Aligning Generalisation Between Humans and Machines

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]

[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, Canada.
[21]Vector Institute, Toronto, 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.

*Corresponding author(s). E-mail(s): f.ilievski@vu.nl;

Abstract

Recent advances in 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 *generalise*. In cognitive science, human generalisation commonly involves abstraction and concept learning. In contrast, AI generalisation encompasses out-of-domain generalisation in machine learning, rule-based reasoning in symbolic AI, and abstraction in neurosymbolic AI. In this perspective paper, we combine insights from AI and cognitive science to identify key commonalities and differences across three dimensions: notions of, methods for, and evaluation of generalisation. We map the different conceptualisations of generalisation 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.

Keywords: generalisation, human-AI teaming, alignment

This manuscript is published as a Perspective paper in Nature Machine Intelligence with DOI: 10.1038/s42256-025-01109-4 [65].

Contents

1 Introduction

Recent advances in artificial intelligence (AI) enable meaningful support for humans in complex tasks, such as scientific discovery and decision-making [71]. Conversely, AI can also potentially disrupt democracies and target individuals [39]. The responsible use of AI increasingly motivates *AI alignment*, which aims to *"make AI systems act according to our preferences"* [21]. AI alignment is essential for effective human-AI teaming in complex scenarios where neither humans nor AI perform well on their own [99]. For example, AI can help invent novel biomedical application hypotheses following the objectives and guidance of a scientist [51]. Alternatively, humans can iteratively edit samples provided by an AI model to improve its accuracy and trustworthiness, for example, in classifying skin cancer [35]. Besides human-AI teaming, AI alignment is necessary for its safe use [8] and demonstrable adherence to accountability, privacy, and transparency requirements in legal frameworks such as the EU's AI Act [6].

A crucial, yet often overlooked aspect of this alignment in interactive scenarios is the complementary ways in which humans and machines *generalise* (Figure 1). Generalisation is typically defined as *"the process of transferring knowledge or skills from specific instances or exemplars to new contexts"* [123]. Following cognitive science, human generalisation commonly involves concept learning and the abstraction of general characteristics to a collection of entities [56]. Humans excel at generalising from a few examples, compositionality, and robust generalisation to noise, shifts, and Out-Of-Distribution (OOD) data [60]. Humans can learn from little data and seemingly generalise 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 characterise physical principles in nature and human behaviour in interactions.

In sharp contrast, data-driven (*statistical*) AI systems struggle to generalise beyond their training distribution and to abstract effectively. Although some neural architectures might display better alignment with physical laws [88], the generalisability of statistical machines, averaged over all distributions and in the absence of prior knowledge, is constant (no-free-lunch theorem [138]).

The goal of *human-machine teaming* [128] is that each side addresses the limitations of the other while aligning on similar goals. For example, some generalisation 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 overgeneralisation errors ("hallucinations" [68]), 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") (Figure 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 [77], whereas instance-based models are robust to distributional shifts when adequate representation is available [91]. Combining the strengths of various machine methods has inspired emerging research directions under the umbrella of neurosymbolic AI [59].

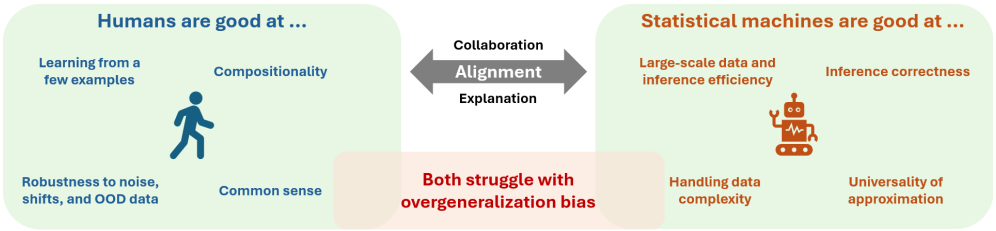


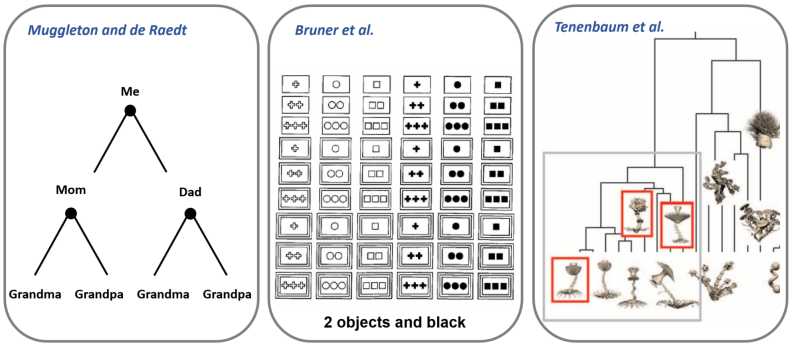
Fig. 1: Comparison of the strengths of humans and statistical machines, illustrating their complementary *generalisation* in human-AI teaming scenarios. Humans excel at compositionality, common sense, abstraction from a few examples, and robustness. Statistical ML excels at large-scale data and inference efficiency, inference correctness, handling data complexity, and the universality of approximation. Overgeneralisation 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.

This perspective paper draws on insights about the generalisation 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 generalisation. We focus on the following questions: What are the known notions of generalisation? What are the strengths and weaknesses of the generalisation of AI methods? How is generalisation evaluated today? What is the impact of current trends in AI, such as foundation models, on generalisation 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 generalisation.

