Current Directions in Psychological Science
1–8

DOI: 10.1177/09637214251336212
www.psychologicalscience.org/CDPS

How Can We Characterize Human Generalization and Distinguish It From Generalization in Machines?

Mirko Thalmann and Eric Schulz
Institute for Human-Centered AI at the Helmholtz Center for Computational Health – Helmholtz Munich, Neuherberg, Germany

Abstract

People appear to excel at generalization: They require little experience to generalize their knowledge to new situations. But can we confidently make such a conclusion? To make progress toward a better understanding, we characterize human generalization by introducing three proposed cognitive mechanisms allowing people to generalize: applying simple rules, judging new objects by considering their similarity to previously encountered objects, and applying abstract rules. We highlight the systematicity with which people use these three mechanisms by, perhaps surprisingly, focusing on failures of generalization. These failures show that people prefer simple ways to generalize, even when simple is not ideal. Together, these results can be subsumed under two proposed stages: First, people infer what aspects of an environment are task relevant, and second, while repeatedly carrying out the task, the mental representations required to solve the task change. In this article, we compare humans to contemporary AI systems. This comparison shows that AI systems use the same generalization mechanisms as humans. However, they differ from humans in the way they abstract patterns from observations and apply these patterns to previously unknown objects—often resulting in generalization performance that is superior to, but sometimes inferior to, that of humans.

Keywords

generalization, mental representations, cognitive processes, memory, abstraction

We never observe the same object in the same circumstances again, yet we easily recognize a known object in a new scene. Although relatively simple for humans, deep convolutional neural networks for object recognition struggle to do so, for example, when human-imperceptible noise is added to an image (Geirhos et al., 2018). Given the relative ease with which humans solve such tasks, Shepard (1987) famously argued that the first law of psychology should be the law of generalization. Here, we define generalization broadly as the use of previously acquired knowledge when responding to previously unobserved objects (see also Taylor et al., 2021). The psychological literature on generalization can be well characterized by separately analyzing (a) generalization within a known feature space (also called "in-domain generalization") and (b) out-of-category generalization (also called "out-of-domain generalization"). Although most branches of cognitive psychology have been interested in in-domain generalization, the fields of language, memory, and category learning have additionally focused on out-of-domain generalization. This distinction not only defines the scope of generalization but also allows us to highlight key differences between humans and contemporary AI systems in their abilities to generalize.

Generalization Within a Known Feature Space

In the area of *absolute identification*, which studies how people learn to identify objects and discriminate one object from other objects, Shepard (1987) showed that the probability of perceiving two stimuli as the same increases monotonically with the psychological similarity between the two stimuli. The same monotonic relationship may not hold without a transformation of the features describing the object from physical space to psychological space. For example, the mental representation of pitch is well described by a helical structure that integrates a linear dimension of pitch height

Corresponding Author:
Mirko Thalmann, Institute for Human-Centered AI at the Helmholtz Center for Computational Health – Helmholtz Munich, Neuherberg, Germany
Email: mirkothalmann@hotmail.com

with a circular dimension of pitch chroma, which explains the increased generalization between tones separated by an octave (see also Shepard, 1987). Although identifying individual stimuli can be considered a categorization task, the field of *category learning* has traditionally focused on much broader categories. Numerous studies have demonstrated that people can learn to categorize individual training stimuli into their respective categories based on their object properties. Importantly, stimuli not observed during training are categorized during a transfer test, with accuracy well above chance (e.g., Johansen & Palmeri, 2002; Nosofsky, 1986).

The results are again similar when the discrete case of learning categories is extended to *learning functions* that relate values from continuous feature dimensions (e.g., dosage of a poison) to continuous outcomes (e.g., symptom severity). It has been shown that people generalize to unobserved stimuli with high accuracy in the intrapolation range (i.e., values between observed values) and with performance well above chance in the extrapolation range (i.e., values outside the observed range; e.g., DeLosh et al., 1997). People behave similarly in *reinforcement learning tasks*, in which they are instructed to collect as many rewards as possible from a limited number of choices. For example, Jagadish et al. (2023) showed that participants learned to relate response keys to rewards according to linear and periodic functions. They furthermore showed that people could generalize to the composition of these two functions (i.e., adding a periodic function to a linear function) with remarkable accuracy on the first trial, with practice only on the individual functions but no practice on the composite function.

