The Cake that is Intelligence and Who Gets to Bake it: An AI Analogy and its Implications for Participation

MARTIN MUNDT, University of Bremen & Queer in AI, Germany

ANAELIA OVALLE, Meta & Queer in AI, United States of America

FELIX FRIEDRICH, Technical University of Darmstadt & hessian.AI, Germany

A PRANAV, University of Hamburg & Queer in AI, Germany

SUBARNADUTI PAUL, University of Bremen, Germany

MANUEL BRACK, Technical University of Darmstadt & DFKI, Germany

KRISTIAN KERSTING, Technical University of Darmstadt & hessian.AI & DFKI, Germany

WILLIAM AGNEW, Carnegie Mellon University & Queer in AI, United States of America

Abstract

In a widely popular analogy by Turing Award Laureate Yann LeCun, machine intelligence has been compared to cake —where unsupervised learning forms the base, supervised learning adds the icing, and reinforcement learning is the cherry on top. We expand this "cake that is intelligence" analogy from a simple structural metaphor to the full life-cycle of AI systems, extending it to sourcing of ingredients (data), conception of recipes (instructions), the baking process (training), and the tasting and selling of the cake (evaluation and distribution). Leveraging our re-conceptualization, we describe each step's entailed social ramifications and how they are bounded by statistical assumptions within machine learning. Whereas these technical foundations and social impacts are deeply intertwined, they are often studied in isolation, creating barriers that restrict meaningful participation. Our re-conceptualization paves the way to bridge this gap by mapping where technical foundations interact with social outcomes, highlighting opportunities for cross-disciplinary dialogue. Finally, we conclude with actionable recommendations at each stage of the metaphorical AI cake's life-cycle, empowering prospective AI practitioners, users, and researchers, with increased awareness and ability to engage in broader AI discourse.

1 Introduction

> *"If intelligence is a cake, the bulk of the cake is unsupervised learning, the icing on the cake is supervised learning, and the cherry on the cake is reinforcement learning."* Yann LeCun, NeurIPS Conference 2016

By now an illustrious analogy in the field of machine learning (ML) and artificial intelligence (AI), the quote's original claim was that the bulk of the cake —i.e., the majority of what supposedly composes intelligence— does not rely on turning over exorbitant amounts of labeled data. On the contrary, the bulk is conjectured to consist of unsupervised learning (i.e., learning from data without labels), the icing is created through supervised learning (i.e., learning to match ground-truth annotations), and the "cherry-on-top" is reinforcement learning (i.e., tuning to feedback). In this cake metaphor, this resembles first baking a solid cake, if you so will the delightfully moist sponge, to lay the foundation for it to later be perfected to one's individual taste, i.e., a specific narrow AI task. If at first resulting in a skeptical smile, we will learn in this paper that the design of AI systems indeed has a lot of parallels to baking a cake. As such, we posit that the metaphor is valuable not for its original reasons, but rather because the creation of AI systems entails many of a real cake's associations and implications on ingredients, recipes, baking processes, taste, and consumers.

In particular, we can follow the typical ML trends and may immediately ask ourselves whether this cake can be perfected? In turn, asking what the glazing on the cake would symbolize? In fact, perhaps it would even be possible to

bake a cherry-chocolate cake from the start, reducing the number of steps along the way? Much of what is considered core AI advancement seems to revolve around these types of questions. Metaphorically speaking, it is refining the cake analogy to revise its recipe —the algorithmic components to bake it—- and conversely contemplate its ingredients —the data that goes into AI systems[1]. From an ML perspective, baking the best possible cake may sound like an admirable goal. It certainly appears like a goal worthy of making progress if it were not for a substantial set of problems. For one, our goal implies measurability, the basis for it to later be optimized, despite being notoriously subjective in nature. As such, we can ask ourselves further questions, questions that are, however, of no less importance: *"How do we acquire our ingredients?"*, *"Who knows and understands the recipe?"* and *"Who gets to bake the cake?"*. Thinking about these questions then creates a ripple effect to critically reflect on *"Who decides if it is delicious?"*, *"Who is allowed to indulge in it?"*, or *"Who profits from its consumption?"*. Worst of all, if we do not like the current answer to any of these questions, is there anything we can do to change? For instance, what if we have personal reasons to object to the ingredients (say intolerance or ethical considerations) or do not like the final taste? Indeed, our proposition is not that the cake analogy is highly accurate because of its initial distinction between learning paradigms. Instead, we posit that it is valuable because once we "stir ingredients and bake it", there presently exist little tools to alter the final product.

Following the "intelligence as a cake" analogy, this paper critically questions how the metaphorical AI cake is presently being baked, shared, and eaten, where its ingredients originate from, and who is enabled to understand the recipe. To this end, we first re-conceptualize the analogy to describe how the process of translating baking ingredients to a cake relates to AI workflows. Having established accurate parallels, we create opportunities for cross-disciplinary dialogue by dissecting entailed issues and linking them to technical foundations. As such, we complement a rich history of works highlighting socio-ethical concerns by mapping where they are intertwined with fundamental technical challenges in AI design. Finally, we provide first actionable recommendations at each stage of the metaphorical AI cake's lifecycle, and in doing so, suggest avenues towards fostering participation and sustainability in AI design processes.

2 AI cake: an accurate analogy

If the introduction left the reader hungry for more, they will hopefully see their hunger stilled in a transition to how the cake analogy translates to AI systems. To this end, let us first re-conceptualize the analogy to describe the various stages of the AI lifecycle and their social ramifications before proceeding to discuss underpinnings in technical limitations.

2.1 Ingredients and their origin

At the beginning of (baking) any cake are its ingredients. Depending on the type of cake and the baker's exact location, some of these ingredients may be locally available, i.e., acquired through trade or bought from a market, whereas others originate from far away. For instance, the use of cocoa seeds is prevalent in cakes of the northern hemisphere, yet its growth is restricted to equatorial (predominantly African) regions. Although the "Western world" now largely acknowledges that respective local cultures have been, and are still continuously, subject to exploitation, it generally remains challenging for the consumer to thoroughly understand the origin of ingredients and respective implications. There may exist nutrition labels, but they abstract away most information. More extensive pushes for transparency and ethical considerations, like the supply chain act [115] recently passed by the European Parliament, remain controversial

---

[1]See for instance the below medium blog post for an overview of how the AI cake analogy was first contested by Andrychowicz et al. [7] to be composed of cherries (hindsight experience replay in reinforcement learning) and later revised to an AI cake 2.0 recipe by LeCun at ISSCC in 2019 by replacing fully 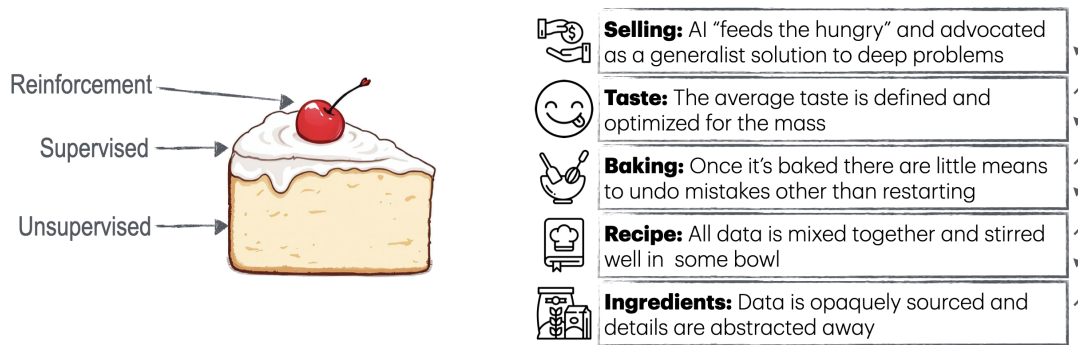unsupervised learning with the idea of self-supervision https://medium.com/syncedreview/yann-lecun-cake-analogy-2-0-a361da560dae

Fig. 1. An illustration of the AI–cake analogy. (Left) Traditionally, the cake was used to provide a structural metaphor for machine intelligence, relating unsupervised (bulk), supervised (icing), and reinforcement learning (cherry) paradigms. (Right) Our re-conceptualized analogy extends the original metaphor by drawing parallels to the way AI's ingredients are sourced, recipes crafted, and ultimately how the metaphorical cake is baked, tasted and sold.

among involved governments. At large, the pattern of obfuscation is exacerbated by ingredients having been processed before they reach the baker —say, cocoa seeds turned into chocolate— or even because the ingredients have been packaged into a ready-to-bake form in a chain of steps by big corporations.

