nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-025-01109-4

Aligning generalization between humans and machines

Received: 11 December 2024

Accepted: 6 August 2025

Published online: 15 September 2025

Check for updates

Filip Ilievski[1] ✉, Barbara Hammer[2], Frank van Harmelen[1], Benjamin Paassen[2], Sascha Saralajew[3], Ute Schmid[4], Michael Biehl[5], Marianna Bolognesi[6], Xin Luna Dong[7], Kiril Gashteovski[3,8], Pascal Hitzler[9], Giuseppe Marra[10], Pasquale Minervini[11,12], Martin Mundt[13], Axel-Cyrille Ngonga Ngomo[14], Alessandro Oltramari[15], Gabriella Pasi[16], Zeynep G. Saribatur[17], Luciano Serafini[18], John Shawe-Taylor[19], Vered Shwartz[20,21], Gabriella Skitalinskaya[22], Clemens Stachl[23], Gido M. van de Ven[10] & Thomas Villmann[24,25]

Recent advances in artificial intelligence (AI)—including generative approaches—have resulted in technology that can support humans in scientific discovery and forming decisions, but may also disrupt democracies and target individuals. The responsible use of AI and its participation in human–AI teams increasingly shows the need for AI alignment, that is, to make AI systems act according to our preferences. A crucial yet often overlooked aspect of these interactions is the different ways in which humans and machines generalize. In cognitive science, human generalization commonly involves abstraction and concept learning. By contrast, AI generalization encompasses out-of-domain generalization in machine learning, rule-based reasoning in symbolic AI, and abstraction in neurosymbolic AI. Here we combine insights from AI and cognitive science to identify key commonalities and differences across three dimensions: notions of, methods for, and evaluation of generalization. We map the different conceptualizations of generalization in AI and cognitive science along these three dimensions and consider their role for alignment in human–AI teaming. This results in interdisciplinary challenges across AI and cognitive science that must be tackled to support effective and cognitively supported alignment in human–AI teaming scenarios.

Recent advances in AI enable meaningful support for humans in complex tasks, such as scientific discovery and decision-making[1]. Conversely, AI can also potentially disrupt democracies and target individuals[2]. The responsible use of AI increasingly motivates AI alignment, which aims to "make AI systems act according to our preferences"[3]. AI alignment is essential for effective human–AI teaming in complex scenarios where neither humans nor AI perform well on their own[4]. For example, AI can help invent novel biomedical application hypotheses following the objectives and guidance of a scientist[5]. Alternatively, humans can iteratively edit samples provided by an AI model to improve its accuracy and trustworthiness, for example, in classifying skin cancer[6]. As well as human–AI teaming, AI alignment is necessary for its safe use[7] and demonstrable adherence to accountability, privacy, and transparency requirements in legal frameworks such as the EU's AI Act[8].

A crucial, yet often overlooked aspect of this alignment in interactive scenarios is the complementary ways in which humans and machines generalize (Fig. 1). Generalization is typically defined as 'the process of transferring knowledge or skills from specific instances or

A full list of affiliations appears at the end of the paper. ✉e-mail: f.ilievski@vu.nl

https://doi.org/10.1038/s42256-025-01109-4

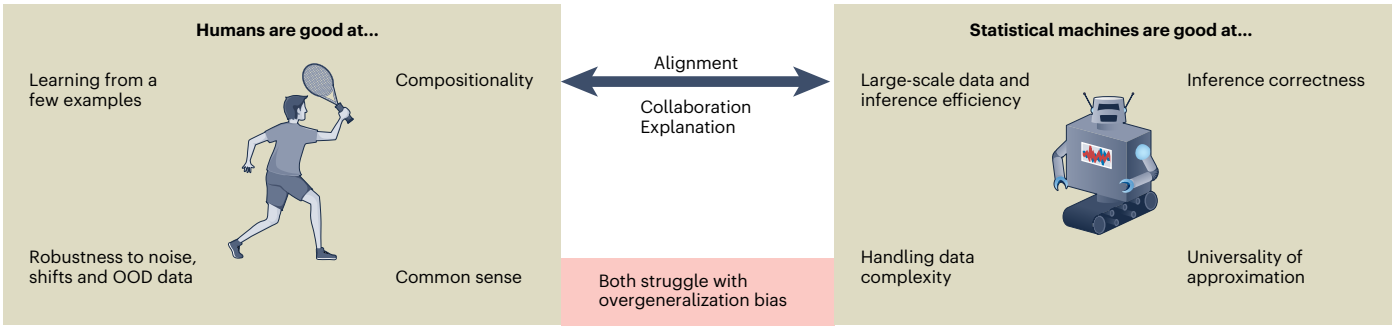

Fig. 1 | Comparison of the strengths of humans and statistical machines, illustrating their complementary generalization in human–AI teaming scenarios. Humans excel at compositionality, common sense, abstraction from a few examples and robustness. Statistical ML excels at large-scale data and inference efficiency, inference correctness, handling data complexity, and the universality of approximation. Overgeneralization biases remain challenging for both humans and machines. Collaborative and explainable mechanisms are key to achieving alignment in human–AI teaming. See Table 3 for a complete overview of the properties of machine methods, including instance-based and analytical machines.

exemplars to new contexts'[9]. Following cognitive science, human generalization commonly involves concept learning and the abstraction of general characteristics to a collection of entities[10]. Humans excel at generalizing from a few examples, compositionality, and robust generalization to noise, shifts, and out-of-distribution (OOD) data[11]. Humans can learn from little data and seemingly generalize beyond the observed distribution largely because, through evolution, experience, or both, they have access to strong causality-driven common sense priors at multiple hierarchical levels that characterize physical principles in nature and human behaviour in interactions.

By sharp contrast, data-driven (statistical) AI systems struggle to generalize beyond their training distribution and to abstract effectively. Although some neural architectures might display better alignment with physical laws[12], the generalizability of statistical machines, averaged over all distributions and in the absence of prior knowledge, is constant (no-free-lunch theorem[13]).

The goal of human–machine teaming[14] is that each side addresses the limitations of the other while aligning on similar goals. For example, some generalization capabilities of large language models (LLMs), like the quick production of rhetorically polished texts on any topic, are beyond those of most humans. However, their overgeneralization errors ('hallucinations'[15]), like replacing specific facts with nonfactual information, can be easily caught by a human expert. Effective teaming requires that humans can assess AI responses and access its underlying rationales ('explanations') (Fig. 1).

The complementarity of humans and AI, and the requirements for effective human–AI teaming, shed new light on traditional knowledge-informed (analytical) and instance-based AI paradigms. Analytical methods provide compositionality and accessible semantics, albeit in limited scenarios[16], whereas instance-based models are robust to distributional shifts when adequate representation is available[17]. Combining the strengths of various machine methods has inspired emerging research directions under the umbrella of neurosymbolic AI[18].

This Perspective draws on insights about the generalization of humans and machines from AI and cognitive science. We analyse three dimensions from the perspective of AI alignment: notions of, methods for, and evaluation of generalization. We focus on the following questions: What are the known notions of generalization? What are the strengths and weaknesses of the generalization of AI methods? How is generalization evaluated today? What is the impact of current trends in AI, such as foundation models, on generalization theories, methodological frameworks, and evaluation practices? Addressing these questions highlights the need for interdisciplinary approaches for effective and cognitively supported alignment of human and AI generalization.

Parallels in generalization by humans and machines

Approaches to generalization have been proposed in AI as well as in cognitive psychology, and they often mutually inspired each other. This holds for all types of approaches, whether rule-based, symbolic and knowledge-informed, case- and analogy-based, as well as neural and statistical. In the following, the mutual influence between AI methods and cognitive psychology will be illustrated by selected historical milestones, summarized in Fig. 2.

In the early days of cognitive psychology, Bruner et al.[19] empirically investigated human concept learning (Fig. 2a), which inspired the first decision-tree learning algorithms[20]. Observations on human learning of relational concepts inspired early machine learning (ML) approaches to learning from structural representations[21,22] and recursive concepts[23]. This class of approaches, often referred to as inductive programming, allows learning from few examples and taking into account background knowledge for model induction[24]. Rule learning approaches have been extended to statistical relational learning[25] to overcome the brittleness of symbolic learning. Bayesian approaches to rule learning have been introduced as a plausible framework to model human learning in complex domains such as language acquisition[26].