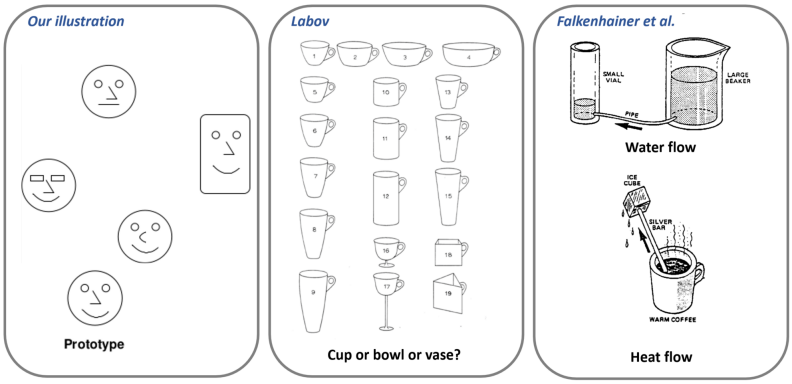
2 Parallels in Generalisation by Humans and Machines

Approaches to generalisation 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, summarised in Figure 2.

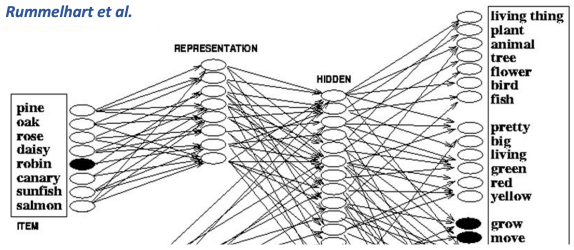
In the early days of cognitive psychology, Bruner et al. [17] empirically investigated human concept learning (Figure 2a), which inspired the first decision tree learning algorithms [63]. Observations on human learning of relational concepts inspired early machine learning (ML) approaches to learning from structural representations [102, 137] and recursive concepts [116]. This class of approaches, often referred to as inductive programming, allows learning from few examples and taking into account background knowledge for model induction [54]. Rule learning approaches have been extended to statistical relational learning [32] to overcome the brittleness of symbolic



(a) Learning the relational rule 'grandparent' using a background theory 'parent' in Muggleton and De Raedt [102], conjunctive rules in Bruner et al. [17], and names of alien objects modelled as Bayesian inference over a tree-structured domain representation in Tenenbaum et al. [126].



(b) Example-based prototypical representations (cf. Rosch and Mervis [113], and Medin et al. [98]), context-effects (Labov [79]), and analogy (Falkenhainer et al. [38]).



(c) Statistical generalisation: neural network model of semantic memory (Rumelhart and Todd [114]; crop of the illustration).

Fig. 2: Illustrative examples of human generalisation and its inspiration of rule-based (top), example-based (middle), and statistical ML approaches (bottom).

learning. Bayesian approaches to rule learning have been introduced as a plausible framework to model human learning in complex domains such as language acquisition [126].

Rule learning as an explicit approach (system 2 [72]) is apparent for domains of high-level cognition where relevant features can be verbalised [80]. 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 maximised and between categories minimised [113], see Figure 2b. This approach is reflected in similarity-based methods to ML, especially k-nearest neighbours [27]. Exemplar theories [108] have been proposed to address the flexibility of human categorisation. For instance, context-dependence of the classification of visual objects has been shown by Labov [79] where a cup might be classified as a bowl or a vase when different contexts, e.g., 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 [38, 42, 46]. In contrast to other methods, the analogy is not based on feature but on structural similarity.

Neural network approaches were proposed by cognitive scientists [115] as a method of generalisation learning that overcomes the brittleness of symbolic approaches; see Figure 2c. Despite strong arguments from researchers in symbolic AI [41], 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 [41] about the relationship between statistical ML and human cognition remain. First, data are separated from a semantic model. While humans who have learnt a concept robustly recognise OOD inputs, ML has only addressed this problem more recently [139]. Second, 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 [1]. Third, human explanations are typically based on the causal history of an event and a causal explanation for the generalisation itself [100]. 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 [132]. In ML, discovering high-level causal variables from low-level observations remains a significant challenge [117].

The combination of implicit, neural learning, and explicit, symbolic approaches is addressed in neurosymbolic AI research [59]. 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 [97]. Combining neural and symbolic approaches to generalisation and explainability within human-centric AI reflects many aspects of human learning flexibility [64]. For effective joint decision-making and problem-solving, the human-AI interface must align with human information processing [67]. Alignment must be established for different aspects, including knowledge state, current information needs, or values [19]. In cognitive modelling, researchers aim to align algorithmic approaches to learning and human learning [82]. 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 [128].

3 Notions of Generalisation

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

3.1 Generalisation as a process

Generalisation 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, generalisation by similarity, or the transformation of schemas from lower to higher stages of cognitive development [25]. More broadly, French [43] distinguishes (1) generalisation of concrete instances into an abstract schema, which we call *abstraction*, (2) generalisation through the application or extension of the schema to various situations, which we call *extension*, and (3) generalisation 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 [11]; extension (2) relates to online, multi-task, few-shot, or continual learning schemes in statistical and instance-based AI [92, 131, 145]; and analogy (3) relates to transfer learning [146] in statistical AI and reasoning by analogy in analytical AI.

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

3.2 Generalisation as a product

Generalisation may also refer to the *products* of a generalisation process, such as categories, concepts, rules, and models, in their various representations. Generalisations of *categories and concepts* may be represented using a symbolic definition, as a list of attributes and their bounds (cognitive science: [66]; or decision trees in AI [10]), as a prototype (cognitive science: [112]; AI: [12]), or as a set of exemplars of the category (cognitive science: [108]; or k-nearest neighbor in AI: [110]). 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 [7]. Beyond categories or concepts, products may also be *rules or relations*, represented, for instance, via functions or graphs in parametric or non-parametric form.