As much as cake relies on its ingredients, so do AI systems depend on their underlying data. Unfortunately, the analogy also extends from consumers remaining unaware of how ingredients have reached their plates, to AI system designers typically lacking full visibility into the sourcing and creation of training datasets. In both cases, companies (or an in-circle of stakeholders) embed their own biases into the gathering pipeline, dictating sourcing decisions that ultimately carry significant social ramifications. The rapid expansion of machine learning datasets (e.g., Schuhmann et al. [140]) and increasingly opaque closed-source practices (e.g., OpenAI's ChatGPT) have only exacerbated these issues. On the one hand, unchecked scaling of AI systems has thus led to increased exploitative practices. This is evidenced, for instance, by the outsourcing of psychologically harmful content to African workers for moderation and annotation [134], a malpractice labeled as "ethics dumping" by the EU Horizon Research Council [53, 139]. Similarly, ChatGPT's "surprisingly" frequent use of terms like "delve" was later attributed to the terminology's prevalence in Nigeria [74]. On the other hand, the lack of transparency and access to the data pipeline has severe implications for data ownership or copyright infringement. As an example, the work "Does CLIP know my face" [75] has raised attention to the fact that people's personal data may indeed have seeped into the metaphorical AI cake as an ingredient without consent, but it remains very challenging to find out about it. Consequently, such model applications have the potential to violate privacy laws by retaining personally identifiable information, resulting in the first legal actions against the corporations behind them [113].

Clearly, both real cake and its AI analogy thus hinge on their ingredients. Where colonialism has facilitated the exploitation of natural ingredients, amassing data seems to follow a reminiscent pattern based on digital imperialism. As the origins of AI's "ingredients" remain obscure and processes remain opaque, the lack of transparency accelerates the privatization of knowledge [55]. The underlying pattern of abstracting away the contextual complexity and ecosystem removes a sense of responsibility —understanding that humans are behind— as it gets easier in practice to exploit because we can neither see nor retrace.

2.2   The "best" cake recipes

Baking a cake does not only require a set of ingredients, it generally needs to follow a recipe. The latter typically ensures that chosen ingredients "interact" sufficiently well. Although there exists an abundance of instructions on how to combine a cake's ingredients, the individual ingredients eventually all fall victim to the same blending process when mixed together into dough. Remarkably, this process seems simultaneously accessible and demanding. As long as the recipe is followed closely, almost anyone can successfully bake a cake, provided they have the massive ovens to bake it (see baking process section). If, however, one wishes to include a particular new ingredient in the mix, then baking almost becomes a precision science. It is a delicate balance of the modifications being noticeable, e.g., in color, texture, or taste of the cake, or potentially overpowering and thus ruining the mix —a balance only few may claim to truly understand. It certainly becomes a futile effort if one wishes to turn cake ingredients into another dessert when only generic cake recipes are available.

As much as cake recipes ultimately combine the majority of ingredients together, so do AI systems mix their training data. Unfortunately, our analogy extends from the blending procedure relying on the predominant belief that adding more data to the metaphorical dough is sufficient for inclusion, to the fact that few experts, if any, understand the consequences of respective interleaving. In both parts of the analogy, the assumption that simply adding more data ensures inclusivity is flawed. Delgado et al. [47] have pointed out that *"we cannot just add diverse end-users and stakeholders and stir"*. On the one hand, it is unclear which combinations of data may advance model capabilities and which data additions entail social consequences; their nature hinges on complex interplay with the rest of the opaque data mixture. For instance, from the perspective of purely instrumental improvement, it is heavily debated when and how the addition of synthetically generated data is indeed beneficial, e.g. Phi-2 & 3 [4, 79] or Dall-e-3 [15], or whether its inclusion in the mix is useless or downright adverse to performance [69]. From a complementary normative angle, initially sincere efforts to "debias" models through increased addition of non-anglocentric languages have similarly concluded that stirring these ingredients has contributed fairly little towards the desired goal of broad applicability in globally inclusive perspectives [60]. On the other hand, the recent repercussions [127] surrounding Google DeepMinds' Gemini image generation [65] have rendered the challenge of highlighting ingredients in a recipe and unwillingly overshooting perfectly visible[2]. In short, an effort to "diversify" the generated images through naive integration attempts of broader coverage of society has further resulted in the formation of adverse relations [63, 131], e.g. now picturing strong racial diversity in the generation of Nazi Germany's Wehrmacht soldiers. This contrived effect is not only highly illogical but further contributes to perpetuating harmful AI content and fostering oppression in complete contradiction to the initial amiable aim [112].

Clearly, both real cake and its AI analogy thus hinge on combining their ingredients. Whereas careful measuring and stirring of ingredients is a challenging task already, complex interplay of data makes careful weighting only one imperative element of a recipe. A lack of understanding in interleaving arbitrary data in AI systems yields mixtures where inclusion of ingredients, even if added with the best of intentions, can range anywhere from completely inconsequential, to instrumentally beneficial, or resulting in adverse fallout. The seemingly accessible and effortless aggregation of all data to streamline the ensuing baking thus also obfuscates ingredients' uniqueness, even suppresses their critical nuances, and in consequence gives rise to counter-intuitive ramifications. As such, homogenization of the recipe, i.e. blending all ingredients together to follow the instructions for cake dough (training a deep machine learning model), entails accessibility advantages that simultaneously make it excessively hard to bake anything else.

---

[2]and further lacking sufficient input of how people would in principle like to be depicted - explored in-depth in the "what makes for a tasty cake" section

2.3   Cake foundation and the baking process

Once settled on ingredients and a recipe, baking a cake is very much a unidirectional process. Once in the oven, there is no turning back. Forgotten an ingredient, bake a new one. Took a slight misstep in following instructions, bake a new one. Do not like the outcome, bake a new one. Once baked, a cake's composition becomes fixed and permits few revisions. Incorporating any raw ingredients post-baking is impractical and far from appetizing — think for instance of pouring milk over or adding raw egg to a baked cake. At best, we can superficially tune the cake, e.g. adorning it with fruit, candy, or a touch of glazing to make it more appealing.

As hard as it is to change a baked cake, so is modifying an AI model after it has been trained. Within the preference alignment literature, the "superficial alignment hypothesis" posits that reinforcement learning from human feedback primarily affects a foundation model's [126, 151] textual output style, rather than its core capabilities [165]. In fact, respectively finetuned models result in near identical performance of tuned and base model versions (i.e., they trade off one dimension for another), at best resulting in superficial improvement [97]. This is analogous to how we can only decorate a finished cake rather than change its composition, resulting in an unreasonable amount of baking repetitions for every fundamental change. In this context, however, such modification attempts (i.e. re-training) cost millions of dollars [88, 155] and consume extraordinary amounts of energy for minor improvements [144, 146][3]. For example, while presumably being updated to improve certain model abilities in Spring 2023, ChatGPT actually lost proficiency in basic tasks like identifying prime numbers or writing simple code [37] by June of the same year. This reveals a fundamental tension with current regulatory frameworks, such as the Biden-Harris AI executive order [77, 114] and the United Kingdom's "pro-innovation" approach [143], which rely heavily on post-training interventions. While model auditing and red-teaming are crucial to identifying problems in AI systems, resolving them thus remains challenging due to the difficulty of modifying a model after training — much like trying to fix a single ingredient in an already baked cake. These effects are further exacerbated by the fact that, following the prior section's arguments, we lack understanding of which data corresponds to the metaphorical egg or milk, and which ones are decorative in nature.

Overall, both real cake and its AI analogy thus hinge on a single-cycle baking process. Whereas baking cannot be fully changed and undone once completed, AI training largely dictates the final utility of the system. Attempts at fine-tuning are either superficial or entail strong trade-offs. In particular, later attempts at adding entirely new ingredients can interfere catastrophically, depending on the nature of the ingredient. Contrary to intelligence also being described as *"the ability to adapt to change"* (commonly attributed to Stephen Hawking, see Strauss [145]), the current inflexibility of the "cake-like" training pipeline entails excessive process repetition. As each training run requires exorbitant amounts of computational resources [100, 122], this seems akin to baking cakes by having thousands of ovens emit carbon, to ultimately re-do the entirety of produced cakes any time a non-superficial adjustment needs to be added.

2.4   What makes for a tasty cake?

A cake may have been baked successfully, but is it also tasty? Some flavors may widely be assumed to be "safe bets", like chocolate seemingly enjoying popularity, but who gets to provide this assessment? Imperial history may have imposed particular cuisine aspirations upon us, but what is realistically considered delicious will vary drastically. When considering geographical and cultural influences, it will inevitably be impossible to single out one taste. Resorting to

---

[3]We note that the baking of the metaphorical AI cake is not only unsustainable because of the baking process itself, but is also subject to sustainability considerations regarding the equipment required for baking, similar to our sustainability concerns regarding data ingredients. The scarcity of essential hardware, particularly GPUs, concentrates power among a few actors. Furthermore, manufacturing these components requires rare minerals, often sourced from conflict regions, adding another layer of sustainability concerns to AI development.

stereotypical hyperbole for clarity, strong sweetness may for instance be desirable in portions of the world, but much less strongly desired in other parts. Certainly, the notion of an average delightful cake itself is a poor simplification even within one region. After all, every human has their own flavor preferences; preferences that are uniquely shaped by their upbringing and further influenced by constraints such as food intolerances or allergies. And in the end, different parts of the world may also prefer entirely different deserts or be more cautious of a potentially unhealthy diet.

