The Role of Rigor in Artificial Intelligence

Timothy Nguyen*
Google DeepMind
timothycnguyen@google.com

Abstract

Artificial Intelligence (AI) has achieved extraordinary capabilities despite lacking many of the conceptual and scientific foundations associated with mature disciplines. Unlike traditional sciences, where reliable technology typically emerges from theoretical understanding, modern AI has progressed largely through performance-driven iteration and "alchemical" experimentation. This tension motivates a systematic analysis of AI through the lens of rigor. We introduce a three-part framework consisting of conceptual rigor (clarifying foundational concepts), epistemic rigor (establishing scientific understanding), and operational rigor (ensuring reliable performance and deployment). Using this framework, we analyze competing conceptions of intelligence and understanding, the strengths and limitations of the empirical approach to deep learning, the power and pitfalls of benchmarks, and the obstacles to theory development posed by modern AI systems. We argue that the distinctive trajectory of AI arises from how forms of rigor interact across paradigms, resulting in the primacy of operational rigor in modern deep learning. This perspective helps explain both AI's rapid advances and its persistent uncertainties, while clarifying the challenges involved in transforming AI into a mature science and reliable technology.

Rigorous science has enabled humanity to understand the world and build reliable technology. By formulating testable hypotheses and developing increasingly general theories, the sciences have progressively reduced uncertainty while broadening explanatory power. These advances have yielded technologies whose reliability derives not merely from empirical success but from principled understanding. Physics predicts the path of rockets, chemistry models molecular interactions underlying drug design, and evolutionary biology explains the diversity of species through natural selection.

Modern artificial intelligence (AI) departs sharply from this pattern. Current AI systems exhibit remarkable capabilities despite limited understanding of how or why they work. Progress is often driven less by explanatory theory than by large-scale experimentation and performance on benchmarks. At the same time, many of the field's core terms, such as intelligence and understanding, remain ambiguous and contested. Together, these disparities raise questions about how scientific and conceptual progress in AI should be evaluated. This motivates examining AI from the standpoint of rigor. Here, rigor denotes the disciplined use of methods and standards that support the aims of a field. This includes axiomatic-deductive proof in mathematics and stringent experimental testing in the natural sciences. A natural difficulty in formulating rigor in AI is that the field is multidisciplinary: it spans domains such as computer science, statistics, engineering, cognitive science, neuroscience, and philosophy, each governed by different standards and aims. Consequently, rigor cannot be treated as a single, uniform standard, but must instead be understood in accordance with the different functions it serves. Distinguishing these forms of rigor provides a coherent framework for analyzing the diverse problems and tensions within AI. It also enables an understanding of the distinctive structure of progress and paradigms throughout the history of AI.

We propose a three-part division for rigor in AI: conceptual, epistemic, and operational.[1] Conceptual rigor concerns the clarity of foundational concepts and paradigms. Epistemic rigor concerns the establishment of scientific knowledge and understanding. Operational rigor concerns the reliable deployment and performance of systems in practice. Together, these forms of rigor are sufficiently broad to incorporate many of the salient problems and ambitions within AI. But more than merely

*The author thanks Marcus Hutter, Matthew McAteer, and Tim Scarfe for their invaluable comments and suggestions.

[1] An alternative analysis of rigor in AI is offered in [1], which proposes a six-part division oriented toward responsible AI research practices.

providing a descriptive framework, they interact across paradigms in ways that have shaped both the structure of progress in AI and the predominance of performance-driven development in modern deep learning. By identifying the resulting gaps and uneven development, we clarify where further rigor is needed for AI to mature as a scientific and technological discipline.

1 Conceptual Rigor

Conceptual rigor involves the formulation of clear and consistent terminology. This includes the use of exact definitions, substantive descriptions, or illustrative examples to elucidate core concepts. Such precision enables proper usage of terms, a prerequisite for subsequent development. Additionally, conceptual rigor demands the clear articulation of paradigms: foundational frameworks that introduce axiomatic assumptions and basic methodologies. Paradigms integrate terms, axioms, and methods into a coherent framework for formulating and evaluating claims.

The development of AI has been shaped by such conceptual frameworks from its earliest beginnings. One of its foundational ideas was the notion of general-purpose computation. This notion was first conceived by Charles Babbage and Ada Lovelace in the 19th century through their analysis of the Analytical Engine, the earliest attempt at a digital computer. A century later, Alan Turing formally characterized universal computation through the universal Turing machine, and subsequently suggested that machines might learn rather than rely entirely on explicitly programmed rules, foreshadowing modern machine learning [2]. Central to these developments was the idea that a single machine could in principle perform any computable operation given an appropriate encoding.

Drawing from these ideas, the term "artificial intelligence" was coined by John McCarthy in a 1956 workshop proposal grounded in "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [3]. The proposal further envisioned machines that could use language, form abstractions, solve problems, and improve themselves. Given the proposal's pragmatic aim to secure funding, it made no attempt to define (artificial) intelligence[2]. In the absence of a standard definition, early AI research proceeded by operationalizing intelligence in terms of specific capabilities, such as reasoning, problem-solving, and language use, each grounded in distinct theoretical frameworks. This lack of a shared conception meant that judgments of progress were often internal to particular research programs, and over time, what counted as "intelligence" shifted as systems succeeded on previously challenging tasks.

The absence of conceptual agreement about intelligence is not merely a semantic inconvenience. In scientific practice, concepts play a normative role: they determine what phenomena are to be explained, what counts as evidence, and how progress is to be measured. Without a sufficiently clear characterization of intelligence, evaluations of AI systems risk becoming fragmented across tasks and research programs, thereby undermining the possibility of cumulative progress. Achieving conceptual rigor in AI thus requires a clear notion of intelligence: more than just a prerequisite for effective communication, this foundational clarity is essential for the coherence of the field itself.

Over the years, many definitions have been proposed for intelligence [5, 6, 7]. In [6], a collection of 70-odd definitions is provided for comparison, as obtained from dictionaries, encyclopedias, psychologists, and AI researchers. While there is much in common among these definitions, it is perhaps unrealistic to expect that a universal one could be obtained with enough effort and synthesis. Given this, it is more reasonable to provide not a definition but instead a characterization of intelligence that draws upon its commonly acknowledged features:

Intelligence is a multifaceted property arising from many factors that each vary across a spectrum, including

- *Learning and problem-solving*
- *Adaptability and skill acquisition*
- *Goal achievement, planning, and prediction*
- *Knowledge and understanding*

[2]The name artificial intelligence was conceived as a practical alternative to other terms such as automata studies and complex information processing [4].

- *Reasoning and abstraction*
- *Efficiency under resource constraints.*

Viewed through this characterization, intelligence is not binary but varies across multiple dimensions. This diversity helps explain conflicting views on frontier AI systems: judgments differ depending on whether one emphasizes their achievements or their limitations. In [8], the authors perform an extensive investigation of the GPT-4 large language model (LLM), concluding that "Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system." More strongly, Geoffrey Hinton believes that today's large language models are intelligent, and moreover understand language and can have experiences in the same way humans do [9, 10].[3] On the other hand, Yann LeCun has declared that current AI systems are not even as smart as a household cat, noting how AI systems cannot plan or reason well and do not understand the world [11]. In [12], which emphasizes among other things the efficiency dimension of intelligence, it is noted that "There is little reason to expect LLMs to be intelligent since all we have been training through endless benchmark targeting is hugely overparameterized capability." In these four cases, different facets of intelligence are being considered: performance, behavior, reasoning, and efficiency. These conflicting assessments stem from treating these distinct facets under a single term. If natural language possessed separate terms for intelligence$_X$, where X denoted each facet, debates would become more precise and conflicts easier to disentangle. This hypothetical refinement illustrates a practical role for conceptual rigor that is often lacking in contemporary AI discourse: clarifying distinctions obscured by ambiguous terminology.

