What would be Hilbert's problems of AI?

Michael Felsberg*, Fredrik Heintz*, Fredrik Johansson†, Danica Kragic‡, Fredrik Lindsten*, Amy Loutfi*§, Alexandre Proutiere‡, Serge Belongie**, Virginia Dignum^x, Fredrik Kahl†, Kathlén Kohn‡, and Bernhard Schölkopf††

*Linköping University, Center of Excellence AI4x, corresponding author: michael.felsberg@liu.se,
†Chalmers University of Technology, ‡Royal Institute of Technology, §Örebro University, ^xUmeå University,
**Pioneer Centre for AI, ††Max Planck Institute for Intelligent Systems & ELLIS Institute Tübingen

March 31, 2026, version 0.04

Abstract

Hilbert's 23 problems in mathematics published in 1900 proved to be very influential for 20th-century mathematics. The attempt to solve these problems did not only consolidate the scientific field, but also led to new areas of research. Consolidation and structuring is definitely required in the field of AI, which has been growing in an untamed way during the past dozen years. Much of this research has been driven by large companies, which even when supporting the basic research are primarily interested in solving problems from application domains rather than addressing the fundamental questions related to the advancement of a scientific field. Those questions have to be addressed by publicly funded basic research that is focused on contributions to the scientific field that might enable long-term benefits of their applications. This continuously evolving white paper addresses the goal of starting an initiative that will formulate the program equivalent to Hilbert's problems for the field of AI. It will be open to public comments, forming the basis for future revisions and extensions. Also, a public website with the status of the problems is available. The long-term impact of the initiative will be an urgently needed complement to the predominantly short-term results produced by applied research.

I. Introduction

Hilbert's problems were one of the main drivers for the field of mathematics during the 20th century [1]. They are an excellent example that formalizing scientific questions is an efficient way to catalyze progress and consolidation of a scientific field. Thus, rather than formulating grand challenges for artificial intelligence (AI) [2], this continuously updated white paper aims to have an impact on the field comparable to that of Hilbert's problems on mathematics.

AI drives a technological revolution with previously unseen pace of progress in a multitude of application domains. For instance, foundation models (FMs) such as BERT [3], ViT [4], and CLIP [5] re-define the state-of-the-art on many machine learning tasks, taking over after previous models based on convolutional neural networks [6] and fully connected networks [7]. Documented evidence suggests that AI models are getting progressively bigger and benefit from access to data and computational resources [8], which is practically easier to provide in corporate research than publicly funded academic research[1]. However, companies are more likely to focus on fast turnarounds and need-driven research to supersede their competitors regarding results in their application domains. This pragmatic focus often leads to FMs trained as black-boxes that successfully predict in the hull of training samples, but might fail on compositional tasks [9] or generate visually appealing yet biologically implausible results[2], limiting their extrapolation capabilities and thus relevance for, e.g., scientific applications.

Little attention is spent on the theoretical aspects, including the use of model knowledge [10], originating from mathematics [11], [12], [13] or the problem domain, such as constraints [14], proper treatment of uncertainty [15], [16], [17], causality [18], [19], [20], [21], and calibration [22], [23], guarantees for the accuracy of outputs [24], and understanding learned representations [25] and failure cases. However, those aspects are essential for many applications, e.g., in safety [26], weather science [27], and for embodied AI [28], [29]. Similarly the interaction with humans [30] and long-term societal consequences need attention as the technology matures.

This white paper aims to strengthen the focus on theoretic aspects and to support the structuring process of AI by means of scientific questions, analogous to Hilbert's problems in mathematics that shaped the field during the past century:
The goal of this white paper is to advance AI as a scientific field beyond an application-driven technology by formulating the fundamental problems and making an attempt to their solutions, while incorporating societal, ethical, and security dimensions.

Hilbert's problems, 23 in number, are usually structured into several groups [1, p. 287pp]: well-formulated problems, many of them solved by today, vaguely-formulated problems that are considered unaddressable, and problems that even Hilbert himself disregarded (e.g. Hilbert's 24th problem, [31]). The ambition here is to formulate problems sufficiently concrete to facilitate progress, and continuous revisions are foreseen to add or improve problem formulations by the authors and other peers that help broadening the perspective to the field.

[1]For example, the 2026 budget request for the U.S. National Science Foundation's Directorate for Computer and Information Science and Engineering is about $0.35 billion, while several major tech companies are each projected to spend roughly $100 billion on AI-related capital expenditures in 2026.
[2]https://neurips.cc/virtual/2024/invited-talk/101129, from 12:08

Many topics in AI are now on the precipice of reaching formulation, e.g.,

- the model identifiability problem [32, p. 281], [33], see also II-D,
- the irreducibility of aleatoric uncertainty [34], see also II-F, and
- the origin of double descent and grokking and the relation between the two [35], [36], see also II-G,

but many topics have not reached a state of maturity. Effort has to be spent on accurate problem formulations, e.g., pointing to our limited understanding of non-linear models (sum of parts differs from the whole) such as in the relation between Volterra series and kernel methods [37], ways of materializing kernel methods in machine learning models [38], and the role of Pareto theory in multi-objective learning [39], [40].

a) Why the applied perspective to AI is not enough: Looking at failures of AI renders the consolidation of AI methodologies purposeful from fundamental and applied perspectives. For instance, observing that LLMs hallucinate references in essays or fail on basic math tasks [9] has led to extensions of models with reasoning units [41], but first after the discovery of these failures.

In a consolidated field, the goal is to predict failures before they occur, instead of relying on empirical observations that are manually analyzed after failure. The principle to verify systems and to have preventive technical checks on devices and vehicles has been established and accepted in society for centuries. For instance, critical software components are subject to automatic testing before committing changes and cars need to do pass regular safety checks. Nobody would accept to use an untested control software in a nuclear power plant or enter a taxi without working brakes. For some reason, it is commonly accepted to do exactly that in applications of AI: Correctness seems to be less important than pleasure. For instance, people use LLM-generated text just because it reads well or people believe in fancy AI-generated images and videos with implausible configurations.

