THE WHITE ELEPHANT IN THE ROOM – HISTORICAL ANALOGIES FOR AI DISCUSSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Research in the fields of Machine Learning (ML) and Artificial Intelligence (AI) has changed in the past years. Financial support, hardware, software, data, and scientific processes have improved substantially. These developments contributed to scientific advances across many disciplines and to broader societal effects. How these events will unfold is still unclear. But gaining a better understanding of this situation and responsible usage of AI technology likely requires an interdisciplinary discourse of informed citizens, scientists, and policy makers. We argue that one obstacle to this discourse is structural: many questions raised by AI are not well captured by the field's conventional evaluation apparatus. Benchmarks report performance on specified tasks and metrics, but broader questions about understanding, downstream effects, measurement validity, and investment require additional ways of reasoning. Since a comprehensive theory of how AI will impact our lives remains the subject of an often polarized debate, we here propose a complementary approach familiar to ML researchers: learning from examples. We explore four historical analogies: Alchemy, Plastic Industry, Free Will, and White Elephants. Each analogy foregrounds a different question that benchmarks cannot answer on their own: The Alchemy analogy asks whether we understand what we have built, the analogy on Plastic asks what it will cost us later, the Free Will example asks whether we measure what we think we measure, and the White Elephants analogy asks whether the investment is worth it. Because these narratives are comparatively accessible, they may help structure discussions across disciplinary and professional boundaries, including discussion of where each analogy holds and where it breaks. Ultimately, this may support a more differentiated discussion of possible trajectories for AI.

1 INTRODUCTION

Researchers in Artificial Intelligence (AI) have achieved and witnessed developments in the past decades that have had profound impact on science and society. Novel hardware, new software, better model architectures, unprecedented volumes of data, and ever improving scientific processes enabled AI researchers to increase scientific progress in their own field and other scientific disciplines. These developments culminated in the massive popularity of Large Language Models (LLMs), with implications on all levels of society.

While most ML researchers probably see the advantages of the spike in popularity and financial support, there are some aspects of these recent developments that other researchers perceive as concerning. More importantly, some of these concerns are likely not fully independent of economic interests. Some researchers even argue that the criticism of AI and its fast progress is itself part of AI companies' marketing strategy or at least diverging attention from more important problems around AI (Vinsel, 2021; Gebru et al., 2023; Kapoor & Narayanan, 2023). Independent of the intentions that different actors have in these developments, one thing is certain: Research and applications related to AI have changed. In science, education as well as many parts of society, these changes and their implications are poorly understood. This uncertainty leads to questions that can only be answered in an inclusive transdisciplinary effort. This text aims at contributing to such an informed discourse by reviving some perspectives helpful to think about AI and its implications.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To summarize the current situation, public opinion towards AI technology appears to have reached a state of polarization: Some rely heavily on that technology and invest time, money, and attention. Other parts of society have shifted recently to a more skeptical stance (Kennedy et al., 2025). Advocates of AI argue that scientific progress has been speeding up in terms of the half-life time of benchmarks and the funding in academic and industrial research. But AI skeptics on the other hand argue that the productivity increases with AI throughout the past years have remained below expectations (Becker et al., 2025) or are aiming for solutions to irrelevant problems and were thus marginalizing more important problems (Bender et al., 2021; Raji et al., 2021). But there are more: a third group is neither invested in nor arguing against the underlying technology, but has simply grown tired of its omnipresence, an affective reaction described in the literature as "AI fatigue", encompassing cognitive overload and emotional exhaustion from the sheer volume of AI in products, marketing and everyday conversation (Ragolane & Patel, 2025). Understanding these developments is crucial to argue about responsible usage of AI. But comprehensive unifying accounts of these developments have not emerged yet.

Instead of aiming for a comprehensive theoretical account on a technical, ethical, and sociological level, this text follows an approach that ML researchers are familiar with: Learning from examples, or analogies, of technological developments that are similar to the recent developments in AI. Not all of these analogies are helpful in all aspects. But some are similar enough to contribute to an informal but informed public discourse, at scientific conferences, personal conversations, blogs, and anywhere we need transdisciplinary exchange between scientists of different fields, civil society and policy makers. This view is inspired in part by the ideas put forward in Feyerabend (2018). Feyerabend argues that myths, analogies, narratives, and other non-formal modes of explanation need not be regarded simply as inferior precursors to scientific reasoning, but can contribute to how knowledge is communicated and interpreted.

Each analogy foregrounds a question that conventional benchmark-centered evaluation does not by itself settle. A benchmark compares systems on a fixed proxy metric, which is what makes it precise about what it measures and silent about everything else. There are important extensions of benchmarks aiming for a more holistic quantitative evaluation (Liang et al., 2023). But where a field has no measurement for a question, the argument gets carried by comparison to something else. We thus consider our approach an informal complement to more technical or theoretical accounts of holistic evaluation of AI systems (Weidinger et al., 2023). The goal is to foster broader discussions that ultimately could lead to novel quantitative measures of success or failure.

The sections follow a simple structure around four questions and analogies: The first analogy is due to Ali Rahimi and his test of time award speech at NeurIPS where he compared AI with Alchemy. The second analogy is comparing the AI developments with the rise of the plastic industry. The third analogy is on the neurophilosophical discussions on human free will. The last analogy is inspired by the work of Dirk van Laak on White Elephants (van Laak, 1999), which addresses some of the implications of overreliance on technology. We do not claim that these four analogies are the only suitable ones. We chose them because, in our view, together they cover the questions that recur most in the current debate. Other historical episodes could certainly serve a similar purpose, for instance the steam engine, nuclear power, the internet, or social media, and some of them resurface within our sections as supporting examples. Each of these sections is self-contained and does not require technical expertise. We hope that discussion around these examples fosters interdisciplinary exchange and contributes to a broader, less polarized discourse on recent developments in AI.