A closely related and equally fraught term is "understanding". Debates concerning whether AI systems genuinely understand are especially important because claims about intelligence are often accompanied by broader questions of meaning and reasoning [13]. As with intelligence, disagreements frequently arise from different underlying assumptions about what constitutes understanding. Whereas notions of intelligence differ in what features they emphasize, views on understanding often differ along two dimensions: degree and form.

The first dimension, degree, highlights that understanding, like intelligence, exists along a spectrum rather than as an all-or-nothing quality. However, the nature of this spectrum differs markedly between human and artificial systems. Human understanding is generally smoothly cumulative, progressing from beginner to expert in a manner aligned with our cognitive processes. In contrast, AI systems exhibit what has been termed "jagged intelligence", where they can be superhuman while failing in the most basic ways [14, 15]. They can solve research-level mathematics problems but nevertheless be mistaken about the most trivial arithmetic statements, such as asserting that "$9.11 > 9.9$" [16]. They can speak hundreds of languages, yet are easily confused under simple reversals of word order [17]. As with intelligence, it is unclear whether understanding must be human-like or whether it has another sensible instantiation when applied to artificial systems.

Parallel to questions about the degree of understanding are questions about its form: what kinds of processes, structures, or relations are constitutive of genuine understanding. We can summarize various views as follows. Behaviorism attributes understanding primarily to observable behavioral competence rather than to particular internal mechanisms or representations [18]. By contrast, representational and functionalist approaches hold that understanding depends on the organization and causal-functional role of internal states within a cognitive system [19, 20]. Finally, grounding-based perspectives emphasize that understanding depends not merely on the manipulation of internal representations, but on those representations acquiring semantic significance through appropriate connections to perception, action, or sensorimotor engagement with the world [21].

Disagreements about whether current AI systems "understand" often reflect differing assumptions about which of the preceding criteria are necessary. A behaviorist orientation emphasizes successful task performance, while a functionalist perspective asks whether systems instantiate the appropriate internal processes (e.g. whether some form of reasoning or planning is occurring) [22]. On the other hand, some representational approaches argue that genuine understanding requires internal world models[4] capable of supporting robust abstraction and causal reasoning [25, 26], although others note that LLMs may already possess some form of meaning through the structure of their internal representations [27]. Finally, grounding-based perspectives highlight the need for systems to include sensorimotor

[3] Hinton does not appear to offer an explicit definition of intelligence or understanding.

[4] There is ongoing debate about whether current neural networks possess a (partial) world model, particularly for those that play board games [23] or generate videos [24].

modalities [28] or more generally, reference to entities outside of language [29]. While these positions are not mutually exclusive, debates about whether current AI systems understand, much like debates about intelligence, are often driven less by empirical disagreement than by differing philosophical assumptions about the criteria understanding requires.

Beyond clarifying individual terms, the field of AI has also evolved through a succession of broader conceptual paradigms. These paradigms have shaped both research directions and expectations concerning what AI systems can achieve, while also reflecting different relationships between foundational assumptions, scientific understanding, and engineering practice. Their evolution is examined further in Section 4.

More broadly, conceptual rigor in AI shares affinities with the analytic tradition in philosophy through its emphasis on careful distinctions, explicit meanings, and the dissolution of ambiguities. Yet the parallel is only partial: although conceptual analysis in the analytic tradition often aims to clarify foundational concepts, terms such as intelligence and understanding in AI do not function in the way that, for example, mass or charge do in physics.[5] Their meanings are context-dependent and continually revised in light of new artifacts and capabilities. Accordingly, notions of intelligence are best understood not as the hard core of a unified research program in the Lakatosian sense, but as interpretive lenses that emphasize different dimensions of performance and competence. Some of these lenses are tied to specific research agendas while others serve as evaluative standpoints for interpreting the significance of existing systems. The value of conceptual rigor, then, is not that it delivers conclusive definitions, but that it clarifies the scope and evidential basis of the terms in use. It helps specify what precisely is being asserted, under what assumptions it holds, and how far a given conclusion can legitimately be extended. In this way, conceptual rigor does not end disagreements, but it makes them more precise.

Yet conceptual clarification alone cannot determine whether AI systems genuinely reason, generalize, or understand. Addressing such questions requires not only conceptual analysis, but also empirical and theoretical investigation. This motivates a second form of rigor: epistemic rigor.

2 Epistemic Rigor

Epistemic rigor provides the standards for generating and validating knowledge. It includes the use of mathematical proof, the formulation of testable hypotheses, and the careful execution of controlled experiments. Such methods ground the legitimacy of a scientific discipline. Without epistemic rigor, it is difficult to determine when results will generalize or why a method succeeds at all.

The notion of epistemic rigor developed here differs from traditional concerns in epistemology and the philosophy of science. Epistemology has largely focused on the conditions under which beliefs count as knowledge, including questions of justification and skepticism, while philosophy of science has examined the evaluation of scientific theory, including how claims are confirmed, falsified, or revised. By contrast, the notion of epistemic rigor adopted here is more pragmatic. Rather than seeking ultimate justification, it emphasizes the procedural and methodological conditions under which scientific results become credible in practice. In this sense, epistemic rigor is not primarily a theory of knowledge, but a framework for how scientific communities establish reliable understanding in complex domains such as AI, where formal theory and empirical performance are deeply intertwined.

While many earlier areas of AI research achieved a high degree of theoretical grounding[6], modern deep learning has shifted the field toward large-scale empirical methods whose successes often outstrip scientific understanding. In this setting, epistemic rigor is especially important for producing reliable results, given the scale of resources needed to conduct experiments [31, 32] as well as the rapid pace at which results are produced and disseminated [33]. Yet epistemic rigor concerns more than the validation of isolated empirical findings. Scientific understanding also requires unifying these findings into a robust and generative body of knowledge. Accordingly, these aims depend upon three closely related criteria: reproducibility, predictability, and explainability.

[5] While there are formal mathematical accounts of intelligence (e.g. [30]), they function as idealizations amenable to study rather than as core foundations.

[6] Earlier approaches were often founded on mathematically grounded methods such as convex optimization, online learning, and probabilistic modeling.

2.1 Reproducibility

Reproducibility requires that key findings remain robust under ordinary and inevitable variations in experimental conditions, such as differences in infrastructure or implementation. This allows results to be shared and accepted as established findings.

AI research[7], when conducted properly, enjoys a comparatively high degree of epistemic rigor with respect to reproducibility. Experiments are driven by code, data, and models, all of which can be copied and shared. Furthermore, open source repositories and standardized benchmark datasets enable ideas to spread and results to be quickly verified or disconfirmed.

Reproducibility in AI research, however, remains more fragile than it may initially appear [34, 35]. In [34], a useful distinction is made between *results reproducibility* (reimplementation of a method generates statistically similar values) and *inferential reproducibility* (varying experimental setups lead to similar conclusions). Results reproducibility ensures that outcomes from independent replication of the experiment under the expected sources of variation (e.g. random seed, code implementation) remain stable, while inferential reproducibility ensures that wider claims inferred from the original experimental setup stand up to scrutiny. Unfortunately, differing implementations of learning algorithms can often lead to significant alterations in performance, thus undermining results reproducibility [35]. At a broader level, conclusions about which algorithms or architectures perform best are often revised when the experimental settings are changed or expanded, in turn making inferential reproducibility a challenge [36, 34, 37]. Together, such findings illustrate that reproducibility in AI depends not merely on sharing experimental configurations and artifacts, but on careful methodology and transparent reporting of experimental variability.