3.3 Generalisation as an operator

Finally, we refer to generalisation as an *operator*, namely the application of a product to new data. The successful application to new data is at the core of generalisation in statistical AI [119]. Under the assumption that training data and new data are independently sampled from the same distribution (IID), one can mathematically prove generalisation via one of three theories: (1) The *Probably Approximately Correct (PAC)* framework analyses whether a model (i. e., a product) derived via a machine learning algorithm (i. e., 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 [119]. (2) *Statistical physics of learning* aims to understand the typical properties of

learning algorithms (i. e., processes) with many adaptive parameters [33]. (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 [130].

Generalisation theory in ML is limited in several ways. It typically predicts that generalisation only occurs if the available data is large enough to not just be memorised [28]. In contrast, humans can generalise from a few samples for a specific task, as generalisation 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 [16, 103].

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 generalisation across compositions —for instance, in language —which has been addressed in analytical AI [48], but is limited by the undecidability or high complexity of inference [144]. Despite such fundamental restrictions, humans operate with compositions similar to those found in language and vision.

3.4 Alignment of human and machine notions of generalisation

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

4 Machine Methods for Generalisation

Humans excel in systematic generalisation across representations, contexts, and tasks based on a few observations. Although specific machine methods have shown remarkable results in solving compositional tasks [81], the underlying mechanisms of human generalisation are not sufficiently understood to mimic them in artificial systems [126].

AI methods are usually structured according to algorithmic aspects rather than their generalisability (see Table 1). Although these impact generalisation behaviour —for instance, symbolic methods often implement compositionality —there is no simple mapping to the form of generalisation that an AI model can achieve. Therefore, we focus on another categorisation, the interplay of observational data (i. e., single instances) and models (i. e., principles that apply to a whole population), as this correlates to the generalisability of the model and its suitable evaluation (Section 5). Three categories can be distinguished: (1) The transfer of individual observations to a population is the basis for *statistical generalisation 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 generalisation.

Table 1: Common categories to structure AI methods algorithmically centred. These categories are not uniquely related to their type of generalisation.

Category	Attributes
Training signal	supervised, unsupervised, reinforcement, semi-supervised, self-supervised
Data type	tabular data, data structures (e. g., 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 choices have different characteristics in terms of their generalisability and alignment with human generalisation.

4.1 Statistical generalisation methods in AI

Many modern methods, including deep learning, aim at statistical generalisation: observational data (i. e., training data) serve as input to an inference mechanism that extracts a model for the entire population (i. e., the underlying distribution). Generalisation refers to the ability of the inferred patterns to be successfully applied to new data (Section 3.3). Typically, algorithmic methods are expressed as optimisation 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 minimisation* [129]. Although evaluation in an independent test set constitutes an unbiased estimator of generalisation ability (Section 5), the empirical loss in the training set systematically underestimates the model loss —*generalisation needs to be accounted for explicitly*. Popular strategies regularize models towards a better generalisation behaviour, such as maximum margin or stability [14]. Even heavily over-parametrised deep models display surprising generalisation capabilities due to intrinsic regularisation [52].

Statistical methods excel in inference correctness and efficiency. Yet, they typically require large training data sets. 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 [103]. Still, empirical risk minimisation has a fundamental limitation compared to human generalisation: *generalisation can only be expected in areas covered by observations*, but not for out-of-sample events, novel contexts, or distributional changes [142]. Indeed, machine behaviour for OOD settings can significantly deviate from human expectation, with adversarial attacks as prominent examples of this phenomenon [109].

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 generalisability arises [16]. However, *the emerging generalisation abilities are only partially understood and do not necessarily align with human expectation*, necessitating a downstream evaluation (Section 5) 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 [21].

Statistical generalisation 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 [70]. Yet, the product is typically a black box, which does not reveal insight into its generalisation behaviour; hence, partially unintended generalisation behaviour can easily occur. Recently developed post hoc *explanation methods* allow for a closer inspection of the underlying rationale and its impact on the generalisation behaviour of the model [29].

4.2 Knowledge-informed generalisation methods in AI

Knowledge-informed generalisation methods aim to find empirical evidence of a theory, resulting in a meaningful representation confirmed by the data. Popular methods include mechanistic models [5], causal models [141], or functional programs [76]. As their semantics is directly accessible, *humans can inspect how these models generalise to previously not encountered scenarios*. Generalisation is often well aligned with human expectations. Yet, model parameters require semantic grounding, which is challenging to realise with low-level sensor data. Neurosymbolic integration can partially overcome such limitations [96].

Learning the optimal model structure is demanding, and fundamental limitations such as non-identifiability of structural components might exist [77]. 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 [20]. Knowledge-based approaches enable the explicit inspection and manipulation of information, allowing generalisation based on a few examples, as accompanying rules can ensure valid generalisation. Yet, they are restricted to domains where a theory can be formalised 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 generalise 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 generalisation [41]. Compositionality seems to be a universal principle in nature, since it has been observed in many species [9]. In recent years, significant progress has been made in improving compositional generalisation in deep learning, typically by adding components that mirror the compositional structure of the domain, such as structure-processing neural networks [55] or metalearning for compositional generalisation [81]. Although these efforts provide a pathway for neural networks to generalise systematically, most of the results are only empirical [136]. There remains a *significant gap between the systematic generalisation capabilities of knowledge-informed models*

and the representation learning techniques of deep models, with neurosymbolic AI promising a viable bridge [59].

4.3 Instance-based translation in AI

Instance-based non-parametric techniques, such as nearest neighbour methods or case-based reasoning [2], 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 [75]. Since they adjust their complexity as needed, they offer universal approximation capability. Suitable data structures enable efficient training and inference in large data sets. 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 [112].

Instance-based models have shown great promise for incremental learning of distributional shifts [91]: 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 [22]. 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 generalisation* [78], 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 generalisations across tasks or domains [57].

Context has a unique role as generalisation 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 [120]. ML techniques such as transfer learning, prompting, or retrieval augmented generalisation mimic parts of this process [45]. In this realm, LLMs have demonstrated remarkable capabilities for few-shot or in-context learning [16], 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.