If we want to avoid that AI accelerates the general societal trend to believe in rather fake-facts from social media than scientific evidence, we need to turn AI into an area where correctness and trustworthiness of results are predictable or verifiable, where AI is used if it is reliable and where it is capable of indicating when it has failed. To achieve this is not a scholastic exercise, but essential to the stability of democratic systems, individuals, and possibly even mankind.

b) The quest for answering the problems will lead to groundbreaking findings: While the societal aspects of the potential outcomes are of pragmatic value, the findings are also expected to provide structured input to AI strategies, regulation, policy, and education. This goes beyond Hilbert's original endeavor, as we anticipate that advances in methodology, achieved by means of basic AI research, will inform both public discourse and policymaking.

While this is not the first attempt to cast the concept of Hilbert problems onto AI, see e.g., Jitendra Malik's talk "Hilbert Problems of Computational Vision" 2004[3] or Nicklas Berild Lundblad's talk during the WASP Academia and Industry Days 2025[4], neither of them aimed to produce a list of fundamental AI problems as a community effort. Malik formulated three fundamental computer vision challenges (early vision, static scene understanding, and dynamic scene understanding) and Lundblad outlined four areas of potential academic impact on industry (the next big problem, curation of questions and data, trustworthy safety testing, and governance for public AI).

c) From Turing to Hilbert: Imitation vs. Understanding: A natural point of reference for AI is the formulation of intelligence by Alan Turing, who proposed the imitation game as an operational criterion: a machine is intelligent if its behavior is indistinguishable from that of a human [42]. This perspective has been highly influential and aligns with much of modern AI practice, where success is often measured by performance on observable tasks, which constitutes one element of Donoho's concept of "frictionless reproducibility" [43], and may be one of the factors underlying the rapid growth of AI research.

However, behavioral success alone does not necessarily imply that a system is reliable, robust, or operates for the right reasons. As AI systems are increasingly deployed in settings where failures have significant consequences, it becomes important to move beyond imitation as the sole criterion. In analogy to David Hilbert's program of identifying foundational questions in mathematics, this motivates the formulation of problems that address the underlying principles of intelligent systems, including representation, learning, generalization, and uncertainty.

II. Suggested problems

In the effort to formulate Hilbert problems within AI, contributions from the co-authors have been collected and reviewed. This white paper deliberately covers those AI problems that admit, at least in principle, a concrete formulation. This is what distinguishes a Hilbert problem in AI from a general open problem. It is not a claim that these are AI's most important problems, many of the hardest and most consequential ones, concerning values, accountability, and human-AI interaction, resist formalization by nature. Complementary efforts covering those dimensions are equally necessary.

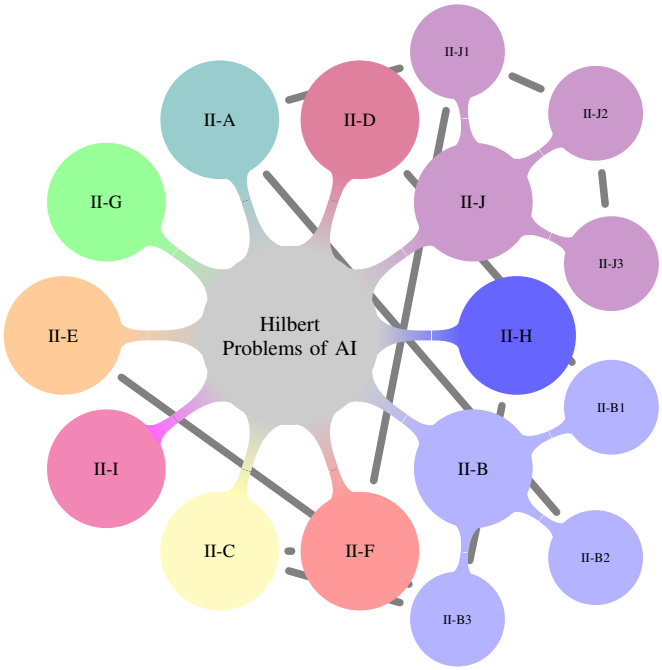
The "Hilbert problems of AI" should be sufficiently concrete so that it is possible to determine whether or not they have been adequately addressed. This could imply a mathematical problem formulation, or it could involve some form of test that would determine the status of the problem.

[3]http://www.cs.berkeley.edu/~malik/talks/hilbert.ppt

[4]https://wasp-sweden.org/wasp-academia-industry-days-2025-bold-ideas-practical-pathways-shared-ambition/

Although desirable in general, to translate the broad AI-challenges to more formal/mathematical problems is actually one of the major open problems (where some would argue that every formalization will miss the core issue). By requiring that there are mathematical formulations, we only cover a small subset of AI and many interesting problems are about multi-objective trade-offs where the specific trade-off in a particular situation is highly context and application dependent.

This section lists a selection of problems formulated by the authors, see Fig. 1, that are sufficiently detailed so that they are ready for discussion in a wider audience, although most of them still need to be made more concrete, or *Hilberty*. Also, this list is far from being finalized – if that is possible at all – and we are continuously working on adding more problems, some of them already suggested but in the need for more details, and some them yet to be suggested.



II-A : Architectural universality
II-B : Learning in dynamical systems
II-B1: The principle of decoupling
II-B2: The universal representation problem
II-B3: Satisfying the physics prior, provably
II-C : The balance between transfer, meta-, and on-line learning
II-D : Problem classes of real-world learning
II-E : Tracking back uncertainty during inference
II-F : Multimodality and selfsupervised learning
II-G : Implicit biases and generalization
II-H : Structure, symmetry, and constraints in learning systems
II-I : Computationalize epistemology
II-J : The meaning of actions
II-J1 : Understanding beyond imitation
II-J2 : Human-AI interaction
II-J3 : Normative specification

Fig. 1. Overview of (sub-) problems in AI and how some are linked. The links are illustrated by gray edges and they indicate potential for future consolidation.