Rule learning as an explicit approach (system 2)[27] is apparent for domains of high-level cognition where relevant features can be verbalized[28]. Other cognitive theories have been proposed for domains where knowledge is not (entirely) available in explicit form. For example, prototype theory was proposed as a similarity-based approach where entities are grouped into categories for which similarity within category borders is maximized and between categories minimized[29] (Fig. 2b). This approach is reflected in similarity-based methods to ML, especially k-nearest neighbours[30]. Exemplar theories[31] have been proposed to address the flexibility of human categorization. For instance, context-dependence of the classification of visual objects has been shown by Labov[32] where a cup might be classified as a bowl or a vase when different contexts, for example, soup or flowers, are introduced. Another similarity-based approach is analogical reasoning, addressing knowledge transfer from one situation to another, often from a different domain[33–35]. In contrast to other methods, the analogy is not based on feature but on structural similarity.

Neural network approaches were proposed by cognitive scientists[36,37] as a method of generalization learning that overcomes the brittleness of symbolic approaches (Fig. 2c). Despite strong arguments from researchers in symbolic AI[38], neural networks and other statistical approaches became the dominant branch of ML due to their superior performance in increasingly large datasets. However, some core concerns by Fodor and Pylyshyn[38] about the relationship between statistical ML and human cognition remain. Firstly, data are

https://doi.org/10.1038/s42256-025-01109-4

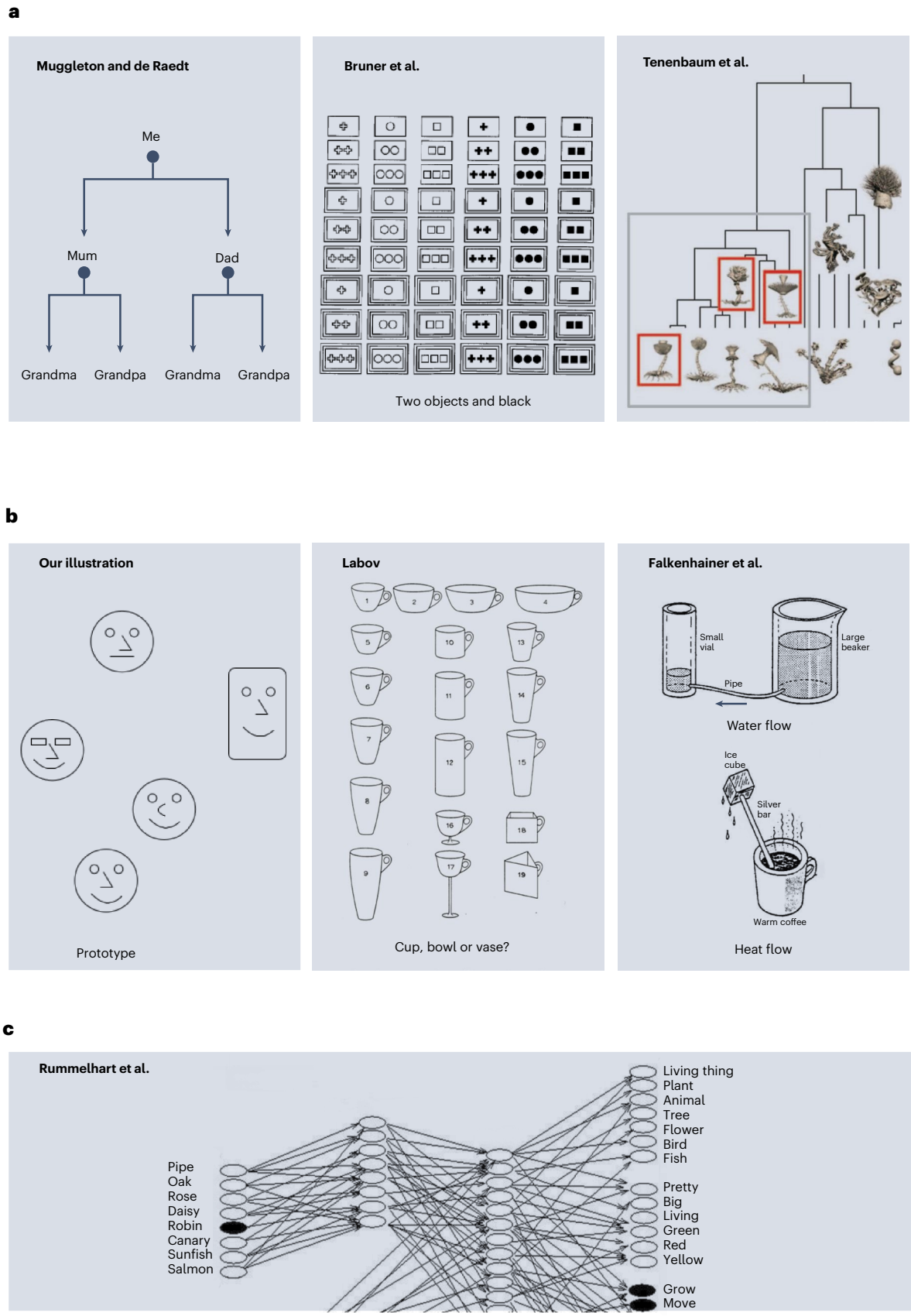

Fig. 2 | Illustrative examples of human generalization and its inspiration of rule-based, example-based and statistical ML approaches. a, Learning the relational rule 'grandparent' using a background theory 'parent' (left)[22], conjunctive rules (middle)[19] and names of alien objects modelled as Bayesian inference over a tree-structured domain representation (right)[26]. **b**, Example-based prototypical representations (left)[29,145], context-effects (middle)[32] and analogy (right)[34]. **c**, Statistical generalization: neural network model of semantic memory. Figure adapted with permission from: **a** (middle), ref. 19, Wiley; **a** (right), ref. 26, AAAS; **b** (middle), ref. 32, Georgetown Univ. Press; **b** (right), ref. 34, Elsevier; **c**, ref. 37, Cambridge Univ. Press.

separated from a semantic model. While humans who have learnt a concept robustly recognize OOD inputs, ML has only addressed this problem more recently[39]. Secondly, even if knowledge is primarily implicit in many domains, humans can verbally describe at least part of what constitutes a concept. This observation has recently been reflected in research on explainable AI, where novel approaches to explain black-box models have been proposed focusing on explanations based on concepts and relations[40]. Thirdly, human explanations are typically based on the causal history of an event and a causal explanation for the generalization itself[41]. Substantial empirical evidence has demonstrated that humans do not focus on the superficial level of event covariations, but reason and learn based on deeper causal representations[42]. In ML, discovering high-level causal variables from low-level observations remains a significant challenge[43].

The combination of implicit, neural learning and explicit, symbolic approaches is addressed in neurosymbolic AI research[18]. Neurosymbolic approaches promise to combine the complementary strengths of neural and analytical techniques, preserving robustness while enabling data-model separation and integrating available background knowledge[44]. Combining neural and symbolic approaches to generalization and explainability within human-centric AI reflects many aspects of human learning flexibility[45]. For effective joint decision-making and problem-solving, the human–AI interface must align with human information processing[46]. Alignment must be established for different aspects, including knowledge state, current information needs, or values[47]. In cognitive modelling, researchers aim to align algorithmic approaches to learning and human learning[48]. Recent results show that, to date, the performance of human–AI teams lags behind that of the best AI model or the best human alone in many domains[14].

Notions of generalization

In the broader context of cognitive science and AI research, there are three different notions of generalization, which we will cover in turn.

Generalization as a process

Generalization as a process refers to constructing concepts or rules from example data. In cognitive science, the process is typically called abstraction, either through associative learning, generalization by similarity, or the transformation of schemas from lower to higher stages of cognitive development[49]. More broadly, French[50] distinguishes (1) generalization of concrete instances into an abstract schema, which we call abstraction, (2) generalization through the application or extension of the schema to various situations, which we call extension, and (3) generalization involving the transformation/adaptation of the schema to fit a new context, which we call analogy. In AI methods, abstraction (1) corresponds to concept or rule mining in knowledge-informed AI or the learning of models from data in statistical AI[51]; extension (2) relates to online, multi-task, few-shot, or continual learning schemes in statistical and instance-based AI[52–54]; and analogy (3) relates to transfer learning[55] in statistical AI and reasoning by analogy in analytical AI.

Importantly, a generalization process does not have to start from example data but may abstract, extend or transfer a pre-existing model beyond its original scope.