4.4 Aligning machine generalisation methods and human expectations

While *statistical methods* offer powerful universal approximation, their generalisation behaviour does not match human generalisation well, lacking generalisation to OOD samples and compositionality. Another challenge is their black-box nature, where post hoc explanations provide solutions for specific cases. In 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.

Table 2: Characterisation of AI generalisation methods.

Pros	Cons
Statistical: generalisation from observations to a population	
universal approximation, surprising generalisation of deep models	black boxes, generalisation failures outside of training distribution
Knowledge-informed: confirm/adapt hypothesis based on observations	
meaningful models, identifiable parameters, generalisation 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

Instance-based methods try a different approach, focusing on generalisation from and to individual instances. This principle is well-aligned with human generalisation 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 generalise from a few examples, much like humans. Ultimately, making claims about the generalisation properties of various machine approaches requires a meaningful evaluation, which we discuss next.

5 Evaluation of Generalisation

The theoretical generalisation strengths and weaknesses of the machine method families are summarised 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*, *memorise training samples* reliably, and can *learn from a few samples*.

Deriving provable robustness and generalisation guarantees is necessary to define the theoretical limits of models. Meanwhile, empirically evaluating the machine's generalisation is also desirable. From a statistical learning perspective, evaluating the generalisation of supervised approaches estimates their applicability to new data (Section 3.3). This formalisation 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 generalisability assumptions: IID and task-representative data. For example, Li and Flanigan [86] 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 generalisation 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 generalisation [64, 85]. Namely, while the emergence of foundation models have enabled evaluation of *learning from a few examples* via zero- and few-shot tests [103], the risk of test set memorisation means that test data may appear partially in their training set, invalidating the findings (data contamination [34]).

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

5.1 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 [90]. 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) machine learning models allow for gauging relations through intermittent embeddings. Learning evaluation in the context of *drift* can be done using tailored benchmarks [26]. 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 [109]. In contrast, counterfactual techniques create data samples that alter target prediction with minimal input changes [85].

5.2 Determining under- and overgeneralisation

AI models are created to provide value to humans and are thus assessed against human generalisation, often using the human-centric concepts of under- and overgeneralisation. *Undergeneralisation* occurs when a change in the input, perceptible or imperceptible, causes a considerable modification within a model. Examples of undergeneralisation include model performance degradation under various natural changes, such as environmental perturbations in computer vision [58]. For foundation models, prompt engineering substantially affects performance [49]. In contrast, models *overgeneralise*, which means that they over-confidently make false predictions for (known or novel) concepts because critical differences are ignored in prediction [13]. A well-known overgeneralisation phenomenon is hallucination, which refers to models that deviate from the source of the information [68]. Other overgeneralisations 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 [61], and logical fallacies [124].

Characterising the model's under- and overgeneralisation requires an appropriate metric, defining its use, and establishing a mechanism to interpret the metric's

score in terms of generalisability beyond the particular test examples. This procedure is susceptible to three caveats. First, the choice between discriminative and generative models determines which representational basis is used to infer similarity [104]. For instance, a discriminative model will only learn representations useful to optimise classification accuracy, whereas a generative approach would additionally learn representations necessary to describe the data distribution. Second, deep models are prone to learning decision shortcuts and ignoring meaningful features [83], sometimes also called simplicity bias [118]. Third, 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 [143], and to exhibit *compositionality* in tasks such as commonsense reasoning [30], analogies [125], and concept induction [106].

5.3 Distinguishing memorisation and generalisation

In AI, *memorisation* refers to learning details from the training data, including facts and noise. Memorisation may be beneficial in some cases (e. g., Paris is the capital of France), but detrimental in others (e. g., Biden is the president of the United States). This observation raises a question: *When should models generalise, when should they memorise, and how can this distinction be evaluated?* Whether the models should generalise or memorise is set a priori. When learning from experience, generalisation is crucial, for instance, in recognising a new manifestation of a vase [79]. Consequently, generalisation setups include cross-domain validation and robustness testing. In contrast, factual knowledge is often memorised: Paris is the capital of France, and mosquitoes fly. Tasks such as answering factual questions [133] and reasoning about legal precedents [53] require *memorisation*. While evaluating memorisation and generalisation separately is informative, many tasks, including causal reasoning [107], argumentation [4], and theorem proving [140], require holistic integration of generalisation and memorisation, centred around *explicit knowledge manipulation*. The explicit knowledge use allows testing a model's *explainable predictions* through human studies and faithfulness evaluation [105].

5.4 Alignment of machine evaluation of generalisation to humans

Effective alignment of AI with humans requires a principled evaluation of its strengths and weaknesses in generalisation. Evaluating AI generalisability 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 generalisability measures its performance concerning distribution shifts, determining its undergeneralisation (adaptability to task variations such as camera perturbations)

and overgeneralisation (e. g., hallucinations), and its ability to memorise and generalise 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 summarise typical approaches for evaluating key generalisation properties in Table 3. In the next section, we discuss open challenges for measuring generalisation in the context of human-AI alignment.

Table 3: Desired properties of generalisation that emerge Sections 3 and 4. The properties are listed in order of their appearance in Section 5, split by a subsection using a horizontal line. Each of the properties is supported by statistical (S), analytical (A), and instance-based (I) AI methods, as discussed in Section 4. The evaluation methods for each property are discussed and substantiated by the references in Section 5. Achievement is indicated by "+" and non-achievement by "−"; we use a strict evaluation to avoid partial scoring.