A. Architectural universality

One of the open problems is to characterize architectural universality for intelligence, or said differently: does there exist a (technical) learning architecture that can, in principle, acquire and integrate perception, action, abstraction, reasoning, and self-modeling across open-ended tasks without task-specific redesign under realistic constraints on data, computation, and even embodiment? The goal is to understand whether there are architecture-level equivalents of homomorphism and invariance: structures that allow learning, reasoning, and acting to remain coherent as the agent's internal models and goals evolve, so that intelligence can be understood not as a collection of tasks, but as stable processes operating over changing representational spaces. While universal approximation theorems guarantee that many model classes can approximate arbitrary functions, such results are probably insufficient for intelligence. We should not view "Intelligence" as merely a mapping from inputs to outputs, but an ongoing process of model construction, revision, and deployment under intervention. Universality in this sense requires not only representational capacity, but the ability to adapt internal structure, and modes of inference as the agent encounters qualitatively novel situations.

A variant of the problem would be to study if there are any general principles for how neural architectures could evolve to tackle universal problems. One could imagine a system that starts from a core that could handle initial input/output (or basic sensorimotoric processing) and then gradually changes through structural transformations. The key open problem here would be to determine what principles should be used in such transformations. How can such transformations balance stability and plasticity, ensuring that new capabilities could be gained without losing previous learned capabilites/functions?

To contrast, standard learning fixes issues like a hypothesis class, optimization procedures and it also standardizes the notion of loss, but essentially, this presupposes that the space of problems is known in advance. What we would like to ask is whether there is a universal (technical) intelligent architecture that can somehow be capable of learning new problem formulations, discovering new state variables, inventing abstractions and essentially reorganizing itself while still remaining a coherent agent/entity over time. Technically, this can potentially be reduced to defining what architectural universality means beyond function approximation, or under what conditions a single architecture can support multiple forms of representation, be it statistical, causal, symbolic, or procedural. This relates to the challenge of bounded rationality by Herbert Simon [44]. It also needs to address the challenge of identifying structural constraints on memory, time, energy etc. The interesting problem for AI is whether we can develop architectures that quite simply re-architects itself while preserving both agency and identity.

B. Learning in dynamical systems

Overarching question is: What is the most data-frugal way to learn to control dynamical systems?

1) The principle of decoupling: When should dynamics and control be learned separately?

A dominant trend today is to learn world models that predict system dynamics and subsequently use planning or reinforcement learning (RL) to derive control policies. This mirrors classical control theory, which largely relies on a separation between system identification and optimal control. However, under distribution shifts induced by control, accurate prediction may become impossible or even ill-defined, particularly when the system is driven outside its nominal operating regime. This explains why regularization, limiting exploration close to current policies (on-line learning) or those represented in the data (offline learning), is essential in RL algorithms. An important problem is to characterize the classes of dynamical systems, task families, and data regimes for which decoupling system identification and control is provably data-efficient, and to identify regimes in which such a separation must necessarily fail.

2) The universal representation problem: What must be learned to control?

Learning-based control hinges not on learning dynamics per se, but on learning a representation of the system state that renders prediction, planning, and decision-making feasible. World models, latent representations that support rollouts and planning, are widely believed to represent the current state of the art. An important question is whether universal or near-universal representations exist for broad classes of dynamical systems, representations that are minimal, predictive, and actionable under control, and how such representations can be learned from data using principled approaches, including self-supervised (contrastive) and predictive learning objectives.

In current practice, there is a growing trend to learn large, generic world models from diverse and heterogeneous data sources, and subsequently adapt them to specific domains through fine-tuning. These pretrained world models appear to provide strong initializations, enabling rapid specialization with comparatively little domain-specific data. However, it remains poorly understood why such world models offer effective starting points: what structural, statistical, or representational properties make them broadly transferable? Fundamental questions arise regarding the appropriate balance between broad pretraining and domain-specific adaptation, the most effective strategies for fine-tuning, and the limits of this two-stage paradigm. It is possible that a richer, multi-level training hierarchy is required, involving several intermediate stages of adaptation rather than a single generic-to-specific transition. A central problem is therefore to characterize transferable latent representations that span multiple classes of dynamical systems, to understand how they should be progressively specialized across domains and sensor modalities, and to quantify the minimal data required at each stage of adaptation.

3) Satisfying the physics prior, provably: Physical systems obey on the one hand conservation laws, symmetries, and invariants, whose incorporation into models is in principle reasonably well understood [12], [45] (see problem II-H), and on the other hand more complicated physical laws, e.g., in the form of differential equations, which are currently only incorporated as weak inductive bias [14], [46] or, in generic world models, implicitly through large-scale training on massive datasets in which physical regularities are merely latent. The omission of domain constraints leads to fundamental data inefficiency and uncontrollable model inaccuracies. A central challenge is to understand how physical laws can be incorporated as hard or semi-hard constraints in representation learning and world models. More fundamentally, to what extent are the data actually required to learn physically consistent models, as opposed to being used to select among physically admissible hypotheses? An important direction is to investigate automated verification-style approaches, analogous to those developed for formal reasoning and machine-checked proofs, as a means to guide the construction and validation of physically valid models with minimal reliance on data. Furthermore, can such tools be used to certify predictive validity under control? In particular, an important objective is to determine how these certificates might bound prediction error, detect extrapolation beyond the training regime, and inform downstream planning and control.