2 ALCHEMY

In December 2017, Ali Rahimi (then at Google) and Ben Recht (UC Berkeley) received the NeurIPS 2017 (then called NIPS) Test of Time Award for their 2007 paper Random Features for Large-Scale Kernel Machines (Rahimi & Recht, 2007). Rahimi used the acceptance speech not merely to celebrate the paper but to deliver a pointed critique of the entire field. Rahimi opened the critique by contrasting the common ML hype often referred to as "the new electricity" with his own alternative metaphor:

> "Machine learning has become alchemy."
>
> Rahimi & Recht (2017)

He elaborated on that analogy in stating when alchemy actually worked and when it did not work as well. On the plus side, it produced metallurgy, glass, dyes, and medicines. But on the other hand it also produced two thousand years of false theories (transmutation, immortality) that the scientific revolution had to dismantle. The concern was that ML produces positive results without a mechanistic understanding and that the field had stopped demanding such understanding. Rahimi gave examples of the weak understanding the ML community had about popular tools:

- Gradient descent on a trivial problem: A two-layer linear network optimizing a badly conditioned linear problem. A classical optimization algorithm like Levenberg–Marquardt solves it to machine precision in a few hundred iterations. This shows the optimizer itself is poorly understood on even trivial deep architectures.
- Batch Normalization (Ioffe & Szegedy, 2015), a stabilizing component in neural network training: Widely used, yet "we know almost nothing about it."
- Engineering fragility: A change in TensorFlow's rounding behavior caused an internal Adam parameter to become invalid and produced a dramatic error-rate jump; Rahimi and Recht later identified this case as a software bug rather than evidence about the optimizer itself (Rahimi & Recht, 2017).

In their addendum, Rahimi and Recht argued that superficial understanding is tolerable for low-risk applications, but many applications, such as health care, require more empirical rigor in the model development (Rahimi & Recht, 2017). The alchemy speech fueled a heated debate around the state of AI and scientific methods. Yann LeCun (then Director of Facebook AI Research) posted an immediate rebuttal on Facebook to the claim that AI was alchemy. One argument of LeCun is that of the erosion of trust in a technology that has repeatedly demonstrated its usefulness. The skeptical attitude Rahimi promoted (demanding theory before accepting empirical results) is

> "exactly this kind of attitude that lead [sic] the ML community to abandon neural nets for over 10 years, despite ample empirical evidence that they worked very well in many situations."
>
> LeCun (2017)

LeCun lived through the second AI winter (roughly 1995–2010), when neural networks were dismissed by the mainstream ML community as theoretically unfounded and for over a decade other methods such as kernel machines took over the main stages at conferences. He views this as a cautionary tale about premature demands for rigor. LeCun connects this idea to a historical anecdote: Empirical progress has repeatedly preceded formal theory in ML and other disciplines. The steam engine came before thermodynamics. Theory often follows invention; it does not always guide it. Both theory and empirical engineering are valuable, but dismissing an "incredibly successful" technology for lacking complete theory is counterproductive.

Investment, scale, and better tooling. Now, almost 10 years after these discussions, it seems that the problems Rahimi and Recht addressed have not been solved. But many more astounding achievements were made. And just like alchemy produced valuable novel tools as a byproduct, also AI research has brought us tooling, frameworks, data sets, and overall improved research conditions. This tension between empirical productivity and epistemic rigor is not merely a historical footnote: a 2024 position paper argues that most empirical ML research is still conducted and interpreted as if it were confirmatory science, when in practice it remains exploratory, a mismatch that continues to drive non-replicable results (Herrmann et al., 2024).

How much understanding is enough? The disagreement between Rahimi and LeCun leaves a practical question for AI researchers. How well do we need to understand a method before using it? Calling it alchemy gives us no answer, since useful empirical work can precede a theory. We can, however, improve the empirical rigor, as suggested by Rahimi. Better evaluation protocols are an essential part of this effort (Liang et al., 2023; Srivastava, 2023). How much evidence is enough depends on the application. Rahimi and Recht already made this distinction when discussing low-risk applications and health care.

Isaac Newton – the last magician. The founding father of modern physical theory himself wrote more on alchemy than on physics, which led John Maynard Keynes to call him "the last of the magicians" (Kuehn, 2012). The fact that Newton's main contribution on gravitational forces might

have been inspired by ideas around occult forces demonstrates how difficult it is to draw a clean historical boundary between alchemy and the science that followed.

Alchemy actually works. In hindsight the alchemy analogy took a surprising turn when physicists reported the production of fleeting gold nuclei from lead at the Large Hadron Collider (Acharya et al., 2025). This technical transmutation does not vindicate alchemical methods, but it illustrates how Rahimi's and LeCun's views can be reconciled: The experiment was possible only through decades of scientific exchange between empirical and theoretical work.

3 PLASTIC

When generative AI applications emerged, one analogy has become popular:

> "AI is the new plastic waste."
>
> cf. Furze (2023); Gewaltig (2024)

The metaphor was used in several blog articles around 2023 (Furze, 2023; Gewaltig, 2024)[1] and has since been formalized into a peer-reviewed framework for AI literacy in education (Roe et al., 2025). One idea here is that creators and artists were afraid of how cheap it would be to generate new content with AI. Others were concerned about the similarity with microplastics and how AI would be absorbed by our thinking.

Plastics are among the most consequential materials and comprise a large family of moldable polymers. These long-chain molecules are characterized by a structure that can be engineered for strength, flexibility, transparency, insulation, chemical resistance, or low cost. Their history is a progression from natural materials, to chemically modified natural polymers, to fully synthetic materials made largely from petrochemical feedstocks. A short summary of the steps from natural materials similar to plastics to modern forms is listed in Table 1.

Table 1: Historical periods in the development of plastics

Period	Development	Importance
Before 1850	Natural polymeric materials, including rubber, shellac, cellulose, horn, and ivory	Used for tools, coatings, ornaments, clothing, and household goods.
1850s–1870s	Semisynthetic materials such as Parkesine, celluloid and rubber	Replaced scarce ivory, tortoiseshell, silk, and wood in combs, buttons, photographic film, and toys.
1907–1930s	Fully synthetic resins, including Bakelite and PVC	Enabled electrical insulation, radios, telephones, appliances, coatings, and industrial components.
1930s–1950s	Nylon, polyethylene, polypropylene, PET, PTFE, and other high-volume polymers	Supported wartime production, synthetic fibers, packaging, machinery, and new consumer goods.
After 1950	Petrochemical expansion, injection molding, extrusion, film production, and automated mass manufacturing	Made plastics cheap, abundant, and globally ubiquitous.
Late twentieth century–today	Engineering plastics, polymer composites, conductive polymers, biomedical polymers, 3D-printing materials, and bioplastics	Extended plastics into aerospace, electronics, medicine, renewable energy, and advanced manufacturing.