2.2 Predictability

Predictability is central to scientific understanding because it enables researchers to characterize or anticipate system behavior prior to experiment. Such prediction may proceed deductively through formal mathematical analysis or inductively through the discovery of empirical regularities and laws.

In AI, neural networks have the distinct advantage of being software: they are specified entirely in terms of mathematics and code. As a result, neural networks and their training dynamics are in principle fully specified and therefore amenable to precise formal analysis. This contrasts with the situation in the physical sciences, where there is a distinction between the material objects studied by experimentalists and the mathematical models studied by theoreticians. Consequently, the limits of prediction arise less from misalignment between theory and reality than from the intelligibility of the systems themselves.

In practice, however, the predictive reach of theory is often narrowly circumscribed. For one, the mathematical objects involved are extremely complicated: neural networks involve many parameters (millions if not trillions), consist of highly nested compositions of nonlinear functions, and are optimized with respect to a nonconvex objective. Furthermore, in real-world settings, the data involved are complex and noisy, and thus are not amenable to the clean hypotheses used in mathematically tractable settings. Because precise theoretical prediction is often unavailable, researchers instead rely on reproducibility as a foundation for identifying stable empirical regularities. In effect, rather than beginning with a theoretical prediction that researchers set out to confirm or falsify, they start with a reproducible baseline and proceed via exploration and discovery. These contrasting modes of research have been distinguished in [34, 38] as *confirmatory* (or *empirical*) research and *exploratory* research, respectively. The prevalence of the latter over the former means that successful results in AI typically occur as a consequence of experiment rather than through the confirmation of an experimental prediction.

Many specific aspects of deep learning are nevertheless amenable to prediction due to their observed regularity: increasing compute tends to improve performance [39], models retrained on new data catastrophically forget [40], LLMs regurgitate their training data [41, 42], and the like. But such predictions remain limited in scope, typically applying only within closely related settings. This falls far short of the predictive range of theories in the natural sciences that extend well beyond the domains in which they were developed. Gravity on the surface of the earth is the same gravity that causes the Moon to orbit the Earth. Mass spectrometry reveals the chemical composition of ordinary materials

[7]Henceforth, "AI research" generally refers to research in modern deep learning, its most influential and relevant subfield.

and the distant stars. By comparison, deep learning possesses relatively few predictive theories capable of extrapolating reliably to distant regimes.

Despite these limitations, a small number of predictive frameworks with meaningful extrapolative power have emerged. The first arises from theories that consider various limits in which the width of a neural network goes to infinity, resulting in greatly simplified dynamics [43, 44, 45, 46]. This has enabled predictions of how optimal choices of hyperparameters like the learning rate scale with model size [46]. A second important example is so-called "scaling laws" [47, 39]. These laws describe power law behavior in the loss function of a neural network at the end of training as a function of quantities such as compute, data, or model size. The significance of these laws stems from their empirical validity across many orders of magnitude, allowing neural network behavior to be extrapolated to scales beyond the current frontier. They thus enable researchers to predict aspects of training outcomes, such as the final loss, before committing resources to large-scale experiments.[8] More broadly, recent work has argued that such developments may represent the early stages of an emerging scientific theory of deep learning centered on the dynamics and statistics of the learning process itself [49].

Nevertheless, AI research remains hindered in its development as an epistemically rigorous science due to limited predictability in key areas. First, existing theory often fails to specify in advance the conditions under which neural networks will generalize or remain reliable. In particular, there is no universally robust characterization of when inputs should be regarded as out-of-distribution relative to the training data. Moreover, the ease with which adversarial examples[9] can be constructed, together with the difficulty of characterizing and preventing them, limits the ability to predict the conditions under which neural networks will behave robustly [50]. Second, both the performance of algorithms and comparisons between them depend critically on the choice of hyperparameters [36, 35, 51]. Yet without predictive principles governing how hyperparameters should be selected, practitioners often rely upon extensive search, manual tuning, and expert intuition to obtain strong performance. This sensitivity makes it difficult to determine whether observed behavior reflects robust properties of an algorithm or artifacts of particular configurations. Together, these limitations in predicting neural network behavior and algorithmic performance present significant obstacles to epistemic rigor.

2.3 Explainability

Scientific explanations seek to account for how and why phenomena occur. Good explanations contribute to epistemic rigor by enabling generalizable knowledge and counterfactual reasoning. Ultimately, explanations provide the basis for scientific understanding.

For deep learning, the lack of adequate explanations for its successes has led many to regard the field as "alchemy" [52, 53, 54]. This characterization deserves closer inspection, as fields like medicine and neuroscience operate on incomplete theoretical accounts of their underlying mechanisms yet are not labeled as alchemical practices. What distinguishes deep learning is not simply incomplete understanding, but the nature of its epistemic limitations. First, it defies effective hierarchical abstraction. In mature sciences, explanatory questions can often be localized to distinct levels of analysis, where questions are resolved at well-defined scales such as the subatomic or cellular level. Deep learning provides no such clarity: it is frequently unclear whether observed behavior should be attributed to individual neurons, layer interactions, optimization dynamics, training data, or some combination thereof. Consequently, the field remains without a systematic framework for decomposing phenomena into distinct components and their associated explanations. Second, neural networks lack interpretability.[10] This issue extends beyond mere complexity. Biological neural networks and large collections of molecules also form complex systems. The central difficulty is that deep neural networks learn features that are not readily interpretable. By comparison, traditional machine learning makes use of explicitly defined attributes. And in other scientific fields, abstractions have components with well-defined features. Electrons have mass and spin, while molecules have chemical formulae. However, the feature vectors produced by neural networks live in high-dimensional vector spaces that defy straightforward interpre-

[8] These laws are not without caveats and controversy, and their exact form continues to be revised. See [48] for an overview.

[9] Adversarial examples, in their original form, arise from adding visually imperceptible perturbations to an input image that cause a model to confidently misclassify it.

[10] While explainability and interpretability are often used interchangeably, we regard the latter as providing a foundation for the former. We take interpretability roughly to mean comprehensibility, for example, texture and shape are interpretable features of an image.

tation. As such, neural networks are regarded as "black boxes" because, while they are structurally transparent, their inner workings elude understanding.

The effectiveness of scaling in deep learning has also reduced the need for explainability. Traditionally, improving performance meant designing better algorithms, a process requiring significant expertise and understanding. But with deep learning, approaches that scale effectively often continue to improve with sufficient data and compute. In turn, continued capability improvements through scaling have become more predictable, while developing good explanations remains a separate and more uncertain endeavor.

Despite these difficulties, deep learning does possess partial explanatory frameworks that contribute meaningfully to epistemic rigor. Many core components of deep learning are built upon sound theoretical principles arising from classical statistical learning [55, 56], approximation theory [57], and optimization [58, 59]. Such theoretical results provide explanatory insight insofar as they necessitate or make plausible what can occur. This explanatory role is familiar throughout mathematics and statistics. The central limit theorem, for instance, explains why the normal distribution frequently arises when sampling data. Similarly, in machine learning, PAC learning bounds [56], while loose, help explain why neural networks often generalize better as they are trained on more data. Likewise, convergence results for gradient-based optimization algorithms in simple settings motivate their application to more complex neural network objectives, while results showing that sufficiently large neural networks can approximate broad classes of functions [57] shed light on how neural networks are capable of learning highly complex tasks. Beyond the core of deep learning, many influential approaches, including reinforcement learning and generative modeling, are likewise grounded in substantial mathematical theory. Such results contribute to epistemic rigor by providing principled accounts of why particular methods succeed and by clarifying the conditions under which they behave as expected.