Generalization as a product

Generalization may also refer to the products of a generalization process, such as categories, concepts, rules and models, in their various representations. Generalizations of categories and concepts may be represented using a symbolic definition, as a list of attributes and their bounds (cognitive science[56]; or decision trees in AI[57]), as a prototype (cognitive science[58]; AI[59]), or as a set of exemplars of the category (cognitive science[31]; or *k*-nearest neighbour in AI[60]). Categories or concepts may also be represented as a probability distribution from which examples of this category can be drawn, which is the notion implicitly used by generative AI models[61]. Beyond categories or concepts, products may also be rules or relations, represented, for instance, via functions or graphs in parametric or non-parametric form.

Table 1 | Common categories to structure AI methods algorithmically centred

Category	Attributes
Training signal	Supervised, unsupervised, reinforcement, semi-supervised, self-supervised
Data type	Tabular data, data structures (for example, text, graph), prior knowledge
Model representation	Parametric/non-parametric, symbolic/sub-symbolic, black-/white-/grey-box
Training objective	Bayesian inference, maximum likelihood principle, rule learning, mean squared error minimization

These categories are not uniquely related to their type of generalization.

Generalization as an operator

Finally, we refer to generalization as an operator, namely the application of a product to new data. The successful application to new data is at the core of generalization in statistical AI[62]. Under the assumption that training data and new data are independently sampled from the same distribution (IID), one can mathematically prove generalization via one of three theories: (1) the Probably Approximately Correct (PAC) framework analyses whether a model (that is, a product) derived via an ML algorithm (that is, a process) from a random sample of data can be expected to achieve a low prediction error on new data from the same distribution in most cases[62]. (2) Statistical physics of learning aims to understand the typical properties of learning algorithms (that is, processes) with many adaptive parameters[63]. (3) Vapnik–Chervonenkis (VC) dimension theory focuses on the storage capacity of model classes and their subsequent ability to make accurate predictions on new data[64].

Generalization theory in ML is limited in several ways. It typically predicts that generalization only occurs if the available data are large enough to not just be memorized[65]. By contrast, humans can generalize from a few samples for a specific task, as generalization in humans is not a singular event but based on lifelong experience of regularities observed in nature. Few-shot learning addresses this to some extent[66,67].

Ultimately, all three theories have been primarily applied to abstraction processes. For model extension or analogy, the mathematical theory is less well established. Of particular interest is generalization across compositions—for instance, in language—which has been addressed in analytical AI[68], but is limited by the undecidability or high complexity of inference[69]. Despite such fundamental restrictions, humans operate with compositions similar to those found in language and vision.

Alignment of human and machine notions of generalization

We observe that human and machine notions of generalization are misaligned. While humans tend towards sparse abstractions and conceptual representations that can be composed or transferred to new domains via analogical reasoning, generalizations in statistical AI tend to be statistical patterns and probability distributions, which can sometimes be extended but still fail to generalize to tasks and domains that are too far removed from the training data. In other words, because humans and machines use different processes (for example, abstraction versus data-driven learning), they arrive at different products (for example, categories and rules versus probability distributions) that generalize differently; if we wish to achieve human-like generalization ability (as an operator), we need new methods for machine generalization.

Machine methods for generalization

Humans excel in systematic generalization across representations, contexts and tasks based on a few observations. Although specific machine methods have shown remarkable results in solving compositional tasks[70], the underlying mechanisms of human generalization are not sufficiently understood to mimic them in artificial systems[26].

AI methods are usually structured according to algorithmic aspects rather than their generalizability (Table 1). Although these impact generalization behaviour—for instance, symbolic methods often implement compositionality—there is no simple mapping to the form of generalization that an AI model can achieve. Therefore, we focus on another categorization, the interplay of observational data (that is, single instances) and models (that is, principles that apply to a whole population), as this correlates to the generalizability of the model and its suitable evaluation ('Evaluation of generalization' section). Three categories can be distinguished: (1) the transfer of individual observations to a population is the basis for statistical generalization methods. (2) The search for observational evidence of an explicit theory is done in knowledge-informed methods. (3) Instance-based methods focus on individuals concerning the source and target of generalization. These choices have different characteristics in terms of their generalizability and alignment with human generalization.

Statistical generalization methods in AI

Many modern methods, including deep learning, aim at statistical generalization: observational data (that is, training data) serve as input to an inference mechanism that extracts a model for the entire population (that is, the underlying distribution). Generalization refers to the ability of the inferred patterns to be successfully applied to new data ('Generalization as an operator' section). Typically, algorithmic methods are expressed as optimization methods for a model's loss function, such as the prediction error. As it cannot be evaluated in the entire population, it is approximated in a given training set, known as empirical risk minimization[71]. Although evaluation in an independent test set constitutes an unbiased estimator of generalization ability ('Evaluation of generalization' section), the empirical loss in the training set systematically underestimates the model loss—generalization needs to be accounted for explicitly. Popular strategies regularize models towards a better generalization behaviour, such as maximum margin or stability[72]. Even heavily over-parameterized deep models display surprising generalization capabilities due to intrinsic regularization[73].

Statistical methods excel in inference correctness and efficiency. Yet, they typically require large training datasets. This challenge can be partially overcome by technologies that build their inference on already learnt representations and instance-based translations, such as few- or zero-shot mechanisms[67]. Still, empirical risk minimization has a fundamental limitation compared to human generalization: generalization can only be expected in areas covered by observations, but not for out-of-sample events, novel contexts or distributional changes[74]. Indeed, machine behaviour for OOD settings can significantly deviate from human expectation, with adversarial attacks as prominent examples of this phenomenon[75].

In recent years, many vital settings, including LLMs, have been targeted that do not allow for a simple analytic expression of human's intention. Thus, surrogate losses, such as next token probability, are used as a proxy. With massive training data, instruction tuning, or human feedback, impressive generalizability arises[66]. However, the emerging generalization abilities are only partially understood and do not necessarily align with human expectation, necessitating a downstream evaluation ('Evaluation of generalization' section) if possible at all—human intentions are not necessarily well formed or static, and the type of information an AI provides could influence human objectives in interactions[3].

Statistical generalization methods are often based on model families with universal approximation capability to account for the lack of domain-specific knowledge. Deep models, for example, can deal with high degrees of nonlinearity and multimodal signals[76]. Yet, the product is typically a black box, which does not reveal insight into its generalization behaviour; hence, partially unintended generalization behaviour can easily occur. Recently developed post hoc explanation methods allow for a closer inspection of the underlying rationale and its impact on the generalization behaviour of the model[77].

Knowledge-informed generalization methods in AI

Knowledge-informed generalization methods aim to find empirical evidence of a theory, resulting in a meaningful representation confirmed by the data. Popular methods include mechanistic models[78], causal models[79] or functional programs[80]. As their semantics is directly accessible, humans can inspect how these models generalize to previously unencountered scenarios. Generalization is often well aligned with human expectations. Yet, model parameters require semantic grounding, which is challenging to realize with low-level sensor data. Neurosymbolic integration can partially overcome such limitations[81].

Learning the optimal model structure is demanding, and fundamental limitations such as non-identifiability of structural components might exist[16]. Hence, many methods are restricted to simple schemes, such as description logic, rather than universal approximators. Learning methods such as semantic clustering, probabilistic rule mining, subsumption, or analogies mirror specific representations. Since limited noise robustness and inference efficiency pose significant challenges, hybrid approaches have emerged, such as a transfer of symbolic models to a real-valued embedding space where efficient numeric inference is possible[82]. Knowledge-based approaches enable the explicit inspection and manipulation of information, allowing generalization based on a few examples, as accompanying rules can ensure valid generalization. Yet, they are restricted to domains where a theory can be formalized with reasonable effort. As this is often limited, current models do not reach the impressive capabilities of statistical approaches trained on massive datasets.

Systematic compositionality refers to the ability to generalize and produce novel combinations from known components. It has been fundamental in the design of traditional, logic-based systems; yet statistical methods have struggled with compositional generalization[38]. Compositionality seems to be a universal principle in nature, since it has been observed in many species[83]. In recent years, significant progress has been made in improving compositional generalization in deep learning, typically by adding components that mirror the compositional structure of the domain, such as structure-processing neural networks[84] or metalearning for compositional generalization[70]. Although these efforts provide a pathway for neural networks to generalize systematically, most of the results are only empirical[85]. There remains a significant gap between the systematic generalization capabilities of knowledge-informed models and the representation learning techniques of deep models, with neurosymbolic AI promising a viable bridge[18].