The economic success of plastics and cheap production encouraged a model of extract, produce, use, and discard. Global plastic production rose from about 2 million tonnes in 1950 to 460 million tonnes in 2019 (OECD, 2022). Plastic waste increased from 156 million tonnes in 2000 to 353 million tonnes in 2019, while only about 9% was ultimately recycled. Plastic pollution now affects land, rivers, oceans, and the atmosphere, while additives and microplastics raise concerns about ecological and human exposure. These burdens are not distributed equally, because communities with weak

[1] Ironically, one of these authors used ChatGPT to generate the article (Gewaltig, 2024).

waste systems and informal waste workers often face greater exposure and fewer resources (United Nations Environment Programme, 2024).

Mass production and prices. With AI, ideas, texts and images can now be mass-produced. Thus AI appears to replicate for cognitive work what plastic represented in the physical world. Mass production made many items affordable for a broad range of citizens. But interestingly the impact on prices was not as straightforward. Take for instance a standard commodity product like chairs. There are plastic chairs available for a few dollars. At the same time, designer chairs such as the Panton Chair, made with the same process, are still sold for more than ten times as much.

Scarcity of human provenance. One possible consequence of such abundance is that scarcity shifts from the artifact to its provenance. As texts, images, and other creative products become increasingly cheap to generate, demonstrably human effort may itself acquire value. Experiments on visual art, for example, find that the same AI-generated works are judged more positively when labeled as human-created instead of AI-created, particularly on dimensions such as profundity and worth (Bellaiche et al., 2023). A fictional version of this tension appears in Spike Jonze's *Her* (Jonze, 2013), where the main character Theodore earns a living by composing intimate letters on behalf of other people in a world increasingly populated by highly capable AI. The scarce commodity may therefore not always be the text or image itself, but the belief that human effort, intention, or experience stands behind it.

Recycling genAI content? The idea that AI content is invading all cultural artifacts might be concerning. The accumulation of low-effort AI output across the web has itself become an object of study, with typologies such as the "7Vs of AI slop" attempting to characterize this new form of generative waste (Madsen & Puyt, 2025). In education, researchers investigate the ways in which "AI-giarism" (in reference to plagiarism) is perceived by students (Chan, 2025). Recursive effects are possible: indiscriminate training on generated data can degrade model quality (Shumailov et al., 2024). This result motivates attention to data provenance and curation, but it does not establish that generated content is irreversibly inseparable from human content or that degradation must compound in deployed systems. Of course the plastic waste analogy is misleading, as deleting files is not as difficult or harmful as disposing of physical waste. However, what deletion cannot undo is the effect on people who relied on AI when writing texts and the effect on models trained on generated content.

Impact on human cognition. Some aspects of the microplastic analogy can be associated with accumulating evidence around the impact of AI on human cognition. Performance of students in exams degrades when relying on AI assistance for homework (Stromberg et al., 2026). In education and grading, the effects of genAI are an active field of research (Chan, 2025; 2026) and universities are taking measures to counteract some of the negative effects on learning and exams (Horowitch, 2026). But assessing the broader impact on human cognitive functions is a challenging research area. A preliminary EEG study reports weaker neural engagement and lower self-reported ownership among writers relying on an AI assistant compared to those using a search engine or no tool at all (Kosmyna et al., 2025).

Feedback into human behavior. The microplastics analogy may also extend beyond synthetic artifacts accumulating in the information environment: algorithmically altered representations can feed back into the humans who repeatedly encounter them. Image filters provide a concrete example: more frequent use of filters has been associated with greater dysmorphic concern and greater acceptance of cosmetic surgery (Ateq et al., 2024), while clinicians have described patients seeking cosmetic procedures in order to resemble filtered versions of themselves (Rajanala et al., 2018). In that sense, an algorithmically shaped information environment may not only contain synthetic material, but also influence human preferences, self-perception, and behavior.

Design for life cycles. When it comes to AI software components and maintainability the insights from plastic fall in line with what other researchers have emphasized in the ML systems communities. Replacing durable components in physical machines with less durable plastic parts might be cheap in the short term, but lead to higher cost or more waste in the long run. Similarly, researchers have argued in the ML systems field that ML is like a *credit card for technical debt* (Sculley et al., 2015). It can be easy to add functionality to a software product with ML. But maintainability often suffers if data is not tested (Schelter et al., 2018b), infrastructure components are not scalable, or novel security issues are not taken care of (Schelter et al., 2018a). Researchers in AI could learn from the plastics revolution to design systems for the entire lifecycle and evaluation should cover

data collection, training, deployment, monitoring, updating, and retirement, just as plastic policy increasingly considers production, use, and disposal. It is encouraging that many of these ideas are being considered in the literature, for instance in the increased care for documentation of data (Gebru et al., 2021) and models (Mitchell et al., 2019).

Accountability for hidden cost and safety. The amount of plastic waste leads to high costs that can be difficult to hold relevant stakeholders accountable for. Analogously, the implications of AI technology can be difficult to predict, but many initiatives are aiming for an early mitigation of risks associated with AI systems (OECD, 2023; Smuha, 2025). An example from pharmaceuticals, the sulfanilamide disaster of 1937, demonstrated the limitations of a primarily function-oriented approach to product development: The preparation initially met the manufacturer's requirement to provide an active ingredient in a liquid dosage form, but the safety of the finished product for users had not been sufficiently tested. The resulting mass poisoning of more than 100 consumers directly drove passage of the 1938 Federal Food, Drug, and Cosmetic Act (Ballentine, 1981; U.S. Congress, 1938). Studies such as MacLeod et al. (2020) demonstrate how material development can be significantly accelerated through closed, autonomous experimental cycles; similar studies show this for molecular synthesis in pharmaceutical manufacturing (Koscher et al., 2023). However, occupational health and safety must also receive adequate attention, even if regulatory frameworks are not yet in place. Potential psychiatric harms associated with chatbot interactions have also prompted calls for longitudinal study and safeguards (Hudon & Stip, 2025). With increasing autonomy, safety, and user-related evaluation criteria must be better integrated in evaluation cycles, such that the product not only fulfills its specified function but also takes into account the consequences and risks for consumers.