As much as it is impossible to define a best-tasting cake, so is it impossible to define the best-performing AI. Grappling with how to integrate these preferences is a recurring exercise throughout machine learning research, as demonstrated by the subjective nature of human feedback in learning preferences [86]. For instance, in large language modeling, a model's performance could typically be determined through a set of prescriptive benchmarks, which predominantly center common-sense reasoning [1, 148, 162], reading comprehension [41, 42, 128], or mathematical abilities [43, 96]. While these evaluations provide valuable insights into the capabilities of AI models, they fail to capture the full complexity of human preferences and the socio-technical implications of deploying these models in real-world contexts, including the impact on historically targeted, marginalized, or economically underprivileged groups. As such, gender bias literature is typically restricted to a binary lens, limiting critical discourse on AI harms deriving from a focus on binary gender [118, 125] and primarily male perspectives in AI systems [51, 117]. Likewise, the famous GenderShades audit [30] revealed that commercial AI systems often misclassified darker-skinned females [17]. A recent WIRED article [132] further illustrates the collapse to an expected average, showing how OpenAI's video generator Sora [27] defaults to depicting bisexual, asexual, or transgender persons with pink hair — highlighting the need for increased demographic representation and intersectional measurement practices in the baking process. These challenges are compounded by the influence of those who decide which AI technologies are put into practice and how acceptable performance is defined — i.e., what the baker determines as the average taste gives rise to power [22]. On the other hand, expanding a baker's palette to make a tastier "AI" cake often presents a challenging and sometimes contradictory undertaking, as evidenced by the more than 21 definitions of fairness [111]. Achieving the latter simultaneously proves impossible, especially when group and individual fairness conflict [16, 87]. At the same time, the social implications of each are unclear, as each method brings its own unique set of fairness challenges [35, 101]. In turn, the research community has proposed over 70 AI fairness evaluation guidelines [12].

Overall, both real cake and its AI analogy thus hinge on subjective and even incompatible notions of being delightful. Just as a bakery attempts to cater to the average customer's preferences, AI systems often rely on a limited understanding of typicality or correctness, leading to a collapse of diversity into oversimplified norms. Faced with complexity and abundance of taste assessment, AI bakers may opt for measures that are easy to satisfy or superficially appealing. This facilitates "ethics shopping" or "ethics bluewashing" [58], respectively cherry-picking what can be satisfied and using superficial measures in favor of positive appearance.

2.5 "If they do not have bread, let them eat cake!" — Sharing and (over-)selling

The reader may finally chuckle when reading about the metaphor's last connection to the above infamous quote, attributed to Marie Antoinette (and likely stated by Jean-Jacques Rousseau in 1765). But like any product, cake is eventually shared. More likely, it will be sold in order to make up for the initial cost and earn enough profit to make a living. Initially, this may have a naive benevolent intention, much like in our 18th-century anecdote, but there certainly is little use to suggest a starving population bake cake. After all, as elaborated in the prior sections, they will lack access to ingredients, lack understanding of the recipe, lack the tools to bake, or derive little nutritional value from the cake.

As much as the starving population of the 18th century had more basic needs than cake, so does much of the current population not actually profit from claimed AI progress. On the one hand, this may partially be due to a severe mismatch between the generally sold capability, and what is practically specified in the AI creation process. For instance, "diverse representation" is frequently advocated for greater inclusion of underrepresented groups in datasets, yet oversimplified notions lead to objectification or exploitation [14, 40]. Similarly, common error rate measures often lead to misleading promises on fair systems, given that assessment does not account for ultimately entailed effects [45]. This leads to treatment and impact disparity, where correlations arise unintentionally and outcomes start to differ across groups [99]. On the other hand, falsely promised profit may also be due to a growing belief in technosolutionism: the belief that technology is always the solution [28]. In this belief, an algorithm's role in selective targeting may simply be neglected at the prospect of later improvement. As such, periodic examples of how tech exacerbates inequality can be found. In fact, it generally seems that many AI contributions that were sold with some initial notion of good in mind, ultimately ended up fostering undesired population surveillance [81], all the way to the latter being used to AI-enabled direct persecution [70]. Sometimes the technosolutionist narrative's harm may be invisible on the surface, especially in scenarios where the AI cake is oversold to users that presently don't require it at all. An intuitive example of this pattern is the frequent marketing of AI as an opportunity to establish food security and enhance health care, in particular on the African continent [8]. Even if the potential benefits of AI are enormous, it is also challenging to reap them while ignoring the structural inequalities that need to be overcome before AI can actually live up to its sold promise [6]. Dynamic power relations between countries are not easily resolved by AI and general power imbalance cannot be overlooked. For instance, the US and EU heavily subsidize agriculture to export to African consumers [89] and big tech corporations own an overwhelming amount of infrastructure. Even when local grassroots progress is thus made, external parties and companies end up reaping many of its benefits - a pattern referred to as cooptation [108].

Overall, both distributing cake to a starving population and overselling AI systems hinge on an initially amicable incentive that fails to deliver on its intended goal. Whereas suggesting to feed people with cake is a poor nutritional and mismatched solution, so too are AI systems frequently oversold beyond challenges they are capable of solving. Technosolutionism and the resulting practice to create terminology to fuel this belief, e.g., "foundation models" or "frontier AI" [71], shift away focus from actual current capabilities to long-term speculation. In turn, a habit of "ethics shirking" [2, 58] is facilitated, where less work on ethical aspects is conducted if less hypothetical reward is expected.

3 Technological underpinnings: Why it is difficult to change the AI cake

The previous sections have substantiated the AI cake analogy and have linked it to several concerning practices. We now revisit these ramifications and map where social outcomes are intertwined with and exacerbated by technical underpinnings. In particular, we posit that even if consensus on instrumental and normative angles to AI development existed, the present technical foundations make it tremendously challenging to translate benevolent aims into practice.

This is not to say that researchers should be exempt from any responsibility. On the contrary, we precisely wish to empower them with a thorough understanding of how their present technical choices imbue constraints, foster biases, and why fundamental (mathematical) properties at the heart of (statistical) machine learning imply significant barriers. Mirroring our earlier paper structure, we begin each subsection with a quote from a recent publication to exemplify one key technical hurdle underpinning the previously described challenges, before proceeding to disentangle its technical caveats. We respectively finalize each subsection with a set of technical recommendations for future work, to highlight cross-disciplinary opportunities alongside existing social and ethical research advances. Figure 2 provides a summary.
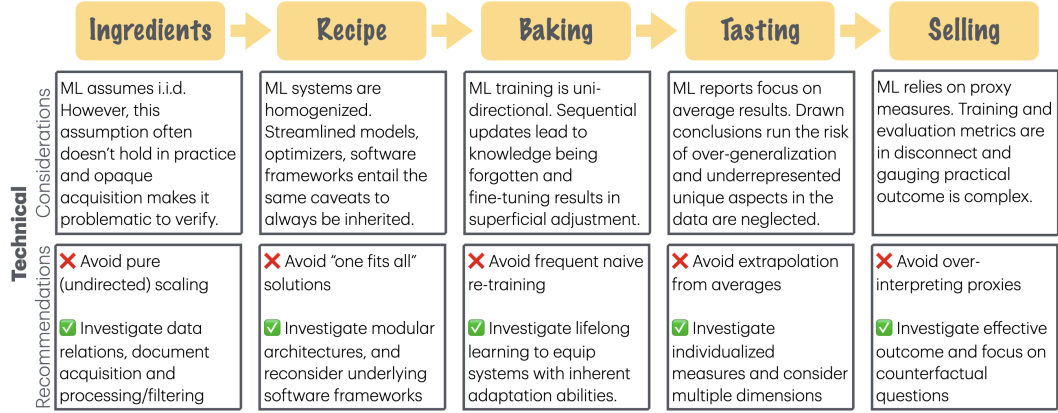
Fig. 2. An overview of section three's considerations with respect to the AI-cake's technical foundation and our future recommendations for each process stage from ingredients and recipes, to baking, tasting and selling.

3.1 "Ingredients": the challenge of data "i.i.d.-ness"

> *"Our results show that hate content increased by nearly 12% with dataset scale, measured both qualitatively and quantitatively using a metric that we term as Hate Content Rate.... This, as we hypothesize, may be a consequence of rich non-i.i.d. inter-sample correlations emerging from a graph-structured prior for CommonCrawl"* - Birhane et al. [19]

The above quote refers to the scaling of data examples from the LAION-400m [141] to the LAION-5b [140] dataset (respectively containing 400 million and 5 billion data points), attributing increasingly hateful content to the lack of understanding of non-i.i.d. correlations in the data selection and filtering mechanisms. More formally, i.i.d. refers to "independent and identically distributed". The U.S. National Institute of Standards and Technology (NIST) provides a concise definition: "A quality of a sequence of random variables for which each element of the sequence has the same probability distribution as the other values, and all values are mutually independent" [153]. In other words, we assume each data point to be different from and unaffected by others, while all data points are expected to originate from a common data generation process. Intuitively, this further entails "exchangeability", i.e., the notion that we can exchange the order of our data points in practice — a property that will also be central to our later learning recipes.