Theoretically, two mechanisms have been proposed for how people generalize in these tasks. First, people use simple rules to partition a feature space (e.g., Ashby & Townsend, 1986). Such rules can often be represented by a simple or conjunctive conditional. For example, a person may consider a new face to belong to Family A if the nose is small or if the nose is small *and* the hair color is red (Johansen & Palmeri, 2002). Second, people compare the new face to all previously encountered faces and use the overall similarity to previously encountered faces to categorize the new face. Research has shown that people likely use both mechanisms but tend to start generalizing using a simple rule and then slowly shift to using similarity to stored examples as more of the same examples are presented (Johansen & Palmeri, 2002; Nosofsky et al., 1994). Although similarity in these examples focused on the similarity of features of a single object, similarity can be defined more broadly to include similarity in associations between objects. For example, it has been argued that similarity-based generalization occurs in certain associative memory tasks even before people are given a generalization test. Shohamy and Wagner (2008) showed that when two faces (F1 and F2) share one but not all associations with a set of stimuli (S1 and S2; i.e., F1-S1, F1-S2, and F2-S1 are associated) and are therefore similar, the nonoverlapping association (F2-S2) is still associated in memory. In addition to these two mechanisms, studies have shown that the precision of representations within a feature space becomes more accurate with practice (Goldstone & Steyvers, 2001; Thalmann et al., 2024). More accurate representations help with generalization, especially for objects close to a category boundary.

Neither of these mechanisms can be said to be unique to humans. Although neural network models have a bias toward classifying novel objects on the basis of their similarity to previously observed examples, they also use rule-based mechanisms for categorization (Dasgupta et al., 2022). Moreover, modern artificial neural network models achieve superhuman performance in each of the four introduced tasks (e.g., in visual object recognition; van Dyck et al., 2021). Although these observations show that AI systems outperform humans in each modality (e.g., visual, verbal), there are two differences between human and machine generalization. First, humans perform reasonably well on all tasks involving stimuli from multiple modalities (e.g., object recognition, language understanding), but most AI systems are modality-specific; for example, systems that are good at object recognition often cannot easily process language. Second, humans and neural networks likely have different internal representations. For example, although humans learn an object's representation by observing it from different angles and interacting with it in various task contexts, an AI system may excel at object recognition for entirely different reasons. For instance, a neural network might "recognize" cows simply because they are consistently presented against a green background (Ilyas et al., 2019). In this sense, it has been argued that modern AI systems excel at extracting statistical regularities from data but lack the ability to form internal models of the world (Vafa et al., 2024; but for a different perspective, see Gurnee and Tegmark, 2024). Another way to look at this debate is that AI systems learn an internal model to perform well at the task they are trained on. However, this model does not necessarily match the human model, for example, because humans interact with the same object in several different tasks and contexts.

Out-of-Category Generalization

In the domain of language, it has been suggested that abstract knowledge (e.g., structuring a sentence in terms of noun units and verb units) facilitates both language comprehension and production. Supporting this claim, G. F. Marcus et al. (1999) demonstrated that infants as young as 7 months old can recognize abstract sequential patterns (i.e., an ABA sequence in which A and B represent syllable placeholders) in a transfer sequence composed of previously unobserved syllables. Abstract sequential patterns also play a crucial role in memory. Wu et al. (2023) showed that humans extract abstract knowledge from patterned sequences, which enhances their short-term memory performance for sequences following the same pattern but consisting of novel items. Similarly, in the field of category learning, Goldwater et al. (2018) found that people can learn to infer categories based on the relationships between stimuli. For instance, participants learned that three bars of different lengths formed distinct categories when aligned monotonically in increasing length compared with when they were aligned nonmonotonically. Crucially, the participants also generalized this rule when categorizing a new set of previously unobserved objects (i.e., the luminance of circles). The results from these three areas, together with anecdotal evidence, such as mathematicians deriving general laws by referring to variables rather than individual data points, show that people can abstract patterns from a set of concrete observations and apply them to previously unobserved objects. Although traditional neural networks were unable to apply abstracted knowledge to feature domains that were not varied during training (G. Marcus, 2020), newer approaches mimic human-like generalization and sometimes generalize better than humans in such cases (Lake & Baroni, 2023).

