 10032

Author contributions: S.L., K.K., and L.F.A. designed research; S.L. performed research; S.L. analyzed data; and S.L., K.K., G.J., V.P.F., and L.F.A. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: samuel.lippl@columbia.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2314511121/-/DCSupplemental.

Published July 5, 2024.

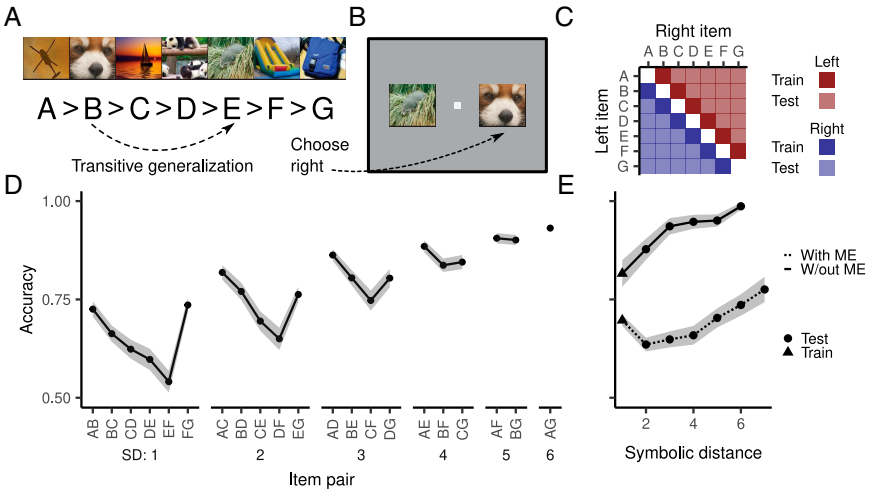

Fig. 1. Transitive inference and behavioral patterns observed in subjects performing this task. (*A*) Example stimuli taken from ref. 23. An example generalization ($B > E$) is highlighted. (*B*) In a given trial, the subject is asked to choose between the two presented items. They are rewarded if the chosen item is larger according to the underlying hierarchy. This panel depicts a test trial where successful performance would consist in picking the item on the right. (*C*) Schematic of the training and test cases. (*D*) Example accuracy on all training and test cases (in this case by rhesus macaques). Terminal item, symbolic distance, and asymmetry effect are apparent in the plot. In the subsequent figures, we leave off the item pair labels but use the same ordering. Symbolic distance (SD) is the separation in the rank hierarchy. The data are reproduced from ref. 23. (*E*) Symbolic distance effect with and without a memorization effect (ME). Data without ME reproduced from ref. 23 and data with ME reproduced from ref. 21. Shaded regions in both panels indicate mean ± one SE.

Importantly, subjects are not informed of this task structure (cf. refs. 21 and 22) and, further, only receive feedback about the correct response on adjacent pairs (*AB, BC, . . . , FG*). Hence, they must infer the underlying relation and use transitivity to determine the correct response to nonadjacent pairs ($B > E$, Fig. 1*C*).

Notably, humans (24) and a variety of animals [ranging from monkeys (25) and rodents (26) to wasps (27) and fish (28)] perform TI successfully. Moreover, they show consistent behavioral patterns (in their reaction time and accuracy, Fig. 1*D*), which may be informative of the underlying neural implementation (19, 20). First, subjects' performance improves with increasing separation in the hierarchy ("symbolic distance effect," e.g., 24, 29, 30). Second, subjects' performance tends to be better for trials involving more terminal (e.g., A and G) rather than more intermediate (e.g., D and E) items ("terminal item effect," e.g., refs. 31 and 32). Third, some studies indicate better performance on training trials (which have a symbolic distance of 1) than on test trials with a symbolic distance of 2 (21, 33, 34, Fig. 1*E*), a limited violation of the symbolic distance effect which we here refer to as the memorization effect. Finally, subjects' performance is often better for item pairs toward the beginning of the hierarchy than item pairs toward the end (the "asymmetry effect," which we do not address but return to in the discussion) (e.g., ref. 21).

Many simple learning models have been shown to perform TI. However, the inductive biases that give rise to this ability are not well understood. Various learning models associate a numerical "rank" with each presented item and choose whichever item has a higher rank (21, 23, 35). Such models are preconfigured to generalize transitively and leave unclear how transitive generalization could arise from more basic learning principles. They also leave unclear how the brain, which is not constrained in this way (36), could implement TI. Intriguingly, several studies have found that generic neural networks can generalize transitively, suggesting that statistical learning principles can sometimes give rise to a suitable relational inductive bias (22, 37–39). However, TI in neural networks has largely been studied through simulations, rather than analytically (cf., ref. 40), raising the question of when and how statistical learning models can implement TI.

Here we show, via both analytical approaches and simulations, that a broad range of biologically relevant learning models generalize transitively and recapitulate the symbolic distance, terminal item, and memorization effect. We first consider models with "additive" representations that represent the two presented items independently and show that they are constrained to implementing a transitive relation. If models additionally represent nonlinear conjunctions between items (as is important for many other tasks) but use norm minimization to determine their readout weights, they also generalize transitively. Remarkably, the same learning principle that underpins many instances of successful near transfer also enables this instance of successful far transfer. We further show that, for TI, the effect of a particular choice of internal representation can be characterized by a single scalar "conjunctivity factor." Finally, we consider models which adapt their internal representation to a given task, an ability thought critical to human and animal cognition. Surprisingly, we find that this impairs performance on TI and leads to behavioral patterns that deviate from those in living subjects. Notably, this anomalous behavior is explained by a different form of norm minimization, namely one over all weights in the network rather than just the readout weights.

At first glance, TI appears to be a complex task involving relational, rule-based cognition. Nevertheless, we can characterize a broad range of learning models performing this task in exact analytical terms, explaining how they give rise to the rich behavioral patterns observed in living subjects. In doing so, our investigation clarifies systematically how a learning principle that has largely been considered in the context of near transfer, also implements an important instance of far transfer and relational abstraction.

Model Setup

We represent individual items as high-dimensional vectors. A trial input is a concatenation of the two vectors X, Y corresponding to the two presented items. We generally consider a learning model f that represents this input as a numerical vector $g(X, Y)$ (in the simplest case, this could be the input vector itself). The model then computes a linear readout from that representation: $f(X, Y) = w \circ g(\mathrm{X}, \mathrm{Y})$. A positive model output ($f(X, Y) > 0$) corresponds to $X > Y$, whereas a negative output ($f(X, Y) < 0$) corresponds to $X < Y$. We generally assume that w is learned from the training trials, whereas $g(X, Y)$ remains fixed (for example arising from the representation in a neural network with random or prelearned weights). In the final two sections, we investigate models that also learn $g(X, Y)$.

Inputs where $f(X, Y) = 0$ lie on the model's decision boundary. Accordingly, a higher magnitude of the margin ($f(X, Y)$ if $X > Y$ or $-f(X, Y)$ if $X < Y$) corresponds to

a larger distance from the decision boundary. We take this to indicate better performance, corresponding to higher choice accuracy and lower reaction time. This is a standard assumption in decision-making models (22, 41). For example, in a drift diffusion model (42), the output $f(X, Y)$ determines the model's drift rate. A higher drift rate makes the model less susceptible to noise (improving accuracy) and makes it cross its threshold faster (improving reaction time) (43).

Results

An Additive Representation Yields Transitive Generalization. To perform TI, or indeed any kind of relational inference, a learning model's representation of items X and Y, $g(X, Y)$, should reflect the fact that X and Y are separate items. In the most extreme case, this would amount to an additive representation, where $g(X, Y)$ is a sum of two separate representations $g_1(X)$ and $g_2(Y)$: $g(X, Y) = g_1(X) + g_2(Y)$. A simple instance of this is a model architecture where nodes respond exclusively either to X or Y (Fig. 2*A*).

A change in one of the two items will leave half of the additive representation unchanged (Fig. 2*B*), implementing a kind of compositionality (44). A linear readout from an additive representation, $f(X, Y) = w \circ g(X, Y)$, is a sum of responses to each individual item, $w \circ g_1(X) + w \circ g_2(Y)$, and therefore also additive. The consequences of this are especially clear in the case of a model without a choice bias, i.e., if $f(X, X) = w \circ g_1(X) + w \circ g_2(X) = 0$ (*SI Appendix*, section S1.A considers the biased case). In this case, we know that $-w \circ g_1(X) = w \circ g_2(X)$, and, as a result, the model's decision function can be expressed as

$$f(X, Y) = w \circ g_1(X) - w \circ g_1(Y).$$

This means that the model necessarily learns to assign a scalar rank $r(X) = w \circ g_1(X) = -w \circ g_2(X)$ to each item, and computes its decision by comparing the two ranks (Fig. 2*C*):

$$f(X, Y) = r(X) - r(Y).$$

Note that $f(X, Y) > 0$, i.e., the decision to choose the left item, is equivalent to $r(X) > r(Y)$.