Naturally, there are various ways in which the i.i.d. assumption won't hold in the real world. Unfortunately, violations will mostly occur in unknown ways, with each respectively obscuring our understanding of the gathered data. An easy example is if the acquired dataset selectively contains (near) duplicates. This has recently occurred for up to 30% of LAION-2b [160], but can be dealt with effectively through various statistical tests [78]. It becomes significantly more challenging when a) there are complex inter-dependencies between subsequent data points, b) the distribution changes over time and becomes non-stationary, c) the data is adversarially crafted [46]. Although the latter scenarios are well-known to be realistic, the i.i.d. assumption is rooted deeply in our algorithms for pattern recognition and in statistical learning theory [156]. It is a key requirement to render many algorithms practical, in particular in the prevalent machine learning angle to AI. Specifically, i.i.d.-ness of data is presumed because it provides a required simplification of the likelihood function, the essential tool underlying learning/optimization procedures:

$$\mathcal{L}(\theta) = p(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}|\theta) = p(\mathbf{x_1}|\theta)p(\mathbf{x_2}|\theta)\ldots p(\mathbf{x_N}|\theta) \tag{1}$$

Here, the i.i.d. assumption is critical because it allows to convert an initially complicated to calculate function, where we wish to infer parameters θ that describe a full set of data points $\mathbf{x}_n$, into a product of terms based on respective individual data points. In fact, if we move to a logarithmic space, the product even becomes a simple sum:

$$\log \mathcal{L}(\theta) = \log p(\mathbf{x_1}|\theta) + \ldots + \log p(\mathbf{x_N}|\theta) \tag{2}$$

In turn, computation becomes manageable at the cost of neglecting data inter-dependencies. In fact, even in scenarios where such inter-dependencies are directly obvious, for instance, in the naturally temporally ordered data streams of reinforcement learning, a host of tricks are introduced to leverage i.i.d.-ness. For the particular example, it is typically experience replay that is used, where a subset of old observations is stored in a memory buffer that then resembles a typical i.i.d. dataset [133]. As such, the oversimplification of datasets and the equal mixture of all data ingredients is a direct result of the steps seemingly necessary to render computation feasible. The lack of technical tools to discover independence and discover causal relations at scale [138], coupled with the fact that real-world datasets are frequently inaccessible or in-transparently acquired, exacerbate the opaqueness.

Recommendations — *towards understanding and tracing of ingredients:*
A central recommendation is to avoid over-reliance on the "unreasonable effectiveness of data set scale" [147]. As evident from the critical history of, e.g., ImageNet [49, 50], datasets require change over time. Whereas outliers in the form of mislabeled examples may pass as i.i.d. by reducing their prediction variance through large training sets, systematic correlation will reciprocate systemic bias at any scale [52]. It is, thus, essential to approach datasets more carefully than random scraping and to meticulously document the processes surrounding its acquisition, e.g., using resources such as data statements [13, 61]. If the i.i.d. assumption is computationally necessary, then additional processing - such as clustering and removal - should be performed to warrant the assumption. These early initiatives should be extended to encompass detailed analysis of non-i.i.d. relations and enable digital data tracing. This will improve understanding of included data in a system, both in terms of technical assumptions and to expose malicious practices, such as stealing.

3.2 "Recipes": Homogenization of instructions

> "Though foundation models are based on standard deep learning and transfer learning, their scale results in new emergent capabilities, and their effectiveness across so many tasks incentivizes homogenization. Homogenization provides powerful leverage but demands caution, as the defects of the foundation model are inherited by all the adapted models downstream." - Bommasani et al. [23]

The above quote emphasizes the increasing homogenization as a result of deep learning's success at scale, but also advises caution with respect to always inheriting the same caveats. The latter is primarily driven by the seeming possibility to employ a heavily standardized framework of deep models, nowadays transformer [157] based foundation models, and optimization through backpropagation [94, 135, 161]. As deep neural networks are known to be universal approximators [76], this fosters a narrative of "one model to learn them all" [80, 130]. Although homogenization offers several benefits in terms of accessibility and ease of use, it suffers from oversimplifying and locking in the recipe.

From a technical perspective, it is a result of modeling any modern neural network as a cascade of layers $l = 1, \ldots, L$ that map from an initial dimension $d_{l-1} \in \mathbb{N}_{l-1}$ to a resulting arbitrary dimension $d_l \in \mathbb{N}_l$ through a progressive set of transformations $T_l : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}$ to assemble a "hypothesis" (model) h_θ:

$$h_\theta(\mathbf{x}) = T_L(T_{L-1}(\ldots(T_1(\mathbf{x})))), \quad x \in \mathbb{R}^d \tag{3}$$

This hypothesis is then learned to match the data evidence with a common set of optimization algorithms belonging to the (stochastic) gradient descent family, such as the tremendously popular Adam optimizer [83]. Concerns over the limitations, for instance, the fear of overfitting to the data, i.e., the technical equivalent of "learning by heart", seem to be surmounted by homogenization at scale through the double descent phenomenon [109][4]. In addition to the social implications of the homogenized recipe —recall "we cannot just add diverse end-users and stakeholders and stir" [47]— the entailed technical limitations of this homogenization can best be understood in its manifestation in programming frameworks. Over time, the original differences of pioneering machine learning software frameworks, such as Theano [149] and Torch7 [44], have converged in terms of functionality. As such, whereas there may be syntactical differences and functionality nuances, current age software —e.g. the prevalent Pytorch [121], Tensorflow [3], and Jax [26]— is severely streamlined towards deep neural networks. This makes breaking out of homogenized recipes more than just challenging, as noted recently by a work with the provocative thesis "ML is stuck in a rut" [11]. In this work, they show that an attempt at more modular, more brain-like architectures, for instance, the seminal Capsule network [136], underperformed the homogenized recipes. Alas, the catch is that this is not necessarily due to inferior design, but rather a direct function of programming frameworks being excessively tailored to homogenized deep learning recipes. Despite employing very similar mathematical functions (i.e. convolutions), Capsule networks thus appeared significantly worse than they practically should be. This tight coupling between a narrative that a homogenized recipe is "all we need" [23, 157] with focused tailoring of software frameworks exacerbates the challenge of creating any other solutions. As such, overcoming the challenges of our homogenized recipe is not only conceptually challenging but also exacerbated on a technical level through excessive convergence of underlying tools.

Recommendations —*towards unique and customizable recipes:*
A central recommendation is to avoid over-reliance on homogenized foundation models with streamlined software. Although there is potential for rapid applications, the underlying convergence disincentivizes the development of breakthroughs in understanding or effective small-scale solutions. As such, we recommend research on modular layering and inspectable representations, for instance by furthering theoretical understanding of information flow [137, 142] instead of hoping for "emergence" of capabilities. For a more radical recommendation, we recommend questioning whether a homogenized model is suited to individual needs and considering whether auxiliary non-data-driven methods may be advantageous. Any such alternative approaches are not easy to implement in existing software frameworks, similar to our earlier Capsule reference. To give but one example, we can draw inspiration from neurogenesis in mammalian brains [67, 154] - shrinking and growing neurons on the fly - and translate this ability to dynamically adapt neural structures to the task at hand [10, 54, 103], requiring diversified development of programming frameworks.

3.3 "Baking process": the interference dilemma

> "Human learning has evolved to thrive in dynamic learning settings. However, this robustness is in stark contrast to the most powerful modern machine learning methods, which perform well only when presented with data that are carefully shuffled, balanced, and homogenized." - Hadsell et al. [68]

The above quote emphasizes how fluid intelligence allows humans to excel in dynamic environments through sequential adaptation and progressive refinement of their knowledge [56, 57]. In contrast, the strength of (large) machine learning models is only apparent through crystallization of carefully balanced and homogenized data. Whereas it is significantly

---

[4]The observation that training with substantial data amounts for prolonged periods of time eventually overcomes any initial performance deterioration

easier for humans to continually learn "the n-th thing" after learning "$n-1$ things" [68, 150], human-like knowledge transfer in machines remains a challenging desideratum [91, 119].

From a technical perspective, the lack of ability to learn continually entails an exorbitant amount of retraining. Yet, this re-training cost is willingly embraced, as the unfortunate alternative is induced catastrophic forgetting [102, 129]. The latter refers to the phenomenon of abruptly losing acquired information when sequentially tuning to new data. Unfortunately, the reasons for this phenomenon are deeply ingrained in our optimization toolkit, where one culprit is the iterative nature of our optimizers [152]. At the example of our homogenized recipe's stochastic gradient descent, steps τ of updates to parameters θ are conducted in a loss function across observed data points with a "learning rate" η:

$$\text{for } \tau = 1, 2, \ldots, t \text{ do}: \quad \theta \leftarrow \theta - \eta \nabla \mathcal{L}_\tau(\theta) \tag{4}$$

Intuitively, this works well if we present the optimizer with all the concepts we ultimately care about. The optimizer will "move in directions" that satisfy all observed concepts and progressively improve. However, it is now similarly unsurprising that if presented with a novel fraction of data at a later time step, this optimizer will move uni-directionally to improve on only what it has recently seen. This challenge is significantly exacerbated by the semi-distributed nature of representations in neural models [59]. Although entangled representations foster generalization across concepts by moving away from a look-up table, they also imply that most any update has an influence on every learned concept so far. The field of continual machine learning [68, 105] aims to overcome this central limitation. Alas, it is presently bounded by our homogenized model and optimizer recipes. If we cannot leave the frame, the metaphorical cake can only be changed superficially post-hoc or re-baked fully when we wish to make major additions or changes. Learning becomes a (Markov) chain that only takes into account the most recent past. It does not explicitly take into account all prior observations, making any addition either superficial, or potentially catastrophic, if the desired change was not part of the original data mix.