C. The balance between transfer, meta-, and on-line learning

In embodied AI, perhaps the most relevant aspect will be the balance between transfer, meta-, on-line, and multimodal learning [47], [48], with respect to the following questions: *What to transfer?* Addressing prior knowledge that should be ignored or forgotten and what should be transferred when dealing with new environments, objects, tasks, and contexts. Here, "knowledge", as in transferable facts and patterns (e.g., LLMs), and "skill", as in the ability to quickly adapt (meta-learn) to new training data (e.g., Prior-fitted Networks (PFNs)), need to be considered as these address the ability to perform transfer on different cognitive and capabilities levels [47]. *How to transfer?* Addressing the contrast between learning on offline and simulated data and methods starting from the transferred knowledge, while filling in the gaps by conducting "targeted" search or by autonomously exploring sensory and proprioceptive data specific to the new context. This relates to active representation learning. *When to transfer?* Methods that establish if and what transfer is viable or if learning from scratch is needed. *Cross-modal transfer?* To what extent can knowledge be transferred between modalities? For example, under what conditions can old sensors be replaced by new ones with no/minimal tuning?

Most of the above require safe exploration: methods for identifying if, and what type of data is needed to ensure performance with guarantees, as well as methods that confirm that something cannot be performed, etc.; methods for consolidating learned representations, removing redundant and obsolete models to ensure efficiency.

D. Problem classes of real-world learning

A well-known concept from computer science is the classification of problem classes according to the asymptotic complexity of their solution: P (solvable in polynomial time) versus NP (nondeterministic polynomial time, requires exponential time in practice) [49]. While the conjecture that P is a proper subclass of NP has still not been proven, similar questions arise in context of machine learning and data-driven approaches in general, both for compute and data. Here, we focus on the second: Assume a model trained on a dataset that is continually extended vs. a model trained using an embodied system interacting with the world. *How do the two cases differ and can both models learn the same functionality?*[5]

Irrespectively resolution and quantization levels, a finite digital dataset is in the end a finite set of discrete values. If a model is trained continuously on data, this set potentially becomes (countable) infinite, e.g., in case of generated data [50], and the trained model becomes increasingly more powerful. However, the relation to training an embodied system is less obvious and the question in the previous paragraph can be reformulated as: Does embodiment effectively go beyond discrete, countable infinite sets? Or: *Is embodied learning more powerful than learning from data?*

Also an embodied system acquires quantized data in a clocked way, thus learning from the same data manifold as a continually learning system without embodiment. Still, it can be of a different functionality than a purely data-driven model if the new data is conditioned on the system, i.e., the system explores its environment [51]. Since the current system itself cannot be known at the time of training data acquisition – the system is learned *after* the respective data has been acquired – the model cannot learn functionalities that depend on the effect of the current embodied system to the real world [52]. On the other hand, if we assume having access to exactly the same training data for the system without embodiment as for the embodied system, we implicitly assume complete knowledge about the embodied system at *all* stages.

This is only possible if we have access to the real system, a model is not sufficient. Even the most advanced simulation-based approaches [50] make simplifying assumptions about the real world and is subject to modeling errors. Thus, irrespectively the same dimensionality of data, the conclusion is that embodied learning is a proper super class of data-driven learning, even in a continual setting. However, several questions remain: 1. What is the difference in functionality between data-driven and embodied learning? 2. How can this difference be measured without confusing it with functionality variation induced by other causes? 3. The assumption that embodied learning uses more resources (e.g. compute and time) than data-driven learning, implies that a certain functionality is achieved more efficiently by data-driven learning, but where is the break-even point?

E. Tracking back uncertainty during inference

Given a specific output of a model during inference, how much uncertainty is originating from the model prior, how much from the input uncertainty? While the uncertainty during training of models is extensively discussed, less attention has been directed to the uncertainty when the model is in use, i.e., during inference, in particular when looking at specific samples, beyond expectation.

The expected uncertainty of a model describes the distribution of output errors in expectation sense and if the predicted uncertainty is a good estimate of the empirical uncertainty, the model is calibrated [53]. The expected uncertainty is composed of two parts, the aleatoric uncertainty and the epistemic uncertainty [54]. While these are commonly called irreducible and reducible, modifications of the input space can lead to a reduced aleatoric uncertainty for the price of a higher epistemic uncertainty [54]. Also, recent results seem to indicated that they are inherently entangled [55].

The training data distribution also implies a bias in the model. This bias can be explicit, e.g. in a classification problem different classes have different amount of training data [56], or implicit, e.g. by one class having a single mode in the latent representation while another one (with the same amount of training data) has multiple modes in the latent space and thus having less support for each mode in the training data or because different imbalances across modalities [57]. If the distribution during test time differs from the training distribution (explicitly or implicitly), the output accuracy is reduced and the expected uncertainty is increased [56].

However, the expected uncertainty is not always sufficient to explain why a certain prediction might fail – which is a highly relevant case, e.g. for investigating failures of AI systems and the resulting responsibilities – or whether it succeeds for the right reasons [58]. Besides the expected uncertainty of the trained model and its potential bias, also the input data might contain errors [59]. If these errors are correctly modeled by the input error distribution during training, one can presumably determine – in expectation value sense – what uncertainty was more likely to be the reason for a failure. However, if the input data is an outlier compared to the input distribution in the training data, outputs become uncorrelated to the uncertainties.

If the input data is an outlier compared to the input data in the training data, the robustness of the model is challenged. The model output should not be affected by the value of the outlier, it should only influence the output uncertainty prediction. Some techniques, such as drop-out [60], systematically improve robustness and other methods propagate the input uncertainty to the prediction [15]. In the latter case, the knowledge about an input being an outlier could be exploited, but it is still difficult to determine whether a certain prediction during inference is incorrect because of an outlier in the input data, because of a bias in the training data, or because of the stochasticity of the model itself. Can the latter causality be excluded by training several models with the same training data or the second one without knowing the data distributions?

[5]This question, formulated by Vladlen Koltun, was debated during the CVPR 2023 AC/SAC workshop “Foundation Models and Embodiment”.