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
memorisation of training samples	−	−	+	factuality datasets, precedents
explicit knowledge manipulation	−	+	−	causality, argumentation, theorem proving
explainable predictions	−	+	−	human studies, faithfulness

6 Emerging Directions

The previous sections addressed the challenges of aligning human and machine intelligence, emphasising AI's potential to enhance human generalisation. Table 3 summarises the properties, methods and evaluation practices, and highlights the complementarity of the three methods in achieving important properties for aligning human and AI generalisation. Statistical approaches enable universality of approximation and inference correctness, instance-based methods enable robustness and memorisation, 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 generalisation theories, hybrid methods, evaluation practices, and alignment in future human-AI teams.

Generalisation theory in the era of foundation models

Recent zero-shot and in-context learning approaches in LLMs and LRMs (large reasoning models, [73]) implicitly generalise to tasks unrelated to their training without explicit similarity [18]. In other words, model builders assume that LLMs have implicitly generalised (process, Section 3.1) to generalisations (product, Section 3.2) that allow generalisation (operator, Section 3.3) to entirely new tasks and domains. However, this assumption remains unsubstantiated, leading to an overestimation of the generalisability of foundation models [73]. This highlights the need for further research. First, new generalisations *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 [24], as used in complex architectures such as AlphaFold; or through cognitively inspired representations, such as prototypes, which have proven efficient for domain generalisation [87]. Second, 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 generalisability of few-shot learners and over-parametrised deep networks [127].

Generalisable neurosymbolic methods

Neurosymbolic AI combines robust, data-driven latent models with the precision of explicit compositional models [59]. However, several challenges remain: Defining *provable generalisation 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 generalisations [95]. 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 generalisation. 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 generalisation [122].

While a theory about the *compositionality* of neurosymbolic *systems* is emerging [31], a theory on how to compose the generalisations themselves is lacking. For symbolic representations, a theory on abstractions [47] exists, but the question of composing latent representations, such as embeddings, remains open.

Finally, handling the *context dependency* of generalisations remains a challenge. How do we measure the distance between contexts and apply generalisation across contexts? How do we know when a context is too novel and we overgeneralise? A possible direction is formal modelling of contextual dimensions such as time and space, following prior research on axiomatising commonsense knowledge [50]. CYC's notion of hierarchical microtheories, each containing a collection of axioms [84], can be revisited from a neurosymbolic perspective, for instance, by expressing axioms of microtheories in flexible natural language representations [134].

Generalisation 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 [93].

The relation between continual learning and generalisation is complex: continual learning methods are designed to combat catastrophic forgetting when generalising to novel tasks, while within-task generalisation facilitates faster learning and improved performance in subsequent continual learning tasks [121]. 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, formalising 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 generalisation capabilities of the system. Recent work on graph streams [94] employs 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 memorisation of important information, albeit at increased computational costs [62]. More efficient alternatives aim for architectural solutions such as the incorporation of instance-based representations into statistical models [44]. However, theoretical insight on the effect of overparametrisation or task similarity on the generalisability and forgetting of a model currently exists for very simple models only [89].

Evaluation of generalisation 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 [23], analogy [125], and lateral thinking [69] are gaining popularity. Crowdsourcing can be used to create and scale benchmarks, but it can also introduce cognitive and cultural biases by annotators [36], which remains poorly understood. On the other hand, evaluation servers and public leaderboards with private test datasets prevent overfitting but lack standardisation and are costly to maintain. A final direction is simulation environments and synthetic data generators [37], though they often suffer from a sim-to-real gap. To address reproducibility, researchers proposed the model [101] and data cards [111] to report the details of the experiment, and reproducibility checklists based on a broad consensus [74], albeit with limited coverage of generalisability.

Aligning generalisation in future human-AI teams

The goal of effective human-AI teaming and the appearance of legal frameworks such as the EU AI Act [6, 8] require transparent *collaboration* workflows, with *explanations* bridging the gaps between human and AI reasoning [3, 64]. Section 2 discussed that this alignment must occur at the *output level*. However, when misalignments occur, (e. g., 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 [82]. 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 [135] and those considering relationships [40] 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 (e. g., in manual data labelling), AI that helps humans (e. g., conversational question answering), and balanced collaborations where both contribute equally (e. g., medical decision-making) [15], there is little research on evaluating the generalisation capabilities of such teams.

Acknowledgements. The manuscript resulted from the May 2024 Dagstuhl seminar: Generalization by People and Machines (24192). Ken Forbus, Piek Vossen, Dafna Shahaf, Wael Abd-Almageed, and Michael Waldmann provided valuable insights during the seminar. FI is funded by the NWO AiNed project "Human-Centric AI Agents with Common Sense". BH, BP, A-CNN 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.

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From Attribution Maps to Human-Understandable Explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, 5:1006–1019, 2023.

[2] D.W. Aha. *Lazy Learning*. Springer Netherlands, 2013.

[3] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen V. Hindriks, Holger H. Hoos, and et al. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *IEEE Computer*, 53(8):18–28, 2020.

[4] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. Towards Artificial Argumentation. *AI Magazine*, 38(3):25–36, 2017.

[5] Ruth E Baker, Jose-Maria Peña, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*, 14(5), 2018.

[6] Alejandro Bellogín, Oliver Grau, Stefan Larsson, Gerhard Schimpf, Biswa Sengupta, and Gürkan Solmaz. The EU AI Act and the Wager on Trustworthy AI. *Communications of the ACM*, 67(12):58–65, 2024.

[7] Staphord Bengesi, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access*, 2024.

[8] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme AI risks amid rapid progress. *Science*, 384 (6698):842–845, 2024.

[9] M. Berthet, M. Surbeck, and S. W. Townsend. Extensive compositionality in the vocal system of bonobos. *Science*, 388(6742):104–108, 2025.

[10] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106:1039–1082, 2017.

[11] Michael Biehl. *The Shallow and the Deep: A biased introduction to neural networks and old school machine learning.* University of Groningen Press, 2023.

[12] Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4):2403–2424, 2011.