As shown in previous work (e.g., refs. 21, 23, and 31), learning systems that are preconfigured to have such a ranking system yield both transitive generalization and the symbolic distance effect. This is because, to learn the training set, the model's ranks must be monotonically decreasing:

$$r(A) > r(B) > \cdots > r(G).$$

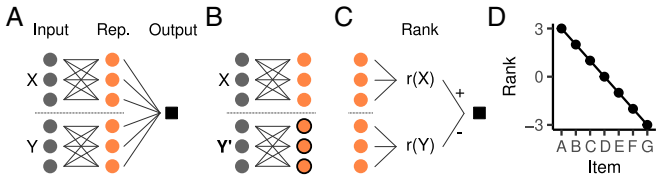

Fig. 2. An additive representation yields transitive generalization and the symbolic distance effect. (*A*) Schematic illustration of an additive representation (rep.). The first three orange nodes represent the first item (X), and the latter three represent the second item (Y). (*B*) Replacing the second item only results in changes to the highlighted half of the units. (*C*) The readout weights of the model can be grouped into those pertaining to item X and those pertaining to item Y. The model's output can be understood as assigning a rank $r(X)$ and $r(Y)$ to each item and then computing $r(X) - r(Y)$. (*D*) Example of a model's rank assignment.

As an example, if the model output has a margin of one for all inputs composed of adjacent items, the rank must decrease by one for each successive item (Fig. 2*D*). A monotonically decreasing rank directly implies that nonadjacent items are also ordered correctly and, consequently, that the model will generalize transitively. Further, item pairs with larger symbolic distance will have a larger difference in their ranks, giving rise to a symbolic distance effect.

Critically, and in distinction to previous work, the above model structure (additive representation) does not explicitly preconfigure or assume a ranking system. Rather, the above analysis shows that any learning model that implements an additive input representation necessarily implements a ranking system and thus encodes a transitive relation. Further, the above analysis makes no assumptions about how the model learns. As long as the model correctly classifies the training cases, it will correctly classify all test cases, i.e., transitively generalize, and show a symbolic distance effect.*

A Single Scalar Fully Characterizes a Broad Range of Relational Representations. The above analysis indicates that an additive representation would enable the brain to perform TI. However, neural representations are not thought to be fully additive (45). Indeed, nonadditive representations (Fig. 3*A*) are important for learning relevant tasks across a broad range of domains and, with some differences in implementation, are known under a correspondingly broad range of names, including conjunctive (46) or configural (47) representations, nonlinear features, and representations with mixed selectivity (45). We next asked whether and how nonadditive representations can support transitive generalization.

To begin, consider the most extreme conjunctive case: a one-hot representation in which each composition of items is represented by a different hidden unit (Fig. 3*B*). In this case, a change in one of the two items yields a completely different representation. Such a model is able to memorize the training cases, but cannot generalize transitively.

Many representations, whether in the brain or in other learning systems, are neither fully conjunctive nor fully additive, but rather lie in between these two extremes (45, 48). To characterize this spectrum formally, we considered the representational similarity between two trials (X, Y) and (X', Y'), as measured by their dot product, $\langle g(X, Y), g(X', Y') \rangle$ (leaving the representations $g(X, Y)$ fixed). We assumed that the representational similarity between two trials only depends on whether these trials are distinct ((X, Y) and (X', Y')), overlapping ((X, Y) and (X', Y) or (X, Y')), or identical ((X, Y) and (X, Y)) (Fig. 3*C*). We call this the "exchangeability assumption."

The exchangeability assumption captures the fact that model behavior should not depend on the particular (i.e., arbitrary) hierarchy in which the set of items is arranged (49). To promote exchangeability, we assumed that all input items are equally correlated with each other. In this case, most commonly used nonlinear representations of that input satisfy exchangeability as well. As a paradigmatic learning model, we considered a neural network with a ReLU nonlinearity and random weights. By determining the expected value of the representational similarity analytically (*SI Appendix*, section S1.C), we found that the network's hidden layer, in expectation, satisfies the exchangeability assumption (Fig. 3*C*). This is because even

*A related paper (39) has come to our attention, in which the authors are concerned with a linear item representation. This is a special case of the additive representations here. Further, the authors derive an analytic solution for a specific learning algorithm, whereas we demonstrate a transitive constraint for a broader class of models.

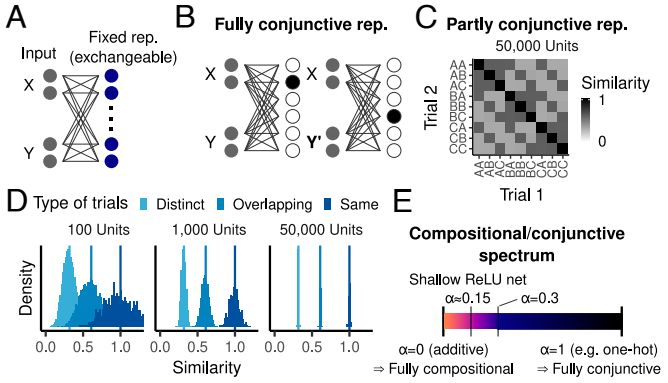

Fig. 3. Nonadditive representations of TI cases. (*A*) A nonadditive representation encodes nonlinear interactions between items. (*B*) A one-hot representation represents each combination of items by a distinct node. (*C*) Representational similarity (cross-correlation) between a subset of trials in a ReLU network with 50,000 units. (*D*) Representational similarity between all possible trials in networks with different numbers of hidden units, organized according to the type of trials. (*E*) The conjunctivity factor characterizes a given representation according to how similarly it represents overlapping trials. Additive representations lie at one end of the spectrum, whereas one-hot representations lie at the other end.

though the hidden layer computes a nonlinear transformation, it partially inherits the input's similarity structure (50). In network simulations, the empirical representational similarity exhibits some variance around the expected value due to the random weight initialization. However, this variance vanishes as the network's hidden layer becomes wider (Fig. 3*D*).

Under the exchangeability assumption, we found that for learning models that only modify their readout weights, a single scalar, which we call the conjunctivity factor $\alpha \in [0, 1]$, fully determines their TI task behavior, i.e., all models that have representations with the same α exhibit the same behavior on TI (*SI Appendix*, section S1.B). The conjunctivity factor is given by

$$\alpha := 1 - 2\frac{\kappa_o - \kappa_d}{\kappa_s - \kappa_d}, \quad \kappa_s = \langle g(X, Y), g(X, Y)\rangle,$$
$$\kappa_o = \langle g(X, Y), g(X, Y')\rangle, \quad \kappa_d = \langle g(X, Y), g(X', Y')\rangle. \qquad [1]$$

Here κ_s, κ_o, and κ_d are the similarity between identical, overlapping, and distinct pairs, respectively. Notably, we can extend the definition of the conjunctivity factor to nonexchangeable representations by taking the average over all identical, overlapping, or distinct pairs.

For an additive representation, half the units share their activation between different overlapping pairs. As a result, the similarity between overlapping pairs is halfway between that of distinct and that of identical pairs, corresponding to $\alpha = 0$. At the other extreme, $\alpha = 1$ indicates that overlapping pairs are encoded in the model with equal similarity to each other as completely distinct pairs (as is the case for a one-hot representation). Consequently, α systematically characterizes the spectrum from fully additive to fully conjunctive representations (Fig. 3*E*).

In most commonly used representations, overlapping trials are more similar to each other than distinct trials and therefore have an intermediate value for α ("partly conjunctive" representations). This is because the input space represents overlapping trials as more similar than distinct trials and, as noted above, the hidden layer partly inherits the input's similarity structure. Through analytical computation, we found that a random ReLU network with one hidden layer, for example, has $\alpha \approx 0.15$ (*SI Appendix*, Eq. **S28**; Fig. 3 *C* and *D*).

Importantly, the network giving rise to the representation could have an arbitrary number of layers; all we need to know is the conjunctivity factor of the network's final layer (see also *SI Appendix*, section S1.C).

For TI, the conjunctivity factor raises two questions. First, how does $\alpha > 0$ affect TI behavior? Second, if the learning model's internal representation is modifiable rather than fixed, how does this affect the conjunctivity factor and subsequent TI behavior?

Norm Minimization and Partly Conjunctive Representations Yield Transitive Generalization. Unlike an additive representation, a representation with $\alpha > 0$ is not constrained to implementing a transitive relation. To understand how models with partly conjunctive representations perform on TI, we need to consider additional constraints. In particular, we analyzed models in which the learning of readout weights implements norm minimization (Fig. 4*A*), a paradigmatic statistical inductive bias (Fig. 4*B*). Under the exchangeability assumption, we were able to characterize model behavior on TI through exact analytical solutions. As noted above, to perform this analysis, we do not need to know the particular representation implemented, only its conjunctivity factor α.

On the training cases, the minimal norm model necessarily assigns a margin of ± 1, as dictated by the desired output. On the test cases, however, our analysis revealed an intriguing emergent behavior: the model's response to item pair (i, j) invariably reflects a ranking system:

$$f(i, j) = r_i(\alpha) - r_j(\alpha),$$
$$\text{with } r_i(\alpha) = \frac{\sinh((\frac{n+1}{2} - i)\lambda(\alpha))}{\sinh(\frac{n+1}{2}\lambda(\alpha)) - \sinh(\frac{n-1}{2}\lambda(\alpha))} \qquad [2]$$
$$\text{and } \lambda(\alpha) := \operatorname{arccosh}\left(\frac{1}{1-\alpha}\right),$$

where n is the total number of items (see *SI Appendix*, section S1.D.3 for the derivation of this result). This is remarkable as the model architecture is not constrained or preconfigured to implement a ranking system. Rather, the behavior is a consequence of the principle of norm minimization, operating on a partly conjunctive representation. Importantly, as implied by its ranking system, the model generalizes transitively (as long as $\alpha < 1$), and exhibits a symbolic distance effect.