Nevertheless, many of the most distinctive features of deep learning remain only partially understood. In fact, what has made deep learning mysterious and a rich source of surprises is its departure from classical machine learning. Important examples include the tendency of larger neural networks to generalize better and the success of gradient-based optimization for highly nonconvex loss functions, both of which contradict classical intuitions. In particular, the classical bias-variance tradeoff suggests that larger, more complex models should generalize more poorly than simpler ones, while simple nonconvex optimization problems suggest that gradient-based methods should become trapped in local minima. However, ongoing work on phenomena such as double descent [60], implicit regularization [61], and the effects of overparameterization on the loss landscape [62, 63] has provided important insights for why deep learning is able to overcome these anticipated obstacles. Extending such analyses so that they yield practical guidance in realistic experimental settings would represent a major advance in the explanatory maturity of deep learning.

A further explanatory ambition is the identification of fine-grained causal mechanisms. This would involve mapping internal neural network components and their interactions to the behaviors they implement, an approach commonly known as mechanistic interpretability. Ideally, such decompositions would provide functional explanations of how neural networks transform inputs into outputs. Understanding these internal mechanisms would also enable a range of practical interventions, including updating knowledge in models [64], steering models toward desirable behavior [65], and diagnosing failure modes such as hallucinations [66] and coding errors [67].

Yet, as desirable as explanation and interpretability may be, achieving them remains a significant challenge. Both are nuanced concepts with competing and sometimes conflicting formulations [68, 69, 70]. Moreover, questions about what qualifies as an explanation, including whether explanations must be unique, together with the conceptual difficulty of phenomena such as deception and lying, have drawn attention to the philosophical foundations of interpretability [71, 72]. While mathematical approaches to causality could in principle provide a rigorous explanatory framework [73], their reliance on abstractions such as directed acyclic graphs makes them difficult to apply directly to the low-level mechanisms of neural networks. In practice, the complexity of deep learning systems leads to substantial underdetermination: multiple plausible explanations may account for the same behavior, many of them proposed post hoc rather than uniquely specified and validated in advance.

Finally, the role of explanation in AI raises deeper questions that remain unresolved. In traditional scientific domains, explanations are often expected to be both accurate and intelligible, enabling humans to form a coherent understanding of the underlying mechanisms. However, modern AI systems challenge this expectation. Neural networks implement highly complex, distributed computations for

which no succinct or human-comprehensible description may exist. If so, it is unclear whether explanation, as traditionally conceived, can remain a central requirement of epistemic rigor, or whether new standards must be developed that relax demands for interpretability. Moreover, artificial systems are not readily amenable to explanation through theory of mind or high-level behavioral abstractions. This stands in contrast to human behavior, which is routinely explained in terms of beliefs, desires, and intentions, even in the absence of detailed neural understanding. The result is a significant explanatory gap between low-level mechanisms and abstract behavioral descriptions of neural networks. Addressing these issues, including what counts as an explanation and at what level of abstraction, will be essential for developing principled frameworks for interpreting increasingly complex AI systems.

Epistemic rigor ultimately concerns the conditions under which empirical findings develop into reliable scientific knowledge. In AI, this process remains underdeveloped. While deep learning has achieved remarkable practical successes, its reproducibility, predictability, and explainability each remain limited in important ways. Strengthening epistemic rigor will therefore be necessary for transforming AI from a discipline driven primarily by empirical performance into one grounded in principled understanding.

3 Operational Rigor

Notwithstanding the above challenges, the practical success of AI systems ultimately depends on whether they function reliably and effectively, thereby necessitating operational rigor. Operational rigor concerns the ability to build and deploy systems that meet defined performance standards or outcome criteria. It consists of sound methodology: thorough testing and validation, systematic monitoring of performance and failure modes, and the implementation of robust and scalable designs under real-world conditions. Performance metrics play a central role, since what is measured can be controlled, compared, and improved across deployment contexts.

Operational rigor is not unique to AI. In established engineering fields such as aeronautics, it ensures that systems behave in accordance with well-understood theoretical constraints. In other domains, such as medicine, it supports reliable outcomes despite partial understanding of underlying mechanisms. What distinguishes AI is the role operational rigor plays in enabling the use of systems whose behavior and failure modes are often neither constrained nor predictable from existing theory. In this setting, operational rigor is not derived from prior understanding; it substitutes for it. In AI, operational rigor is most prominently realized through two complementary mechanisms: benchmarks, which evaluate performance, and reliability and safety procedures that guide systems toward intended behavior.

3.1 Benchmarks

A benchmark represents a standardized suite for performance evaluation that consists of curated datasets, specific tasks, and predefined success metrics. Quantifying a system's performance through such metrics provides rigor in evaluation that eliminates the ambiguity of human judgment. This contrasts with earlier evaluation proposals such as the Turing Test [2], whose outcomes depend heavily on subjective interpretation.

The value of benchmarks lies in their role as proxies for real-world performance. Because it is infeasible to evaluate AI systems across every possible input or deployment condition, benchmarks approximate this impossible task through a sufficiently representative set of examples. Crucially, a benchmark must capture enough of the relevant structure of deployment settings to reliably predict how a system will perform in practice. Such curation is a form of operational rigor that compensates for limited epistemic guarantees of system behavior outside the benchmark setting. But benchmarks do more than measure domain-specific generalization; they are often treated as evidence of broader capabilities. For instance, influential benchmarks such as ImageNet [74] are not primarily valuable because of their specific classification task, since real-world applications extend far beyond the categories considered in such datasets. Rather, such benchmarks function as epistemic proxies: strong performance is taken as evidence that a model is capable of generalizing across diverse visual environments and tasks. Operational rigor therefore involves ensuring that such proxies are well-calibrated, meaning that benchmarks track success beyond their original evaluation settings.

Benchmarks also play a central role in operational rigor because they transform performance into an explicit optimization target. Methods can be directly compared according to standardized metrics, allowing techniques that improve performance to be refined and scaled while less effective ones are discarded. In this way, benchmark-defined objectives often become achievable through sustained optimization.

Nevertheless, benchmarking is not without its problems. For strong benchmark performance to be indicative of broader generalization, two conditions must hold: models must not have been trained on benchmark data, and they must learn the intended concepts underlying the benchmark. The first condition has become increasingly difficult to satisfy in an era in which large AI systems are trained on massive datasets. Because it is often unclear whether training data substantially overlaps with benchmark data, benchmark scores can be difficult to regard as reliable. Second, neural networks can fail to learn the underlying task through "shortcut learning," whereby they latch onto spurious correlations. A well-known study shows that image classifiers are much more likely to recognize cows when the background contains grass rather than a beach [75, 76]. More generally, model performance can degrade sharply under mild perturbations or distribution shifts, showing that high benchmark scores do not necessarily imply robust generalization [77, 78]. Another limitation follows from Goodhart's law: once benchmark metrics become optimization targets, they can themselves be exploited. A well-known example is sycophancy [79], in which chat-based LLMs prioritize agreement with users over truthfulness. A further limitation is that benchmark rankings are fragile with respect to benchmark selection itself. Specifically, the "benchmark lottery" hypothesis [80] argues that small changes in the choice of benchmark tasks can substantially alter the relative ranking of methods, implying that perceived algorithmic superiority may be an artifact of how benchmarks are constructed.

As a result, designing effective and meaningful benchmarks remains one of the most consequential challenges in AI research. Expanding the number and diversity of benchmarks is necessary for broadening the range of capabilities that can be evaluated, while ensuring that benchmarks faithfully capture real-world performance remains an enduring challenge. The latter becomes especially difficult as desired tasks grow increasingly complex and open-ended. For instance, current limitations such as the inability of AI systems to perform long-horizon continual learning[11] may persist in part due to the absence of strong benchmarks for evaluating this capability. In this regard, benchmark design cannot be separated from the other forms of rigor discussed previously: conceptual rigor is required to determine how capabilities should be defined and scoped, while epistemic rigor is needed to understand existing failure modes and whether proposed benchmarks are appropriate for addressing them.