Although the former cases referred to the abstraction of feature domains within one modality (e.g., visual), multimodal large language models can process information from multiple modalities. For example, they can answer textual questions about images. Therefore, they have the potential to surpass human performance in multiple modalities and to generalize abstract patterns across modalities, as humans do. Schulze Buschoff et al. (2025) set out to test the latest multimodal models on three tasks that require the use of an abstract rule to be successful: logical reasoning, intuitive physics, and intuitive psychology. However, the models performed significantly worse than humans on all three tasks.

In summary, humans use simple rules, similarity to previously encountered objects, and abstract rules to generalize. Modern AI systems use the same mechanisms, perform better at a given task within a particular modality, but are mostly modality-specific. More recent multimodal models, however, do not yet generalize on par with humans. Thus, because humans generalize reasonably well in most of these tasks and across different feature domains, the evidence presented so far sets high expectations for human generalization abilities. However, there are some circumstances in which humans systematically fail to generalize, which we discuss below.

Systematic Failures of Generalization

Studies in the area of category learning have shown that humans oversimplify structure. For example, Vermaercke et al. (2014) trained humans and rats on a rule-based and an information-integrating category structure (see Fig. 1). The latter cannot be solved with a simple one- or two-dimensional rule. Given sufficient training, humans and rats learned both structures equally well. When rats were prompted to categorize unobserved transfer stimuli, performance remained roughly the same for both structures. Performance, however, dropped substantially for humans on the information-integration structure but not on the rule-based structure. In a similar information-integration category learning task, about a third of the participants in a study by Donkin et al. (2015) relied on a rule-based categorization strategy, even though this strategy was clearly not the best representation of the category structure. Taken together, the Donkin et al. (2015) and Vermaercke et al. (2014) studies suggest that people make systematic errors when asked to infer the category of unobserved objects: They rely too heavily on rules.

Evidence from the field of *function learning* specifies this systematicity: People tend to prefer simple rules to complicated ones (see also Chater & Vitányi, 2003). For example, they learn linear functions faster than quadratic functions (Brehmer, 1974), and they simplify more complicated functions, for example, by approximating a quadratic function with a linear function (DeLosh et al., 1997; Little & Shiffrin, 2009).

In the area of *cognitive training*, a large corpus of studies has shown that despite large performance gains on trained tasks, performance on similar untrained tasks does not improve. For example, despite large improvements in performance on a working memory task, performance on very similar untrained working memory tasks is unaffected (De Simoni & von Bastian, 2018). This lack of generalized training gains (e.g., Melby-Lervåg et al., 2016) is somewhat surprising given the large positive correlation between working memory and fluid intelligence (e.g. De Simoni & von Bastian, 2018). It has been assumed that training gains in working memory tasks should therefore generalize to other tasks requiring reasoning or fluid intelligence.

These systematic failures of generalization cast doubt on the proposition that mastery of a task seamlessly generalizes to performance on new objects within the same task or to performance on similar tasks. So how