F. Multimodality and selfsupervised learning

Selfsupervised learning is a key to foundation models that need gigantic datasets to be trained where annotation is infeasible. The most common principle is the "next-token-prediction" where tokens are masked out and the task is to predict that token, used both in natural language processing [3] and vision [4]. While the meaning of "next" is well-defined in language and can be generalized to neighborhoods in images, video, and other spatial data, it is less clear how to do this across modalities.

In general we cannot assume simple correlation between modalities. For example, in contact-rich tasks in robotics, it is common that force and torque feedback varies significantly, despite high sensing frequency [61]. Another example may be the olfactory perception, where predicting chemical compounds or even describing various odors may not be consistent between individuals [62].

When training foundation models across modalities, the most commonly used principle is to apply contrastive learning, e.g. in CLIP [5]. The training happens in several steps where first foundation models for the individual modalities are learned, e.g. by masking, and in a second step, tokens are aligned according to the contrastive loss.

Notably, this approach deviates significantly from biological learning, where modalities are trained using support from other, often previously trained modalities [63]. In addition, the relationships can be highly causal, like detected force may mean something moves in the visual field, activations from proprioception or muscle neurons are used to train visual representations [64], or the intricate redundancy between smell, taste and vision [65]. Learning language comes as one of the last steps in biological learning and is heavily supported by other modalities [66], [67].

Hypothetically, selfsupervised learning might become much more powerful by a deep integration of multiple modalities instead of a post-hoc binding of tokens, in particular if there are strong correlations across modalities. In remote sensing applications, this has partly been demonstrated by joint encoder learning, in sensor specific settings [68] or sensor agnostic [69]. However, neither of these works train the foundation models easy-to-hard, i.e., in a progressively more difficult sequence.

If we assume that we add one modality to an existing model that has already been trained on $N-1$ modalities, we consider two different settings, depending on the redundancy of the modality:

a) Redundant case. The new modality does not add new concepts, i.e., the latent space is not extended. In this case, the previous modalities act as a generator for ground truth and training on unlabeled data is effectively supervised learning. Interesting questions here are: 1. How to determine the redundancy of the modality? And 2. What data or how much data is redundant and what are the methods to test this?

b) Nonredundant case. The new modality requires an extension of the latent space. In this case, we can use the approach from single modality selfsupervised learning to introduce the new concepts. Key questions that arise are: 1. Will these new concepts need to be connected to the previous modalities in some way? 2. How to determine the overlap with existing concepts? 3. How does training effort and performance scale in comparison to independent selfsupervised learning plus contrastive learning? And finally and also related to section II-D: 4. What is the optimal order for adding modalities?

G. Implicit biases and generalization

Deep neural networks learn and generalize well, even in the overparametrized regime, i.e., when the model is large with respect to the dataset size. This challenges the bias-variance tradeoff from classical statistical learning theory, that would predict the model to suffer from overfitting. Large neural learning systems seem to exhibit some form of *implicit bias*, that is present in the untrained model [70] or emerges during the training process [71], [72], and prevents them from overfitting.

This surprising behavior is regarded a crucial ingredient of the success of deep learning. It has gained attention in recent years due to the ubiquity and success of extremely large models, such as FMs. In fact, implicit biases are closely related to several phenomena observed and studied in recent years. These include *grokking* [73] – i.e., the tendency of the loss to drop abruptly during training – and *feature learning* [74] – i.e., the emergence of meaningful semantics in specific neurons, but also the learning of short-cuts [58].

A particularly interesting consequence is that the test loss of deep neural networks is often observed to decrease twice, in relation to both model size and training time. While the first drop is a mere consequence of early fitting, the second one is a natural attributed to implicit biases. This phenomenon is referred to as *double descent*, and has been popularized by, among others, Belkin [75].

Clarifying the mathematics behind the emergence of implicit biases of deep neural networks is a fundamental problem, whose solution would shed light on the inner-workings of modern learning systems. It is well-understood for linear, kernel, or random features regression [76], [77], [78], [79], where gradient descent initialized at the origin converges to a minimum-norm solution, resulting in a form of ridge regularization.

However, for deep networks, the scenario is more complex and mysterious, with partial attempts to explain this phenomenon via architectural depth [80], or stochasticity of the dynamics [81]. All in all, there is no general and agreed theory, and a mathematical explanation of implicit biases of neural networks remains open.

H. Structure, symmetry, and constraints in learning systems

Physical systems obey structural constraints such as conservation laws, symmetries, and invariants. Some of these, such as translation symmetry, can be incorporated into model architectures, for example through convolutional neural networks (CNNs) [82]. However, most modern large-scale systems rely on learning such structure implicitly from data, as in DINOv3 [83] and AlphaFold 3 [84]. In particular, AlphaFold 3 illustrates a broader trend: compared to earlier approaches such as AlphaFold 2 [85], which employed equivariant components to address coordinate ambiguities, recent models tend to abandon explicit equivariance in favor of scaling and data-driven learning.

The omission of symmetry constraints can lead to data inefficiency and reduced control over learned representations. Methods from geometric deep learning [86] show that incorporating symmetries as hard constraints can improve generalization and robustness. There is also emerging evidence that symmetry can, in some cases, yield computational and memory efficiency at scale while remaining competitive with standard architectures [87]. At the same time, widely used large-scale models such as DINOv3, AlphaFold 3, and Stable Diffusion [88] do not explicitly enforce such symmetries, but instead rely on learning invariances from large amounts of data.

A central challenge is therefore to determine whether symmetry constraints should be incorporated as hard or semi-hard constraints in representation learning and large-scale models, or learned implicitly from data. More fundamentally, under what conditions does each approach lead to improved data efficiency, generalization, or computational efficiency? Are there classes of symmetries for which explicit incorporation is inherently beneficial, or conversely fundamentally difficult to scale? More broadly, what role should symmetry play in representation learning at scale? This connects to problem II-B3.