[13] Terrance E. Boult, Steve Cruz, Akshay R. Dhamija, Manuel Gunther, James Henrydoss, and Walter J. Scheirer. Learning and the Unknown : Surveying Steps Toward Open World Recognition. *AAAI Conference on Artificial Intelligence*, 2019.

[14] Olivier Bousquet and André Elisseeff. Algorithmic Stability and Generalization Performance. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[15] Marvin Braun, Maike Greve, and Ulrich Gnewuch. The New Dream Team?: A Review of Human-AI Collaboration Research From a Human Teamwork Perspective. In *International Conference on Information Systems*, 2023. Association for Information Systems.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. In *Advances in*

Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[17] Jerome Bruner, Jacqueline J. Goodnow, and George A. Austin. *A Study of Thinking*. Wiley, 1956.

[18] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 2023.

[19] Patrick Butlin. AI Alignment and Human Reward. In *2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 437–445, 2021.

[20] Jiahang Cao, Jinyuan Fang, Zaiqiao Meng, and Shangsong Liang. Knowledge Graph Embedding: A Survey from the Perspective of Representation Spaces. *ACM Computing Surveys*, 56(6), 2024.

[21] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. AI alignment with changing and influenceable reward functions. In *41st International Conference on Machine Learning*. JMLR, 2024.

[22] Hung-Jen Chen, An-Chieh Cheng, Da-Cheng Juan, Wei Wei, and Min Sun. Mitigating Forgetting in Online Continual Learning via Instance-Aware Parameterization. In *Advances in Neural Information Processing Systems*, volume 33, pages 17466–17477. Curran Associates, Inc., 2020.

[23] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[24] Taco Cohen and Max Welling. Group Equivariant Convolutional Networks. In *33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, 2016. PMLR.

[25] Eliana Colunga and Linda B Smith. The emergence of abstract ideas: Evidence from networks and babies. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1435):1205–1214, 2003.

[26] Andrea Cossu, Davide Bacciu, Alessio Bernardo, Emanuele Della Valle, Alexander Gepperth, Federico Giannini, Barbara Hammer, and Giacomo Ziffer. Don't Drift Away: Advances and Applications of Streaming and Continual Learning. *European Symposium on Artificial Neural Networks*, 2025.

[27] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[28] Thomas M Cover. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on*

Electronic Computers, (3):326–334, 1965.

[29] Abhilekha Dalal, Rushrukh Rayan, Adrita Barua, Eugene Y. Vasserman, Md. Kamruzzaman Sarker, and Pascal Hitzler. On the Value of Labeled Data and Symbolic Methods for Hidden Neuron Activation Analysis. In *Neural-Symbolic Learning and Reasoning Proceedings, Part II*, volume 14980 of *Lecture Notes in Computer Science*, pages 109–131, 2024. Springer.

[30] Ernest Davis. Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Computing Surveys*, 56(4):1–41, 2023.

[31] Maaike de Boer, Quirine Smit, Michael van Bekkum, André Meyer-Vitali, and Thomas Schmid. Design Patterns for LLM-based Neuro-Symbolic Systems. *Neurosymbolic Artificial Intelligence (to appear)*, 2025.

[32] Luc De Raedt and Kristian Kersting. Probabilistic inductive logic programming. In *Probabilistic inductive logic programming: theory and applications*, pages 1–27. Springer, 2008.

[33] Aurélien Decelle. An introduction to machine learning: a perspective from statistical physics. *Physica A: Statistical Mechanics and its Applications*, page 128154, 2022.

[34] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, 2021. Association for Computational Linguistics.

[35] Jon Donnelly, Zhicheng Guo, Alina Jade Barnett, Hayden McTavish, Chaofan Chen, and Cynthia Rudin. Rashomon Sets for Prototypical-Part Networks: Editing Interpretable Models in Real-Time. In *Computer Vision and Pattern Recognition Conference (CVPR)*, pages 4528–4538, 2025.

[36] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A Checklist to Combat Cognitive Biases in Crowdsourcing. In *AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 48–59, 2021.

[37] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A Survey of Embodied AI: From Simulators to Research Tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[38] Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1):1–63, 1989.

[39] Emilio Ferrara. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, pages 1–21, 2024.

[40] Bettina Finzel, Patrick Hilme, Johannes Rabold, and Ute Schmid. When a Relation Tells More Than a Concept: Exploring and Evaluating Classifier Decisions with CoReX. *arXiv preprint arXiv:2405.01661*, 2024.

[41] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[42] Kenneth D Forbus, Ronald W Ferguson, Andrew Lovett, and Dedre Gentner. Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, 41 (5):1152–1201, 2017.

[43] Robert M. French. *The Subtlety of Sameness: A Theory and Computer Model of Analogy-Making*. MIT Press, 1995.

[44] Neil De La Fuente, Maria Pilligua, Daniel Vidal, Albin Soutiff, Cecilia Curreli, Daniel Cremers, and Andrey Barsky. Prototype Augmented Hypernetworks for Continual Learning. *arXiv preprint, arXiv:2505.07450*, 2025.

[45] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint, arXiv:2312.10997*, 2024.

[46] Dedre Gentner. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7(2):155–170, 1983.

[47] Fausto Giunchiglia, Adolfo Villafiorita, and Toby Walsh. Theories of abstraction. *AI Communications*, 10(3-4):167–176, 1997.

[48] E Mark Gold. Language identification in the limit. *Information and Control*, 10 (5):447–474, 1967.

[49] Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. Demystifying Prompts in Language Models via Perplexity Estimation. In *Findings of the Conference on Empirical Methods in Natural Language Processing*, pages 10136–10148, 2023. Association for Computational Linguistics.

[50] Andrew S Gordon and Jerry R Hobbs. *A Formal Theory of Commonsense Psychology: How People Think People Think*. Cambridge University Press, 2017.

[51] Juraj Gottweis, Wei-Hung Weng, Alexander N. Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an AI co-scientist. *CoRR*, abs/2502.18864, 2025.