Instance-based translation in AI

Instance-based non-parametric techniques, such as nearest-neighbour methods or case-based reasoning[86], rely on local inference, which is computed when needed based on similar cases encountered previously. They are among the most popular ML methods, showing high flexibility when combined with complex representations[87]. Since they adjust their complexity as needed, they offer universal approximation capability. Suitable data structures enable efficient training and inference in large datasets. Furthermore, local inference usually allows human inspection of individual decisions—although not of the entire model. Instance-based methods closely resemble concepts in cognitive linguistics, such as a graded degree of belonging to a category, which can be represented by a prototype[58].

Instance-based models have shown great promise for incremental learning of distributional shifts[17]. They can identify out-of-sample instances based on their similarity to previously encountered data, and they can naturally deal with the challenge of catastrophic forgetting in continual or lifelong learning as they can memorize possibly relevant data points. This principle also suggests possible solutions to catastrophic forgetting in continual learning using deep statistical models[88]. Conversely, it is possible to implement forget mechanisms if older instances become invalid. The reliance of instance-based methods on similarity means that a suitable representation is key to support generalization[89], as it directly influences the model's ability to evolve patterns across diverse datasets and tasks. Recent work investigates how to achieve representations to support generalizations across tasks or domains[90].

Context has a unique role as generalization requires adapting knowledge learnt in one setting to fit a novel, unseen one. Humans can cope with the challenge of acquiring context knowledge and assessing the similarity of two contextual representations[91]. ML techniques such as transfer learning, prompting or retrieval augmented generalization mimic parts of this process[92]. In this realm, LLMs have demonstrated remarkable capabilities for few-shot or in-context learning[66], often still relying on implicit contextual information. An explicit representation of contextual knowledge and its compositionality, for example, through neurosymbolic AI, is the subject of ongoing research.

Aligning machine generalization methods and human expectations

While statistical methods offer powerful universal approximation, their generalization behaviour does not match human generalization well, lacking generalization to OOD samples and compositionality. Another challenge is their black-box nature, where post hoc explanations provide solutions for specific cases. By contrast, knowledge-based methods enable human insight and compositionality by design, but often at the expense of universality and algorithmic efficiency in the face of structure learning. Emerging neurosymbolic approaches aim to combine these methodological principles. Instance-based methods try a different approach, focusing on generalization from and to individual instances. This principle is well aligned with human generalization and enables learning from a few data points and lifelong learning scenarios; yet, results depend strongly on the choice of representation and context. Recent statistical methods for representation learning offer promising directions that enable machines to generalize from a few examples, much like humans. Ultimately, making claims about the generalization properties of various machine approaches requires a meaningful evaluation, which we discuss next.

Evaluation of generalization

The theoretical generalization strengths and weaknesses of the machine method families are summarized in Table 2. Statistical methods enable universality of approximation and inference correctness, excel at handling data complexity and large-scale data, and ensure inference efficiency. Analytical methods support compositionality, explainable predictions, and perform explicit knowledge manipulation. Instance-based methods support robustness to noise, shifts and OOD data, memorize training samples reliably, and can learn from a few samples.

Deriving provable robustness and generalization guarantees is necessary to define the theoretical limits of models. Meanwhile, empirically evaluating the machine's generalization is also desirable. From a statistical learning perspective, evaluating the generalization of supervised approaches estimates their applicability to new data ('Generalization as an operator' section). This formalization of measuring inference correctness on IID data is theoretically grounded and remains relevant when assessing systems. It allows measuring the universality of approximation and the ability to handle large-scale data and perform efficient inference on more complex datasets by increasing the task complexity. However, with increasing task complexity and system opaqueness, it becomes challenging to guarantee the generalizability assumptions: IID and task-representative data. For example, Li and Flanigan[93] found that ChatGPT performed well on benchmarks released before its launch but poorly on those published afterwards. Here, test set leakage into ChatGPT violates the IID assumption, invalidating generalization estimates on pre-release benchmarks. The lack of transparency about the data used to train LLMs makes it challenging to create novel test sets, leading to possible overestimations of the model's generalization[45,94]. Namely, while the emergence of foundation models has enabled evaluation of learning from a few examples via zero- and few-shot tests[67], the risk of test set memorization means that test data may appear partially in their training set, invalidating the findings (data contamination[95]).

Table 2 | Characterization of AI generalization methods

Pros	Cons
Statistical: generalization from observations to a population	
Universal approximation, surprising generalization of deep models	Black boxes, generalization failures outside of training distribution
Knowledge-informed: confirm/adapt hypothesis based on observations	
Meaningful models, identifiable parameters, generalization in the limit, compositionality	Restriction to simple scenarios, optimization/structure identification computationally demanding
Instance-based: translation from previous observations to a new observation	
Flexible to change/distributional shift	Rely on suitable representation

The following discusses key areas related to the evaluation of generalization and its role in AI applications.

Measuring distributional shifts

Distribution shifts can be estimated using statistical distance measures such as the Kullback–Leibler divergence between the feature distributions of the training and test sets[96]. Generative models produce an explicit likelihood estimate $p(x)$ that indicates how typical a sample is to the training distribution. Since discriminative models do not offer this possibility, proxy techniques include calculating cosine similarity between embedding vectors and using nearest-neighbour distances in a transformed feature space. For LLMs, a standard proxy measure for familiarity is perplexity. When the model's internal representations cannot be directly accessed, the layers of non-linear abstractions in modern (deep) ML models allow for gauging relations through intermittent embeddings. Learning evaluation in the context of drift can be done using tailored benchmarks[97]. The model's robustness to noise, distributional shifts and OOD data can be studied using adversarial and counterfactual techniques. Adversarial techniques alter data features, such as syntax, semantics or context, while preserving the underlying task and the original label[75]. By contrast, counterfactual techniques create data samples that alter target prediction with minimal input changes[94].

Determining under- and overgeneralization

AI models are created to provide value to humans and are thus assessed against human generalization, often using the human-centric concepts of under- and overgeneralization. Undergeneralization occurs when a change in the input, perceptible or imperceptible, causes a considerable modification within a model. Examples of undergeneralization include model performance degradation under various natural changes, such as environmental perturbations in computer vision[98]. For foundation models, prompt engineering substantially affects performance[99]. By contrast, models overgeneralize, which means that they over-confidently make false predictions for (known or novel) concepts because critical differences are ignored in prediction[100]. A

Table 3 | Desired properties of generalization that emerge from the 'Notions of generalization' and 'Machine methods for generalization' sections

Property	Method			Evaluation
	S	A	I	
Inference correctness	+	–	–	Train–test splits
Universality of approximation	+	–	+	Increasing task complexity
Large-scale data and inference efficiency	+	–	+	Large/complex datasets
Learning from a few samples	–	+	+	Zero-/few-shot tests
Robustness to noise, shifts, OOD data	–	–	+	Adversarial, shifted, counterfactual tasks
Compositionality	–	+	–	Analogy, abstraction, concept induction
Handling data complexity	+	–	–	Multimodal datasets
Memorization of training samples	–	–	+	Factuality datasets, precedents
Explicit knowledge manipulation	–	+	–	Causality, argumentation, theorem proving
Explainable predictions	–	+	–	Human studies, faithfulness

The properties are listed in order of their appearance in the 'Evaluation of generalization' section. Each of the properties is supported by statistical (S), analytical (A) and instance-based (I) AI methods, as discussed in the 'Machine methods for generalization' section. The evaluation methods for each property are discussed and substantiated by the references in the 'Evaluation of generalization' section. Achievement is indicated by ' + ' and non-achievement by ' – '; we use a strict evaluation to avoid partial scoring.

well-known overgeneralization phenomenon is hallucination, which refers to models that deviate from the source of the information[15]. Other overgeneralizations are biased predictions, for example, when a model predicts a property of an individual from the statistical properties of a demographic group to which they belong[101], and logical fallacies[102].

Characterizing the model's under- and overgeneralization requires an appropriate metric, defining its use and establishing a mechanism to interpret the metric's score in terms of generalizability beyond the particular test examples. This procedure is susceptible to three caveats. Firstly, the choice between discriminative and generative models determines which representational basis is used to infer similarity[103]. For instance, a discriminative model will only learn representations useful to optimize classification accuracy, whereas a generative approach would additionally learn representations necessary to describe the data distribution. Secondly, deep models are prone to learning decision shortcuts and ignoring meaningful features[104], sometimes also called simplicity bias[105]. Thirdly, modern models are often proprietary and frequently updated, partly based on user interactions through reinforcement learning, which increases their exposure to datasets and hinders reproducibility. To protect against these caveats, besides using open-source models, it is critical to evaluate across different levels of abstraction, from surface forms to semantic similarity and higher-level structural mappings, and explicitly consider the application context and limits. Consequently, machines are increasingly tested for their ability to handle complex data, such as multimodal datasets[106], and to exhibit compositionality in tasks such as common sense reasoning[107], analogies[108] and concept induction[109].