4 Free Will

The goals of alchemists were clearly defined. The goals of AI are different as there is no consensus on what intelligence is and how it can be measured (Marcus, 2021; Chollet, 2019), a disagreement lively enough that a 2025 position paper still argues general intelligence requires a fundamentally different training paradigm than the one behind current large language models (Han et al., 2025). This is surprising, as many would argue that the concept of intelligence is intuitively understood by everyone. And most of us can provide at least anecdotal evidence about how some version of a chatbot is better than the previous one, or vice versa. But when asking for the evidence and how this insight could be generalized, answers are as disappointing as the usual joke on the intelligence quotient (IQ):

> "Intelligence is what the tests test."
>
> Boring (1923)

Leaving aside the difficulties with the Flynn effect (Flynn, 1984; 1987), the reverse Flynn effect (Bratsberg & Rogeberg, 2018), and the problems with benchmarks (Geirhos et al., 2020; Church et al., 2021; Raji et al., 2021), which can be considered the IQ of AI systems, there is a striking similarity between this problem of intelligence definitions and another episode of science on human cognition: the research around free will. Like debates about intelligence, these findings quickly acquired implications beyond their experimental scope. Interdisciplinary scrutiny then showed that results about the timing of simple actions did not by themselves settle the broader concept of free will (Schurger et al., 2012; 2021; Brass et al., 2019; Triggiani et al., 2023).

Empirical finding: Our brain decides before we do? In a first line of neuroscientific experiments, subjects were instructed to move their arms at a time of their choosing, while the experimenters recorded brain activity throughout (Kornhuber & Deecke, 1965; Kornhuber, 1984; Libet et al., 1982). A later protocol added introspection: subjects watched a revolving clock and reported the moment they became aware of the decision to move (Libet et al., 1983). Well *before* the subjects reported they had made the decision, the brain showed characteristic activity indicative of the movement, often referred to as Readiness Potential (RP). But if our self-initiated actions only follow the brain states that initiate the movement, is free will merely an epiphenomenon (Libet, 2005)?

At the heart of the problem with free will was probably an epistemological misunderstanding. Just like with intelligence, the concept of *free will* was seemingly clear to everyone, at least on a level

of understanding often referred to by certain philosophers as "folk psychology" (Sellars, 1956). At the same time, the scientific discourse suggested that the community did not actually agree on what exactly they referred to with the term *free will* until a transdisciplinary effort between neuroscientists, psychologists and philosophers started to take a more differentiated view on the issue (Schurger et al., 2012; 2021; Brass et al., 2019; Triggiani et al., 2023).

Tempting Empiricism. A striking similarity between the free will and intelligence discourse was that scientists and laymen were excited, because with a new empirical finding we could suddenly argue about a concept that is very relevant to society as a whole. Free will is a central element of our relationships and legal systems, with implications for responsibility for our decisions. Being able to do science on that central concept of our cognition merely by sticking electrodes onto our scalp and obtaining quantitative measures of brain activity appeared tempting. Libet's experiments made the timing of action feel experimentally tractable (Libet, 2002; Schurger et al., 2021). Intelligence tests, psychometrics, neuroimaging, and later AI benchmarks made intelligence feel measurable and comparable across people and systems (Colom et al., 2010; Legg & Hutter, 2007; Gignac & Szodorai, 2024). In both cases, the measurements were scientifically fruitful, but they also tempted researchers to treat a narrow operational handle as if it settled the broader concept.

Interdisciplinary discourse helps differentiate tension in empirical results. The research on free will bears some potential for the AI debate, as it demonstrates how a field can mature without discarding its often contradictory empirical findings. The results on the readiness potential could be replicated and remained, but their meaning changed through interdisciplinary scientific discourse. For instance, researchers argued that RP onset may not mark an unconscious decision at all, as reported intention times are highly method dependent (Schurger et al., 2012; Lau et al., 2006; Schurger et al., 2021; Triggiani et al., 2023). The result of this joint effort was not the collapse of the concept of free will, but refinement. Philosophy, psychology, and neuroscience together made the target concept more precise and the empirical claims more modest and specific (Brass et al., 2019; Triggiani et al., 2023).

The proxy is not the target. A lesson AI research could learn from the free will debate is that as tempting as it might be to believe in empirical results such as benchmark scores or anecdotal evidence, scientists should be careful when mistaking any empirical proxy for the target concept, especially when it is not clearly defined. Free will research improved once it differentiated concepts such as intention, veto, arbitrary picking, and deliberate choice (Brass et al., 2019; Triggiani et al., 2023). AI research needs the same discipline with intelligence, understanding, generality, robustness, and agency (Gignac & Szodorai, 2024; Mitchell & Krakauer, 2023). This call is now echoed from inside the ML community itself: recent position papers argue that current benchmarking practice conflates literal, human-inspired tests with the underlying capability they are meant to measure (Riemer et al., 2025). In addition to more holistic benchmarks that complement classical predictive performance metrics with aspects such as fairness (Liang et al., 2023), creativity (Srivastava, 2023) or energy efficiency (Strubell et al., 2019), there are researchers advocating that AI evaluation should adopt the same item-level rigor used in testing humans (Zhuang et al., 2025; Wei et al., 2025), and that benchmark scores are increasingly untrustworthy as a proxy for capability altogether (Cheng et al., 2025).

Two recent episodes illustrate the pattern directly. The widely publicized claim that GPT-4 showed early 'sparks' of general intelligence relied largely on diverse, hand-designed qualitative probes rather than a controlled measurement framework (Bubeck et al., 2023). The evidential scope was therefore narrower than the conceptual language of general intelligence. A number of striking "emergent abilities" reported for large language models were later argued to be largely an artifact of the choice of evaluation metric, not a genuine discontinuity in model behavior (Schaeffer et al., 2023). In both cases, a proxy, an anecdote or a benchmark curve, was mistaken for the target concept it was meant to measure.