I. Computationalize epistemology

Another open problem is to computationalize epistemology: how can an agent, artificial or natural, whose only access to the world is through sensorimotor data, learn internal models that are structurally related to a hypothetical reality that exists independently of the agent?

Standard statistical representations capture correlations but not how actions change the world. Causal representations, in contrast, are mappings from observations to latent variables, and from actions to transformations of those latents, such that the effect of any real-world intervention can be consistently computed in the latent space, leading to a commutative "intervention diagram" [89], [90], [91]. This notion generalizes ideas from Helmholtz and Hertz about perception as unconscious inference, and internal "symbols" whose consequences, computed in the representation space, mirror those observed in the external world. It builds on the idea of homomorphism: i.e., that certain structures are preserved when mapped into another domain, e.g., from the world into a suitable representation.

Physically, every agent has certain interventions at their disposal, and these interventions should inform the nature of the representation. A human equipped with a shovel sees sand as a continuous, manipulable mass, whereas an ant perceives individual grains. Moreover, the representation should be structured such that it best supports planning in terms of the agent's available actions, and thinking as "acting in an imagined space" (K. Lorenz). In this sense, the units of perception should align with units of intervention: what constitutes an "object" to us is something that we can meaningfully intervene upon.

The technical challenge is to define and learn such representations from data: when do they exist, when are they identifiable, when can they be made sparse, disentangled, homomorphic with respect to the composition of interventions, or consistent with an interventional calculus with respect to a causal graph? Can we characterize the conditions under which learned causal representations provably reflect aspects of the modular and causal structure of the world, even without positing a unique ground-truth generative model—with the goal of understanding thinking as "acting in an imagined space" via interventions in the learned representation?

These questions partly connect back to problem II-B1.

J. The meaning of actions

1) Understanding beyond imitation: Can notions such as understanding, reasoning, or reliability be defined in a way that goes beyond behavioral indistinguishability? In particular, is it possible to formulate operational or verifiable criteria that determine whether an AI system succeeds for the right reasons, rather than producing convincing outputs alone? What forms of evidence, structure, or guarantees would be required to certify such properties, especially in settings where ground truth is unavailable or infeasible to verify?

This problem complements Problem A (architectural universality): while Problem A concerns the existence of architectures capable of general intelligence, the present problem asks how such capabilities can be defined and verified. It is also closely related to Problem G (computationalize epistemology), as meaningful notions of understanding are likely to depend on whether learned representations capture the causal and structural properties of the world.

2) Human-AI interaction: If AI systems gain a full understanding of their actions, the aspect of interaction with humans needs to be revisited. Under what conditions does human agency remain meaningful in systems where AI mediates decisions, recommendations, or actions? The challenge is to formalize notions of meaningful human control, contestability, and reliance in ways that are operationalizable, without reducing agency to a binary on/off property or ignoring the social and organizational contexts in which it is exercised. Those contexts are considered in the third sub-problem on the meaning of actions.

3) Normative specification: A core open problem is whether normative requirements, values, fairness criteria, social norms, and contextual obligations, can be specified in ways that are both formally tractable and socially meaningful. Any formalization risks losing the contextual and relational nature of what makes norms binding; yet without some degree of formalization, normative requirements cannot be verified or enforced. The challenge is to characterize what can and cannot be preserved in the translation from social to formal, and under what conditions such translations are even valid.

III. Conclusions

The list of problems above is far from being final and, furthermore, much work remains to turn these initial formulations into concrete and well-formulated problems. We still need to explore the field more deeply for existing questions and related activities, we need to integrate suggested problems that we received, and we are hoping for many constructive comments. Also, the goal of this white paper is not to produce a list of finalized problem formulations, but, by means of the attempt to formulate a list of problems, to trigger discussions in the field.

This white paper, including the graph in Fig. 1, is continuously updated, its version number converging to $\frac{1}{23}$. Substantial suggestions for new problems will be considered for future versions. Also, propositions for modifications and consolidation of problems are welcome. If the resulting changes are substantial and approved, the proposers will be considered for co-authorship. All published problems will, together with references pointing to their origin, be published on a dedicated website[6].

Acknowledgment

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, by the Swedish Research Council through the project with Registration Number 2025-07498, by Linköping University through the Center of Excellence AI4x, and by the Pioneer Centre for AI, DNRF grant number P1.

We acknowledge the contributions of all peers that participated in discussions, in particular for this version Karl Åström, Samuel Kaski, Denis Kleyko, Marcus Liwicki, Giovanni Luca Marchetti, Simon Olsson, Thomas Schön, Jonas Unger, and Anders Ynnerman. Should we have forgotten to acknowledge anyone or missed to invite for co-authorship, please contact us.

References

[1] J. J. Gray and D. Rowe, *The Hilbert Challenge*. Oxford University Press, 2000.

[2] R. Reddy, "Foundations and grand challenges of Artificial Intelligence: AAAI presidential address," *AI Magazine*, vol. 9, no. 4, p. 9, 1988.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2019, pp. 4171–4186.

[4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021, pp. 1–21.

[5] A. Radford et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[6] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[7] F. Rosenblatt, "The Perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[8] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[9] N. Dziri et al., "Faith and fate: Limits of transformers on compositionality," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023, pp. 1–40.

[10] J. Ok, A. Proutiere, and D. Tranos, "Exploration in structured reinforcement learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 1–9.

[11] G. L. Marchetti, V. Shahverdi, S. Mereta, M. Trager, and K. Kohn, "Position: Algebra unveils deep learning – an invitation to neuroalgebraic geometry," in *International Conference on Machine Learning (ICML)*, 2025, pp. 1–15.

[12] P. Melnyk, M. Felsberg, M. Wadenbäck, A. Robinson, and C. Le, "On learning deep O(n)-equivariant hyperspheres," in *International Conference on Machine Learning (ICML)*, 2024, pp. 1–16.