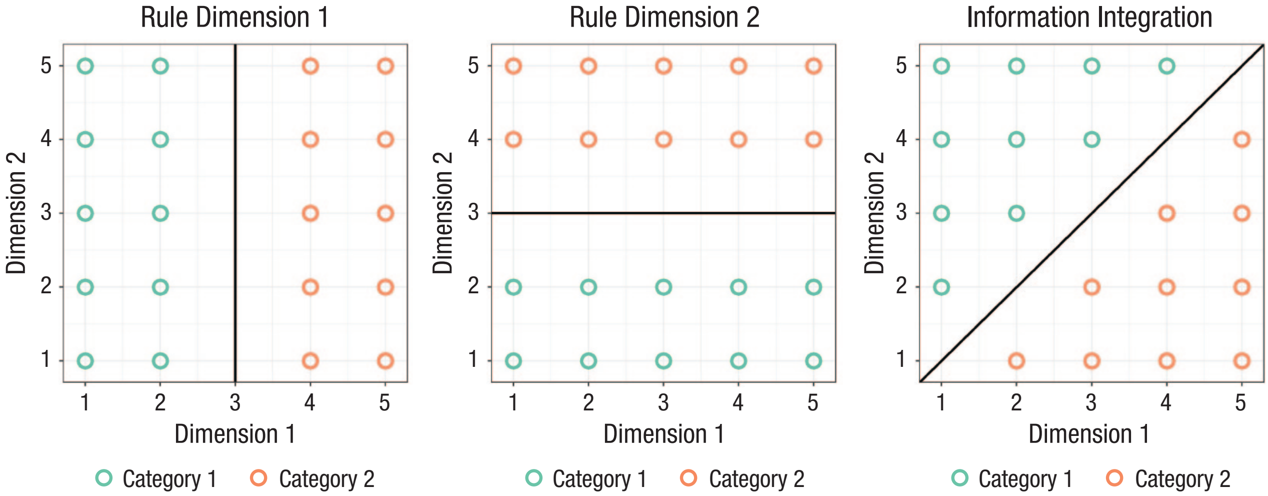

Fig. 1. Three commonly studied category learning structures. Note that both dimensions in the information-integration structure must be considered at the same time for the accurate categorization of an object. Sequentially applying two simple rules will result in suboptimal performance.

can we understand these divergent results about remarkable generalization abilities and systematic generalization failures (for a summary, see Table 1)? In what follows, we introduce two stages that determine how well humans generalize and whether practice in one task can be expected to transfer to other tasks. These two stages will moreover allow us to integrate evidence for and against good human generalization abilities.

Two Stages of Human Generalization

We propose that generalization is primarily a function of mental representations acquired through engagement with a task. These representations emerge at both the task level and the level of individual objects or feature dimensions. We further suggest that these representations are formed during learning episodes that can be roughly divided into two partially overlapping stages (see Fig. 2).

Constraining the task representation

When people are given a task, they need to constrain the dimensionality of the problem space. Although experimenters often try to constrain the possible task representations through instructions, there are typically several ways to derive them. For example, Mason et al. (2022) argued that people generate hypotheses about which aspects of a task environment are relevant. They showed that small changes in instructions and stimulus presentation affected people's task representations. This

Table 1. Evidence for and Against Good Generalization Abilities

	Evidence	
Domain	For	Against
Absolute identification	Psychological similarity predicts whether stimuli are considered to be the same stimulus	—
Category learning	Above-chance generalization of category knowledge to unobserved stimuli	Preference for rules, even if not favorable
Function and reinforcement learning	Above-chance generalization of function knowledge to unobserved stimuli (inter- and extrapolation)	Preference for simple rules (e.g., representing a quadratic function with a linear function)
Memory and language	Use of abstracted patterns in concrete space (e.g., words, chunks) and in abstract space (e.g., nouns vs. verbs, motifs)	—
Cognitive training	—	Absence of transfer in the face of large performance gains in trained tasks