For an intuition as to why a ranking system emerges, consider a particular class of representations having one-hot representations of the first and second item individually as well as their conjunction (Fig. 4*C*). This representation will have a conjunctivity factor of α if the item-wise units are weighted by $\sqrt{\frac{1-\alpha}{2}}$ and the conjunctive units are weighted by $\sqrt{\alpha}$. Changes in the item-wise unit weights correspond to changes in the model's rank, as they generalize to overlapping trials. In contrast, changes in the conjunctive unit weights correspond to memorization, as they have no effect on overlapping trials. In principle, the model could learn the training set through changes in the conjunctive unit weights alone. However, because more distributed weights tend to have a smaller norm, norm minimization causes the model to learn the training trials by changing both the conjunctive unit weights (resulting in memorization) and the item-wise unit weights (resulting in transitive generalization). Thus, partial conjunctivity is necessary for the existence of an

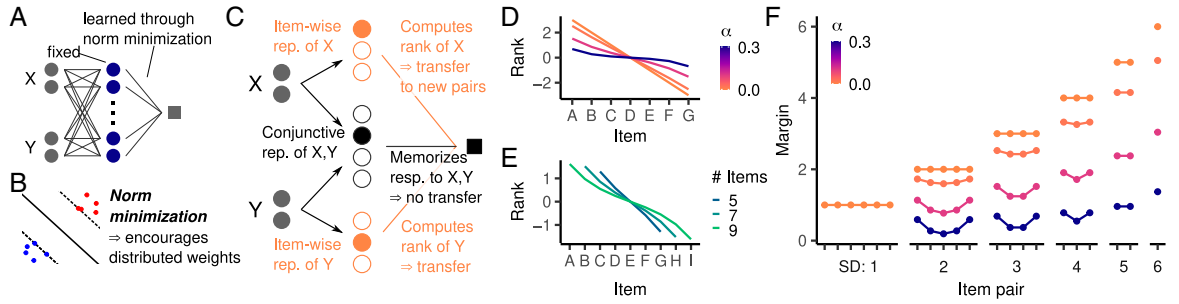

Fig. 4. We analyze the behavior of models having readout weights trained with norm minimization. (*A*) Schematic illustration of the setup. (*B*) Norm minimization implements a useful inductive bias for generalization to nearby data points. On categorization tasks, it determines the hyperplane separating the two categories with the maximal margin. (*C*) Intuitively, a partly conjunctive representation is given by an item-wise representation of *X* and *Y* concatenated with a fully conjunctive representation of *X* and *Y*. The readout from the item-wise representations computes a rank for each item that transfers to overlapping pairs. The readout from the fully conjunctive representation memorizes a response to a given pair and does not transfer to overlapping pairs. Because norm minimization encourages distributed weights, it finds a solution that partly uses the item-wise representation and hence computes a rank. This leads to transitive generalization. (*D* and *E*) The emergent rank representation at the end of training (Eq. **2**) for (*D*) seven items and different values for α; and (*E*) $\alpha = 0.1$ and different numbers of items. (*F*) For seven items and different values for α, the corresponding margin for all trials. Item pairs are arranged by their position in the hierarchy, as in Fig. 1*E*.

item-wise population and norm minimization ascertains that this population is implicated in the learning process.

Beyond transitive generalization, the principle of norm minimization gives rise to several empirically observed behavioral patterns. The hyperbolic sine making up the rank expression compresses more intermediate items more strongly than more terminal items (Fig. 4*D*), thus giving rise to a terminal item effect (Fig. 4*F*). This effect becomes stronger for higher values of α. A higher α also compresses the ranking more strongly overall, leading to lower margins on the test set. As α approaches one (the fully conjunctive case), the ranking becomes entirely flat and therefore no longer supports transitive generalization.

The form of the ranking also depends on the total number of items. Specifically, intermediate items are compressed more strongly when there are more items in total, an effect that is moreover dependent on α. For $\alpha = 0$, the ranking grows linearly with the number of items, whereas at higher values of α, the ranking's overall range (between the first and last item) is nearly invariant to the number of items (Fig. 4*E*).

Finally, when $\alpha > 0$, the model assigns a larger margin to the training cases than specified by the ranking. Intuitively, this is because the conjunctive unit weights contribute to model behavior on the training cases but do not transfer to the test cases. Since a higher α compresses the ranks further, it leads to a higher discrepancy between training and test behavior. At a sufficiently high α, the margin of the training cases is larger than that of the test cases with a small symbolic distance (Fig. 4*F*), giving rise to a memorization effect. While Ciranka et al. (21) explained this effect by fitting an explicit memorization parameter, our analysis reveals that it is an emergent consequence of having a nonlinear representation with sufficiently high conjunctivity factor. Notably, the fact that the memorization effect only arises in a subset of models (i.e., those with relatively conjunctive representations) may explain why it is only occasionally observed in living subjects. In contrast, the symbolic distance and terminal item effect arise across the full spectrum of relational representations (except the fully conjunctive case) and indeed are also observed more consistently in living subjects.

The above analytical solutions depend on the exchangeability assumption. In practice, the trials might be represented in a manner that violates this assumption. Indeed, we already saw that though randomly sampled features satisfy exchangeability in expectation, a model with insufficiently many of those features will have a nonexchangeable representation due to finite samples (Fig. 3*D*). In simulations (*SI Appendix*, section S2.B), we found that a slight violation of exchangeability (e.g., resulting from a representation with many random features) does not change model behavior substantially. For larger violations (e.g., resulting from a representation with fewer random features), model behavior deviates from our theoretical account, but is still well approximated by it. Further, our account captured the average behavior across many models with a small number of random features almost exactly. This suggests that models with nonexchangeable representations behave differently from those with exchangeable representations but our analytical solutions can still be useful for understanding their behavior.

Our analysis demonstrates that norm minimization can explain not only transitive generalization, but also the symbolic distance effect, the terminal item effect (though only on test cases), and the memorization effect. We next characterize two popular statistical learning models implementing norm minimization: ridge regression and gradient flow, as applied to the learning of network readout weights.

Learning through Gradient Flow or with Weight Regularization Smoothly Approaches the Minimal Norm Solution. To see how the principle of norm minimization governs TI behavior across learning, we analyzed models with readout weights learned through either ridge regression (Fig. 5*A*; ref. 11) or gradient flow, using mean squared error as the loss function $L(w)$ (with the target response on the training cases being ± 1). Ridge regression minimizes the sum of this loss and the squared L_2-norm of the model weights, i.e., $L(w) + \frac{1}{c}\|w\|_2^2$.

Here, the regularization coefficient c balances the weight penalty $\|w\|_2^2$ with the minimization of the loss function $L(w)$. Gradient flow, on the other hand, assumes that a model minimizes $L(w)$ by following the pointwise gradient exactly. This approximates gradient descent with a small learning rate and is more amenable to formal analysis (51). In this case, the learning duration t determines how well the model has learned to minimize the loss function in the allotted time. At initialization, the model should be agnostic to all choices (i.e., output zero) and we therefore assumed that the weights are initialized at zero.

The regularization coefficient c and the learning duration t play a similar role in the two learning models. With $c = 0$, the weight penalization is infinitely more important than the task-based

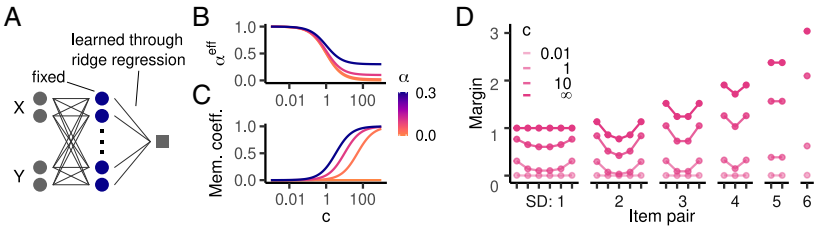

Fig. 5. TI behavior for models with a fixed representation and readout weights trained using regularized regression. (*A*) Illustration of the setup. (*B* and *C*) The (*B*) effective conjunctivity factor and (*C*) memorization coefficient as a function of the inverse regularization coefficient *c*. (*D*) Generalization behavior for $\alpha = 0.1$ and different values of *c*. The margins overall become larger as *c* increases.

component of the loss function and therefore the model weights are all zero. Similarly, at $t = 0$, the weights are initialized to zero. On the other hand, in the limit of infinite training ($t \to \infty$), the model converges to the minimal norm solution determined in the section above (52). This is also the case for the limit of models trained with increasingly small weight penalization ($c \to \infty$) (53). Ridge regression and gradient flow are therefore two instances of common learning models which can perform TI by implementing the principle of norm minimization.

Going beyond these limits, we obtained exact solutions to model behavior for arbitrary t or c. For ridge regression (*SI Appendix*, section S1.D.3), we found that the test behavior of a model with weight regularization and a given conjunctivity factor α is equivalent to that of a model without weight regularization but with a different, "effective" conjunctivity factor α^{eff}, which depends on both α and c (*SI Appendix*, Lemma S1.4): $\alpha^{\text{eff}} = (\alpha + \frac{1}{c})/(1 + \frac{1}{c})$. α^{eff} is generally higher (i.e., more conjunctive) than α and, as c becomes larger, gradually approaches α from above (Fig. 5*B*). Thus, a model with smaller c has a more compressed rank and a more pronounced terminal item effect.