Recommendations — *towards adaptive and collaborative learning:*
A central recommendation is to avoid unsustainable re-training advice, e.g. Google Cloud's [66] "Developing a model is a process of experimentation and incremental adjustment. You should expect to spend a lot of time refining and modifying your model to get the best results." Although deep learning representations have historically been argued to rival primate IT Cortex (for vision) [33], their ability to "embrace change" [68] certainly falls far behind that of any human [39, 56]. In turn, we recommend drawing inspiration from lifelong learning mechanisms to equip AI with efficient adaptation capabilities, absolving us from high re-training costs. On the one hand, this requires translating the plethora of biological capabilities known to contribute to lifelong learning to artificial systems [91] and establishing respective AI theory [124]. On the other hand, a revisit to neuro-symbolic systems, that in parts were able to adapt sustainably throughout their entire life-cycles by interfacing learning with reasoning [34, 38, 104], may be warranted.

3.4 "Taste": limits of averages and aggregates

> "Some pitfalls are only visible when examining the dataset as a whole or the proposed aggregating metrics. Since what the benchmarks aim to measure is not well articulated, it can be difficult to distinguish whether and when the pitfalls we list below suggest a poor conceptualization of stereotyping or instead call into question the way it is operationalized" - Blodgett et al. [21]

The above quote, originally written in the context of the misattribution of stereotypes, points to issues in the interpretation of averaged assessment. It implies that an AI system's pitfalls may become obfuscated, for instance, an

underrepresented group not being adequately covered, and that important nuances may typically be neglected, such as through aggregating fairness metrics [35]. The present over-reliance on aggregate measures [31] respectively makes it challenging to predict practical behavior in real situations outside benchmark datasets.

From a technical perspective, the popularity of average assessment is not only driven by the scientific literature's seeming necessity to compare single benchmark numbers, but the challenges are once more rooted deeply in the algorithmic underpinnings themselves. Following earlier equation 2, which highlighted the significance and prevalence of the i.i.d. assumption, evaluation losses ($\mathcal{L}$) and measures are averaged in equal weighting of all data points $n = 1, \ldots, N$:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_n(\theta) \tag{5}$$

This assumption is crucial when coupled with the prevalent iterative learning algorithms. In particular, it makes textbook machine learning analysis readily available [20], where training, validation, and test data splits are compared to understand whether the model "generalizes" in evaluation. Unfortunately, such assessment of generalization is also limited to what is well represented and practically assessable in the split [163]. Similarly, popular algorithms such as variational inference in autoencoders [84, 85] rely on mean-field theory, e.g., measuring divergences to the mean and standard deviation of a (Normal) distribution to learn about the data generating process.

The rooting of averages in these machine learning foundations entails several problematic technical factors. First, relying on aggregate measures exacerbates overconfident (false) predictions. In fact, models themselves are typically trained to give maximal predictions (or minimum losses) and as such seldom give out anything that is far away from the observed average value (which in labeled scenarios is typically a 100% confidence of a category) [72, 116] [5]. Second, relying too much on averaging in generative modeling can either "smooth out" the diversity represented in the data, e.g., we create an envelope around two distribution modes, or approximate a particular mode well at the expense of dropping another, i.e., mode collapse [90]. Finally, averaged measures as targets for evaluations incentivize the conception of systems that leverage shortcuts. For instance, the popular example of CleverHans predictors [93] has shown that an average accuracy measure does not allow us to distinguish whether images of a horse are classified correctly because they indeed contain a horse, or because they contain a different, potentially unidentifiable common feature (here, a photographer's tag). These confounders are particularly problematic in non-i.i.d. scenarios, where certain groups of features can become over or underrepresented at specific points in time [32], relating back to our section two's Gemini example. Aggregate measures thus make machine learning seem straightforward to evaluate and allow us to compare models, but the fact that these are imbued in our fundamental technological stack also severely limits prospective assessment.

Recommendations —*towards individualized socio-technical assessment:*
A central recommendation is to avoid drawing general conclusions from assessments relying on averaged metrics. Although it seems convenient to report a single metric and be able to compare it to other works in the literature, the assertion of the average as being correct legitimizes the agenda of the majority [73]. For AI to find diverse applications, it must shift power to include what is different from the average. From a technical perspective, this may include steering away from classically employed statistics [159], to the analysis of individual samples or extreme values [25], or to re-investigate tilted loss functions [95]. The latter exist in theory but are seldom used. We additionally recommend the use of multi-dimensional evaluation strategies, e.g., in the form of "compasses" [62, 106]. These tools allow inspection of several dimensions of interest and should continue to be developed for transparent socio-technical assessment.

---

[5]The observation of misleading overconfident predictions and the entailed false sense of evaluation correctness has empirically been made for both discriminative and generative models [110] and further for model types beyond neural networks, such as probabilistic circuits [158].

3.5 "(Over-)Selling": the abundant surrogate impasse

> "Machine learning models routinely achieve perfect performance in one dataset even when that dataset is a
> large international multisite clinical trial. However, when that exact model was tested in truly independent
> clinical trials, performance fell to chance levels. Even when building what should be a more robust model
> by aggregating across a group of similar multisite trials, subsequent predictive performance remained poor."
> Chekroud et al. [36], editor summary - Peter Stern.

The above quote highlights strong concerns over practical AI system deployability, despite large-scale trials. Although the quote also hints at our earlier section's issue with aggregates, we posit that an additional technical aspect is overlaid. We term this challenge the "abundant surrogate impasse", pointing to a lack of understanding of how optimization goals relate to practical measurement. In turn, we posit that even the most rigorous approach is technologically challenged in its assessment of practical implications, contributing to the overselling of our AI systems beyond human intent.

From a technical perspective, almost any machine learning system needs to be optimized via a proxy. This is both motivated by the fact that we require smooth and differentiable loss functions to obtain learning signals [107] (e.g. turning a categorical 0 or 1 signal in classification into a spectrum between 0-1), and the fundamental limitation that we can rarely express our goal directly mathematically. Take for instance two prominent recent advances, generative vision models and large language models. In the former, we wish to train a model that is capable of faithfully generating a diverse set of images, yet it is unclear how to express this goal directly. As such, it is frequent, for instance in auto-encoding based models, to minimize the discrepancy in pixel values of an original $\mathbf{x}$ and a reconstructed image $\hat{\mathbf{x}}$:

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} ||\hat{\mathbf{x}}_n - \mathbf{x}_n||_2^2 \quad \text{where} \quad \hat{\mathbf{x}}_n = h_\theta(\mathbf{x}_n) \tag{6}$$

Similarly, in large language models, we wish to accurately model language, yet we don't explicitly encode linguistic rules or semantic coherence as training objectives. Instead, the standard approach involves tokenizing text into discrete units, treating each token from the vocabulary as a distinct class, and training the model to predict the next token ($\mathbf{x}_{t+1}$) in a sequence ($\mathbf{x}_{t:1}$) through maximum likelihood estimation. This seemingly simple objective of next-token prediction serves as a proxy for learning deeper linguistic patterns and relationships.

$$\mathcal{L}(\theta) = -\sum_t \log p_\theta(\mathbf{x}_{t+1}|\mathbf{x}_{t:1}) \tag{7}$$

In both cases, these training objectives and loss functions are used to optimize the system and evaluate its performance on held-out test data. However, when assessing real-world generation capabilities, we typically sample new outputs and evaluate their quality. This assessment requires fundamentally different metrics - perceptual scores for images [82], reference-based [98, 120] and reference-free metrics for languages [164].

Thus, it strikes us as unsurprising that AI systems are commonly oversold. Some real-world concepts remain unmeasurable or their complexity cannot be measured through a single value. A single loss proxy, where we are unable to express our true goal, is used for training, and a set of different measures is brought out in evaluation. In turn, the narrow optimization focus lacks relation to the actual world, but our often desired abstract concepts are hard, if not impossible, to conceptualize mathematically. The discrepancy between what is being optimized for, what is desired as the outcome, and what is being evaluated, can lead to questionable conjectures of a system's capabilities.

Recommendations —*towards meaningful assessment of outcome and impact:*
Our final central recommendation is to avoid generalized conclusions originating from evaluation of limited proxies.