Generalization with care. The readiness potential could be replicated, but researchers disagreed about what it meant for free will (Schurger et al., 2012; Triggiani et al., 2023). AI researchers face a similar difficulty when interpreting benchmark results. Unlike free will, intelligence has a measurement tradition of its own (Gignac & Szodorai, 2024), and the analogy cannot tell us which tests are appropriate. It does remind us to explain the connection between a test and the claim made on its basis. If a paper claims that a system can perform a task in everyday use, its evaluation should provide evidence for that setting. A benchmark score may leave much of this claim untested.

5 WHITE ELEPHANTS

Another analogy that comes to mind when discussing AI is that of a *white elephant*. The idiom derives from a Siamese legend recorded by European travelers from at least the seventeenth century. In the kingdoms of mainland South-East Asia, an albino elephant was held to be sacred, a living token of royal power and divine favor. Such an animal could neither be put to work nor slaughtered; it had to be housed, fed, and venerated at enormous expense. The King of Siam was therefore said to give a white elephant to someone he wished to ruin: the gift was simultaneously a mark of distinction and a financial death-sentence.

The German historian Dirk van Laak appropriated this legend as the organizing metaphor for hubris associated with technological investments (van Laak, 1999). Van Laak argues that many large-scale technological projects share the logic of the Siamese gift: they are initiated with great fanfare, consume resources on a spectacular scale, and ultimately have little use in relation to their associated cost. Unlike ordinary cost-overrun projects, the defining feature of a true white elephant is that the catastrophic outcome was, in retrospect, foreseeable, and was often foreseen, from near the outset. What kept them alive was not engineering optimism alone but what van Laak calls an *ideology of technological gigantism*: a culturally and politically rooted conviction that the sheer scale of a project was itself proof of national vitality, scientific mastery, and modernity. In this worldview, to cancel a megaproject was tantamount to admitting civilizational weakness.

Van Laak traces this ideology across three broad political contexts: the utopian planning cultures of the interwar period, the Cold War competition for technological prestige, and the late-industrial search for energy security. He illustrates it through a series of case studies drawn from European history. The three examples, Atlantropa, Concorde, and the Kalkar nuclear plant, each illuminate a different facet of the syndrome: visionary hubris with no political constituency, national prestige pursued at ruinous cost, and energy-security anxiety elevated into a monument that never generated a single kilowatt-hour. In Table 2, the projects are compared against each other on key dimensions.

Table 2: Selected megaprojects discussed by van Laak (1999).

Project	**Start**	**Primary motivation**	**Final cost (approx., USD)**
Atlantropa	*Never*	Pan-European hydroelectric power and land reclamation via Mediterranean sea-level reduction; interwar civilizational utopianism	*No estimate possible*; project never left the design stage
Concorde	1965	Franco-British national prestige; civil supersonic air travel as index of Western European technological parity with the superpowers	$\approx$ \$2.7–3.8 bn (converted from £1.5–2.1 bn at 1976 prices; development costs only; no commercial return sufficient to offset outlay)
Kalkar	1973	West German / Belgian / Dutch energy-security strategy; fast-breeder technology as solution to long-term uranium scarcity	$\approx$ \$3.9–4.1 bn (converted from $\approx$ 7 bn DM; plant completed 1985, *never generated electricity*, canceled 1991)

Note: Currency conversions use approximate contemporary mid-market rates (\$1.75–1.80/£ for 1976; 1.7–1.8 DM/\$ for mid-1980s) and are indicative only. Concorde figure covers government expenditure; manufacturer losses and airline subsidies are excluded. Kalkar figure incl. construction costs to completion plus partial cancellation compensation; full liability including long-term decommissioning is higher.

Each of the three projects described above attracted capable, often brilliant, technical personnel; each was supported at its inception by broadly plausible reasoning about energy, transport, or civilizational need. And after all: many of these white elephants did work, including the Concorde. But the *ideological* commitment to scale as such was the conviction that a sufficiently large and technically audacious project was self-justifying, and that the political will to cancel it, once the problems became apparent, was a form of weakness that no responsible government could display.

When overinvestment remains useful. The history of railways, a popular analogy to the current AI and data center investments, complicates this picture. Britain's "Railway Manias" of the nineteenth

century involved speculative excess and substantial overinvestment, yet the physical infrastructure did not disappear when financial returns disappointed. Railways became part of the durable infrastructure of industrial Britain, a transformation that features prominently in Hobsbawm's account of British industrialization (Hobsbawm, 1968). More recent economic history makes the paradox particularly explicit: speculative railway investment left an excessive infrastructure legacy even as expected financial returns failed to materialize (Lewis & Offer, 2022). Lewis and Offer point to a similar pattern in the late-1990s broadband boom: after much of the invested money was lost, the physical network assets remained (Lewis & Offer, 2022). Yet when comparing these developments with AI, the analogy requires a more differentiated view. Unlike trackbeds, tunnels, and fiber conduits, accelerators and memory may become economically obsolete within a few years, although physical failure and economic obsolescence should not be conflated. Data centers also contain longer-lived assets such as buildings, grid connections, and substations, so the analogy depends on which layer of the stack receives the investment. The current build-out of AI data-center infrastructure, with hyperscaler capital expenditure reaching hundreds of billions of dollars in aggregate, has prompted economists to ask whether investment in some layers is outpacing observed monetization (Wang & Chen, 2026). This raises, but does not answer, the question of whether any current project exhibits van Laak's self-justifying gigantism.

Who cannot stop. Van Laak's projects were held in place by states, where cancellation reads as a public admission of failure and the budget constraint is soft. The current build-out is financed privately, and a firm that has misjudged demand writes the asset down and moves on. The lock-in mechanism is not the same. But the most striking similarity is probably the fact that the central question raised in van Laak's research

> "How much should a society spend on a technological project?"
>
> paraphrased from van Laak (1999)

remains difficult to answer for AI. The railway and broadband examples show why poor financial returns alone are insufficient grounds for calling a project a white elephant. Some benefits take time to emerge or are difficult to price. This also makes the comparison difficult to apply to ongoing AI investments. We would need to examine what a project is supposed to achieve, who benefits from it and what else could be done with the resources. Researchers can help by checking whether improvements on their benchmarks translate into the benefits used to justify the investment.