On the training cases, the model's performance is boosted compared to the rank difference, just as in the case of norm minimization. Specifically, its behavior is given by

$$f(i, i \pm 1) = m \cdot \pm 1 + (1 - m) \cdot (r_i(\alpha^{\text{eff}}) - r_{i\pm1}(\alpha^{\text{eff}})),$$
$$m := \frac{\alpha}{\alpha + \frac{1}{c}}, \qquad [3]$$

i.e., a mixture of memorization (assigning a constant margin of ± 1) and reliance on the ranking system. The balance between the two behaviors is specified by the "memorization coefficient" $m \in [0, 1]$. m starts out at zero, indicating no memorization and full reliance on the ranking system. As c grows larger, m increases as well (Fig. 5*C*). The smaller m, the more strongly model behavior relies on the ranking. In particular, this partial reliance leads to a terminal item effect on the training cases in addition to the test cases (Fig. 5*D*). For $c \to \infty$, m converges to 1, indicating full memorization as observed for the minimal norm solution.

For gradient flow, we found qualitatively similar solutions for model behavior (*SI Appendix*, section S1.D.4). In particular, the model's behavior on test cases can be described by a ranking system throughout learning. Further, the model ranks gradually approach the minimal norm solution and, at earlier stages of training, have a more pronounced terminal item effect. Finally, the model has a transient terminal item effect on the training cases that vanishes as $t \to \infty$.

Humans and animals generally exhibit a terminal item effect on the training set. Our analysis suggests that this could be caused by either weight regularization or incomplete training (or both), arising from a mechanism that is related to but distinct from the mechanism giving rise to the terminal item effect on the test cases.

The Conjunctivity Factor Exposes a Tradeoff between Learning Transitive and Nontransitive Relational Tasks. Our analysis thus far indicates that a higher conjunctivity factor generally yields worse performance on TI. Indeed, if maximal generalization performance on TI were the sole aim, models with fully additive representations would be ideal. However, as noted above, humans and animals learn a broad range of relational tasks, not all of them transitive. Because models with fully additive representations are constrained to implementing a transitive relation, this makes partially conjunctive representations necessary (*SI Appendix*, Fig. S1*A*).

To investigate how different representational geometries change model behavior beyond TI, we considered transverse patterning ("rock, paper, scissors" with more than three items, e.g.: $A > B, B > C, \ldots, F > G, G > A$). This task exemplifies a nontransitive relation and can be learned by both humans and various animals (36, 54, 55).

As with TI, we considered models with fixed, exchangeable representations with readout weights trained through gradient flow (*SI Appendix*, Fig. S1*A*). We found that as long as $\alpha > 0$, models were able to learn the task by relying on the conjunctive population. Further, by solving the learning dynamics of gradient descent analytically (*SI Appendix*, section S1.F), we found that a higher α leads to faster learning of the training trials (*SI Appendix*, Fig. S1 *B* and *C*). In contrast, on TI, a higher α causes such models to have a smaller test margin. These behaviors highlight a potential tradeoff that α imposes across different kinds of relations. In particular, our analysis predicts that subjects who are better at TI should be slower to learn transverse patterning and vice versa.

One strategy for avoiding the tradeoff described above is representation learning. If models were able to adapt their internal representations to a given task, they could in principle learn an additive representation for TI and a nonadditive representation for nontransitive relational tasks such as transverse patterning. We now investigate this hypothesis.

Neural Networks with Adaptive Representations Show Anomalous TI Behavior. Deep neural networks, which learn by updating their internal weights, have become increasingly relevant both as artificial learning systems (56) and models of cognitive processes (57, 58). Importantly, recent work has revealed that their generalization behavior fundamentally depends on the magnitude of their initial weights. For large initial weights, learning dynamics can be approximated by gradient flow on a particular fixed-feature model called the neural tangent kernel (NTK; 59). Thus, even though the network updates its internal weights, it effectively still relies on a fixed representation. Accordingly, this regime is often called "lazy" (60). In contrast, neural networks initialized from sufficiently small values learn truly task-specific representations ("rich regime"). Broadly considered, adapting a model's internal representation to a given task could address the competing demands imposed by the wide range of tasks subjects need to perform.

Indeed, the rich regime is seen as essential to the remarkable generalization capabilities of deep neural networks (60–62). There is also widespread evidence that subjects adapt their representation to a given task (63) and may do so similarly to deep neural networks trained in the rich regime (64). In relational tasks, in particular, neural representations change to reflect how different items are related to each other (65–67), and this may be essential for successful generalization on certain tasks (22, 68). In light of this lazy vs. rich distinction and its potential relevance to biological learning, we investigated through simulations whether deep neural networks are suitable as a model of relational representation learning (Fig. 6*A*). Specifically, we trained neural networks from different scales of initialization using gradient descent (for details on training, see *SI Appendix*, section S2.C). In light of the important role played by the initialization scale, we covered a broad range of potential values, focusing on three representative values: 1 (resulting in lazy behavior), 10^{-3} (resulting in rich behavior), 10^{-16} (to characterize model behavior in the limit of small initialization).

In the lazy regime, we computed the NTK's conjunctivity factor analytically (*SI Appendix*, section S1.E). For large initialization, the learning trajectory of neural networks trained with full weight updating is approximated by gradient flow on the fixed NTK representation. Hence we were able to use the gradient flow solutions determined in the previous sections to predict the network behavior over the course of training, finding a virtually perfect match with simulations (Fig. 6 *B* and *E*, green line). Our account may therefore be able to explain why previous studies (37, 38) found empirically that feedforward neural networks generalize transitively.

Surprisingly, we found that networks trained in the rich regime performed worse at TI. They had a smaller test margin (Fig. 6*C*) and also made systematic errors on the cases CE, BE, and CF for a sufficiently small initialization scale (Fig. 6 *D* and *E*, hollow points). Further, and in contrast to all learning models considered thus far, the behavior of neural networks in the rich regime was not consistent with a ranking system. This is apparent from the fact that the networks' margin at a symbolic distance of three was smaller than at a symbolic distance of two (Fig. 6*E*). Importantly, in contrast to the previous limited violation of the symbolic distance effect, this cannot be explained by memorization as none of these cases are in the training set. Finally, the networks in the rich regime exhibited a violation of the terminal item effect at a symbolic distance of 4, a somewhat surprising pattern that no models examined thus far have exhibited. The networks' unconventional behavior was not due to the specific setup considered here: we observed similar behavior for alternative activation functions (*SI Appendix*, Fig. S6), alternative loss functions (*SI Appendix*, Fig. S8*A*), and deeper networks (*SI Appendix*, Fig. S8*B*). Further, the networks exhibited even more overtly idiosyncratic behavior for larger numbers of items, one striking behavioral pattern being a periodic (rather than monotonically increasing) symbolic distance effect (*SI Appendix*, Fig. S7). These findings indicate that, for TI, representation learning in standard neural networks does not necessarily confer the benefits of representation learning suggested in prior work.

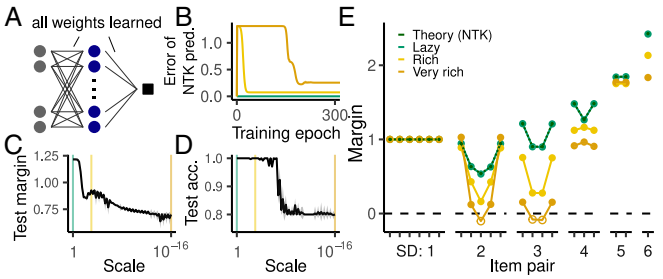

Fig. 6. TI behavior in ReLU networks with one hidden layer and 50,000 units that are trained through backpropagation. Shaded regions (sometimes too small to be visible) indicate mean ± one SD across twenty instances. (*A*) Illustration of network training. In contrast to the previous setups, the hidden layer weights were also trained using backpropagation. (*B*) The mean squared error of the prediction made by the NTK at three different initialization scales (lazy: 1; rich: 10^{-3}; very rich: 10^{-16}). (*C* and *D*) The (*C*) average test margin and (*D*) test accuracy as a function of initialization scale. The colored lines highlight the three representative values analyzed in more detail in panels *B* and *E*. (*E*) TI performance according to our NTK-based prediction as well as of networks trained with backpropagation at the three representative scales.

Rich-Regime Networks Implement a Cooperative Code that Lacks a Transitive Inductive Bias. Given that the rich regime has been found previously to improve generalization on other tasks, we sought to understand why it yields anomalous behavior on TI. To this end, we leveraged previous work indicating that lazy and rich regimes implement different forms of norm minimization: the lazy regime minimizes the ℓ_2-norm of the network's readout weights, whereas the rich regime approximately minimizes the ℓ_2-norm of all weights in the network (69, 70; but see refs. 71 and 72). For the networks studied here, we found that the norm of all weights in a fully trained network is dramatically smaller for smaller scales of initialization (Fig. 7*A*). To clarify why norm minimization over all network weights is associated with an anomalous inductive bias (unlike norm minimization over readout weights, analyzed in the previous sections), we directly analyzed the computations performed by the rich-regime neural network.

Prior studies (70, 73, 74) have found that minimization of the ℓ_2-norm over all network weights induces the networks' hidden units to "specialize" into a low number of functionally distinct clusters with low ℓ_2-norm. To assess this possible structure, we performed k-means clustering with respect to the normalized weight vectors across all units, finding that in the rich regime, all 50,000 units of the network fall into just six clusters (Fig. 7*B*; details described in *SI Appendix*, section S3.C). In contrast, such a compact description was not apparent in the lazy regime. Remarkably, the six cluster centroids were highly consistent across random initializations (Fig. 7*C*).

We found that three of the centroids ("units") had positive readout weights, whereas the other three units had negative readout weights. This is because the network has rectified activations in its hidden layer, which caused different units to specialize for trials with positive and negative labels. In examining hidden-layer weights, we focused on the positive units, which we denote by E_{1+}, E_{2+}, and E_{3+} (negative units are analogous, *SI Appendix*, Fig. S11). Because we presented the network with concatenated one-hot vectors, each weight entry corresponds to a different presented item and we identify each weight by its corresponding item. Note that items presented in the first and second position correspond to entirely distinct weights; we denote item X in the first or second position by $X^{(1)}$ and $X^{(2)}$, respectively.