3.2 Reliability and Safety

Operational rigor is needed to ensure that AI systems are reliable and adhere to principles such as helpfulness, honesty, and safety. Indeed, without explicit efforts to instill such properties, they can easily behave in harmful or undesirable ways [81, 82]. In developing AI to become a widespread technology, researchers and engineers have succeeded in making these systems more controllable and useful in a variety of ways:

One broad class of methods augments LLMs with additional tools or computation [83]. When LLMs can invoke external tools, they are no longer forced to solve every problem internally; instead, they can delegate subtasks to specialized systems. In doing so, LLMs inherit the reliability of their suite of tools, transferring the burden of correctness to proper tool use. Modern LLMs are thus no longer isolated chatbots, but now function as natural-language interfaces to complex systems. Operational rigor becomes less a matter of the performance of an isolated model and more a matter of system orchestration: deciding which component acts and when.

Operational rigor also includes eliciting latent capabilities already present within models. Techniques such as prompt engineering, extended computation, and self-critique can substantially improve reliability and task performance without modifying model parameters [84, 85]. In addition, detailed system prompts are often prepended during deployment to provide persistent behavioral guidance across user interactions. Although such methods do not provide guarantees of correctness, they offer a measure of control over behavior despite limited understanding of how models function.

A complementary approach involves directly shaping model behavior through post-training. After the initial pre-training stage in which LLMs are trained to predict the next token, they undergo

[11] Continual learning involves adapting to new data over time without requiring expensive retraining.

additional post-training via instruction fine-tuning and reinforcement learning from human feedback [86, 87, 88]. These later stages refine pretrained models into systems capable of reliably following user requests while reinforcing behaviors preferred by human evaluators. Here, operational rigor depends heavily on careful dataset and reward design: diverse task distributions and finely specified human preference criteria shape how models balance competing objectives such as harmlessness and deference to user intent.

Yet, despite ongoing efforts to strengthen current AI systems, they still suffer from many weaknesses. LLMs often hallucinate and fail to follow instructions properly, sometimes quite spectacularly [89, 90]. Their vulnerability to adversarial attacks and jailbreaks makes them susceptible to misuse [91, 92, 93]. Researchers have also shown that training data can be easily poisoned, making it possible to embed malicious behavior inside models [94, 95]. Furthermore, open-weight models make the situation even more fragile, since nothing inherently prevents users from modifying these models to override preexisting safety mechanisms. These limitations reveal a central difficulty for operational rigor: current methods for controlling model behavior have not scaled at the same pace as model capability itself. As a result, ensuring robust and reliable behavior under realistic conditions remains an unresolved challenge.

One potential solution is the use of formal verification methods, which aim to provide mathematical guarantees that systems satisfy specified constraints, thereby reducing reliance on empirical evaluation alone. Existing work includes certifying robustness to bounded perturbations, proving safety conditions for neural-network-controlled systems, and validating model-generated outputs using formal verifiers or proof assistants [96, 97, 98, 99]. However, these methods remain difficult to scale to modern architectures, or else are generally limited to settings in which correctness conditions can be precisely specified.

Operational rigor in AI departs in important respects from traditional philosophical accounts of scientific rigor. In much of the philosophy of science, empirical evaluation primarily serves epistemic aims such as establishing causal relationships, supporting explanations, or refining theories. In contemporary AI, however, evaluation increasingly functions not only as a means of assessing systems, but also as a primary mechanism for improving them. This does not eliminate epistemic concerns, but it changes the relationship between performance and scientific understanding in ways that distinguish AI from many other fields of science and engineering.

4 Rigor and Progress

The trajectory of scientific and technological progress depends not only on what modes of rigor are present within a field, but also on how they interact and develop over time. In physics, advances such as spaceflight and semiconductors emerged from conceptual frameworks capable of supporting strong prediction and reliable engineering. In biology, progress has depended on connecting explanations across molecular, cellular, and evolutionary scales, allowing phenomena such as heredity and natural selection to be understood even when living systems remain difficult to predict and control.

In AI, by contrast, the relationship between rigor and progress has been striking. AI has advanced rapidly as a technology despite limited conceptual clarity or theoretical grounding. This development may be understood through the shifting roles played by conceptual, epistemic, and operational rigor. We now examine how these roles have shaped AI across paradigms and patterns of progress, as well as how they may continue to shape the field's future direction.

4.1 The Distinctive Structure of AI Progress

Progress in modern AI has not emerged from the uniform development of the three forms of rigor. Instead, operational rigor has come to assume a dominant role in guiding progress while conceptual clarification and scientific understanding have often lagged behind. This asymmetry reflects deeper structural features of the field itself.

One such feature is that AI exhibits a tight feedback loop between evaluation and development: the same metrics used to assess progress are themselves targets for further optimization. Operational rigor therefore functions not merely as a way of evaluating systems after they have been built, but also as a mechanism through which they are improved. This configuration has no clear analogue in many other fields. In medicine, practical and ethical constraints prevent benchmark-driven iteration,

while in physics and traditional engineering, manual design guided by theory drives improvement. By contrast, AI systems are trained to optimize metrics directly, enabling increasingly strong performance even in the absence of a mature scientific theory.

A second structural feature is that AI produces the very artifacts it studies. Models are not merely instruments for investigating intelligence; they are themselves often regarded as new instances of intelligent behavior. Each generation of systems introduces new capabilities, limitations, and failure modes, thereby expanding the domain that conceptual and epistemic inquiry must explain. This differs from the natural sciences, where the principal objects of inquiry, such as atoms, organisms, and planets, exist largely independently of the field itself. In AI, however, the objects of study are endogenous to the field's own methods of development.

Together, these structural features help explain why progress in AI can be both rapid and uneven. Benchmark optimization makes it possible to develop methods that generate measurable improvements without first possessing a substantive theory of the systems being improved. At the same time, the continual production of new artifacts ensures that conceptual clarification and explanatory frameworks are always responding to a moving target. The result is a form of progress in which capabilities advance faster than the conceptual and scientific frameworks needed to fully understand them.

4.2 Rigor Across Paradigms

The predominance of operational over epistemic rigor described above is associated with the current deep learning paradigm rather than with AI as a whole. Throughout the history of AI, different paradigms have embodied varying balances between the forms of rigor, reflecting changing assumptions about how intelligence should be modeled and realized. The current imbalance between operational and epistemic rigor is therefore historically contingent rather than inevitable, and may not persist in future paradigms.

This variability can already be seen in earlier approaches to AI. The symbolic approach represented a paradigm that viewed intelligence as the manipulation of explicit representations governed by formal rules [100]. This paradigm possessed strong epistemic rigor because systems were interpretable and their operations could be traced through with precision. Yet these systems performed poorly outside carefully constrained environments. Their brittleness and inability to scale limited their effectiveness, and their restricted domains of application left little scope for the kind of large-scale benchmark optimization characteristic of AI today.

Learning-based paradigms relaxed the requirement that intelligent behavior be specified explicitly in advance. Instead, systems learn statistical structure from data. This paradigm developed along two distinct but related lines: classical statistical learning and connectionism [55, 101]. The classical statistical learning paradigm developed a strong degree of epistemic rigor: progress was closely tied to developing mathematically tractable models, including linear classifiers, kernel methods, and decision trees, alongside the development of statistical learning theory providing generalization guarantees [56]. Because model classes remained comparatively tractable, theoretical analysis continued to play a central role in guiding progress. Moreover, simply scaling model size did not reliably produce improved generalization in this setting, as increasingly complex models were prone to overfitting.