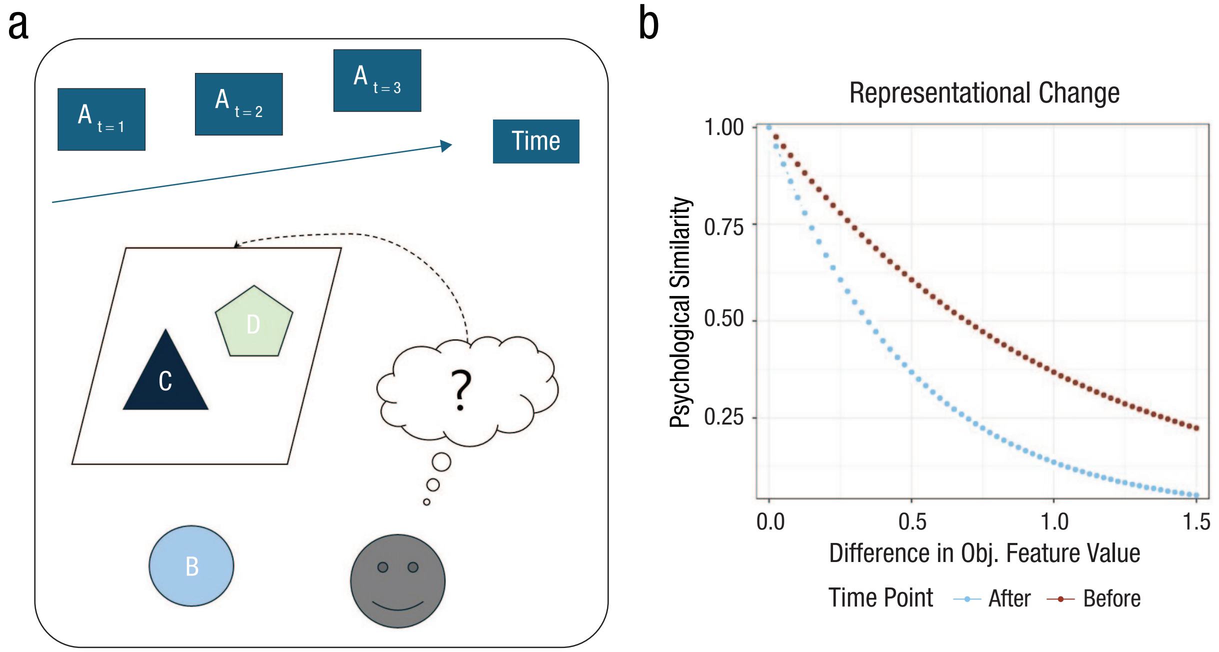

Fig. 2. Two stages of human generalization. A participant in a category learning experiment considers (a) Features A through D to be potentially relevant in the experiment. Feature A represents sequential information about the presented stimuli (e.g., whether every n^{th} object belongs to a certain category). The person considers Feature B irrelevant and Features C and D likely relevant. After extensive learning, representations of (b) feature values on Features C and D become more precise. That is, two stimuli that differ from each other according to a fixed distance in objective feature space are psychologically less similar after learning.

idea is consistent with Feldman's (1992) proposal that people focus on dimensions that they expect to be task relevant given their background knowledge but ignore dimensions that they expect to be task irrelevant. As a consequence, individual differences in task representations arise because people differ in the features they consider task relevant. Difficulties in finding task-relevant feature dimensions can be exacerbated in real-world tasks. For example, John Snow's discovery that cholera outbreaks were associated with the use of a particular well in London is similar to finding a needle in a haystack. Although much research has focused on how people make decisions on a small set of feature dimensions, more work is needed to understand how people generate hypotheses about which dimensions are relevant in a task.

This tendency to generate hypotheses about what information is relevant in a given task and how that information helps solve a task can slow down learning or even lead to what are known as *learning traps*. For example, when there is a change in which stimulus dimensions are task relevant, humans have a hard time disengaging from the previously relevant dimensions (Kruschke, 1996). This differs from AI systems, which can instantly assign an experimenter/engineer-defined error signal to the relevant feature dimension(s). AI systems can directly exploit statistical regularities in the data without needing to form hypotheses about which features are important, and they do so across a myriad of dimensions without human limitations of attentional capacity.

Representational change

The more experience people have with a task, the more they tend to solve it and generalize using information about individual stimuli (i.e., exemplars; Johansen & Palmeri, 2002). In particular, people use the perceived similarity between a representation of a novel stimulus and representations of previously observed stimuli to respond to the novel stimulus (Nosofsky, 1986). Furthermore, as people perform a task, their representations of task-relevant objects change over time. First, representations of individual stimuli become more precise with repeated exposure (Goldstone & Steyvers, 2001; Thalmann et al., 2024). Second, repeated exposure to the same sequences of items (e.g., F-B-I) leads to the emergence of chunks, and these chunks are used in novel situations and tasks to deal more efficiently with limited working memory capacity (Thalmann et al., 2019). Finally, people learn more abstract representations, for example, to categorize nouns as subjects and objects. They use sequential regularities at the abstract level (e.g., subject-verb-object) to generalize efficiently. That is, they handle previously unobserved sequences that follow the same regularities with relative ease (e.g., G. F. Marcus et al., 1999; Wu et al., 2023).