[13] L. A. Pérez Rey, G. L. Marchetti, D. Kragic, D. Jarnikov, and M. Holenderski, "Equivariant representation learning in the presence of stabilizers," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2023, pp. 693–708.

[6] https://isy.gitlab-pages.liu.se/staff/micfe03/en/#hilbert-problems-of-ai

[14] D. Hansen, D. C. Maddix, S. Alizadeh, G. Gupta, and M. W. Mahoney, "Learning physical models that can respect conservation laws," *Physica D: Nonlinear Phenomena*, vol. 457, 2024.

[15] A. Eldesokey, M. Felsberg, K. Holmquist, and M. Persson, "Uncertainty-aware CNNs for depth completion: Uncertainty from beginning to end," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 015–12 023.

[16] H. Govindarajan, P. Sidén, J. Roll, and F. Lindsten, "DINO as a von Mises-Fisher mixture model," in *International Conference on Learning Representations (ICLR)*, 2023, pp. 1–19.

[17] A. Olmin and F. Lindsten, "Robustness and reliability when training with noisy labels," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022, pp. 1–21.

[18] F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag, "Generalization bounds and representation learning for estimation of potential outcomes and causal effects," *Journal of Machine Learning Research*, vol. 23, no. 166, pp. 1–50, 2022.

[19] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[20] P. Spirtes, C.N. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2000, vol. 81.

[21] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, 2017.

[22] E. S. Aimar, A. Jonnarth, M. Felsberg, and M. Kuhlmann, "Balanced product of experts for long-tailed recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 967–19 977.

[23] D. Widmann, F. Lindsten, and D. Zachariah, "Calibration tests in multi-class classification: A unifying framework," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 1–11.

[24] F. Zheng and A. Proutiere, "Conformal prediction under Markovian data," in *International Conference on Machine Learning (ICML)*, 2024, pp. 1–28.

[25] G. L. Marchetti, C. J. Hillar, D. Kragic, and S. Sanborn, "Harmonics of learning: Universal Fourier features emerge in invariant networks," in *Annual Conference on Learning Theory (COLT)*, 2024, pp. 1–23.

[26] Y. Bengio et al., "International AI safety report: First key update capabilities and risk implications," *SuperIntelligence – Robotics – Safety & Alignment*, vol. 2, no. 6, 2025.

[27] J. Oskarsson, T. Landelius, M. P. Deisenroth, and F. Lindsten, "Probabilistic weather forecasting with hierarchical graph neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, pp. 41 577–41 648.

[28] A. Jonnarth, O. Johansson, J. Zhao, and M. Felsberg, "Sim-to-real transfer of deep reinforcement learning agents for online coverage path planning," *IEEE Access*, vol. 13, pp. 106 883–106 905, 2025.

[29] V. Aregbede, S. S. Abraham, A. Persson, M. Längkvist, and A. Loutfi, "Affordance-based goal imagination for embodied AI agents," in *IEEE International Conference on Development and Learning (ICDL)*, 2024, pp. 1–6.

[30] N. Akalin, A. Kiselev, A. Kristoffersson, and A. Loutfi, "A taxonomy of factors influencing perceived safety in human–robot interaction," *International Journal of Social Robotics*, vol. 15, no. 12, pp. 1993–2004, 2023.

[31] R. Thiele, "Hilbert's twenty-fourth problem," *The American Mathematical Monthly*, vol. 110, no. 1, pp. 1–24, 2003.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[33] N. W. Henry, G. L. Marchetti, and K. Kohn, "Geometry of lightning self-attention: Identifiability and dimension," in *International Conference on Learning Representations (ICLR)*, 2025, pp. 1–17.

[34] A. D. Kiureghian and O. D. Ditlevsen, "Aleatory or epistemic? Does it matter?" *Structural Safety*, vol. 31, pp. 105–112, 2009.

[35] X. Davies, L. Langosco, and D. Krueger, "Unifying grokking and double descent," in *NeurIPS Workshop on ML Safety*, 2022, pp. 1–9.

[36] A. Olmin and F. Lindsten, "Towards understanding epoch-wise double descent in two-layer linear neural networks," *arXiv:2407.09845*, pp. 1–48, 2024.

[37] M. O. Franz and B. Scholkopf, "A unifying view of Wiener and Volterra theory and polynomial kernel regression," *Neural Computation*, vol. 18, pp. 3097–3118, 2006.

[38] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2007, pp. 1–8.

[39] P. Vamplew et al., "Scalar reward is not enough: A response to silver, singh, precup and sutton (2021)," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 41, 2022.

[40] P. Ma, T. Du, and W. Matusik, "Efficient continuous Pareto exploration in multi-task learning," in *International Conference on Machine Learning (ICML)*, 2020, pp. 6522–6531.

[41] F. Präntare, H. Appelgren, and F. Heintz, "Anytime heuristic and Monte Carlo methods for large-scale simultaneous coalition structure generation and assignment," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2021, pp. 11 317–11 324.

[42] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 49, pp. 433–460, 1950.

[43] D. Donoho, "Data science at the singularity," *Harvard Data Science Review*, vol. 6, no. 1, pp. 1–35, 2024.

[44] H. A. Simon, "A behavioral model of rational choice," *The Quarterly Journal of Economics*, vol. 69, no. 1, pp. 99–118, 1955.

[45] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[46] M. Raissi, P. Perdikaris, and G. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[47] N. Jaquier et al., "Transfer learning in robotics: An upcoming breakthrough? A review of promises and challenges," *The International Journal of Robotics Research*, vol. 44, no. 3, pp. 465–485, 2025.

[48] A. Billard et al., "A roadmap for AI in robotics," *Nature Machine Intelligence*, vol. 7, no. 6, pp. 818–824, 2025.

[49] L. Fortnow, "The status of the P versus NP problem," *Communications of the ACM*, vol. 52, no. 9, pp. 78–86, 2009.