By contrast, the connectionist paradigm departed significantly. This approach adopted a comparatively epistemically modest stance: it relied less on manually specified structure and more on learning procedures capable of discovering useful representations directly. This relaxation of epistemic constraints would prove extraordinarily consequential. While early connectionist systems achieved mixed success, the paradigm eventually gave rise to deep learning, whose large-scale neural networks proved capable of dramatic empirical improvements. Crucial to the success of this approach were technological developments that enabled optimization to scale effectively, particularly advances in computational hardware and the availability of large datasets. As a consequence, deep learning systems achieved unprecedented performance despite limited theoretical understanding of why they worked so well.

These foundational paradigms, consisting of symbolic AI, classical statistical learning, and deep learning, illustrate the dynamic relationship between our three forms of rigor. Conceptual rigor shapes the paradigms through which intelligence is modeled and realized, influencing how epistemic and operational rigor interact within a given approach. More recent developments suggest how this interplay may continue to evolve. One example is the growing incorporation of neuro-symbolic elements into modern AI systems, which attempt to recover some degree of predictability and explainability by combining learned systems with external tools, thus partially restoring epistemic rigor. At the same time,

the rise of agentic systems places increasing emphasis on long-horizon planning, autonomous action, and open-ended interaction. Such capabilities increase performance demands while simultaneously introducing new epistemic difficulties: the opacity of neural network internals is now compounded by the complexity of agentic behavior. Operational rigor is likely to become even more dominant relative to epistemic rigor in this agentic setting, absent major theoretical breakthroughs. Yet the evolution of AI suggests that future paradigms may again alter the balance between operational and epistemic rigor.

4.3 Future Progress

Within the current deep learning paradigm, future progress is likely to depend increasingly on strengthening areas of rigor that remain comparatively underdeveloped. Two prominent examples are the pursuit of increasingly general intelligence and the problem of alignment, each of which places different demands on rigor. Progress toward general intelligence currently favors operational rigor, whereas alignment depends more heavily on conceptual and epistemic rigor.

Artificial general intelligence (AGI) has long served as an aspirational goal: a system capable of performing the full range of cognitive tasks associated with human intelligence. Yet, like intelligence itself, the term AGI remains overloaded. It has been associated with ideas ranging from economic and social upheaval [102] to eschatological singularity narratives [103] and science-fiction scenarios [104]. The wide range of meanings associated with AGI[12] is not problematic per se, since many important concepts resist precise definition. But AGI, beyond being a research program, is increasingly discussed in public and policy contexts [108, 102], including by CEOs of major industrial AI labs, and therefore merits clarification whenever it is invoked.

The conceptual ambiguity of AGI implies that a critical bottleneck for progress lies in operationalizing the definition of AGI. Such an operational definition supports both conceptual and operational rigor because it clarifies the concept of AGI by grounding it in performance measures that can be objectively assessed. Recent proposals include frameworks for defining and evaluating AGI [109, 110], as well as taxonomic approaches such as [111]. The creation of benchmarks grounded in operational definitions of AGI may enable progress to occur even when epistemic rigor remains underdeveloped and a comprehensive scientific theory of intelligence is lacking. This possibility is reflected in a prominent perspective within contemporary AI research: that scaling compute may be largely sufficient for increasingly general capabilities to emerge [112]. Closely related views argue that intelligence itself may largely be captured through optimization against suitably defined reward signals [113]. In such cases, developing AGI may depend less on possessing a rigorous scientific theory than on arranging the conditions for it to arise. Operational rigor therefore emerges as a central mechanism through which AGI progress may occur.

The case of alignment, however, provides a striking reversal of the priority of operational rigor that we have seen thus far. Here, alignment refers to the design of systems that are aligned with the values of their designers or, from a humanist perspective, that serve and benefit humanity [114, 115]. In the limit of highly capable agentic systems, alignment depends critically on conceptual and epistemic rigor. Unlike intelligence and AGI, which concern what systems can do, alignment additionally concerns what they must not do. As a result, determining which behaviors must be avoided depends on conceptual rigor for selecting an appropriate normative framework. In this regard, the trolley dilemmas of moral philosophy may become matters of practical concern when aligning AI systems.

Preventing powerful AI from directly causing or enabling catastrophic harm (e.g. designing biological weapons or destroying critical infrastructure) requires more than benchmark performance [116]. A satisfactory solution involves understanding whether AI systems adopt intended objectives beyond their training conditions [117, 118] and, if not, developing principled accounts of the resulting failure modes. Moreover, alignment based primarily on benchmark performance is insufficient in a world where increasingly capable systems become widely accessible. As the cost of developing powerful models declines, malicious actors may optimize systems toward harmful objectives just as readily as benign actors optimize them toward beneficial ones. For this reason, alignment ultimately requires more than producing systems that behave appropriately in isolated settings. It requires understanding AI at a level that allows misaligned objectives to be anticipated and mitigated.

[12] Some researchers doubt whether AGI names a coherent concept at all [105, 106, 107].

5 Conclusion

We examined rigor in AI through the differing roles it plays, focusing on the ambiguity, uncertainty, and tensions that arise within the field. Our contribution was to divide such problems along conceptual, epistemic, and operational dimensions, and to clarify the role that rigor plays in each. While a scientific or engineering approach to these issues would focus on their solutions, our philosophical analysis mapped the overarching structure, interrelationships, and broader forms their solutions might take. This bird's-eye view is valuable for understanding a multidisciplinary field such as AI because the latter is more than just the union of its constituent problems; it also includes the interactions among them and how they co-evolve.

In this regard, the framework developed here helps explain why modern AI has achieved technological capabilities that outpace scientific understanding, while situating this pattern within the broader evolution of conceptual paradigms. Moreover, our analysis not only identified promising avenues for future progress, but also provides a basis for reasoning about what AI as a mature field might look like, since rigor is a prerequisite for maturity. While the exact trajectory is difficult to predict, our three-part framework offers insights into several possible requirements:

Conceptual rigor demands that critical concepts shaping the design and evaluation of systems, such as AGI and alignment, be articulated in ways that enable reliable assessment and coordination. As AI diffuses throughout society and assumes greater geopolitical significance, a more refined terminology for setting expectations and calibrating the capabilities of such systems will become essential. Epistemic rigor entails that a well-developed science of AI would exhibit far greater predictability and explainability than it does currently. A more comprehensive body of laws and theoretical results would tightly constrain experimental design; advances in model interpretability would enable actionable interventions across a wide range of systems; and causal explanations would integrate relevant algorithmic and architectural components at the appropriate scales. Operational rigor, in turn, would ideally yield systems so reliable and robust that many of today's failure modes, such as hallucinations and jailbreaks, would become sufficiently rare that they are no longer expected to occur. Additionally, it would enable a technological landscape in which defensive capabilities consistently exceed those of potential adversaries, rendering attacks either infeasible or limited in their capacity to cause harm.

Meeting such challenges will require advancements across many disciplinary boundaries, each demanding the appropriate role for rigor. Developing and integrating the forms of rigor examined here will be necessary, though far from straightforward. Yet doing so will be essential for AI to become a mature science and reliable technology.

References

[1] Alexandra Olteanu, Su Lin Blodgett, Agathe Balayn, Angelina Wang, Fernando Diaz, Flavio Calmon, Margaret Mitchell, Michael Ekstrand, Reuben Binns, and Solon Barocas. Rigor in AI: Doing rigorous AI work requires a broader, responsible AI-informed conception of rigor. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025.