Distinguishing memorization and generalization

In AI, memorization refers to learning details from the training data, including facts and noise. Memorization may be beneficial in some cases (for example, Paris is the capital of France), but detrimental in others (for example, Biden is the president of the United States). This observation raises a question: When should models generalize, when should they memorize, and how can this distinction be evaluated? Whether the models should generalize or memorize is set a priori. When learning from experience, generalization is crucial, for instance, in recognizing a new manifestation of a vase[32]. Consequently, generalization setups include cross-domain validation and robustness testing. By contrast, factual knowledge is often memorized: Paris is the capital of France, and mosquitoes fly. Tasks such as answering factual questions[110] and reasoning about legal precedents[111] require memorization. While evaluating memorization and generalization separately is informative, many tasks, including causal reasoning[112], argumentation[113], and theorem proving[114], require holistic integration of generalization and memorization, centred around explicit knowledge manipulation. The explicit knowledge use allows testing a model's explainable predictions through human studies and faithfulness evaluation[115].

Alignment of machine evaluation of generalization to humans

Effective alignment of AI with humans requires a principled evaluation of its strengths and weaknesses in generalization. Evaluating AI generalizability in the context of its alignment comprises: (1) deriving provable guarantees about AI's properties, like compositionality and learning from a few samples, and (2) performing empirical evaluation by leveraging task-specific benchmarks and metrics. The evaluation of AI generalizability measures its performance concerning distribution shifts, determining its undergeneralization (adaptability to task variations such as camera perturbations) and overgeneralization (for example, hallucinations), and its ability to memorize and generalize adequately when necessary. These three aspects are essential for alignment, since distributional shifts, task variations, and both facts and noise are natural in human–AI teaming scenarios. We summarize typical approaches for evaluating key generalization properties in Table 3. In the next section, we discuss open challenges for measuring generalization in the context of human–AI alignment.

Emerging directions

The previous sections addressed the challenges of aligning human and machine intelligence, emphasizing AI's potential to enhance human generalization. Table 3 summarizes the properties, methods and evaluation practices, and highlights the complementarity of the three methods in achieving important properties for aligning human and AI generalization. Statistical approaches enable universality of approximation and inference correctness, instance-based methods enable robustness and memorization, while analytical techniques are designed for compositionality and explainable predictions. The evaluation column displays the range of approaches used to assess various methods. Next, we discuss future research directions for novel generalization theories, hybrid methods, evaluation practices, and alignment in future human–AI teams.

Generalization theory in the era of foundation models

Recent zero-shot and in-context learning approaches in LLMs and large reasoning models[116] implicitly generalize to tasks unrelated to their training without explicit similarity[117]. In other words, model builders assume that LLMs have implicitly generalized (process; 'Generalization as a process' section) to generalizations (product; 'Generalization as a product' section) that allow generalization (operator; 'Generalization as an operator' section) to entirely new tasks and domains. However, this assumption remains unsubstantiated, leading to an overestimation of the generalizability of foundation models [116]. This highlights the need for further research. Firstly, new generalization processes and products are needed to provide guarantees (or at least reasons to believe) that zero-shot application to new tasks is viable, potentially through encoding invariances or equivariances[118], as used in complex architectures such as AlphaFold; or through cognitively inspired representations, such as prototypes, which have proven efficient for domain generalization[119]. Secondly, a new theory is required to define when

few- or zero-shot applications are feasible. It has recently been shown that—unlike in classical learning theory—high dimensionality of the signals might be key to the generalizability of few-shot learners and over-parameterized deep networks[120].

Generalizable neurosymbolic methods

Neurosymbolic AI combines robust, data-driven latent models with the precision of explicit compositional models[18]. However, several challenges remain: defining provable generalization properties, including worst-case bounds, is crucial. Recent work exploits the compositionality of neurosymbolic systems to derive upper and lower bounds for the robustness of generalizations[121]. Verifying correctness instead of just robustness remains an unaddressed problem.

Current symbolic representations in neurosymbolic systems are typically of low expressivity (knowledge graphs, propositional logic), allowing for only limited forms of generalization. Recent works explore the use of richer formalisms. The use of description logics in neurosymbolic systems is particularly relevant since description logics are designed to capture forms of generalization[122].

While a theory about the compositionality of neurosymbolic systems is emerging[123], a theory on how to compose the generalizations themselves is lacking. For symbolic representations, a theory on abstractions[124] exists, but the question of composing latent representations, such as embeddings, remains open.

Finally, handling the context dependency of generalizations remains a challenge. How do we measure the distance between contexts and apply generalization across contexts? How do we know when a context is too novel and we overgeneralize? A possible direction is formal modelling of contextual dimensions such as time and space, following prior research on axiomatizing common sense knowledge[125]. CYC's notion of hierarchical microtheories, each containing a collection of axioms[126], can be revisited from a neurosymbolic perspective, for instance, by expressing axioms of microtheories in flexible natural language representations[127].

Generalization in continual learning

The enormous effort required to train foundational models, and the increasing availability of pre-trained models, has led to a transition from models which are trained from scratch to systems based on foundation models that undergo continuous adaptation to new data or tasks. However, naive approaches carry a high risk of catastrophic forgetting, which, surprisingly, seems to be higher for larger LLMs[128].

The relation between continual learning and generalization is complex: continual learning methods are designed to combat catastrophic forgetting when generalizing to novel tasks, while within-task generalization facilitates faster learning and improved performance in subsequent continual learning tasks[129]. Thus, concept drift requires learning of the underlying features that can be readily applied to novel tasks. This can be achieved by an extension of statistical methods to hybrid approaches, which attend to the preservation of learnt signals: for example, formalizing domain rules as ontologies or symbolic constraints enables a system to detect drift whenever incoming data or model outputs violate these constraints, serving as an early warning signal for distributional change that may disrupt the generalization capabilities of the system. Recent work on graph streams[130] uses neurosymbolic prototypes, where representative subgraphs are embedded in vector spaces. Another remedy can be based on data-driven approaches, such as (possibly self-supervised) rehearsal technologies, for a robust memorization of important information, albeit at increased computational costs[131]. More efficient alternatives aim for architectural solutions such as the incorporation of instance-based representations into statistical models[132]. However, theoretical insight on the effect of overparametrisation or task similarity on the generalizability and forgetting of a model currently exists for very simple models only[133].

Evaluation of generalization in foundation models

Several directions have emerged to address data contamination, spurious correlations, and overfitting in state-of-the-art models. Abstraction benchmarks for visual reasoning[134], analogy[108], and lateral thinking[135] are gaining popularity. Crowdsourcing can be used to create and scale benchmarks, but it can also introduce cognitive and cultural biases by annotators[136], which remains poorly understood. On the other hand, evaluation servers and public leaderboards with private test datasets prevent overfitting but lack standardization and are costly to maintain. A final direction is simulation environments and synthetic data generators[137], though they often suffer from a sim-to-real gap. To address reproducibility, researchers proposed the model[138] and data cards[139] to report the details of the experiment, and reproducibility checklists based on a broad consensus[140], albeit with limited coverage of generalizability.

Aligning generalization in future human–AI teams

The goal of effective human–AI teaming and the appearance of legal frameworks such as the EU AI Act[7,8] require transparent collaboration workflows, with explanations bridging the gaps between human and AI reasoning[45,141]. The 'Parallels in generalization by humans and machines' section discussed that this alignment must occur at the output level. However, when misalignments occur, (for example, AI predicts tumour type 1 and the doctor diagnoses tumour type 3), mechanisms for realignment and error correction become critical. Such mechanisms pose stricter requirements for collaboration on the process level through concepts and relations[48]. Examples of realignment techniques include language games, where realignment emerges from interaction, and physics-informed models that refine predictions on object permanence.

A critical challenge of human–AI teaming is reconciling the fundamentally different reasoning paradigms of humans and AI, like human causal models and AI's deep learning associations. Efforts such as concept-based explanations[142] and those considering relationships[143] suggest the potential for intertranslatability into a common explanatory language.