[50] Z. Xian et al., *Genesis: A generative and universal physics engine for robotics and beyond*, https://genesis-embodied-ai.github.io/, 2026.

[51] R. Mon-Williams, G. Li, R. Long, W. Du, and C. G. Lucas, "Embodied large language models enable robots to complete complex tasks in unpredictable environments," *Nature Machine Intelligence*, vol. 7, no. 4, pp. 592–601, 2025.

[52] Y. Zhang, J. Tian, and Q. Xiong, "A review of embodied intelligence systems: A three-layer framework integrating multimodal perception, world modeling, and structured strategies," *Frontiers in Robotics and AI*, vol. 12, pp. 1–16, 2025.

[53] J. Bröcker, "Reliability, sufficiency, and the decomposition of proper scores," *Quarterly Journal of the Royal Meteorological Society*, vol. 135, pp. 1512–1519, 2009.

[54] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, no. 3, pp. 457–506, 2021.

[55] B. Mucsányi, M. Kirchhof, and S. J. Oh, "Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, pp. 50 972–51 038.

[56] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang, "Disentangling label distribution for long-tailed visual recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6626–6636.

[57] P. Fu, N. I. R. Ruhaiyem, and J. Wang, "Reweighting balanced representation learning for long tailed image recognition in multiple domains," *Scientific Reports*, vol. 15, pp. 1–17, 2025.

[58] J. Kauffmann, J. Dippel, L. Ruff, W. Samek, K.-R. Müller, and G. Montavon, "Explainable AI reveals Clever Hans effects in unsupervised learning models," *Nature Machine Intelligence*, vol. 7, no. 3, pp. 412–422, 2025.

[59] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1765–1773.

[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[61] W. Xie and N. Correll, "Towards forceful robotic foundation models: A literature survey," *arXiv:2504.11827*, pp. 1–28, 2025.

[62] D. Purves, G. Augustine, D. Fitzpatrick, and et al., "Olfactory perception in humans," in *Neuroscience*, 2nd, Sunderland, MA: Sinauer Associates, 2001, ch. 15.

[63] L. E. Bahrick and G. Hollich, "Intermodal perception," in *Reference Module in Neuroscience and Biobehavioral Psychology*, Elsevier, 2017.

[64] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 37–45.

[65] J. Prescott, "Multisensory processes in flavour perception and their influence on food choice," *Current Opinion in Food Science*, vol. 3, pp. 47–52, 2015.

[66] A. Gopnik, A. N. Meltzoff, and P. K. Kuhl, *The Scientist in the Crib: What Early Learning Tells Us About the Mind*. William Morrow & Co., 1999.

[67] G. Vigliocco, P. Perniss, and D. Vinson, "Language as a multimodal phenomenon: Implications for language learning, processing and evolution," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 369, no. 1651, pp. 1–7, 2014.

[68] Y. Wang et al., "Towards a unified Copernicus foundation model for earth vision," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 9888–9899.

[69] L. Waldmann et al., "Panopticon: Advancing any-sensor foundation models for earth observation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025, pp. 2229–2239.

[70] D. Teney, A. M. Nicolicioiu, V. Hartmann, and E. Abbasnejad, "Neural Redshift: Random networks are not random functions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 4786–4796.

[71] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *Journal of Machine Learning Research*, vol. 19, no. 70, pp. 1–57, 2018.

[72] V. Shahverdi, G. L. Marchetti, and K. Kohn, "Learning on a razor's edge: Identifiability and singularity of polynomial neural networks," in *International Conference on Learning Representations (ICLR)*, 2026, pp. 1–23.

[73] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization beyond overfitting on small algorithmic datasets," *arXiv:2201.02177*, pp. 1–10, 2022.

[74] L. Sharkey et al., "Open problems in mechanistic interpretability," *Transactions on Machine Learning Research*, pp. 1–89, 2025.

[75] M. Belkin, "Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation," *Acta Numerica*, vol. 30, pp. 203–248, 2021.

[76] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, vol. 117, pp. 30 063–30 070, 2020.

[77] T. Liang and A. Rakhlin, "Just interpolate: Kernel "Ridgeless" regression can generalize," *Annals of Statistics*, vol. 48, pp. 1329–1347, 2020.

[78] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and the double descent curve," *Communications on Pure and Applied Mathematics*, vol. 75, no. 4, pp. 667–766, 2022.

[79] B. Adlam and J. Pennington, "The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization," in *International Conference on Machine Learning (ICML)*, 2020, pp. 74–84.

[80] S. Arora, N. Cohen, and E. Hazan, "On the optimization of deep networks: Implicit acceleration by overparameterization," in *International Conference on Machine Learning (ICML)*, 2018, pp. 244–253.

[81] G. Blanc, N. Gupta, G. Valiant, and P. Valiant, "Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process," in *Annual Conference on Learning Theory (COLT)*, 2020, pp. 483–513.

[82] Y. LeCun et al., "Backpropagation applied to handwritten ZIP code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[83] Meta AI, *Dinov3*, Self-supervised vision model, 2024.

[84] J. Abramson et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, 2024.

[85] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[86] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges*. Cambridge University Press, 2021.

[87] G. Bökman, D. Nordström, and F. Kahl, "Flopping for FLOPs: Leveraging equivariance for computational efficiency," in *International Conference on Machine Learning (ICML)*, 2025, pp. 1–16.

[88] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 684–10 695.

[89] B. Schölkopf et al., "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[90] P. K. Rubenstein et al., "Causal consistency of structural equation models," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017, pp. 1–10.

[91] B. Schölkopf, "What is a causal representation?" In *Overparametrization, Regularization, Identifiability and Uncertainty in Machine Learning*, ser. Oberwolfach Reports, N. Cesa-Bianchi, P. Hennig, A. Krause, and U. von Luxburg, Eds., European Mathematical Society Publishing House, 2025, pp. 199–203.