[52] P. Grohs and G. Kutyniok, editors. *Mathematical Aspects of Deep Learning*. Cambridge University Press, 2022.

[53] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023.

[54] Sumit Gulwani, José Hernández-Orallo, Emanuel Kitzelmann, Stephen H Muggleton, Ute Schmid, and Benjamin Zorn. Inductive programming meets the real world. *Communications of the ACM*, 58(11):90–99, 2015.

[55] Barbara Hammer. *Learning with Recurrent Neural Networks*. Springer-Verlag, 2000.

[56] S. Harnad. To cognize is to categorize: Cognition is categorization. In *Handbook of Categorization in Cognitive Science (2nd edition)*, pages 21–54. Elsevier Academic Press, 2017.

[57] Jerry Zhi-Yang He, Zackory Erickson, Daniel S. Brown, Aditi Raghunathan, and Anca Dragan. Learning Representations that Enable Generalization in Assistive Tasks. In *6th Annual Conference on Robot Learning*, 2022.

[58] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *International Conference on Learning Representations (ICLR)*, 2019.

[59] Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors. *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2023.

[60] Andreas Holzinger, Anna Saranti, Alessa Angerschmid, Bettina Finzel, Ute Schmid, and Heimo Mueller. Toward human-level concept learning: Pattern benchmarking for AI algorithms. *Patterns*, 4:100788, 2023.

[61] Dirk Hovy and Shannon L. Spruit. The Social Impact of Natural Language Processing. In *54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016. Association for Computational Linguistics.

[62] Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 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)*, pages 1416–1428, 2024. Association for Computational Linguistics.

[63] Earl B Hunt, Janet Marin, and Philip J Stone. *Experiments in Induction*. Academic Press, 1966.

[64] Filip Ilievski. *Human-Centric AI with Common Sense*. Springer, 2025.

[65] Filip Ilievski, Barbara Hammer, Frank van Harmelen, Benjamin Paassen, Sascha Saralajew, Ute Schmid, Michael Biehl, Marianna Bolognesi, Xin Luna Dong, Kiril Gashteovski, et al. Aligning generalization between humans and machines. *Nature Machine Intelligence*, 7(9):1378–1389, 2025.

[66] Ray S Jackendoff. *Semantics and cognition*, volume 8. MIT press, 1985.

[67] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. AI Alignment: A Comprehensive Survey. *arXiv preprint arXiv:2310.19852*, 2023.

[68] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 2023.

[69] Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. BRAINTEASER: Lateral Thinking Puzzles for Large Language Models. In *Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, 2023. Association for Computational Linguistics.

[70] Tianzhe Jiao, Chaopeng Guo, Xiaoyue Feng, Yuming Chen, and Jie Song. A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications. *Computers, Materials and Continua*, 80(1):1–35, 2024.

[71] John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[72] Daniel Kahneman. *Thinking, Fast and Slow*. Macmillan, 2011.

[73] Subbarao Kambhampati, Kaya Stechly, and Karthik Valmeekam. (How) Do reasoning models reason? *Annals of the New York Academy of Sciences*, 2025.

[74] Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, et al. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10(18): eadk3452, 2024.

[75] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through Memorization: Nearest Neighbor Language Models. In *International Conference on Learning Representations*, 2020.

[76] Emanuel Kitzelmann and Ute Schmid. Inductive Synthesis of Functional Programs: An Explanation Based Generalization Approach. *Journal of Machine Learning Research*, 7(15):429–454, 2006.

[77] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009.

[78] Brian Kulis. Metric Learning: A Survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

[79] William Labov. The boundaries of words and their meanings. In *New Ways of Analyzing Variation in English*, pages 67–90. Georgetown University Press, 1973.

[80] Daniel Lafond, Yves Lacouture, and Andrew L Cohen. Decision-tree models of categorization response times, choice proportions, and typicality judgments. *Psychological Review*, 116(4):833, 2009.

[81] Brenden M. Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023.

[82] Pat Langley and Herbert A Simon. The Central Role of Learning in Cognition. In *Cognitive Skills and Their Acquisition*, pages 361–380. Psychology Press, 2013.

[83] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nature Communications*, 10, 2019.

[84] Douglas B Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[85] Martha Lewis and Melanie Mitchell. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*, 2024.

[86] Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 18471–18480, 2024.

[87] Muxin Liao, Shishun Tian, Yuhang Zhang, Guoguang Hua, Wenbin Zou, and Xia Li. Calibration-Based Multi-Prototype Contrastive Learning for Domain Generalization Semantic Segmentation in Traffic Scenes. *IEEE Transactions on*

Intelligent Transportation Systems, 25(12):20985–21001, 2024.

[88] Henry W. Lin, Max Tegmark, and David Rolnick. Why Does Deep and Cheap Learning Work So Well? *Journal of Statistical Physics*, 168(6):1223–1247, 2017.

[89] Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *International Conference on Machine Learning*. JMLR, 2023.

[90] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[91] Viktor Losing, Barbara Hammer, and Heiko Wersing. KNN Classifier with Self Adjusting Memory for Heterogeneous Concept Drift. In *IEEE International Conference on Data Mining (ICDM)*, pages 291–300, 2016.

[92] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering*, 31:2346–2363, 2019.

[93] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. *arXiv preprint arXiv:2308.08747*, 2025.

[94] Kleanthis Malialis, Jin Li, Christos G Panayiotou, and Marios M Polycarpou. Incremental Learning with Concept Drift Detection and Prototype-based Embeddings for Graph Stream Classification. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2024.

[95] Vasileios Manginas, Nikolaos Manginas, Edward Stevinson, Sherwin Varghese, Nikos Katzouris, Georgios Paliouras, and Alessio Lomuscio. A Scalable Approach to Probabilistic Neuro-Symbolic Verification. *arXiv preprint arXiv:2502.03274*, 2025.