We recommend particular caution in modern systems where optimization objectives and evaluation measures are mismatched. In addition, we raise awareness for the fact that we no longer know if data contamination has already provided the system with answers during training [24], or the reality that much of what is categorized should be considered fluid and complex (social) constructs, i.e., the employed proxy should not have been used from the start. Instead of drawing wide conclusions from measured proxy values, we recommend to focus attention on effective outcomes, for instance, akin to a "social impact dashboard" [62]. This in turn requires a shift away from learning and assessing correlations, to gaining understanding in the form of "Were that factor different, would the outcome change?" [123]. Causality is but one field that attempts to provide answers for such counterfactual questions. However, more research is necessary to understand when conditions for causal analysis and interventions are realizable in practice [92].

4 Limitations and Disclaimer

> *"Most people think the issue is changing social norms. It is, but it's also (. . .) how willing engineers are to change those systems."* Meredith Broussard [29]

Our work does not claim that social challenges will eventually be overcome through advances in AI design alone. On the contrary, resonating with above quote, we believe that making AI systems more socially and technologically sustainable, requires *both* the willingness for people to change social norms and to reconceive technical tools to allow this change. With this in mind, there exist a plethora of societal and ethical aspects that our re-conceptualization of cake as a metaphorical AI system has either referenced only in passing or is unable to cover. This is because our work's focus is on pointing out prevalent ramifications and how they are limited by foundational technical underpinnings. As such our work serves the primary purpose to raise awareness of how social and technical elements are inevitably intertwined and, through mapping of social outcomes to technical foundations, create opportunity for different communities to engage in cross-disciplinary dialogue along actionable dimensions. However, we acknowledge that this approach, despite highlighting important new avenues, is also inevitably bound by the cake analogy as a frame of reference. In particular, we understand that the idea of baking a cake, whether reconceived or not, may not be what is desirable on many occasions. To re-emphasize our portions on (over-)selling, there is a crucial difference between naive over-claiming of capabilities and deliberately deploying AI in unnecessary or harmful places. In a similar spirit, our work remains subject to a host of further factors, including, but not limited to, monetary incentives, pressure to publish, the challenges of the academic reviewing system, and imbalanced power dynamics.

5 Conclusion

We have detailed how the process of making a cake serves as a comprehensive analogy to the design of modern AI systems. Through several drawn parallels across the full AI life cycle, we have highlighted why change towards more sustainable and collaborative AI systems is not merely a social challenge, but how it is also constrained by prevalent technical foundations. Ultimately, our analysis and recommendations thus call for a departure from the traditional ML textbook narrative [20, 64, 107] to equally focus on process improvements and collective exploration [18]. However, such participatory practice [9] that spans *consultation, inclusion, collaboration* and *ownership* seems to presently be difficult to implement in practice [5, 48] and must also be embraced technologically. We believe that our re-conceptualization of the AI cake analogy serves as an opportunity to engage in deeper cross-disciplinary dialogue towards this goal. As such, we envision future research to continue the analysis of the ties between technical limitations and social ramifications, eventually re-imagining the present process of baking our metaphorical AI cake.

References

[1] 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale.

[2] 2024. Shirking. https://www.nasdaq.com/glossary/s/shirking. Last accessed: 2025-01-12.

[3] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI* (2016), 265–283.

[4] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *Tech report arXiv:2404.14219* (2024).

[5] Ada-Lovelace-Institute. 2021. Participatory data stewardship. Technical report available at: https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/.

[6] Abejide Ade-Ibijola and Chinedu Okonkwo. 2023. *Artificial Intelligence in Africa: Emerging Challenges.* Springer International Publishing, 101–117.

[7] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abeel, and Wojciech Zaremba. 2017. Hindsight Experience Replay. *Neural Information Processing Systems (NeurIPS)* (2017).

[8] Emmanuel Ogiemwonyi Arakpogun, Ziad Elsahn, Femi Olan, and Farid Elsahn. 2021. *Artificial Intelligence in Africa: Challenges and Opportunities.* Springer International Publishing, 375–388.

[9] Sherry R. Arnstein. 1969. A ladder of citizen participation. *Journal of the American Institute of planners* 35 (1969), 216–224. Issue 4.

[10] Timur Ash. 1989. Dynamic Node Creation in Backpropagation Networks. *Connection Science* 1, 4 (1989), 365–375.

[11] Paul Barham and Michael Isard. 2019. Machine Learning Systems are Stuck in a Rut. *Proceedings of the Workshop on Hot Topics in Operating Systems (HotOS)* (2019).

[12] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63 (2019).

[13] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[14] A. Stevie Bergman, Lisa Anne Hendricks, Maribeth Rauh, Boxi Wu, William Agnew, Markus Kunesch, Isabella Duan, Iason Gabriel, , and William Isaac. 2023. Representation in AI Evaluations. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2023), 519–533.

[15] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. [n. d.]. OpenAI report: ImprovingImageGenerationwithBetterCaptions.

[16] R. Binns. 2020. On the apparent conflict between individual and group fairness. *Proceedings of the Conference on Fairness, Accountability and Transparency (FAccT)* (2020).

[17] Abeba Birhane. 2022. The unseen Black faces of AI algorithms. *Nature* 610 (2022), 451–452.

[18] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)* (2022).

[19] Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. 2023. Into the LAIONs Den: Investigating Hate in Multimodal Datasets. *Neural Information Processing Systems (NeurIPS)* (2023).

[20] Christoper M. Bishop. 2006. *Pattern Recognition and Machine Learning.* Springer.

[21] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2021), 1004–1015.

[22] Rishi Bommasani. 2022. Evaluation for Change. *Preprint arXiv:2212.11670* (2022).

[23] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *Preprint arXiv:2108.07258* (2021). https://crfm.stanford.edu/assets/report.pdf

[24] Ali Borji. 2023. A Categorical Archive of ChatGPT Failures. *Preprint arXiv:2302.03494* (2023).

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

[26] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. JAX: composable transformations of Python+NumPy programs: http://github.com/google/jax.

[27] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video Generation Models as World Simulators. https://openai.com/index/video-generation-models-as-world-simulators/.

[28] Meredith Broussard. 2019. *Artificial Unintelligence.* The MIT Press.

[29] Meredith Broussard. 2023. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech.* MIT Press.

[30] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the Conference on Fairness, Accountability and Transparency (FAccT)* (2018).

[31] R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, D. Kiela, M. Shanahan, E. M. Voorhees, A. G. Cohn, J. Z. Leibo, and J. Hernandez-Orallo. 2023. Rethink reporting of evaluation results in AI. *Science* 380 (2023).

[32] Florian Peter Busch, Roshni Kamath, Rupert Mitchell, Wolfgang Stammer, Kristian Kersting, and Martin Mundt. 2024. Where is the Truth? The Risk of Getting Confounded in a Continual World. *arXiv preprint arXiv:2402.06434* (2024).

[33] Charles F. Cadieu, Ha Hong, Daniel L.K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. 2014. Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology* 10, 12 (2014).

[34] Andrew Carlson, Justin Betteridge, and Bryan Kisiel. 2010. Toward an Architecture for Never-Ending Language Learning. *AAAI Conference on Artificial Intelligence (AAAI)* (2010).

[35] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Nature Scientific Reports* 12 (2022).

[36] A. M. Chekroud, M. Hawrilenko, H. Loho, J. Bondar, R. Gueorguieva, A. Hasan, J. Kambeitz, P. R. Corlett, N. Koutsouleris, H. M. Krumholz, J. H. Krystal, and M. Paulus. 2024. Illusory generalizability of clinical prediction models. *Science* 383 (2024).

[37] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009* (2023).

[38] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. 2013. NEIL: Extracting visual knowledge from web data. *Proceedings of the IEEE International Conference on Computer Vision (CVPR)* (2013).

[39] Zhiyuan Chen and Bing Liu. 2018. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 12, 3 (2018), 1–207.

[40] Jennifer Chien and David Danks. 2024. Beyond Behaviorist Representational Harms: A Plan for Measurement and Mitigation. *Proceedings of the Conference on Fairness, Accountability and Transparency (FAccT)* (2024), 933–946.

[41] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[42] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL*.

[43] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[44] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. 2011. Torch7: A matlab-like environment for machine learning. *Neural Information Processing Systems (NeurIPS), BigLearn Workshop* (2011).

[45] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The Measure and Mismeasure of Fairness. *Journal of Machine Learning Research (JMLR)* 24 (2023), 1–117.

[46] Trevor Darrell, Marius Kloft, Massimiliano Pontil, Gunnar Rätsch, and Erik Rodner. 2015. Machine Learning with Interdependent and Non-identically Distributed Data (Dagstuhl Seminar 15152). *Dagstuhl Reports* (2015), 18–55.

[47] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2021. Stakeholder Participation in AI: Beyond "Add Diverse Stakeholders and Stir". *Human-Centered AI workshop at NeurIPS* (2021).

[48] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)* (2023).

[49] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition (CVPR)* (2009).

[50] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet. *Big Data and Society* (2021).

[51] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084* (2021).

[52] Murat Dundar, Balaji Krishnapuram, Jinbo Bi, and R Bharat Rao. 2007. Learning Classifiers When the Training Data Is Not IID. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*.

[53] European-Commission. 2015. Horizon 2020 Work Programme 2014-2015, Science with and for Society. https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en.