We noticed two putative aspects of the underlying computation in the network: different units responded predominantly to a nonoverlapping set of trials ("staggered response"; e.g., E_{1+} (pink) responded to AB, BC, EF, and FG, while E_{2+} (blue) responded to CD and E_{3+} (green) responded to DE; Fig. 7 *C*, *Bottom Left*) and each unit encoded its corresponding trials

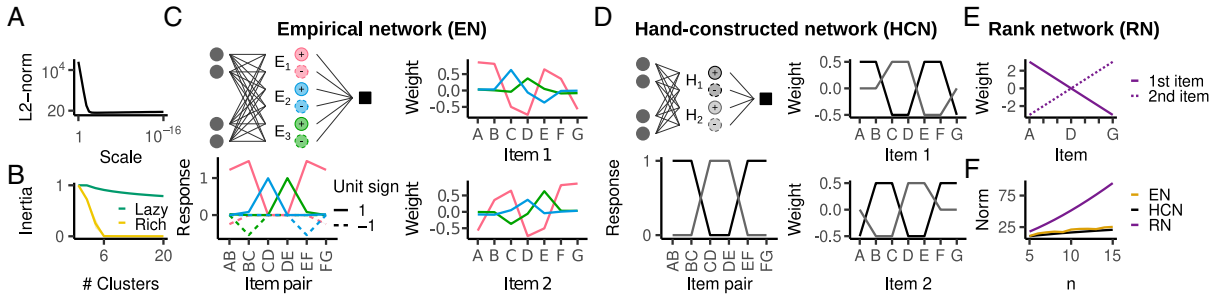

Fig. 7. A mechanistic analysis of the rich regime's inductive bias. All plots show the mean ± SD across twenty random initializations. SD is too small to be visible. (*A*) ℓ_2-norm of all weights in the fully trained networks as a function of initialization scale. The rich regime yields much smaller norm. (*B*) Inertia (i.e., proportion of explained variance) as a function of the number of clusters, for the lazy and rich network. Six clusters only leave 0.001% of the variance unexplained. (*C*) The empirical network is therefore described by a network with six units, three with positive and three with negative responses. The *Right* panels show the weights of the different units and the *Bottom* panel shows how each units responds to the different training trials. Only units with positive readout weights are shown; the units with negative readout weights have the same structure but with Items 1 and 2 reversed. (*D*) Analogous depiction of the hand-constructed network (HCN), which has four units. (*E*) The rank network only has two units, but they span a much wider range. (*F*) Weight norm of the empirical, hand-constructed, and rank network as a function of the number of items. The rank network has a much larger ℓ_2-norm, whereas the norm of the empirical network is similar to that of the HCN.

by distributing positive weights across both items ("cooperative code"; e.g., to encode AB, E_{1+} assigned a positive weight to $A^{(1)}$ and $B^{(2)}$; Fig. 7 *C*, *Right*).

To see whether these two aspects could provide a sufficient explanation for the behavior in the rich regime, we used them to hand-construct a simplified ReLU network (Fig. 7*D*). Specifically, the hand-constructed network (HCN) has four hidden units, two with positive readout weights and two with negative readout weights. Again, we focus on those with positive readout weights, which we denote by H_{1+} and H_{2+} (see *SI Appendix*, Fig. S10 for the analogous negative units). Note that we do not prove that the HCN actually learns the training trials with minimal ℓ_2-norm (though it has the lowest norm among all networks considered here). Rather, it serves as a useful construction to understand why the described computation yields both low norm and a nontransitive inductive bias.

Each unit in our construction implements a cooperative code, i.e., to encode its response to a trial, it distributes its weights equally between the two items. In particular, H_{1+} classifies trials AB and BC by assigning a weight of 0.5 to $A^{(1)}$ and $B^{(2)}$ as well as $B^{(1)}$ and $C^{(2)}$. Because of the positive weights associated with these items, H_{1+} would also respond positively to trials BA, CB, and DC. Preventing this positive response requires negative weights associated with items $A^{(2)}$, $C^{(1)}$, and $D^{(1)}$. However, because a negative weight is already associated with $C^{(1)}$ and $D^{(1)}$, H_{1+} cannot classify trials CD and DE using a cooperative code. Thus, H_{1+} stays silent on these trials, which are instead encoded by H_{2+} (also using a cooperative code). This pattern explains the staggered responses observed in the empirical network. The interference from the negative weights associated with $E^{(1)}$ and $F^{(1)}$ prevents H_{2+} from classifying EF and FG, but these trials are no longer affected by interference with AB and BC and can therefore be encoded by H_{1+}. For TI variants with more than seven items, the two units continue alternating, giving rise to periodic network weights (*SI Appendix*, Fig. S10).

Intuitively, the coding scheme amounts to a set of local rules that support learning of the training trials, but do not generalize to the test trials. For example, because the cooperative code implemented in unit H_{1+} encodes a positive response to trials AB and EF, it also results in an incorrectly positive response on EB (*SI Appendix*, Fig. S10).

Critically, the HCN has a lower ℓ_2-norm because the cooperative code keeps the network weights in a range from -0.5 to 0.5. In contrast, a network using a ranking system has weights between $-\frac{n-1}{2}$ and $\frac{n-1}{2}$ (Fig. 7*E*). As a result, its ℓ_2-norm is not only consistently larger, but also grows faster with an increasing number of items (Fig. 7*F*).

Importantly, the staggered cooperative coding scheme is only norm-efficient because the hidden weights are followed by a rectifying nonlinearity. For example, H_{1+} takes on a value of -1 in response to CD. Without the rectification, H_{2+} would then have to compensate for the negative response to produce a positive label, which would require larger weights and thus would not be norm-efficient. However, because of the rectification following the hidden weights, H_{1+} remains silent on CD and requires no such compensation from H_{2+}. This explains why a cooperative and staggered coding scheme has a low ℓ_2-norm when all weights of the network are trained, but not when only the readout weights are trained.

The HCN illustrates how a constraint that imposes a low ℓ_2-norm on all network weights can give rise to anomalous behavior on TI. As noted above, the empirical network implemented the same computational principles as the HCN, albeit in a slightly different way: E_{1+} approximately implemented H_{1+}, whereas E_{2+} and E_{3+} jointly approximated H_{2+}. These differences resulted in a somewhat higher, but qualitatively similar norm in the empirical network (Fig. 7*F*; see *SI Appendix*, section S3.C.3). In particular, the norm of both the empirical and the HCNs is systematically lower than that of the rank-network (Fig. 7*F*). Remarkably, the mechanism implemented by the empirical network is consistent across not only different random initializations but also different numbers of items n: E_{1+} always approximates H_{1+}, whereas E_{2+} and E_{3+} jointly implement H_{2+} (*SI Appendix*, Fig. S11*F*).

Our findings in this section may appear contrary to those of Nelli et al. (22), who also studied TI performance of neural networks, but found that the rich regime did not impair transitive generalization and further yielded an explicit rank representation in the hidden layer. We found that this was caused by a weight symmetry imposed on the hidden layer that constrains their networks to an additive internal representation (*SI Appendix*, section S2.F). Indeed, our clustering analysis revealed that these networks approximately implemented the rank network (Fig. 7*E*) and the ℓ_2-norm of their weights was much higher than that of

the unconstrained neural network (*SI Appendix*, Fig. S11*A*). This further illustrates the misalignment between the rich regime's norm minimization and a transitive inductive bias: architectural constraints that improve transitive generalization (here, weight symmetry) lead to a higher weight norm.

Discussion

We found that standard statistical learning models can perform TI and recapitulate three empirically observed behavioral patterns (symbolic distance effect, terminal item effect, and memorization effect). The behavior of a given model is sufficiently captured by a single scalar conjunctivity factor α, which characterizes the model's internal representation (of task items) on a spectrum from fully compositional ($\alpha = 0$) to fully conjunctive ($\alpha = 1$). For $\alpha = 0$, the model is constrained to encoding a transitive relation. For partly conjunctive representations ($0 < \alpha < 1$), the model is not constrained in this way, but the principle of norm minimization nevertheless yields transitive generalization. For fully conjunctive representations ($\alpha = 1$), the model cannot generalize transitively. Finally, we found that when representation learning is enabled in hidden layers, networks perform worse on TI and exhibit different behavioral patterns than living subjects. Through hand-constructed networks and a clustering-based analysis of empirical networks, we suggest that this anomalous behavior arises from a different form of norm minimization.

Models of relational cognition often represent relations explicitly and are preconfigured to have a particular relational inductive bias (75). In particular, alternative accounts of TI are either preconfigured to associate a rank (or value) with each item (21, 23, 35) or suggest that humans rely on abstract knowledge of transitivity (e.g., refs. 23, 76, and 77) [as implemented, for instance, in a cognitive map (78–80)]. In contrast to these accounts, we took a "minimal principles" approach, representing nothing but the input itself (i.e., the two presented items). This perspective casts higher-level behavioral capacities (in this case transitive generalization) as emergent from minimally structured learning systems (81). Studies within this paradigm usually rely on simulations, which can leave the mechanisms for emergent behavior, such as generalization, unclear (cf. ref. 82). In contrast, our analytical account identifies the specific model components responsible for transitive generalization. This clarifies how the brain could implement relational generalizations without preconfigured representations or compositional constraints.