6 Conclusion

The analogies discussed here offer a way to bring questions about AI into conversations beyond our own field. Alchemy raises the question of how much we need to understand a method before relying on it. The history of plastics encourages us to follow the costs of a technology beyond its initial use. The free will debate shows how empirical findings can lead public discourse astray. White elephants make us ask what an investment is meant to achieve and whether its benefits justify the resources spent.

These questions have consequences for how we present research. Claims of progress should explain how evaluation protocols capture the proposed use. Where that connection is uncertain, the uncertainty belongs in the discussion of the result. Costs that arise during use and maintenance also deserve attention, including those borne by people who had no part in choosing the technology. Assessing such costs will often require working with the people affected. Their experience may also change which problems researchers consider worth solving.

Historical comparisons can themselves bias this discussion. Choosing a failed megaproject as an example may suggest that an AI investment is bound to fail. Choosing a useful invention whose theory came later may make demands for understanding seem unreasonable. Both readings go beyond what the examples establish. We therefore need to explain why we chose a comparison and where it stops being useful. The railway examples are a reminder to look for cases that complicate our preferred interpretation.

We propose using these analogies to foster discussions about particular AI applications. Researchers can explain what is known about a system, while policy makers and citizens can question whether the evaluation addresses their concerns.

AI USE STATEMENT

We did not use generative AI tools for any of the tasks for which the ICLR 2027 AI policy requires disclosure: we did not use them to develop the conceptual framework or select the analogies, to formulate or refine our claims, to provide feedback on our methodology, or to translate sources, and the remaining required-disclosure tasks (e.g., generating or cleaning data sets, formulating or proving mathematical claims, implementing methods, and analyzing data or interpreting experimental results) are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: searching for information and relevant literature, and editing the manuscript for grammar and readability. We have reviewed all AI-assisted work: we checked every source and factual statement found with AI assistance against the original publication, and we reviewed each suggested language edit individually. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

S. Acharya et al. Proton emission in ultraperipheral pb-pb collisions at $\sqrt{s_{NN}} = 5.02$ tev. *Physical Review C*, 111(5):054906, May 2025. doi: 10.1103/PhysRevC.111.054906. URL https://link.aps.org/doi/10.1103/PhysRevC.111.054906.

Khadijah Ateq, Mohammed Alhajji, and Noara Alhusseini. The association between use of social media and the development of body dysmorphic disorder and attitudes toward cosmetic surgeries: A national survey. *Frontiers in Public Health*, 12:1324092, 2024. doi: 10.3389/fpubh.2024.1324092.

Carol Ballentine. Taste of raspberries, taste of death: The 1937 Elixir Sulfanilamide incident. *FDA Consumer*, 15(5):18–21, June 1981. URL https://www.fda.gov/files/about%20fda/published/The-Sulfanilamide-Disaster.pdf.

Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein. Measuring the impact of early-2025 AI on experienced open-source developer productivity. *arXiv preprint arXiv:2507.09089*, 2025.

Lucas Bellaiche, Rohin Shahi, Martin Harry Turpin, Anya Ragnhildstveit, Shawn Sprockett, Nathaniel Barr, Alexander Christensen, and Paul Seli. Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research: Principles and Implications*, 8:42, 2023. doi: 10.1186/s41235-023-00499-6.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pp. 610–623, 2021. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

Edwin G. Boring. Intelligence as the tests test it. *New Republic*, 35:35–37, 1923. Reprinted in J. J. Jenkins and D. G. Paterson (Eds.), *Studies in Individual Differences: The Search for Intelligence*, pp. 210–214, Appleton-Century-Crofts, 1961.

Marcel Brass, Ariel Furstenberg, and Alfred R. Mele. Why neuroscience does not disprove free will. *Neuroscience and biobehavioral reviews*, 2019. doi: 10.1016/j.neubiorev.2019.04.024.

Bernt Bratsberg and Ole Rogeberg. Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences*, 115(26):6674–6678, 2018. doi: 10.1073/pnas.1718793115.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712, 2023. URL https://arxiv.org/abs/2303.12712.

Cecilia Ka Yuk Chan. Students' perceptions of 'AI-giarism': investigating changes in understandings of academic misconduct. *Education and Information Technologies*, 30(6):8087–8108, April 2025. ISSN 1573-7608. doi: 10.1007/s10639-024-13151-7. URL https://doi.org/10.1007/s10639-024-13151-7.

Cecilia Ka Yuk Chan. Students Know AI Should Not Replace Thinking, but How Do They Regulate It? The TACO Framework for Human-AI Cognitive Partnership, April 2026. URL http://arxiv.org/abs/2604.18737. arXiv:2604.18737 [cs.CY].

Zerui Cheng, Stella Wohnig, Ruchika Gupta, Samiul Alam, Tassallah Abdullahi, João Alves Ribeiro, Christian Nielsen-Garcia, Saif Mir, Siran Li, Jason Orender, Seyed Ali Bahrainian, Daniel Kirste, Aaron Gokaslan, Carsten Eickhoff, and Ruben Wolff. Position: Benchmarking is broken – don't let AI be its own judge. In *Advances in Neural Information Processing Systems (NeurIPS)*, Position Paper Track, 2025.

François Chollet. On the Measure of Intelligence, November 2019. URL http://arxiv.org/abs/1911.01547. arXiv:1911.01547 [cs].

Kenneth Church, Mark Liberman, and Valia Kordoni. Benchmarking: Past, present and future. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pp. 1–7. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.bppf-1.1/.

Roberto Colom, Sherif Karama, Rex E. Jung, and Richard J. Haier. Human intelligence and brain networks. *Dialogues in Clinical Neuroscience*, 2010. doi: 10.31887/DCNS.2010.12.4/rcolom.

Paul Feyerabend. *Naturphilosophie*. Suhrkamp Verlag, 2018.

James R. Flynn. The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, 95(1):29–51, 1984. doi: 10.1037/0033-2909.95.1.29.

James R. Flynn. Massive IQ gains in 14 nations: What iq tests really measure. *Psychological Bulletin*, 101(2):171–191, 1987. doi: 10.1037/0033-2909.101.2.171.