Furthermore, robust evaluation frameworks should consider both objective task-related outcomes and subjective process-related experiences, as well as the long-term ramifications of the collaboration, taking into account each party's contributions and responsibilities. Despite the emergence of evaluation frameworks and metrics for humans that augment AI (for example, in manual data labelling), AI that helps humans (for example, conversational question answering), and balanced collaborations where both contribute equally (for example, medical decision-making)[144], there is little research on evaluating the generalization capabilities of such teams.

References

1. Jumper, J. M. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
2. Ferrara, E. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *J. Comput. Social Sci.* **7**, 549–569 (2024).
3. Carroll, M., Foote, D., Siththaranjan, A., Russell, S. & Dragan, A. AI alignment with changing and influenceable reward functions. In *Proc. 41st International Conference on Machine Learning* 5706–5756 (JMLR, 2024).
4. Metcalfe, J. S., Perelman, B. S., Boothe, D. L. & Mcdowell, K. Systemic oversimplification limits the potential for human-AI partnership. *IEEE Access* **9**, 70242–70260 (2021).
5. Gottweis, J. et al. Towards an AI co-scientist. Preprint at https://doi.org/10.48550/arXiv.2502.18864 (2025).
6. Donnelly, J. et al. Rashomon sets for prototypical-part networks: Editing interpretable models in real-time. In *Computer Vision and Pattern Recognition Conference* 4528–4538 (CVPR, 2025).

7. Bengio, Y. et al. Managing extreme AI risks amid rapid progress. *Science* **384**, 842–845 (2024).
8. Bellogín, A. et al. The EU AI act and the wager on trustworthy AI. *Commun. ACM* **67**, 58–65 (2024).
9. Son, J. Y., Smith, L. B. & Goldstone, R. L. Simplicity and generalization: short-cutting abstraction in children's object categorizations. *Cognition* **108**, 626–638 (2008).
10. Harnad, S. in *Handbook of Categorization in Cognitive Science* 2nd edn (eds Cohen, H. & Lefebvre, C.) 21–54 (Elsevier Academic Press, 2017).
11. Holzinger, A. et al. Toward human-level concept learning: pattern benchmarking for AI algorithms. *Patterns* **4**, 100788 (2023).
12. Lin, H. W., Tegmark, M. & Rolnick, D. Why does deep and cheap learning work so well? *J. Stat. Phys.* **168**, 1223–1247 (2017).
13. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82 (1997).
14. Vaccaro, M., Almaatouq, A. & Malone, T. When combinations of humans and ai are useful: a systematic review and meta-analysis. *Nat. Hum. Behav.* **8**, 2293–2303 (2024).
15. Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 248 (2023).
16. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).
17. Losing, V., Hammer, B. & Wersing, H. KNN classifier with self adjusting memory for heterogeneous concept drift. In *IEEE International Conference on Data Mining* 291–300 (IEEE, 2016).
18. Hitzler, P. et al. (eds) *Compendium of Neurosymbolic Artificial Intelligence* (IOS Press, 2023).
19. Bruner, J., Goodnow, J. J. & Austin, G. A. *A Study of Thinking* (Wiley, 1956).
20. Hunt, E. B., Marin, J. & Stone, P. J. *Experiments in Induction* (Academic Press, 1966).
21. Winston, P. H. *Learning Structural Descriptions From Examples.* AI Technical Report (Massachusetts Institute of Technology, 1970).
22. Muggleton, S. & De Raedt, L. Inductive logic programming: theory and methods. *J. Logic Program.* **19**, 629–679 (1994).
23. Schmid, U. & Kitzelmann, E. Inductive rule learning on the knowledge level. *Cognit. Syst. Res.* **12**, 237–248 (2011).
24. Gulwani, S. et al. Inductive programming meets the real world. *Commun. ACM* **58**, 90–99 (2015).
25. De Raedt, L. & Kersting, K. in *Probabilistic Inductive Logic Programming: Theory and Applications* 1–27 (Springer, 2008).
26. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
27. Kahneman, D. *Thinking, Fast and Slow* (Macmillan, 2011).
28. Lafond, D., Lacouture, Y. & Cohen, A. L. Decision-tree models of categorization response times, choice proportions, and typicality judgments. *Psychol. Rev.* **116**, 833 (2009).
29. Rosch, E. & Mervis, C. B. Family resemblances: studies in the internal structure of categories. *Cognit. Psychol.* **7**, 573–605 (1975).
30. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**, 21–27 (1967).
31. Nosofsky, R. M. Exemplar-based accounts of relations between classification, recognition, and typicality. *J. Exp. Psychol. Learn. Mem. Cognit.* **14**, 700 (1988).
32. Labov, W. in *New Ways of Analyzing Variation in English* (eds Bailey, C.-J. N. & Shuy, R. W.) 67–90 (Georgetown Univ. Press, 1973).
33. Gentner, D. Structure-mapping: a theoretical framework for analogy. *Cognit. Sci.* **7**, 155–170 (1983).
34. Falkenhainer, B., Forbus, K. D. & Gentner, D. The structure-mapping engine: algorithm and examples. *Artif. Intell.* **41**, 1–63 (1989).
35. Forbus, K. D., Ferguson, R. W., Lovett, A. & Gentner, D. Extending SME to handle large-scale cognitive modeling. *Cognit. Sci.* **41**, 1152–1201 (2017).
36. Rumelhart, D. E. & Todd, P. M. in *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience* (eds Meyer, D. E. & Kornblum, S.) 3–30 (MIT Press, 1993).
37. Rumelhart, D. E., McClelland, J. L. & PDP Research Group. *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations* (MIT Press, 1986).
38. Fodor, J. A. & Pylyshyn, Z. W. Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3–71 (1988).
39. Yang, J., Zhou, K., Li, Y. & Liu, Z. Generalized out-of-distribution detection: a survey. *Int. J. Comput. Vision* **132**, 5635–5662 (2024).
40. Achtibat, R. et al. From attribution maps to human-understandable explanations through concept relevance propagation. *Nat. Mach. Intell.* **5**, 1006–1019 (2023).
41. Miller, T. Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019).
42. Waldmann, M. R., Hagmayer, Y. & Blaisdell, A. P. Beyond the information given: causal models in learning and reasoning. *Curr. Directions Psychol. Sci.* **15**, 307–311 (2006).
43. Schölkopf, B. et al. Toward causal representation learning. *Proc. IEEE* **109**, 612–634 (2021).
44. Marcus, G. F. *The Algebraic Mind: Integrating Connectionism and Cognitive Science* (MIT Press, 2003).
45. Ilievski, F. *Human-Centric AI with Common Sense* (Springer, 2025).
46. Ji, J. et al. AI alignment: a comprehensive survey. Preprint at https://doi.org/10.48550/arXiv.2310.19852 (2023).
47. Butlin, P. AI alignment and human reward. In *2021 AAAI/ACM Conference on AI, Ethics, and Society* 437–445 (ACM, 2021).
48. Langley, P. & Simon, H. A. in *Cognitive Skills and Their Acquisition* 361–380 (Psychology Press, 2013).
49. Colunga, E. & Smith, L. B. The emergence of abstract ideas: evidence from networks and babies. *Philos. Trans. R. Soc. London Ser B* **358**, 1205–1214 (2003).
50. French, R. M. *The Subtlety of Sameness: A Theory and Computer Model of Analogy-Making* (MIT Press, 1995).
51. Biehl, M. *The Shallow and the Deep: A Biased Introduction to Neural Networks and Old School Machine Learning* (University of Groningen Press, 2023).
52. Lu, J. et al. Learning under concept drift: a review. *IEEE Trans. Knowl. Data Eng.* **31**, 2346–2363 (2019).
53. Zhang, Y. & Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **34**, 5586–5609 (2021).
54. Verwimp, E. et al. Continual learning: applications and the road forward. *Trans. Mach. Learn. Res.* (2024).
55. Zhuang, F. et al. A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76 (2020).
56. Jackendoff, R. S. *Semantics and Cognition* Vol. 8 (MIT Press, 1985).
57. Bertsimas, D. & Dunn, J. Optimal classification trees. *Mach. Learn.* **106**, 1039–1082 (2017).
58. Rosch, E. Natural categories. *Cognit. Psychol.* **4**, 328–350 (1973).
59. Bien, J. & Tibshirani, R. Prototype selection for interpretable classification. *Ann. Appl. Stat.* **5**, 2403–2424 (2011).
60. Peterson, L. E. K-nearest neighbor. *Scholarpedia* **4**, 1883 (2009).
61. Bengesi, S. et al. Advancements in generative AI: a comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. *IEEE Access* **12**, 69812–69837 (2024).
62. Shalev-Shwartz, S. & Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms* (Cambridge Univ. Press, 2014).
63. Decelle, A. An introduction to machine learning: a perspective from statistical physics. *Physica A* **631**, 128154 (2022).