Although we consider the evidence that mental representations change with practice to be substantial, the literature on cognitive training provides strong evidence that cognitive processes cannot be trained (Melby-Lervåg et al., 2016). We therefore propose that successful generalization between tasks primarily depends on whether participants can use the same representations—rather than merely relying on identical cognitive processes—across both tasks. These representations might include, for instance, a deeper understanding of the musical notes relevant to both flute and piano playing, a series of letters learned as a cohesive chunk that applies to two different memory tasks, or an abstract representation of a sequence of items common to memory tasks that differ in their stimulus domain (e.g., verbal and visual).

How to Compare Generalization in Humans and AI Systems

Over the past decades, the training corpora of AI systems have become larger and larger. Decoupling generalization from experience has therefore become challenging. The same problem, however, also applies to human participants. Psychologists have made an effort to use objects, often artificially created and different from the ones participants already know, to decouple generalization from experience. We are convinced that this principle will continue to play an important role. Decoupling generalization from experience will allow us to make stronger, valid claims about generalization abilities—compared with task performance—in humans, in AI systems, and in how the two differ. We envision a future in which the abilities of humans and AI systems are compared in a more systematic way—founded on the principles of measurement theory. First, an ability should be defined clearly, including a concise statement about its measurement. Second, the measurement of an ability should be based on a set of tasks from different domains to decouple the measurement from domain-specific experience. Third, the tasks should be unknown to the agents to decouple the ability from mere memory recall. A first step in such a direction is the abstract reasoning corpus by Chollet (2019).

Conclusion

Humans have been portrayed as excellent at generalization, especially compared with deep neural networks. In this work, we presented a more detailed view and showed that humans often systematically fail tests of generalization. We then argued that successful generalization depends on two factors: discovering and learning an adequate task representation and learning accurate, efficient, and abstract representations of the objects involved in a task. If an experimenter sets up a complicated, multidimensional, nonlinear task structure, people may end up solving the task with an incorrect simple rule because they have a strong preference for such simple rules; or they may need thousands of training examples and feedback to eventually approximate the complicated function and generalize accordingly. In contrast, when the task requires the execution of a one-dimensional simple rule on a salient feature dimension, humans are likely to perform and generalize well immediately because they can use their lifetime of experience to quickly constrain the problem space. Thus, the true art of human generalization lies not in a universal ability to generalize but in the well-formed craft of using adapted, efficient representations in the face of limited processing capacity.

Recommended Reading

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). (See References). Shows that neural network models pick up those regularities that are most relevant within a given training data set (and must not necessarily be the features humans use).

Johansen, M., & Palmeri, T. J. (2002). (See References). Shows that people initially use rules to generalize to unobserved objects but later on change to exemplar-based processing.

Nosofksy, R. M. (1986). (See References). Explains how the same principles of exemplar-based processing can explain generalization in absolute identification and category learning.

Shepard, R. N. (1987). (See References). Shows that the mapping from representations/descriptions of objects in physical space to psychological space is important when evaluating generalization in absolute identification tasks.

Taylor, J. E., Cortese, A., Barron, H. C., Pan, X., Sakagami, M., & Zeithamova, D. (2021). (See References). Provides an extensive review of generalization phenomena with a particular focus on neurobiology.

Transparency

Action Editor: Robert L. Goldstone
Editor: Robert L. Goldstone
Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
This work was funded by the Institute for Human-Centered AI at the Helmholtz Center for Computational Health.

ORCID iDs

Mirko Thalmann https://orcid.org/0009-0007-5611-7629
Eric Schulz https://orcid.org/0000-0003-3088-0371

Acknowledgments

We thank Marcel Binz, Luca M. Schulze Buschoff, Akshay Kumar Jagadish, and Konstantinos Voudouris for helpful discussions and feedback on a previous version of this manuscript.

References

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154–179. https://doi.org/10.1037/0033-295X.93.2.154

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27. https://doi.org/10.1016/0030-5073(74)90002-6

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22. https://doi.org/10.1016/S1364-6613(02)00005-0

Chollet, F. (2019). *On the measure of intelligence*. arXiv. https://doi.org/10.48550/arXiv.1911.01547

Dasgupta, I., Grant, E., & Griffiths, T. L. (2022). *Distinguishing rule- and exemplar-based generalization in learning systems*. arXiv. https://doi.org/10.48550/arXiv.2110.04328

De Simoni, C., & von Bastian, C. C. (2018). Working memory updating and binding training: Bayesian evidence supporting the absence of transfer. *Journal of Experimental Psychology: General*, *147*, 829–858. https://doi.org/10.1037/xge0000453

DeLosh, E. L., McDaniel, M. A., & Busemeyer, J. R. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968–986.