Leon Furze. Digital plastic: Generative AI and the digital ecosystem. Dr Leon Furze (blog), October 2023. URL https://leonfurze.com/2023/10/25/digital-plastic-generative-ai-and-digital-ecosystem/. Earliest documented, systematic analogy between generative AI content and plastic, framing it as "digital microplastics" that accumulate as low-grade, pervasive residue in the common digital ecosystem. Dated 25 October 2023.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723. URL https://doi.org/10.1145/3458723.

Timnit Gebru, Emily M. Bender, and Angelina McMillan-Major. Statement from the listed authors of Stochastic Parrots on the "AI pause" letter. DAIR Institute, March 2023. URL https://dair-institute.org/blog/letter-statement-March2023/.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. doi: 10.1038/s42256-020-00257-z.

Marc-Oliver Gewaltig. Is AI-generated content the microplastic pollution of our intellect? LinkedIn Pulse, May 2024. URL https://www.linkedin.com/pulse/ai-generated-content-microplastic-pollution-our-gewaltig-gwgue.

Gilles E. Gignac and Eva T. Szodorai. Defining intelligence: Bridging the gap between human and artificial perspectives. *Intelligence*, 2024. doi: 10.1016/j.intell.2024.101832.

Seungwook Han, Jyothish Pari, Samuel J. Gershman, and Pulkit Agrawal. Position: General intelligence requires reward-based pretraining. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, Position Paper Track, 2025.

Moritz Herrmann, F. Julian D. Lange, Katharina Eggensperger, Giuseppe Casalicchio, Marcel Wever, Matthias Feurer, David Rügamer, Eyke Hüllermeier, Anne-Laure Boulesteix, and Bernd Bischl. Position: Why we must rethink empirical research in machine learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Position Paper Track, 2024.

Eric J. Hobsbawm. *Industry and Empire: An Economic History of Britain Since 1750*. Weidenfeld & Nicolson, London, 1968.

Rose Horowitch. How AI killed a 133-year-old princeton tradition. The Atlantic, May 2026. URL https://www.theatlantic.com/ideas/2026/05/princeton-ai-honor-code/687144/.

Alexandre Hudon and Emmanuel Stip. Delusional experiences emerging from AI chatbot interactions or "AI psychosis". *JMIR Mental Health*, 12:e85799, 2025. doi: 10.2196/85799.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Spike Jonze. Her. Motion picture, Warner Bros. Pictures and Annapurna Pictures, 2013. Directed and written by Spike Jonze.

Sayash Kapoor and Arvind Narayanan. A misleading open letter about sci-fi AI dangers ignores the real risks. AI as Normal Technology (normaltech.ai), March 2023. URL https://www.normaltech.ai/p/a-misleading-open-letter-about-sci.

Brian Kennedy, Eileen Yam, Emma Kikuchi, Isabelle Pula, and Javier Fuentes. How americans view AI and its impact on people and society. Technical report, Pew Research Center, September 2025. URL https://www.pewresearch.org/science/2025/09/17/how-americans-view-ai-and-its-impact-on-people-and-society/.

Hans Helmut Kornhuber. Attention, readiness for action, and the stages of voluntary decision: Some electrophysiological correlates in man. In O. Creutzfeldt, R. Schmidt, and W. D. Willis (eds.), *Sensory-Motor Integration in the Nervous System*, pp. 420–429. Springer, Berlin, 1984. doi: 10.1007/978-3-642-69931-3_31.

Hans Helmut Kornhuber and Lüder Deecke. Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pflüger's Archiv für die gesamte Physiologie des Menschen und der Tiere*, 1965. doi: 10.1007/BF00412364. English: Changes in Brain Potentials with Willful and Passive Movements in Humans: The Readiness Potential and Reafferent Potentials.

Brent A. Koscher, Richard B. Canty, Matthew A. McDonald, Kevin P. Greenman, Charles J. McGill, Camille L. Bilodeau, Wengong Jin, Haoyang Wu, Florence H. Vermeire, Brooke Jin, Travis Hart, Timothy Kulesza, Shih-Cheng Li, Tommi S. Jaakkola, Regina Barzilay, Rafael Gómez-Bombarelli, William H. Green, and Klavs F. Jensen. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science*, 382(6677):eadi1407, 2023. doi: 10.1126/science.adi1407. URL https://www.science.org/doi/abs/10.1126/science.adi1407.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 2025.

Daniel Kuehn. Keynes, Newton and the Royal Society: the events of 1942 and 1943. *Notes and Records: The Royal Society Journal of the History of Science*, 67(1):25–36, December 2012. ISSN 0035-9149. doi: 10.1098/rsnr.2012.0053. URL https://doi.org/10.1098/rsnr.2012.0053.

Hakwan C. Lau, Robert D. Rogers, and Richard E. Passingham. On measuring the perceived onsets of spontaneous actions. *The Journal of Neuroscience*, 2006. doi: 10.1523/JNEUROSCI.1138-06.2006.

Yann LeCun. My take on ali rahimi's "test of time" award talk at NIPS. Facebook post, December 2017. URL https://www.facebook.com/yann.lecun/posts/10154938130592143.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 2007. doi: 10.1007/s11023-007-9079-x.

Oliver Lewis and Avner Offer. Railways as patient capital. *Oxford Review of Economic Policy*, 38 (2):260–277, 2022. doi: 10.1093/oxrep/grac004.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, October 2023. URL http://arxiv.org/abs/2211.09110. arXiv:2211.09110 [cs].

Benjamin Libet. The timing of mental events: Libet's experimental findings and their implications. *Consciousness and cognition*, 2002. doi: 10.1006/CCOG.2002.0568.

Benjamin Libet. Do we have free will? In Robert Kane (ed.), *The Oxford Handbook of Free Will*, pp. 551–564. Oxford University Press, 2005. doi: 10.1093/oxfordhb/9780195178548.003.0025.

Benjamin Libet, Elwood W. Wright, and Curtis A. Gleason. Readiness-potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts. *Electroencephalography and clinical neurophysiology*, 1982. doi: 10.1016/0013-4694(82)90181-X.

Benjamin Libet, Curtis A. Gleason, Elwood W. Wright, and Dennis K. Pearl. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain*, 106(3):623–642, 1983. doi: 10.1093/brain/106.3.623.

B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, and C. P. Berlinguette. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20):eaaz8867, 2020. doi: 10.1126/sciadv.aaz8867. URL https://www.science.org/doi/abs/10.1126/sciadv.aaz8867.