More generally, relational cognition likely relies on a wide range of learning mechanisms (83, 84). The models considered here require repeated interleaved presentations of the training trials (85), suggesting learning mechanisms associated with prefrontal or higher-level association cortices. In contrast, other brain regions, such as the hippocampus, support rapid learning without the need for repeated trial presentations, presumably through the operation of memory reactivation (e.g., inferring that $B > D$ by recalling that $B > C$ and $C > D$) (86–88). A recurrent neural network model with Hebbian plasticity ("REMERGE") has been proposed to explain how a reactivation-based mechanism could support transitive generalization (89). While this learning model does not recapitulate behavioral patterns such as the symbolic distance effect, it can explain how subjects may learn TI from a minimal number of trials. Finally, whereas both the reactivation-based and statistical learning models above rely on emergent inductive biases of the underlying learning mechanisms, learning systems can also develop relational inductive biases through structure learning or meta-learning (90, 91), which are associated with both hippocampus and prefrontal cortex (84, 92–94).

Which of these learning mechanisms is implicated in a particular TI task variant likely depends on factors such as the stimulus structure and how training and test trials are presented. Subjects may also rely on a mixture of learning mechanisms on a single task. For example, trials could initially be encoded in the hippocampus but eventually be consolidated in the prefrontal cortex (83, 95). Creating more unified models of relational learning (e.g., fusing the *REMERGE*-model and our similarity-based mechanism, or incorporating structure learning) could shed further light on the interplay between different learning mechanisms on TI and their dependency on different task parameters. Doing so may also allow closer investigation of proposed neural implementations of relational learning (86).

With respect to similarity-based relational learning models, our account could be seen as an endpoint to a series of investigations of TI behavior: expanding upon previous studies (22, 37, 38), we show comprehensively that the principle of norm minimization enables any model with partly conjunctive representations to generalize transitively and further gives rise to naturalistic behavior on TI. Importantly, norm minimization is implicated not only in gradient flow and ridge regression (the examples we consider), but also a much broader range of learning models (96), including reinforcement learning (97). Accordingly, the consistent behaviors many different animals exhibit on the task could be due to this shared, underlying learning principle. This is an alternative to the view that the ubiquity of TI stems from its ecological role in social cognition (98), and the view that TI entails explicit reasoning.

Our results, while providing an alternative explanatory account of TI, should nevertheless be interpreted cautiously with regard to the basis of TI in living subjects, as our model is limited in a number of ways. In terms of behavioral predictions, it cannot account for the asymmetry effect, the observation that performance in living subjects is often better for items toward the start of the hierarchy (e.g., AB) than items toward the end (e.g., FG) (21). Further, our model does not generate predictions for what behavior we might expect from subjects that, after being trained on item pairs, are presented with three items at a time and expected to pick the largest item (25, 99). Finally, our results do not speak to tasks testing for TI in the form of a single question (e.g., "Alice is taller than Bob and Bob is taller than Chris. Who is taller, Alice or Chris?") or after minimal training (18, 100). Such a format likely requires a different mechanism from the one considered here and may be better understood as an instance of explicit reasoning.

These limitations, and extensions of TI, can inspire refining or expanding the statistical modeling approach. For example, Nelli et al. (22) found that conventional statistical learning models could not account for human behavior on a particular variant of TI ("list linking," 101), but that a modification in which uncertainty is encoded could. More broadly, TI, with its rich and well-established set of empirical phenomena (e.g., refs. 20 and 102), may serve as a useful model task to compare statistical learning models (and their failure modes) to human and animal behavior. For example, the lack of an asymmetry effect and the inability to perform the three-item task indicate what is lost by treating TI as a binary categorization task. A model that understands its decision as tied to one of the presented items (rather than as a choice between two arbitrary categories) may be able to better account for these behaviors (21). Notably, better accounts of behavior within the TI framework may not only elucidate how subjects perform relational learning, but

even inform more general models of statistical learning in living subjects.

Despite these potential limitations, our account still makes a set of falsifiable predictions that can produce evidence for or against the biological relevance of our insights. In particular, we identified the conjunctivity factor α as a broadly important parameter for TI task behavior. Our account would predict that changes in α should lead to a set of coordinated changes in behavior. For example, a stronger memorization effect should be associated with a weaker terminal item effect on the training cases (as both are caused by an increase in α). Observing that these different behavioral patterns indeed change in a coordinated manner (for example, between different subjects performing the same task) may indicate that our theory can accurately describe behavior.

Changes in the conjunctivity factor may arise from interindividual differences (different subjects may have representations that are best described by different values of α) but could also be introduced through experimental interventions. In particular, certain task variations may promote a representation of the input as either two separate items or a single conjunctive stimulus; for example, presenting the two items in a common scene rather than as distinct stimuli may result in a higher α (103). More broadly, this suggests that the conjunctivity factor may be a useful parameter for comparing subjects and TI task variants.

Beyond behavioral predictions, the conjunctivity factor could also be used to clarify the neural basis of TI. On the one hand, perturbations in relevant neural areas, for example through lesions (55), may affect the conjunctivity factor. On the other hand, using neural recordings to estimate the empirical representational similarity (104) between distinct, overlapping, and identical trials could enable the estimation of an effective conjunctivity factor in a particular neural area. If differences in this estimated conjunctivity factor (either due to interindividual differences or as a result of lesions) are associated with the corresponding behavioral differences predicted by our model, this may suggest a role for statistical learning as described here, and, in addition, that the recorded neural area is indeed involved in TI.

Moreover, our findings on rich-regime neural networks performing worse on TI may have important implications for machine learning, as deep neural networks have been observed to struggle with tasks involving compositional generalization more broadly (105–108). A common strategy in attempting to address these shortcomings consists in training ever larger models on ever larger datasets (109). Such models have shown impressive results on tasks such as natural language production (110), but the overall scale and complexity of the training data, model, and learning algorithm make it difficult to attribute successful or unsuccessful generalization behaviors to specific components of the model (though see refs. 111 and 112). In contrast, TI is a particularly simple relational task on which deep networks exhibit nonnaturalistic behavior. Our findings explain why the rich regime, which has a useful inductive bias on many different tasks, can give rise to such anomalous behavior on TI. This can help in designing principled modifications to prevent such anomalous behavior—for example, by adding additional training trials, regularizing the hidden layer to prevent a local encoding mechanism, or changing the connectivity structure of the network—and in this way improve our understanding of how relational learning is implemented (113–115). Notably, the tools we employ to analyze the rich-regime behavior may also prove useful for analyzing neural networks trained on other tasks. At the same time, our analysis here was primarily empirical, but future work could provide an analytical characterization of the rich-regime solution (70, 116).

Finally, TI is a simple instance of compositional generalization, i.e., the ability to conceptualize prior experience in terms of components that can be reconfigured in a novel situation (117). More broadly, compositional generalization has long been understood to be a crucial component for human-like learning and generalization, in large part due to the diversity and breadth of its applications: compositional generalization often involves the composition of many rules, relations, or attributes (114). The comparative simplicity of TI enabled us to identify how minimally structured learning systems can implement the inductive biases needed for this task. Our analysis provides a case study for how standard statistical inductive biases determine behavior on compositional tasks, in particular clarifying how representational structure (conceptualized through the conjunctivity factor α) impacts compositional generalization (118). At the same time, TI certainly does not capture the complexity of most compositional task paradigms and future work should extend our formal analysis to a broader scope of compositional tasks. This would clarify the conditions under which standard statistical learning principles can explain compositional generalization and which compositional motifs require additional learning mechanisms.

In summary, we here investigated TI, a task that tests for a fundamental logical capacity and has fascinated researchers across neuroscience and psychology for many decades. We derived exact mathematical equations describing the TI behavior of a large class of statistical learning models. This allowed us to understand systematically how these models can generalize compositionally, using TI as an example. Our theory provides a basis for using TI to investigate the neural structures implementing relational cognition. At the same time, our findings also suggest that TI, in its standard form, may not be sufficient to clarify relational inference abilities that require more than statistical learning. For this purpose, other relational tasks should be considered.

Materials and Methods

Additional methods and results are provided in *SI Appendix*. *SI Appendix*, section S1.A analyzes models with additive representations. *SI Appendix*, sections S1.B and S1.C describe the conjunctivity factor and determine its value across network architectures and nonlinearities. *SI Appendix*, section S1.D describes the mathematical analysis of models whose readout weights are trained using norm minimization, ridge regression, and gradient flow. *SI Appendix*, Section S1.E derives the learning dynamics of neural networks trained in the lazy regime. *SI Appendix*, section S1.F analyzes models whose readout weights are trained on transverse patterning with gradient flow. *SI Appendix*, section S2.A describes the packages used in data analysis and visualization and *SI Appendix*, section S2.B describes experiments analyzing TI behavior in models with non-exchangeable representations. *SI Appendix*, section S2.C describes how neural networks were trained with backpropagation. *SI Appendix*, section S3 describes the analysis of networks related to rich-regime training (empirical and HCNs).

Data, Materials, and Software Availability. The code and data required to reproduce our findings has been stored in a Zenodo database (119).

ACKNOWLEDGMENTS. We are grateful to the members of the Center for Theoretical Neuroscience at Columbia University, the SueYeon Chung lab at Flatiron/NYU, and attendees at Cosyne 2023 and the Gatsby Tri-Centre Meeting 2023 for helpful comments. We thank Taiga Abe, Veronica Bossio, Tala Fakhoury, Ching Fang, Jeff Johnston, Erica Shook, Sharon Su, and Denis Turcu for detailed feedback. The work was supported by NSF 1707398 (Neuronex), Gatsby Charitable Foundation GAT3708, National Institute of Mental Health MH126158-01A1, and NIH-R01MH111703.