[96] Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence*, 298:103504, 2021.

[97] Gary F Marcus. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT press, 2003.

[98] Douglas L Medin, William D Wattenmaker, and Sarah E Hampson. Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, 19(2):242–279, 1987.

[99] Jason S Metcalfe, Brandon S Perelman, David L Boothe, and Kaleb Mcdowell. Systemic oversimplification limits the potential for human-AI partnership. *IEEE Access*, 9:70242–70260, 2021.

[100] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[101] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

[102] Stephen Muggleton and Luc De Raedt. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679, 1994.

[103] Thomas Müller, Guillermo Pérez-Torró, and Marc Franco-Salvador. Few-Shot Learning with Siamese Networks and Label Tuning. In *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8532–8545, 2022. Association for Computational Linguistics.

[104] Martin Mundt, Yongwon Hong, Iuliia Pliushch, and Visvanathan Ramesh. 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.

[105] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 55(13s), 2023.

[106] Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. Bongard-logo: A new benchmark for human-level concept learning and reasoning. *Advances in Neural Information Processing Systems*, 33:16468–16480, 2020.

[107] Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 12(2):e1449, 2022.

[108] Robert M Nosofsky. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(4):700, 1988.

[109] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The Limitations of Deep Learning in Adversarial Settings. In *IEEE European Symposium on Security and Privacy (EuroS&P)*,

pages 372–387, 2016.

[110] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[111] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.

[112] Eleanor Rosch. Natural categories. *Cognitive Psychology*, 4(3):328–350, 1973.

[113] Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605, 1975.

[114] David E Rumelhart and Peter M Todd. Learning and connectionist representations. In *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, pages 3–30. MIT Press, 1993.

[115] David E Rumelhart, James L McClelland, and PDP Research Group. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. MIT press, 1986.

[116] Ute Schmid and Emanuel Kitzelmann. Inductive rule learning on the knowledge level. *Cognitive Systems Research*, 12(3-4):237–248, 2011.

[117] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[118] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The Pitfalls of Simplicity Bias in Neural Networks. *Neural Information Processing Systems*, 2020.

[119] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[120] Roger N. Shepard. Toward a Universal Law of Generalization for Psychological Science. *Science*, 237(4820):1317–1323, 1987.

[121] Zenglin Shi, Jie Jing, Ying Sun, Joo-Hwee Lim, and Mengmi Zhang. Unveiling the Tapestry: The Interplay of Generalization and Forgetting in Continual Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[122] Gunjan Singh, Riccardo Tommasini, Sumit Bhatia, and Raghava Mutharaju. Benchmarking Neuro-Symbolic Description Logic Reasoners: Existing Challenges and A Way Forward. *Neurosymbolic Artificial Intelligence (to appear)*, 2025.

[123] Ji Y Son, Linda B Smith, and Robert L Goldstone. Simplicity and generalization: Short-cutting abstraction in children's object categorizations. *Cognition*, 108(3): 626–638, 2008.

[124] Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418, 2023.

[125] Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. ARN: Analogical Reasoning on Narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086, 2024.

[126] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331 (6022):1279–1285, 2011.

[127] Ivan Y. Tyukin, Alexander N. Gorban, Muhammad H. Alkhudaydi, and Qinghua Zhou. Demystification of Few-shot and One-shot Learning. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2021.

[128] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, pages 1–11, 2024.

[129] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[130] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer, 2015.

[131] Eli Verwimp, Rahaf Aljundi, Shai Ben-David, Matthias Bethge, Andrea Cossu, Alexander Gepperth, Tyler L Hayes, Eyke Hüllermeier, Christopher Kanan, Dhireesha Kudithipudi, et al. Continual learning: Applications and the road forward. *Transactions on Machine Learning Research*, 2024.

[132] Michael R Waldmann, York Hagmayer, and Aaron P Blaisdell. Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, 15(6):307–311, 2006.

[133] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*, 2023.

[134] Nathaniel Weir, Bhavana Dalvi Mishra, Orion Weller, Oyvind Tafjord, Sam Hornstein, Alexander Sabol, Peter Jansen, Benjamin Van Durme, and Peter Clark. From Models to Microtheories: Distilling a Model's Topical Knowledge for Grounded Question Answering. *arXiv preprint arXiv:2412.17701*, 2024.

[135] Cara Leigh Widmer, Md Kamruzzaman Sarker, Srikanth Nadella, Joshua Fiechter, Ion Juvina, Brandon Minnery, Pascal Hitzler, Joshua Schwartz, and Michael Raymer. Towards Human-Compatible XAI: Explaining Data Differentials with Concept Induction over Background Knowledge. *Journal of Web Semantics*, 79:100807, 2023.

[136] Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable Compositional Generalization for Object-Centric Learning. In *The International Conference on Learning Representations*, 2024.

[137] Patrick H Winston. Learning structural descriptions from examples. Technical report, MIT, AI Technical Reports, 1970.

[138] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[139] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pages 1–28, 2024.

[140] Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J Prenger, and Animashree Anandkumar. LeanDojo: Theorem Proving with Retrieval-Augmented Language Models. *Advances in Neural Information Processing Systems*, 36:21573–21612, 2023.

[141] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A Survey on Causal Inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5), 2021.

[142] Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a Theoretical Framework of Out-of-Distribution Generalization. In *Advances in Neural Information Processing Systems*, 2021.

[143] Yuan Yuan, Zhaojian Li, and Bin Zhao. A Survey of Multimodal Learning: Methods, Applications, and Future. *ACM Computing Surveys*, 57(7), 2025.

[144] Thomas Zeugmann. Can Learning in the Limit Be Done Efficiently? In *Algorithmic Learning Theory*, pages 17–38, 2003. Springer Berlin Heidelberg.

[145] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[146] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.