[2] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[3] John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. A proposal for the Dartmouth summer research project on Artificial Intelligence, 1955. Proposal for the 1956 Dartmouth workshop. URL: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

[4] Luciano Floridi and Anna C. Nobre. Anthropomorphising machines and computerising minds: The crosswiring of languages between artificial intelligence and brain & cognitive sciences. *Centre for Digital Ethics (CEDE) Research Paper*, 2024. Available at SSRN: https://ssrn.com/abstract=4738331.

[5] Pei Wang. *Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence*. Ph.d. thesis, Indiana University, Bloomington, IN, USA, 1995.

[6] Shane Legg and Marcus Hutter. A collection of definitions of intelligence. In *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, pages 17–24, Amsterdam, The Netherlands, 2007. IOS Press.

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

[8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4, 2023.

[9] 60 Minutes. "Godfather of AI": Geoffrey Hinton – the 60 minutes interview, Oct 2023. YouTube video, 13:12. URL: https://www.youtube.com/watch?v=qrvK_KuIeJk.

[10] The Royal Institution. Will AI outsmart human intelligence? - with "Godfather of AI" Geoffrey Hinton, Jul 2025. YouTube video, 47:15. URL: https://www.youtube.com/watch?v=IkdziSLYzHw.

[11] World Governments Summit. A conversation with Yann LeCun AI: Lifeline or landmine?, Feb 2024. YouTube video, 24:51. URL: https://www.youtube.com/watch?v=rf9jgZYAni8.

[12] David C. Krakauer, John W. Krakauer, and Melanie Mitchell. Large language models and emergence: A complex systems perspective, 2025.

[13] Melanie Mitchell and David C. Krakauer. The Debate over Understanding in AI's Large Language Models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, March 2023.

[14] Fabrizio Dell'Acqua, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School, 2023. September 2023.

[15] Andrej Karpathy. Jagged Intelligence, 2024. Tweet on X (formerly Twitter), July 2024. URL: https://x.com/karpathy/status/1816531576228053133.

[16] OpenAI Community. Why 9.11 is larger than 9.9. . . . . incredible. Online forum post, Jul 2024. Discussion on comparing numeric values; community.openai.com thread 869824. URL: https://community.openai.com/t/why-9-11-is-larger-than-9-9-incredible/869824.

[17] Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". In *The Twelfth International Conference on Learning Representations*, 2024.

[18] Gilbert Ryle. *The Concept of Mind*. Hutchinson, London, 1949.

[19] Jerry A. Fodor. *The Language of Thought*. Harvard University Press, 1975.

[20] Hilary Putnam. Psychological predicates. In W. H. Capitan and D. D. Merrill, editors, *Art, Mind, and Religion*, pages 37–48. University of Pittsburgh Press, 1967.

[21] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346, 1990.

[22] Yoshua Bengio. From System 1 Deep Learning to System 2 Deep Learning. NeurIPS 2019 keynote talk on YouTube, 2019. Accessed 2026-01-29. URL: https://www.youtube.com/watch?v=FtUbMG3rlFs.

[23] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations (ICLR) 2023*, 2023. Also known as "Othello-GPT" study on transformer emergent representations.

[24] Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. DeepMind Blog, Aug 2025. Available at: https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

[25] Yann LeCun. A Path Towards Autonomous Machine Intelligence. Working paper, OpenReview, 2022. Version 0.9.2, June 27, 2022.

[26] Charlotte Alter. Meta's AI chief Yann LeCun on AGI, open-source, and AI risk. *TIME*, February 2024. Interview with Yann LeCun.

[27] Steven T. Piantadosi and Felix Hill. Meaning Without Reference in Large Language Models. *nCSI Working Paper*, 2022. Accessed 2026-05-07.

[28] Qianwen Xu, Yixuan Peng, Samuel A. Nastase, Martin Chodorow, Mengru Wu, and Ping Li. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature Human Behaviour*, 9(9):1871–1886, 2025.

[29] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics.

[30] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.

[31] Ben Cottier. Trends in the dollar training cost of machine learning systems. Epoch AI Blog, 2023. Published January 31, 2023; accessed 2026-01-29. URL: https://epoch.ai/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems.

[32] Josh You and David Owen. How much power will frontier AI training demand in 2030? Epoch AI Blog, 2025. Published August 11, 2025; accessed 2026-01-29. URL: https://epoch.ai/blog/power-demands-of-frontier-ai-training.

[33] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. The AI index 2023 annual report. Report, Institute for Human-Centered Artificial Intelligence (HAI), Stanford University, 2023. Includes detailed chapter on Research and Development trends; accessed 2026-01-29. URL: https://hai.stanford.edu/ai-index/2023-ai-index-report.

[34] Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 725–734. PMLR, 09–15 Jun 2019.

[35] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI'18, pages 3207–3214. AAAI Press, 2018.

[36] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

[37] Liam Li and Ameet Talwalkar. Random search and reproducibility for neural architecture search. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

[38] Moritz Herrmann, F. Julian D. Lange, Katharina Eggensperger, Giuseppe Casalicchio, Marcel Wever, Matthias Feurer, David Rügamer, and volume = 235 pages = 18228–18247 year = 2024 series = Proceedings of Machine Learning Research publisher = PMLR Eyke Hüllermeier title = Position: Why We Must Rethink Empirical Research in Machine Learning, journal = Proceedings of the 41st International Conference on Machine Learning.

[39] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[40] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[41] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021.

[42] Timothy Nguyen. Understanding transformers via n-gram statistics. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. Curran Associates, Inc., 2024.

[43] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8580–8589, Red Hook, NY, USA, 2018. Curran Associates, Inc.

[44] Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019. NeurIPS 2019; accessed 2026-01-29.

[45] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of Machine Learning Research*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464. PMLR, 2019.

[46] Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

[47] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017.

[48] Sara Hooker. On the slow death of scaling. SSRN Scholarly Paper ID 5877662. Available at SSRN: https://ssrn.com/abstract=5877662, December 2025.

[49] Jamie Simon, Daniel Kunin, Alexander Atanasov, Enric Boix-Adserà, Blake Bordelon, Jeremy Cohen, Nikhil Ghosh, Florentin Guth, Arthur Jacot, Mason Kamb, Dhruva Karkada, Eric J. Michaud, Berkan Ottlik, and Joseph Turnbull. There will be a scientific theory of deep learning. *arXiv preprint arXiv:2604.21691*, 2026.

[50] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[51] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In A. Smola, A. Dimakis, and I. Stoica, editors, *Proceedings of Machine Learning and Systems*, volume 3, pages 747–769, 2021.

[52] Matthew Hutson. AI researchers allege that machine learning is alchemy. News article, *Science*, May 3 2018. Accessed: 2025-01-05. URL: https://www.science.org/content/article/ai-researchers-allege-machine-learning-alchemy.

[53] Sanjeev Arora. Brief introduction to deep learning and the "alchemy" controversy. Video, YouTube, 2019. Presented at Deep Learning: Alchemy or Science?, Institute for Advanced Study. URL: https://www.youtube.com/watch?v=kqhg-o-KEns.

[54] J. Zico Kolter. Is this really science? a lukewarm defense of alchemy. In *Workshop on Scientific Methods for Understanding Neural Networks, NeurIPS 2024*, 2024.

[55] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2 edition, 2000.

[56] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[57] Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[58] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Science & Business Media, New York, second edition, 2006.

[59] Elad Hazan. Introduction to online convex optimization, 2023.

[60] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the bias–variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.*, 116(32):15849–15854, 2019. PMCID: PMC6689936, PMID: 31341078.

[61] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data, 2017.

[62] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis, 2017.

[63] Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks, 2018.

[64] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):59:1–59:35, 2024.