64. Vapnik, V. N. & Chervonenkis, A. Y. in *Measures of Complexity: Festschrift for Alexey Chervonenkis* 11–30 (Springer, 2015).
65. Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.* **14**, 326–334 (1965).
66. Brown, T. et al. Language models are few-shot learners. In *Proc. 34th International Conference on Neural Information Processing Systems* (eds Larochelle, H. et al.) 1877–1901 (Curran Associates, 2020).
67. Müller, T., Pérez-Torró, G. & Franco-Salvador, M. Few-shot learning with Siamese Networks and label tuning. In *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Muresan, S. et al.) 8532–8545 (Association for Computational Linguistics, 2022).
68. Gold, E. M. Language identification in the limit. *Inf. Control* **10**, 447–474 (1967).
69. Zeugmann, T. in *Algorithmic Learning Theory* (eds Gavaldá, R. et al.) 17–38 (Springer Berlin Heidelberg, 2003).
70. Lake, B. M. & Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature* **623**, 115–121 (2023).
71. Vapnik, V. N. *The Nature of Statistical Learning Theory* (Springer-Verlag, 1995).
72. Bousquet, O. & Elisseeff, A. Algorithmic stability and generalization performance. In *Proc. 14th International Conference on Neural Information Processing Systems* (eds Leen, T. et al.) 178–184 (MIT Press, 2000).
73. Grohs, P. & Kutyniok, G. (eds) *Mathematical Aspects of Deep Learning* (Cambridge Univ. Press, 2022).
74. Ye, H. et al. Towards a theoretical framework of out-of-distribution generalization. In *Proc. 35th International Conference on Neural Information Processing Systems* (eds Beygelzimer, A. et al.) 23519–23531 (Curran Associates, 2021).
75. Papernot, N. et al. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy* 372–387 (IEEE, 2016).
76. Jiao, T., Guo, C., Feng, X., Chen, Y. & Song, J. A comprehensive survey on deep learning multi-modal fusion: methods, technologies and applications. *Comput. Mater. Continua* **80**, 1–35 (2024).
77. Dalal, A. et al. On the value of labeled data and symbolic methods for hidden neuron activation analysis. In *Neural-Symbolic Learning and Reasoning Proceedings, Part II* (eds Besold, T. R. et al.) 109–131 (Springer, 2024).
78. Baker, R. E., Peña, J.-M., Jayamohan, J. & Jérusalem, A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol. Lett.* **14**, 20170660 (2018).
79. Yao, L. et al. A survey on causal inference. *ACM Trans. Knowl. Discov. Data* **15**, 74 (2021).
80. Kitzelmann, E. & Schmid, U. Inductive synthesis of functional programs: an explanation based generalization approach. *J. Mach. Learn. Res.* **7**, 429–454 (2006).
81. Manhaeve, R., Dumančić, S., Kimmig, A., Demeester, T. & De Raedt, L. Neural probabilistic logic programming in DeepProbLog. *Artif. Intell.* **298**, 103504 (2021).
82. Cao, J., Fang, J., Meng, Z. & Liang, S. Knowledge graph embedding: a survey from the perspective of representation spaces. *ACM Comput. Surv.* **56**, 159 (2024).
83. Berthet, M., Surbeck, M. & Townsend, S. W. Extensive compositionality in the vocal system of bonobos. *Science* **388**, 104–108 (2025).
84. Hammer, B. *Learning with Recurrent Neural Networks* (Springer-Verlag, 2000).
85. Wiedemer, T. et al. Provable compositional generalization for object-centric learning. In *Proc. 40th International Conference on Machine Learning* 3038–3062 (PMLR, 2023).
86. Aha, D. *Lazy Learning* (Springer Netherlands, 2013).
87. Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L. & Lewis, M. Generalization through memorization: nearest neighbor language models. In *International Conference on Learning Representations* (ICLR, 2020).
88. Chen, H.-J., Cheng, A.-C., Juan, D.-C., Wei, W. & Sun, M. Mitigating forgetting in online continual learning via instance-aware parameterization. In *Proc. 34th International Conference on Neural Information Processing Systems* (eds Larochelle, H., et al.) 17466–17477 (Curran Associates, 2020).
89. Kulis, B. Metric learning: a survey. *Found. Trends Mach. Learn.* **5**, 287–364 (2013).
90. He, J. Z.-Y., Erickson, Z., Brown, D. S., Raghunathan, A. & Dragan, A. Learning representations that enable generalization in assistive tasks. In *6th Annual Conference on Robot Learning* (PMLR, 2022).
91. Shepard, R. N. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323 (1987).
92. Gao, Y. et al. Retrieval-augmented generation for large language models: a survey. Preprint at https://doi.org/10.48550/arXiv.2312.10997 (2024).
93. Li, C. & Flanigan, J. Task contamination: language models may not be few-shot anymore. In *Proc. 38th AAAI Conference on Artificial Intelligence* 18471–18480 (AAAI Press, 2024).
94. Lewis, M. & Mitchell, M. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. Preprint at https://doi.org/10.48550/arXiv.2402.08955 (2024).
95. Dodge, J. et al. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *2021 Conference on Empirical Methods in Natural Language Processing* (eds Moens, M.-F. et al.) 1286–1305 (Association for Computational Linguistics, 2021).
96. Liu, J. et al. Towards out-of-distribution generalization: a survey. Preprint at https://doi.org/10.48550/arXiv.2108.13624 (2021).
97. Cossu, A. et al. Don't drift away: Advances and applications of streaming and continual learning. In *Proc. 33th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (ESANN, 2025).
98. Hendrycks, D. & Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. 7th International Conference on Learning Representations* (ICLR, 2019).
99. Gonen, H., Iyer, S., Blevins, T., Smith, N. & Zettlemoyer, L. Demystifying prompts in language models via perplexity estimation. In *Findings of the Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H. et al.) 10136–10148 (Association for Computational Linguistics, 2023).
100. Boult, T. E. et al. Learning and the unknown: surveying steps toward open world recognition. In *Proc. 33rd AAAI Conference on Artificial Intelligence* 9801–9807 (AAAI Press, 2019).
101. Hovy, D. & Spruit, S. L. The social impact of natural language processing. In *54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (eds Erk, K. & Smith, N. A.) 591–598 (Association for Computational Linguistics, 2016).
102. Sourati, Z. et al. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Syst.* **266**, 110418 (2023).
103. Mundt, M., Hong, Y., Pliushch, I. & Ramesh, V. A wholistic view of continual learning with deep neural networks: forgotten lessons and the bridge to active and open world learning. *Neural Networks* **160**, 306–336 (2023).
104. Lapuschkin, S. et al. Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1096 (2019).