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 933–948. https://doi.org/10.1037/xlm0000083

Feldman, J. (1992). Constructing perceptual categories. In *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 244–250). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/CVPR.1992.223268

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 32*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, *130*(1), 116. https://doi.org/10.1037/0096-3445.130.1.116

Goldwater, M. B., Don, H. J., Krusche, M. J. F., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General*, *147*(1), 1–35. https://doi.org/10.1037/xge0000387

Gurnee, W., & Tegmark, M. (2024). *Language models represent space and time*. arXiv. https://doi.org/10.48550/arXiv.2310.02207

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32 (NeurIPS 2019)*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0f-f2ec19b371514-Paper.pdf

Jagadish, A. K., Binz, M., Saanum, T., Wang, J. X., & Schulz, E. (2023). *Zero-shot compositional reinforcement learning in humans*. PsyArXiv. https://doi.org/10.31234/osf.io/ymve5

Johansen, M., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*(4), 482–553. https://doi.org/10.1016/S0010-0285(02)00505-4

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, *8*(2), 225–248. https://doi.org/10.1080/095400996116893

Lake, B. M., & Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature*, *623*(7985), 115–121. https://doi.org/10.1038/s41586-023-06668-3

Little, D. R., & Shiffrin, R. M. (2009). Simplicity bias in the estimation of causal functions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*, 1157–1162.

Marcus, G. (2020). *The next decade in AI: Four steps towards robust artificial intelligence*. arXiv. https://doi.org/10.48550/arXiv.2002.06177

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80. https://doi.org/10.1126/science.283.5398.77

Mason, A., Szollosi, A., & Newell, B. (2022). *Learning the lie of the land: How people construct mental representations of distributions*. PsyArXiv. https://doi.org/10.31234/osf.io/rdhv8

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": Evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*(4), 512–534. https://doi.org/10.1177/1745691616635612

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Schulze Buschoff, L. M., Akata, E., Bethge, M., & Schulz, E. (2025). Visual cognition in multimodal large language models. *Nature Machine Intelligence*, *7*(1), 96–106. https://doi.org/10.1038/s42256-024-00963-y

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron*, *60*(2), 378–389. https://doi.org/10.1016/j.neuron.2008.09.023

Taylor, J. E., Cortese, A., Barron, H. C., Pan, X., Sakagami, M., & Zeithamova, D. (2021). How do we generalize? *Neurons, Behavior, Data Analysis and Theory*, *1*, Article 001c.27687. https://doi.org/10.51628/001c.27687

Thalmann, M., Schäfer, T. A. J., Theves, S., Doeller, C. F., & Schulz, E. (2024). Task imprinting: Another mechanism of representational change? *Cognitive Psychology*, *152*, Article 101670. https://doi.org/10.1016/j.cogpsych.2024.101670

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 37–55. https://doi.org/10.1037/xlm0000578

Vafa, K., Chen, J. Y., Rambachan, A., Kleinberg, J., & Mullainathan, S. (2024). *Evaluating the world model implicit in a generative model*. arXiv. https://doi.org/10.48550/arXiv.2406.03689

van Dyck, L. E., Kwitt, R., Denzler, S. J., & Gruber, W. R. (2021). Comparing object recognition in humans and deep convolutional neural networks—An eye tracking study. *Frontiers in Neuroscience*, *15*, Article 750639. https://doi.org/10.3389/fnins.2021.750639

Vermaercke, B., Cop, E., Willems, S., D'Hooge, R., Op de & Beeck, H. P. (2014). More complex brains are not always better: Rats outperform humans in implicit category-based generalization by implementing a similarity-based strategy. *Psychonomic Bulletin & Review*, *21*(4), 1080–1086. https://doi.org/10.3758/s13423-013-0579-9

Wu, S., Thalmann, M., & Schulz, E. (2023). *Motif learning facilitates sequence memorization and generalization*. PsyArXiv. https://doi.org/10.31234/osf.io/2a49z