[54] Utku Evci, Bart van Merrienboer, Thomas Unterthiner, Fabian Pedregosa, and Max Vladymyrov. 2022. GradMax: Growing Neural Networks using Gradient Information. *International Conference on Learning Representations (ICLR)* (2022).

[55] Fabian Ferrari, Jose van Dijck, and Antal van den Bosch. 2023. Foundation models and the privatization of public knowledge. *Nature machine intelligence* 5 (2023).

[56] Timo Flesch, Jan Balaguer, Ronald Dekker, Hamed Nili, and Christopher Summerfield. 2018. Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences of the United States of America PNAS* 115 (2018). Issue 44.

[57] Timo Flesch, David G. Nagy, Andrew Saxe, and Christopher Summerfield. 2023. Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *PLoS Computational Biology* 19, 1 (2023).

[58] Luciano Floridi. 2019. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy and Technology* 32 (2019), 185–193.

[59] Robert M. French. 1992. Semi-distributed Representations and Catastrophic Forgetting in Connectionist Networks. *Connection Science* 4, 3-4 (1992), 365–377.

[60] Felix Friedrich, Katharina Hämmerl, Patrick Schramowski, Manuel Brack, Jindrich Libovicky, Kristian Kersting, and Alexander Fraser. 2024. Multilingual Text-to-Image Generation Magnifies Gender Stereotypes and Prompt Engineering May Not Help You. *Preprint arXiv:2401.16092* (2024).

[61] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for Datasets. *Preprint arXiv:1803.09010* (2018).

[62] Avijit Ghosh, Cedric Whitney, Yacine Jernite, and Irene Solaiman. 2024. Social Impact Dashboard. https://huggingface.co/spaces/evijit/SIMPDashboard. Accessed: 2025-01-07.

[63] Chris Gilliard. 2024. The Deeper Problem With Google's Racially Diverse Nazis: Generative AI is not built to honestly mirror reality, no matter what its creators say. The Atlantic: https://www.theatlantic.com/technology/archive/2024/02/google-gemini-diverse-nazis/677575/.

[64] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

[65] Google. 2024. Google Gemini. https://gemini.google.com.

[66] Google-Cloud. 2023. Google Cloud guides: ML solutions overview, the ML workflow. https://cloud.google.com/ai-platform/docs/ml-solutions-overview. 2024-03-20.

[67] Charles G. Gross. 2000. Neurogenesis in the adult brain: Death of a dogma. *Nature Reviews Neuroscience* 1, 1 (2000), 67–73.

[68] Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. 2020. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences* 24, 12 (2020), 1028–1040.

[69] Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, and He Tang. 2024. Synthetic Data in AI: Challenges, Applications, and Ethical Implications. *Preprint arXiv:2401.01629* (2024).

[70] Drew Harwell and Eva Dou. 2020. Huawei tested AI software that could recognize Uighur minorities and alert police.

[71] Gina Helfrich. 2024. The harms of terminology: why we should reject so-called "frontier AI". *AI and Ethics* (2024).

[72] Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)* (2017).

[73] Ferdinand A. Hermens. 1958. The Tyranny of the Majority. *Social Research* 25 (1958), 37–52.

[74] Alex Hern. 2024. TechScape: How cheap, outsourced labour in Africa is shaping AI English. The Guardian: https://www.theguardian.com/technology/2024/apr/16/techscape-ai-gadgest-humane-ai-pin-chatgpt.

[75] Dominik Hintersdorf, Lukas Struppek, Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. 2022. Does CLIP Know my Face? *Preprint arXiv:2209.07341* (2022).

[76] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* (1989).

[77] The White House. 2023. FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. https://resourcecenter.ieee.org/government/usa/cisgovph0110. Last accessed: 2025-02-02.

[78] Marcus Hutter. 2022. Testing Independence of Exchangeable Random Variables. *preprint arXiv:2210.12392* (2022).

[79] Mojan Javaheripi and Sebastien Bubeck. 2023. Phi-2 the surprising power of small language models. Microsoft Blog: https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

[80] Lukasz Kaiser, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. One Model To Learn Them All. *Prerint arXiv:1706.05137* (2017).

[81] Pratyusha Ria Kalluri, William Agnew, Myra Cheng, Kentrell Owens, Luca Soldaini, and Abeba Birhane. 2023. The Surveillance AI Pipeline. *Preprint arXiv:2309.15084* (2023).

[82] Remi Kazmierczak, Gianni Franchi, Nacim Belkhir, Antoine Manzanera, and David Filliat. 2022. A Study of Deep Perceptual Metrics for Image Quality Assessment. *arXiv preprint arXiv:2202.08692* (2022).

[83] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (2015).

[84] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

[85] Diederik P Kingma and Max Welling. 2019. An Introduction to Variational Autoencoders. *Foundations and Trends in Machine Learning* (2019).

[86] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *arXiv preprint arXiv:2404.16019* (2024).

[87] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS)* (2017).

[88] John Koetsier. 2023. ChatGPT Burns Millions Every Day. Can Computer Scientists Make AI One Million Times More Efficient? Forbes: https://www.forbes.com/sites/johnkoetsier/2023/02/10/chatgpt-burns-millions-every-day-can-computer-scientists-make-ai-one-million-times-more-efficient/?sh=24571e676944. Last accessed: 2025-02-02.

[89] Leo Komminoth. 2023. Can AI address Africa's agricultural trade deficit?

[90] Agustinus Kristiadi. 2016. KL Divergence: Forward vs Reverse? https://agustinus.kristia.de/blog/forward-reverse-kl/

[91] Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, and Others. 2022. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence* 4, 3 (2022), 196–210.

[92] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. 2017. Counterfactual Fairness. *Neural Information Processing Systems (NeurIPS)* (2017).

[93] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* 10, 1 (2019).

[94] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. 2012. Efficient BackProp. In *Neural Networks: Tricks of the Trade: Second Edition*. Springer Berlin Heidelberg, Berlin, Heidelberg, 9–48.

[95] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. 2023. On tilted losses in machine learning: theory and applications. *Journal of Machine Learning Research* (2023).

[96] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050* (2023).

[97] Bill Yuchen Lin, Abhilasha Ravichandar, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. *International Conference on Learning Representations (ICLR)* (2024).

[98] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

[99] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. 2019. Does mitigating ML's impact disparity require treatment disparity? *Neural Information Processing Systems (NeurIPS)* (2019).

[100] Kasper Groes Albin Ludvigsen. 2023. The carbon footprint of GPT-4. https://towardsdatascience.com/the-carbon-footprint-of-gpt-4-d6c676eb21ae. Last accessed: 2025-02-02.

[101] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-biasing "bias" measurement. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2022), 379–389.

[102] Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks : The Sequential Learning Problem. *Psychology of Learning and Motivation - Advances in Research and Theory* 24, C (1989), 109–165.

[103] Rupert Mitchell, Robin Menzenbach, Kristian Kersting, and Martin Mundt. 2024. Self Expanding Neural Networks. *Preprint arXiv:2307.04526* (2024).

[104] T Mitchell, W Cohen, E Hruschka, P Talukdar, B Yang, J Betteridge, A Carlson, B Dalvi, M Gardner, B Kisiel, J Krishnamurthy, N Lao, K Mazaitis, T Mohamed, N Nakashole, E Platanios, A Ritter, M Samadi, B Settles, R Wang, D Wijaya, A Gupta, X Chen, A Saparov, M Greaves, and J Welling.

2015. Never-Ending Learning. *AAAI Conference on Artificial Intelligence (AAAI)* (2015).

[105]   Martin Mundt, Yongwon Hong, Iuliia Pliushch, and Visvanathan Ramesh. 2023. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *Neural Networks* 160 (2023), 306–336.

[106]   Martin Mundt, Steven Lang, Quentin Delfosse, and Kristian Kersting. 2022. CLEVA-Compass: A Continual Learning Evaluation Assessment Compass to Promote Research Transparency and Comparability. *International Conference on Learning Representations (ICLR)* (2022).

[107]   Kevin Patrick Murphy. 2012. *Machine Learning: A Probabilistic Perspective.* MIT Press.

[108]   Esther Mwema and Abeba Birhane. 2024. Undersea cables in Africa: The new frontiers of digital colonialism. *First Monday* 29 (2024). Issue 4.

[109]   Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2020. Deep Double Descent: Where Bigger Models and More Data Hurt. *International Conference on Learning Representations (ICLR)* (2020).

[110]   Eric Nalisnick, Akihiro Matsukawa, Yee W. Teh, Dilan Gorur, and Balaji Lakshminarayanan. 2019. Do Deep Generative Models Know What They Don't Know? *International Conference on Learning Representations (ICLR)* (2019).

[111]   A. Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. *Proceedings of the Conference on Fairness, Accountability and Transparency (FAccT)* (2018).

[112]   Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.

[113]   Noyb. 2024. ChatGPT provides false information about people, and OpenAI can't correct it. Noyb: https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it.

[114]   U.S. Deparment of Commerce. 2025. Biden-Harris Administration Announces Regulatory Framework for the Responsible Diffusion of Advanced Artificial Intelligence Technology. https://www.bis.gov/press-release/biden-harris-administration-announces-regulatory-framework-responsible-diffusion. Last accessed: 2025-01-14.