Dag Øivind Madsen and Richard W Puyt. The 7vs of AI slop: A typology of generative waste. SSRN, 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5558018.

Gary Marcus. Ai is harder than we think. Substack: Gary Marcus - The Road to AI We Can Trust, 2021. URL https://garymarcus.substack.com/p/ai-is-harder-than-we-think.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, January 2019. doi: 10.1145/3287560.3287596. URL http://arxiv.org/abs/1810.03993. arXiv:1810.03993 [cs].

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, 2023. doi: 10.1073/pnas.2215907120.

OECD. Global plastics outlook: Economic drivers, environmental impacts and policy options. Technical report, OECD Publishing, Paris, 2022. URL https://doi.org/10.1787/de747aef-en.

OECD. G7 Hiroshima process on generative artificial intelligence (AI): Towards a G7 common understanding on generative AI. Technical report, OECD Publishing, Paris, September 2023. URL https://www.oecd-ilibrary.org/science-and-technology/g7-hiroshima-process-on-generative-artificial-intelligence-ai_bf3c0c60-en.

M. Ragolane and S. Patel. Too much, too fast: Understanding AI fatigue in the digital acceleration era. *International Journal of Arts, Humanities and Social Science*, 6(8):53–60, 2025. doi: 10.56734/ijahss.v6n8a7.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pp. 1177–1184, 2007. URL https://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines.

Ali Rahimi and Benjamin Recht. An addendum to alchemy. arg min blog, December 2017. URL https://archives.argmin.net/2017/12/11/alchemy-addendum/.

Susruthi Rajanala, Mayra B. C. Maymone, and Neelam A. Vashi. Selfies—living in the era of filtered photographs. *JAMA Facial Plastic Surgery*, 20(6):443–444, 2018. doi: 10.1001/jamafacial.2018.0486.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. Position: Theory of mind benchmarks are broken for large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Position Paper Track*, pp. 82091–82130, 2025.

Jasper Roe, Leon Furze, and Mike Perkins. Digital plastic: a metaphorical framework for critical AI literacy in the multiliteracies era. *Pedagogies: An International Journal*, 2025. doi: 10.1080/1554480X.2025.2557491.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.

Sebastian Schelter, Felix Biessmann, Tim Januschowski, David Salinas, Stephan Seufert, and Gyuri Szarvas. On Challenges in Machine Learning Model Management. *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.*, pp. 5–13, 2018a. URL http://sites.computer.org/debull/A18dec/p5.pdf.

Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, 11(12):1781–1794, August 2018b. ISSN 21508097. doi: 10.14778/3229863.3229867. URL http://dl.acm.org/citation.cfm?doid=3229863.3275547.

Aaron Schurger, Jacobo D. Sitt, and Stanislas Dehaene. An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences*, 2012. doi: 10.1073/pnas.1210467109.

Aaron Schurger, Pengbo Hu, Joanna Pak, and Adina L. Roskies. What is the readiness potential? *Trends in cognitive sciences*, 2021. doi: 10.1016/j.tics.2021.04.001.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Adv. Neural Inf. Process. Syst.*, pp. 2503–2511, 2015. URL http://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.

Wilfrid Sellars. Empiricism and the philosophy of mind. In Herbert Feigl and Michael Scriven (eds.), *Minnesota Studies in the Philosophy of Science*, volume 1, pp. 253–329. University of Minnesota Press, Minneapolis, 1956.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, July 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07566-y. URL https://www.nature.com/articles/s41586-024-07566-y.

Nathalie A Smuha. Regulation 2024/1689 of the eur. parl. & council of june 13, 2024 (EU Artificial Intelligence Act). *International Legal Materials*, 64(5):1234–1381, 2025.

Aarohi et al. Srivastava. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, June 2023. URL http://arxiv.org/abs/2206.04615. arXiv:2206.04615 [cs, stat].

David Stromberg, Victor Lei, and Yanhui Wu. The Generative AI Learning Penalty: Evidence from Chinese Secondary Education, June 2026. URL https://papers.ssrn.com/abstract=6868618.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP, June 2019. URL http://arxiv.org/abs/1906.02243. arXiv:1906.02243 [cs].

Antonio I. Triggiani, Gabriel Kreiman, Cara Lewis, Uri Maoz, Alfred Mele, Liad Mudrik, Adina L. Roskies, Aaron Schurger, and Mark Hallett. What is the intention to move and when does it occur? *Neuroscience and biobehavioral reviews*, 2023. doi: 10.1016/j.neubiorev.2023.105199.

United Nations Environment Programme. What is the life-cycle approach and how can it help tackle plastic pollution?, August 2024. URL https://www.unep.org/news-and-stories/story/what-life-cycle-approach-and-how-can-it-help-tackle-plastic-pollution.

U.S. Congress. Federal Food, Drug, and Cosmetic Act of 1938. Public Law 75-717, 52 Stat. 1040, 1938.

Dirk van Laak. *Weiße Elefanten: Anspruch und Scheitern technischer Großprojekte im 20. Jahrhundert*. Deutsche Verlags-Anstalt, Stuttgart, 1999. English: *White Elephants: Claims and Failures of Technical Large-Scale Projects in the 20th Century*.

Lee Vinsel. You're doing it wrong: Notes on criticism and technology hype. *Medium / STS News*, February 2021. URL https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5.

Qianan Wang and Zen Chen. Boom, bubble, or buildout? a multi-method evaluation of whether artificial intelligence is in an ongoing financial bubble. *arXiv preprint arXiv:2606.01575*, 2026.

Kevin Wei, Patricia Paskov, Sunishchal Dev, Michael J. Byun, Anka Reuel, Xavier Roberts-Gaal, Rachel Calcott, Evie Coxon, and Chinmay Deshpande. Position: Human baselines in model evaluations need rigor and transparency (with recommendations & reporting checklist). In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Position Paper Track*, pp. 82265–82325, 2025.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*, 2023.

Yan Zhuang, Qi Liu, Zachary A. Pardos, Patrick C. Kyllonen, Jiyun Zu, and Zhenya Huang. Position: AI evaluation should learn from how we test humans. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Position Paper Track*, pp. 82483–82508, 2025.