1. G. S. Halford, W. H. Wilson, S. Phillips, Relational knowledge: The foundation of higher cognition. *Trend. Cognit. Sci.* **14**, 497–505 (2010).
2. D. L. Cheney, R. M. Seyfarth, J. B. Silk, The responses of female baboons (*Papio cynocephalus ursinus*) to anomalous social interactions: Evidence for causal reasoning? *J. Comp. Psychol.* **109**, 134–141 (1995).
3. A. S. Etienne, K. J. Jeffery, Path integration in mammals. *Hippocampus* **14**, 180–192 (2004).
4. A. Seed, R. Byrne, Animal tool-use. *Curr. Biol.* **20**, R1032–R1039 (2010).
5. M. W. Baldwin, Relational schemas and the processing of social information. *Psychol. Bull.* **112**, 461 (1992).
6. E. C. Tolman, Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
7. D. C. Penn, D. J. Povinelli, Causal cognition in human and nonhuman animals: A comparative, critical review. *Annu. Rev. Psychol.* **58**, 97–118 (2007).
8. T. M. Mitchell, "The need for biases in learning generalizations" (Tech. Rept. CBM-TR-117, Rutgers University, NJ, 1980).
9. A. Krogh, J. A. Hertz, Generalization in a linear perceptron in the presence of noise. *J. Phys. A: Math. Gen.* **25**, 1135 (1992).
10. A. Canatar, B. Bordelon, C. Pehlevan, Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nat. Commun.* **12**, 2914 (2021).
11. A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
12. C. Cortes, V. Vapnik, Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
13. D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, N. Srebro, The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.* **19**, 1–57 (2018).
14. A. Krogh, J. Hertz, "A simple weight decay can improve generalization" in *Advances in Neural Information Processing Systems*, J. Moody, S. Hanson, R. P. Lippmann, Eds. (Morgan-Kaufmann, San Francisco, CA, 1991), vol. 4, pp. 950–957.
15. J. Moody, "The effective number of parameters: An analysis of generalization and regularization" in *Nonlinear Learning Systems in Advances in Neural Information Processing Systems*, J. Moody, S. Hanson, R. P. Lippmann, Eds. (Morgan-Kaufmann, San Francisco, CA, 1991), vol. 4, pp. 847–854.
16. S. M. Barnett, S. J. Ceci, When and where do we apply what we learn? A taxonomy for far transfer. *Psychol. Bull.* **128**, 612–637 (2002).
17. P. W. Battaglia *et al*., Relational inductive biases, deep learning, and graph networks. arXiv [Preprint] (2018). https://arxiv.org/abs/1806.01261 (Accessed 14 June 2024).
18. J. Piaget, *Judgment and reasoning in the child* (Harcourt, Brace, Oxford, England, 1928).
19. M. Vasconcelos, Transitive inference in non-human animals: An empirical and theoretical analysis. *Behav. Proc.* **78**, 313–334 (2008).
20. G. Jensen, "Serial learning" in *APA Handbook of Comparative Psychology: Perception, Learning, and Cognition. APA Handbooks in Psychology*, J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, T. Zentall, Eds. (American Psychological Association, Washington, DC, 2017), vol. 2, pp. 385–409.
21. S. Ciranka *et al*., Asymmetric reinforcement learning facilitates human inference of transitive relations. *Nat. Hum. Behav.* **6**, 555–564 (2022).
22. S. Nelli, L. Braun, T. Dumbalska, A. Saxe, C. Summerfield, Neural knowledge assembly in humans and neural networks. *Neuron* **111**, 1504–1516.e9 (2023).
23. G. Jensen, F. Muñoz, Y. Alkan, V. P. Ferrera, H. S. Terrace, Implicit value updating explains transitive inference performance: The betasort model. *PLoS Comput. Biol.* **11**, e1004523 (2015).
24. P. E. Bryant, T. Trabasso, Transitive inferences and memory in young children. *Nature* **232**, 456–458 (1971).
25. B. O. McGonigle, M. Chalmers, Are monkeys logical? *Nature* **267**, 694–696 (1977).
26. H. Davis, Transitive inference in rats (*Rattus norvegicus*). *J. Comp. Psychol.* **106**, 342–349 (1992).
27. E. A. Tibbetts, J. Agudelo, S. Pandit, J. Riojas, Transitive inference in *Polistes* paper wasps. *Biol. Lett.* **15**, 20190015 (2019).
28. L. Grosenick, T. S. Clement, R. D. Fernald, Fish can infer social rank by observation alone. *Nature* **445**, 429–432 (2007).
29. G. R. Potts, Storing and retrieving information about ordered relationships. *J. Exp. Psychol.* **103**, 431–439 (1974).
30. M. R. D'Amato, M. Colombo, The symbolic distance effect in monkeys (*Cebus apella*). *Anim. Learn. Behav.* **18**, 133–140 (1990).
31. L. Von Fersen, C. D. Wynne, J. D. Delius, J. E. Staddon, Transitive inference formation in pigeons. *J. Exp. Psychol. Anim. Behav. Process.* **17**, 334 (1991).
32. C. D. L. Wynne, Pigeon transitive inference: Tests of simple accounts of a complex performance. *Behav. Process.* **39**, 95–112 (1997).
33. B. D. Acuna, J. N. Sanes, J. P. Donoghue, Cognitive mechanisms of transitive inference. *Exp. Brain Res.* **146**, 1–10 (2002).
34. M. Van Elzakker, R. C. O'Reilly, J. W. Rudy, Transitivity, flexibility, conjunctive representations, and the hippocampus. I. An empirical analysis. *Hippocampus* **13**, 334–340 (2003).
35. C. D. L. Wynne, Reinforcement accounts for transitive inference performance. *Anim. Learn. Behav.* **23**, 207–217 (1995).
36. C. D. L. Wynne, Transverse patterning in pigeons. *Behav. Process.* **38**, 119–130 (1996).
37. C. De Lillo, D. Floreano, F. Antinucci, Transitive choices by a simple, fully connected, backpropagation neural network: Implications for the comparative study of transitive inference. *Anim. Cognit.* **4**, 61–68 (2001).
38. K. Kay *et al*., Emergent neural dynamics and geometry for generalization in a transitive inference task. *PLOS Comput. Biol.* **20**, e1011954 (2024).
39. G. Di Antonio, S. Raglio, M. Mattia, Ranking and serial thinking: A geometric solution. bioRxiv [Preprint] (2023). https://www.biorxiv.org/content/10.1101/2023.08.03.551859v1 (Accessed 14 June 2024).
40. H. O. Carmesin, H. Schwegler, Parallel versus sequential processing of relational stimulus structures. *Biol. Cybernet.* **71**, 523–529 (1994).
41. R. Bogacz, E. Brown, J. Moehlis, P. Holmes, J. D. Cohen, The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
42. R. Ratcliff, A theory of memory retrieval. *Psychol. Rev.* **85**, 59–108 (1978).
43. R. Ratcliff, G. McKoon, The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Comput.* **20**, 873–922 (2008).
44. S. M. Frankland, J. D. Greene, Two ways to build a thought: Distinct forms of compositional semantic representation across brain regions. *Cereb. Cortex* **30**, 3838–3855 (2020).
45. M. Rigotti *et al*., The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
46. J. W. Rudy, R. C. O'Reilly, Conjunctive representations, the hippocampus, and contextual fear conditioning. *Cognit. Affect. Behav. Neurosci.* **1**, 66–82 (2001).
47. R. A. Rescorla, "Configural" conditioning in discrete-trial bar pressing. *J. Comp. Physiol. Psychol.* **79**, 307–317 (1972).
48. V. Mante, D. Sussillo, K. V. Shenoy, W. T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
49. H. Davis, "Logical transitivity in animals" in *Cognitive Aspects of Stimulus Control*, W. K. Honig, J. G. Fetterman, Eds. (Psychology Press, 1992).
50. Y. Cho, L. Saul, "Kernel methods for deep learning" in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, A. Culotta, Eds. (Curran Associates, Inc., Red Hook, NY, 2009), vol. 22, pp. 342–350.
51. A. Saxe, J. McClelland, S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks" in *International Conference on Learning Representations* (2014).
52. S. Gunasekar, J. Lee, D. Soudry, N. Srebro, "Characterizing implicit bias in terms of optimization geometry" in *International Conference on Machine Learning* (PMLR, 2018), pp. 1832–1841.
53. Z. Ji, M. Dudík, R. E. Schapire, M. Telgarsky, "Gradient descent follows the regularization path for general losses" in *Conference on Learning Theory* (PMLR, 2020), pp. 2109–2136.
54. K. W. Spence, The nature of the response in discrimination learning. *Psychol. Rev.* **59**, 89–93 (1952).
55. J. A. Dusek, H. Eichenbaum, The hippocampus and transverse patterning guided by olfactory cues. *Behav. Neurosci.* **112**, 762–771 (1998).
56. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
57. B. A. Richards *et al*., A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
58. A. Saxe, S. Nelli, C. Summerfield, If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* **22**, 55–67 (2021).
59. A. Jacot, F. Gabriel, C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks" in *Advances in Neural Information Processing Systems*, S. Bengio *et al*., Eds. (Curran Associates, Inc., Red Hook, NY, 2018), vol. 31, pp. 8580–8589.
60. L. Chizat, E. Oyallon, F. Bach, "On lazy training in differentiable programming" in *Advances in Neural Information Processing Systems*, H. Wallach *et al*., Eds. (Curran Associates, Inc., Red Hook, NY, 2019), vol. 32, pp. 2937–2947.
61. S. Fort *et al*., "Deep learning versus kernel learning: An empirical study of loss landscape geometry and the time evolution of the Neural Tangent Kernel" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, Inc., Red Hook, NY, 2020), vol. 33, pp. 5850–5861.
62. N. Vyas, Y. Bansal, P. Nakkiran, Limitations of the NTK for understanding generalization in deep learning. arXiv [Preprint] (2022). https://arxiv.org/abs/2206.10012 (Accessed 14 June 2024).
63. J. Poort *et al*., Learning enhances sensory and multiple non-sensory representations in primary visual cortex. *Neuron* **86**, 1478–1490 (2015).
64. T. Flesch, K. Juechems, T. Dumbalska, A. Saxe, C. Summerfield, Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* **110**, 1258–1270.e11 (2022).
65. F. Van Opstal, W. Fias, P. Peigneux, T. Verguts, The neural representation of extensively trained ordered sequences. *NeuroImage* **47**, 367–375 (2009).
66. E. C. Hinton, S. Dymond, U. von Hecker, C. J. Evans, Neural correlates of relational reasoning and the symbolic distance effect: Involvement of parietal cortex. *Neuroscience* **168**, 138–148 (2010).
67. A. O. Constantinescu, J. X. O'Reilly, T. E. J. Behrens, Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
68. J. C. Whittington *et al*., The Tolman–Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183**, 1249–1263 (2020).
69. K. Lyu, J. Li, "Gradient descent maximizes the margin of homogeneous neural networks" in *International Conference on Learning Representations* (2020).
70. L. Chizat, F. Bach, "Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss" in *Proceedings of Thirty Third Conference on Learning Theory* (PMLR, 2020), pp. 1305–1338.
71. N. Razin, N. Cohen, "Implicit regularization in deep learning may not be explainable by norms" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, Inc., Red Hook, NY, 2020), vol. 33, pp. 21174–21187.
72. B. Woodworth *et al*., "Kernel and rich regimes in overparametrized models" in *Conference on Learning Theory* (2020), pp. 3635–3673.
73. P. Savarese, I. Evron, D. Soudry, N. Srebro, "How do infinite width bounded norm networks look in function space?" in *Proceedings of the Thirty-Second Conference on Learning Theory* (2019), vol. 99, pp. 2667–2690.
74. G. Ongie, R. Willett, D. Soudry, N. Srebro, "A function space view of bounded norm infinite width ReLU nets: The multivariate case" in *International Conference on Learning Representations* (2020). https://openreview.net/pdf?id=H1lNPxHKDH. Accessed 21 June 2024.
75. C. Kemp, J. B. Tenenbaum, The discovery of structural form. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10687–10692 (2008).
76. D. J. Gillan, Reasoning in the chimpanzee: II. Transitive inference. *J. Exp. Psychol.: Anim. Behav. Process.* **7**, 150–164 (1981).
77. L. F. Jacobs, From movement to transitivity: The role of hippocampal parallel maps in configural learning. *Rev. Neurosci.* **17**, 99–109 (2006).
78. C. B. De Soto, M. London, S. Handel, Social reasoning and spatial paralogic. *J. Person. Soc. Psychol.* **2**, 513–521 (1965).
79. R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, Y. Niv, Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).