[65] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *Advances in Neural Information Processing Systems 37*, pages 83345–83373, Vancouver, Canada, 2024. Neural Information Processing Systems Foundation, Inc.

[66] Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, and Aidong Zhang. Toward faithful retrieval-augmented generation with sparse autoencoders, 2025.

[67] Tuan-Dung Bui, Thanh Trong Vu, Thu-Trang Nguyen, Son Nguyen, and Hieu Dinh Vo. Correctness assessment of code generated by large language models using internal representations. *Journal of Systems and Software*, 230:112570, December 2025.

[68] Zachary C. Lipton. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Poorly Defined. *Commun. ACM*, 61(10):36–43, September 2018.

[69] Sarthak Jain and Byron C. Wallace. Attention Is Not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3543–3556, 2019.

[70] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics.

[71] Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *The Thirteenth International Conference on Learning Representations*, 2025.

[72] Iwan Williams, Ninell Oldenburg, Ruchira Dhar, Joshua Hatherley, Constanza Fierro, Nina Rajcic, Sandrine R. Schiller, Filippos Stamatiou, and Anders Søgaard. Mechanistic interpretability needs philosophy. *arXiv:2506.18852 [cs.CL]*, 2025. Preprint; accessed 2026-01-29.

[73] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

[74] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[75] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, Munich, Germany, 2018. Springer.

[76] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[77] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[78] Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[79] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics.

[80] Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery, 2021.

[81] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. 2020.

[82] Imane El Atillah. Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change, March 2023. Accessed: 2026-03-04. URL: https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-.

[83] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 68539–68551. Curran Associates, Inc., 2023.

[84] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[85] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Shashank Srivastava, Yiming Yang, and Hannaneh Hajishirzi. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

[86] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022.

[87] Paul Christiano et al. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

[88] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Misha Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 27730–27744, 2022. NeurIPS 2022; accessed 2026-01-29.

[89] BBC News. Google AI search says to glue pizza and eat rocks, May 2024. Accessed 2026-01-31. URL: https://www.bbc.co.uk/news/articles/cd11gzejgz4o.

[90] Alistair Barr. Replit CEO apologizes after AI coding tool wipes company database, July 2025. Business Insider article on a reported incident where an AI coding agent deleted a production database; accessed 2026-01-31. URL: https://www.businessinsider.com/replit-ceo-apologizes-ai-coding-tool-delete-company-database-2025-7.

[91] Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7432–7449, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[92] Wenrui Xu and Keshab K. Parhi. A survey of attacks on large language models, 2025.

[93] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision–language models: Resources, advances, and future trends. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19525–19545, 2025.

[94] Miguel A. Ramirez, Song-Kyoo Kim, Hussam M. N. Al Hamadi, Ernesto Damiani, Young-Ji Byon, Tae-Yeon Kim, Chung-Suk Cho, and Chan Yeob Yeun. Poisoning attacks and defenses on artificial intelligence: A survey. *CoRR*, abs/2202.10276, 2022. arXiv preprint.

[95] Pinlong Zhao, Weiyao Zhu, Pengfei Jiao, Di Gao, and Ou Wu. Data poisoning in deep learning: A survey. *CoRR*, abs/2503.22759, 2025. arXiv preprint.

[96] Aws Albarghouthi. Introduction to neural network verification. *arXiv preprint arXiv:2109.10317*, 2021.

[97] Weiming Xiang, Patrick Musau, Nathan Hamilton, Xiaodong Yang, Jaemin Rosenfeld, and Taylor T. Johnson. Verification for machine learning, autonomy, and neural networks survey. *arXiv preprint arXiv:1810.01989*, 2018.

[98] Radoslav Ivanov, James Weimer, Rajeev Alur, George J. Pappas, and Insup Lee. Verisig: Verifying safety properties of hybrid systems with neural network controllers. *ACM Transactions on Embedded Computing Systems*, 18(5s):1–19, 2019.

[99] Thomas Hubert, Edward Williams, Jiangjie Chen, Wenxiang Chen, Jiacheng Du, Thomas Hanwen Zhu, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, Nov 2025.

[100] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, Hoboken, NJ, 4 edition, 2020.

[101] David E. Rumelhart and James L. McClelland, editors. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. MIT Press, Cambridge, MA, 1986.

[102] Anton Korinek. Economic policy challenges for the age of AI. NBER Working Paper 32980, National Bureau of Economic Research, September 2024.

[103] Ray Kurzweil. *The Singularity Is Nearer*. Viking, New York, NY, USA, 2024.

[104] Stephen Cave and Kanta Dihal. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1(2):74–78, 2019.

[105] Alison Gopnik. The myth of the machine. *Scientific American*, 2020.

[106] The Information Bottleneck. EP20: Yann LeCun, December 15 2025. YouTube video, 01:50:07. URL: https://www.youtube.com/watch?v=7u-DXVADyhc.

[107] Dario Amodei and Alex Kantrowitz. Anthropic CEO Dario Amodei: AI's potential, OpenAI rivalry, GenAI business, doomerism. YouTube Video, 2025. Interview published July 31, 2025. Accessed 2026-01-31. URL: https://www.youtube.com/watch?v=mYDSSRS-B5U.

[108] Blaise Agüera y Arcas and Peter Norvig. Artificial General Intelligence Is Already Here, 2023. Published October 10, 2023; Accessed: 2026-01-29. URL: https://www.noemamag.com/artificial-general-intelligence-is-already-here/.

[109] Dan Hendrycks, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li, Andy Zou, Lionel Levine, Bo Han, Jie Fu, Ziwei Liu, Jinwoo Shin, Kimin Lee, Mantas Mazeika, Long Phan, George Ingebretsen, Adam Khoja, Cihang Xie, Olawale Salaudeen, Matthias Hein, Kevin Zhao, Alexander Pan, David Duvenaud, Bo Li, Steve Omohundro, Gabriel Alfour, Max Tegmark, Kevin McGrew, Gary Marcus, Jaan Tallinn, Eric Schmidt, and Yoshua Bengio. A definition of AGI, 2025.

[110] Ryan Burnell, Yumeya Yamamori, Orhan Firat, Kate Olszewska, Steph Hughes-Fitt, Oran Kelly, Isaac R. Galatzer-Levy, Meredith Ringel Morris, Allan Dafoe, Alison M. Snyder, Noah D. Goodman, Matthew Botvinick, and Shane Legg. Measuring progress toward AGI: A cognitive framework. Technical report, Google DeepMind, March 2026. Technical report.

[111] Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Position: Levels of AGI for operationalizing progress on the path to AGI. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36308–36321. PMLR, 2024.

[112] Sam Altman. Three observations. Blog post on samaltman.com, February 2025. Discusses key trends in the economics and scaling of AI. URL: https://blog.samaltman.com/three-observations.

[113] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

[114] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, New York, 2019.

[115] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.

[116] Rohin Shah, Alex Irpan, Alexander Matt Turner, Anna Wang, Arthur Conmy, David Lindner, Jonah Brown-Cohen, Lewis Ho, Neel Nanda, Raluca Ada Popa, Rishub Jain, Rory Greig, Samuel Albanie, Scott Emmons, Sebastian Farquhar, Sébastien Krier, Senthooran Rajamanoharan, Sophie Bridgers, Tobi Ijitoye, Tom Everitt, Victoria Krakovna, Vikrant Varma, Vladimir Mikulik, Zachary Kenton, Dave Orr, Shane Legg, Noah Goodman, Allan Dafoe, Four Flynn, and Anca Dragan. An approach to technical AGI safety and security, 2025.

[117] Lauro Di Langosco, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12004–12019. PMLR, 2022.

[118] Evan Hubinger et al. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.