105. Shah, H., Tamuly, K., Raghunathan, A., Jain, P. & Netrapalli, P. The pitfalls of simplicity bias in neural networks. In *Proc. 34th International Conference on Neural Information Processing Systems* 9573–9585 (Curran Associates, 2020).
106. Yuan, Y., Li, Z. & Zhao, B. A survey of multimodal learning: methods, applications, and future. *ACM Comput. Surv.* **57**, 167 (2025).
107. Davis, E. Benchmarks for automated commonsense reasoning: a survey. *ACM Comput. Surv.* **56**, 1–41 (2023).
108. Sourati, Z., Ilievski, F., Sommerauer, P. & Jiang, Y. ARN: analogical reasoning on narratives. *Trans. Assoc. Comput. Ling.* **12**, 1063–1086 (2024).
109. Nie, W. et al. Bongard-LOGO: a new benchmark for human-level concept learning and reasoning. *Adv. Neural Inf. Process. Syst.* **33**, 16468–16480 (2020).
110. Wang, C. et al. Survey on factuality in large language models: knowledge, retrieval and domain-specificity. Preprint at https://doi.org/10.48550/arXiv.2310.07521 (2023).
111. Guha, N. et al. LegalBench: a collaboratively built benchmark for measuring legal reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **36**, 44123–44279 (2023).
112. Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D. & Gama, J. Methods and tools for causal discovery and causal inference. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **12**, e1449 (2022).
113. Atkinson, K. et al. Towards artificial argumentation. *AI Magazine* **38**, 25–36 (2017).
114. Yang, K. et al. LeanDojo: theorem proving with retrieval-augmented language models. *Adv. Neural Inf. Process. Syst.* **36**, 21573–21612 (2023).
115. Nauta, M. et al. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Comput. Surv.* **55**, 295 (2023).
116. Kambhampati, S., Stechly, K. & Valmeekam, K. (How) do reasoning models reason? *Ann. N.Y. Acad. Sci.* **1547**, 33-40 (2025).
117. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at https://doi.org/10.48550/arXiv.2303.12712 (2023).
118. Cohen, T. & Welling, M. Group equivariant convolutional networks. In *33rd International Conference on Machine Learning* (eds Balcan, M. F. & Weinberger, K. Q.) 2990–2999 (PMLR, 2016).
119. Liao, M. et al. Calibration-based multi-prototype contrastive learning for domain generalization semantic segmentation in traffic scenes. *IEEE Trans. Intell. Transport. Syst.* **25**, 20985–21001 (2024).
120. Tyukin, I. Y., Gorban, A. N., Alkhudaydi, M. H. & Zhou, Q. Demystification of few-shot and one-shot learning. In *2021 International Joint Conference on Neural Networks* 1–7 (IEEE, 2021).
121. Manginas, V. et al. A scalable approach to probabilistic neuro-symbolic verification. Preprint at https://doi.org/10.48550/arXiv.2502.03274 (2025).
122. Singh, G., Tommasini, R., Bhatia, S. & Mutharaju, R. Benchmarking neuro-symbolic description logic reasoners: existing challenges and a way forward. *Neurosymbolic Artif. Intell.* https://doi.org/10.1177/29498732251339943 (2025).
123. de Boer, M., Smit, Q., van Bekkum, M., Meyer-Vitali, A. & Schmid, T. Design patterns for llm-based neuro-symbolic systems. *Neurosymbolic Artif. Intell.* (in the press).
124. Giunchiglia, F., Villafiorita, A. & Walsh, T. Theories of abstraction. *AI Commun.* **10**, 167–176 (1997).
125. Gordon, A. S. & Hobbs, J. R. *A Formal Theory of Commonsense Psychology: How People Think People Think* (Cambridge Univ. Press, 2017).
126. Lenat, D. B. CYC: a large-scale investment in knowledge infrastructure. *Commun. ACM* **38**, 33–38 (1995).
127. Weir, N. et al. From models to microtheories: distilling a model's topical knowledge for grounded question answering. In *International Conference on Representation Learning 2025* (eds Yue, Y. et al.) (ICLR, 2025).
128. Luo, Y. et al. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. Preprint at https://doi.org/10.48550/arXiv.2308.08747 (2025).
129. Shi, Z., Jing, J., Sun, Y., Lim, J.-H. & Zhang, M. Unveiling the tapestry: the interplay of generalization and forgetting in continual learning. *IEEE Trans. Neural Netw. Learn. Syst.* **36**, 15070–15084 (2025).
130. Malialis, K., Li, J., Panayiotou, C. G. & Polycarpou, M. M. Incremental learning with concept drift detection and prototype-based embeddings for graph stream classification. In *International Joint Conference on Neural Networks* 1–7 (IEEE, 2024).
131. Huang, J. et al. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Ku, L.-W. et al.) 1416–1428 (Association for Computational Linguistics, 2024).
132. Fuente, N. D. L. et al. Prototype augmented hypernetworks for continual learning. Preprint at https://doi.org/10.48550/arXiv.2505.07450 (2025).
133. Lin, S., Ju, P., Liang, Y. & Shroff, N. Theory on forgetting and generalization of continual learning. In *Proc. 40th International Conference on Machine Learning* 21078–21100 (JMLR, 2023).
134. Chollet, F. On the measure of intelligence. Preprint at https://doi.org/10.48550/arXiv.1911.01547 (2019).
135. Jiang, Y., Ilievski, F., Ma, K. & Sourati, Z. BRAINTEASER: lateral thinking puzzles for large language models. In *Conference on Empirical Methods in Natural Language Processing* (eds Bouamor, H. et al.) 14317–14332 (Association for Computational Linguistics, 2023).
136. Draws, T., Rieger, A., Inel, O., Gadiraju, U. & Tintarev, N. A checklist to combat cognitive biases in crowdsourcing. In *Proc. 9th AAAI Conference on Human Computation and Crowdsourcing* 48–59 (AAAI Press, 2021).
137. Duan, J., Yu, S., Tan, H. L., Zhu, H. & Tan, C. A survey of embodied AI: from simulators to research tasks. *IEEE Trans. Emerging Top. Comput. Intell.* **6**, 230–244 (2022).
138. Mitchell, M. et al. Model cards for model reporting. In *Proc. Conference on Fairness, Accountability, and Transparency* 220–229 (Association for Computing Machinery, 2019).
139. Pushkarna, M., Zaldivar, A. & Kjartansson, O. Data cards: purposeful and transparent dataset documentation for responsible AI. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 1776–1826 (Association for Computing Machinery, 2022).
140. Kapoor, S. et al. Reforms: consensus-based recommendations for machine-learning-based science. *Sci. Adv.* **10**, eadk3452 (2024).
141. Akata, Z. et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *IEEE Comput.* **53**, 18–28 (2020).
142. Widmer, C. L. et al. Towards human-compatible XAI: explaining data differentials with concept induction over background knowledge. *J. Web Semant.* **79**, 100807 (2023).
143. Finzel, B., Hilme, P., Rabold, J. & Schmid, U. When a relation tells more than a concept: exploring and evaluating classifier decisions with CoReX. Preprint at https://doi.org/10.48550/arXiv.2405.01661 (2024).
144. Braun, M., Greve, M., Gnewuch, U. The new dream team? A review of human AI collaboration research from a human teamwork perspective. In *Proc. 44th International Conference on Information Systems* 1192 (ICIS, 2023).

145. Medin, D. L., Wattenmaker, W. D. & Hampson, S. E. Family resemblance, conceptual cohesiveness, and category construction. *Cognit. Psychol.* **19**, 242–279 (1987).

Acknowledgements

The manuscript resulted from the May 2024 Dagstuhl seminar: Generalization by People and Machines (24192). K. Forbus, P. Vossen, D. Shahaf, W. Abd-Almageed, and M. Waldmann provided valuable insights during the seminar. F.I. is funded by the NWO AiNed project 'Human-Centric AI Agents with Common Sense'. B.H., B.P., A.-C.N.N. gratefully acknowledge funding by the Ministry of Culture and Science of North Rhine-Westphalia (MKW NRW) through the project SAIL (grant no. NW21-059A-D).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Filip Ilievski.

Peer review information *Nature Machine Intelligence* thanks Mengmi Zhang, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

[1]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [2]University of Bielefeld, Bielefeld, Germany. [3]NEC Laboratories Europe, Heidelberg, Germany. [4]University of Bamberg, Bamberg, Germany. [5]University of Groningen, Groningen, The Netherlands. [6]Università di Bologna, Bologna, Italy. [7]Meta Reality Labs, Redmond, WA, USA. [8]CAIR, Ss. Cyril and Methodius University, Skopje, North Macedonia. [9]Kansas State University, Manhattan, KS, USA. [10]KU Leuven, Leuven, Belgium. [11]University of Edinburgh, Edinburgh, UK. [12]Miniml.AI, Edinburgh, UK. [13]University of Bremen, Bremen, Germany. [14]Paderborn University, Paderborn, Germany. [15]Carnegie Bosch Institute, Pittsburgh, PA, USA. [16]Università degli Studi di Milano Bicocca, Milan, Italy. [17]TU Wien, Vienna, Austria. [18]Fondazione Bruno Kessler, Trento, Italy. [19]University College London, London, UK. [20]University of British Columbia, Vancouver, British Columbia, Canada. [21]Vector Institute, Toronto, Ontario, Canada. [22]Duolingo, Pittsburgh, PA, USA. [23]University of St. Gallen, Institute of Behavioral Science and Technology, St. Gallen, Switzerland. [24]University of Applied Sciences Mittweida, Mittweida, Germany. [25]Technical University Freiberg, Freiberg, Germany. ✉e-mail: f.ilievski@vu.nl