80. T. E. J. Behrens *et al*., What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
81. J. L. McClelland *et al*., Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trend. Cognit. Sci.* **14**, 348–356 (2010).
82. A. M. Saxe, J. L. McClelland, S. Ganguli, A mathematical theory of semantic development in deep neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 11537–11546 (2019).
83. J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
84. D. Kumaran, D. Hassabis, J. L. McClelland, What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trend. Cognit. Sci.* **20**, 512–534 (2016).
85. G. Jensen, F. Munoz, A. Meaney, H. S. Terrace, V. P. Ferrera, Transitive inference after minimal training in rhesus macaques (*Macaca mulatta*). *J. Exp. Psychol.: Anim. Learn. Cognit.* **47**, 464 (2021).
86. D. Zeithamova, M. Schlichting, A. Preston, The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Front. Hum. Neurosci.* **6**, 70 (2012).
87. D. Shohamy, N. D. Daw, Integrating memories to guide decisions. *Curr. Opin. Behav. Sci.* **5**, 85–90 (2015).
88. Z. Kurth-Nelson *et al*., Replay and compositional computation. *Neuron* **111**, 454–469 (2023).
89. D. Kumaran, J. L. McClelland, Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychol. Rev.* **119**, 573–616 (2012).
90. J. Russin, M. Zolfaghar, S. A. Park, E. Boorman, R. C. O'Reilly, Complementary structure–learning neural networks for relational reasoning. *Annu. Conf. Cognit. Sci. Soc.* **2021**, 1560–1566 (2021).
91. T. Miconi, K. Kay, An active neural mechanism for relational learning and fast knowledge reassembly. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.07.27.550739 (Accessed 14 June 2024).
92. J. X. Wang *et al*., Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
93. J. C. R. Whittington, D. McCaffary, J. J. W. Bakermans, T. E. J. Behrens, How to build a cognitive map. *Nat. Neurosci.* **25**, 1257–1272 (2022).
94. L. Ambrogioni, H. F. Ólafsdóttir, Rethinking the hippocampal cognitive map as a meta-learning computational module. *Trend. Cognit. Sci.* **27**, 702–712 (2023).
95. G. Winocur, M. Moscovitch, B. Bontempi, Memory formation and long-term retention in humans and animals: Convergence towards a transformation account of hippocampal–neocortical interactions. *Neuropsychologia* **48**, 2339–2356 (2010).
96. M. Telgarsky, Y. Singer, A primal-dual convergence analysis of boosting. *J. Mach. Learn. Res.* **13**, 561–606 (2012).
97. A. Kumar *et al*., "DR3: Value-based deep reinforcement learning requires explicit regularization" in *International Conference on Learning Representations* (2022).
98. A. B. Bond, A. C. Kamil, R. P. Balda, Social complexity and transitive inference in corvids. *Anim. Behav.* **65**, 479–487 (2003).
99. B. McGonigle, M. Chalmers, Monkeys are rational! *Q. J. Exp. Psychol., Sec. B* **45**, 189–228 (1992).
100. G. S. Halford, Can young children integrate premises in transitivity and serial order tasks? *Cognit. Psychol.* **16**, 65–93 (1984).
101. F. R. Treichler, D. Van Tilburg, Concurrent conditional discrimination tests of transitive inference by macaque monkeys: List linking. *J. Exp. Psychol.: Anim. Behav. Process.* **22**, 105 (1996).
102. O. F. Lazareva, E. A. Wasserman, Transitive inference in pigeons: Measuring the associative values of Stimuli B and D. *Behav. Process.* **89**, 244–255 (2012).
103. H. Eichenbaum, *The Cognitive Neuroscience of Memory: An Introduction* (Oxford University Press, 2011).
104. N. Kriegeskorte, M. Mur, P. A. Bandettini, Representational similarity analysis–Connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
105. J. A. Fodor, Z. W. Pylyshyn, Connectionism and cognitive architecture: A critical analysis. *Cognition* **28**, 3–71 (1988).
106. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
107. E. M. Bender, A. Koller, "Climbing towards NLU: On meaning, form, and understanding in the age of data" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 5185–5198.
108. C. Summerfield, F. Luyckx, H. Sheahan, Structure learning and the posterior parietal cortex. *Progr. Neurobiol.* **184**, 101717 (2020).
109. J. Kaplan *et al*., Scaling laws for neural language models. arXiv [Preprint] (2020). https://arxiv.org/abs/2001.08361 (Accessed 14 June 2024).
110. T. Brown *et al*., "Language models are few-shot learners" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, Inc., 2020), vol. 33, pp. 1877–1901.
111. N. Kandpal, H. Deng, A. Roberts, E. Wallace, C. Raffel, "Large language models struggle to learn long-tail knowledge" in *International Conference on Machine Learning* (PMLR, 2023), pp. 15696–15707.
112. R. Grosse *et al*., Studying large language model generalization with influence functions. arXiv [Preprint] (2023). https://arxiv.org/abs/2308.03296 (Accessed 14 June 2024).
113. P. Smolensky, On the proper treatment of connectionism. *Behav. Brain Sci.* **11**, 1–23 (1988).
114. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
115. F. Hill, A. Santoro, D. Barrett, A. Morcos, T. Lillicrap, "Learning to make analogies by contrasting abstract relational structure" in *International Conference on Learning Representations* (2019).
116. A. Saxe, S. Sodhani, S. J. Lewallen, "The neural race reduction: Dynamics of abstraction in gated networks" in *Proceedings of the 39th International Conference on Machine Learning* (PMLR, 2022), pp. 19287–19309.
117. S. M. Frankland, J. D. Greene, Concepts and compositionality. In search of the brain's language of thought. *Annu. Rev. Psychol.* **71**, 273–303 (2020).
118. M. Lepori, T. Serre, E. Pavlick, "Break it down: Evidence for structural compositionality in neural networks" in *Advances in Neural Information Processing Systems*, A. Oh *et al*., Eds. (Curran Associates, Inc., Red Hook, NY, 2024), vol. 36, pp. 42623–42660.
119. S. Lippl, K. Kay, G. Jensen, V. P. Ferrera, L. F. Abbott, Code for "A mathematical theory of relational generalization in transitive inference." Zenodo. https://doi.org/10.5281/zenodo.12172070. Deposited on 19 June 2024.