[115]   German Federal Ministry of Labour and Social Affairs. 2023. Supply Chain Act. https://www.bmas.de/EN/Europe-and-the-World/International/Supply-Chain-Act/supply-chain-act.html.

[116]   Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *Neural Information Processing Systems (NeurIPS)* (2019).

[117]   Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* 1246–1266.

[118]   Anaelia Ovalle, Krunoslav Lehman Pavasovic, Louis Martin, Luke Zettlemoyer, Eric Michael Smith, Adina Williams, and Levent Sagun. 2024. The Root Shapes the Fruit: On the Persistence of Gender-Exclusive Harms in Aligned Language Models. *arXiv preprint arXiv:2411.03700* (2024).

[119]   Sinno J. Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 22, 10 (2010).

[120]   Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.*

[121]   Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. *Neural Information Processing Systems (NeurIPS)* (2019).

[122]   David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon Emissions and Large Neural Network Training. *Preprint arXiv:2104.10350* (2021).

[123]   Judea Pearl. 2009. *Causality.* Cambridge Univrsity Press.

[124]   Diana Benavides Prado and Patricia Riddle. 2022. A Theory for Knowledge Transfer in Continual Learning. In *Conference on Lifelong Learning Agents.*

[125]   Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2023), 1882–1895.

[126]   Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and Others. 2019. Language models are unsupervised multitask learners. *OpenAI Blog: https://github.com/openai/gpt-2* (2019).

[127]   Prabhakar Raghavan. 2024. Gemini image generation got it wrong. We'll do better. Google Blog: https://blog.google/products/gemini/gemini-image-generation-issue/.

[128]   Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264 arXiv:1606.05250 [cs.CL]

[129] Roger Ratcliff. 1990. Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions. *Psychological Review* 97, 2 (1990), 285–308.

[130] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A Generalist Agent. *Transactions on Machine Learning Research* (2022).

[131] Adi Robertson. 2024. Google apologizes for missing the mark after Gemini generated racially diverse Nazis. The Verge: https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical.

[132] Reece Rogers. 2024. Here's How Generative AI Depicts Queer People. Wired: https://www.wired.com/story/artificial-intelligence-lgbtq-representation-openai-sora/. Last accessed: 2025-02-02.

[133] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. 2019. Experience Replay for Continual Learning. *Neural Information Processing Systems (NeurIPS)* (2019).

[134] Niamh Rowe. 2023. It's destroyed me completely: Kenyan moderators decry toll of training of AI models. The Guardian: https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai.

[135] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323 (1986), 533–536.

[136] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. *Advances in Neural Information Processing Systems (NeurIPS)* (2017).

[137] Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D. Tracey, and David D. Cox. 2018. On the Information Bottleneck Theory of Deep Learning. *International Conference on Learning Representations (ICLR)* (2018).

[138] B. Schölkopf*, F. Locatello*, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. 2021. Toward Causal Representation Learning. *Proc. IEEE* (2021).

[139] Doris Schroeder, Julie Cook, François Hirsch, Solveig Fenet, and Vasantha Muthuswamy. 2018. Ethics Dumping: Case Studies from North-South Research Collaborations. *SpringerBriefs in Research and Innovation Governance* (2018).

[140] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and Others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[141] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *Data Centric AI NeurIPS Workshop* (2021).

[142] Ravid Schwartz-Ziv and Naftali Tishby. 2017. Opening the black box of Deep Neural Networks via Information. *Preprint arXiv: 1703.00810* (2017).

[143] Innovation Secretary of State for Science and Technology by Command of His Majesty. 2023. Policy paper: A pro-innovation approach to AI regulation. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper. Last accessed: 2025-01-21.

[144] David S. So, Chen Liang, and Quoc V. Le. 2019. The Evolved Transformer. *International Conference on Machine Learning (ICML)* (2019).

[145] Valerie Strauss. 2018. Stephen Hawking famously said, 'Intelligence is the ability to adapt to change.' But did he really say it? The Washington Post: https://www.washingtonpost.com/news/answer-sheet/wp/2018/03/29/stephen-hawking-famously-said-intelligence-is-the-ability-to-adapt-to-change-but-did-he-really-say-it/. Last accessed: 2025-02-02.

[146] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), 3645–3650.

[147] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017).

[148] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4149–4158. https://doi.org/10.18653/v1/N19-1421 arXiv:1811.00937 [cs]

[149] The Theano Development Team, Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, Alexander Belopolsky, Yoshua Bengio, Arnaud Bergeron, James Bergstra, Valentin Bisson, Josh Bleecher Snyder, Nicolas Bouchard, Nicolas Boulanger-Lewandowski, Xavier Bouthillier, Alexandre de Brébisson, Olivier Breuleux, Pierre-Luc Carrier, Kyunghyun Cho, Jan Chorowski, Paul Christiano, Tim Cooijmans, Marc-Alexandre Côté, Myriam Côté, Aaron Courville, Yann N. Dauphin, Olivier Delalleau, Julien Demouth, Guillaume Desjardins, Sander Dieleman, Laurent Dinh, Mélanie Ducoffe, Vincent Dumoulin, Samira Ebrahimi Kahou, Dumitru Erhan, Ziye Fan, Orhan Firat, Mathieu Germain, Xavier Glorot, Ian Goodfellow, Matt Graham, Caglar Gulcehre, Philippe Hamel, Iban Harlouchet, Jean-Philippe Heng, Balázs Hidasi, Sina Honari, Arjun Jain, Sébastien Jean, Kai Jia, Mikhail Korobov, Vivek Kulkarni, Alex Lamb, Pascal Lamblin, Eric Larsen, César Laurent, Sean Lee, Simon Lefrancois, Simon Lemieux, Nicholas Léonard, Zhouhan Lin, Jesse A. Livezey, Cory Lorenz, Jeremiah Lowin, Qianli Ma, Pierre-Antoine Manzagol, Olivier Mastropietro, Robert T. McGibbon, Roland Memisevic, Bart van Merriënboer, Vincent Michalski, Mehdi Mirza, Alberto Orlandi, Christopher Pal, Razvan Pascanu, Mohammad Pezeshki, Colin Raffel, Daniel Renshaw, Matthew Rocklin, Adriana Romero, Markus Roth, Peter Sadowski, John Salvatier, François Savard, Jan Schlüter, John Schulman, Gabriel Schwartz, Iulian Vlad Serban, Dmitriy Serdyuk, Samira Shabanian, Étienne Simon, Sigurd Spieckermann, S. Ramana Subramanyam, Jakub Sygnowski, Jérémie Tanguay, Gijs van Tulder, Joseph Turian, Sebastian Urban, Pascal Vincent, Francesco Visin, Harm de Vries, David Warde-Farley, Dustin J. Webb, Matthew

Willson, Kelvin Xu, Lijun Xue, Li Yao, Saizheng Zhang, and Ying Zhang. 2016. Theano: A Python framework for fast computation of mathematical expressions. *Preprint arXiv:1605.02688* (2016).

[150] Sebastian Thrun. 1996. Is Learning The n-th Thing Any Easier Than Learning The First? *Neural Information Processing Systems (NeurIPS)* (1996).

[151] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and Others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint arXiv:2307.09288* (2023).

[152] Y. Z. Tsypkin. 1966. Adaptation, training and self-organization automatic control systems. *Avtomatika I Telemekhanika* 27 (1966), 23–61.

[153] Meltem Sönmez Turan, Elaine Barker, John Kelsey, Kerry McKay, Mary L. Baish, and Mike Boyle. 2018. Recommendation for the Entropy Sources Used for Random Bit Generation. *NIST Special Publication 800-90B* (2018). https://doi.org/10.6028/NIST.SP.800-90B

[154] Krishna C. Vadodaria and Sebastian Jessberger. 2014. Functional neurogenesis in the adult hippocampus: Then and now. *Frontiers in Neuroscience* 8 (2014).

[155] Jonathan Vanian and Kif Leswing. 2023. ChatGPT and generative AI are blooming, but the costs can be extraordinary. CNBC: https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html. Last accessed: 2025-02-02.

[156] Vladimir N. Vapnik. 1999. *The Nature of Statistical Learning Theory.* Springer New York.

[157] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[158] Fabrizio Ventola, Steven Braun, Zhongjie Yu, Martin Mundt, and Kristian Kersting. 2023. Probabilistic Circuits That Know What They Don't Know. *Uncertainty in Artificial Intelligence (UAI)* (2023).

[159] Richard von Mises. 1964. *Mathematical theory of probability and statistics.* Academic Press, New York, Chapter VIII.9.3.

[160] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. 2023. On the De-duplication of LAION-2B. *preprint arXiv:2303.12733* (2023).

[161] P. J. Werbos. 1982. Applications of advances in nonlinear sensitivity analysis. *Systems Modeling and Optimization: Proc. IFIP* (1982).

[162] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.*

[163] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations.*

[164] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL] https://arxiv.org/abs/1904.09675

[165] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is More for Alignment. *Neural Information Processing Systems (NeurIPS)* (2023).