

Misrepresented Technological Solutions in Imagined Futures: The Origins and Dangers of AI Hype in the Research Community

Savannah Thais

Columbia University Data Science Institute
New York, New York 11221 USA
st3565@columbia.edu

Abstract

Technology does not exist in a vacuum; technological development, media representation, public perception, and governmental regulation cyclically influence each other to produce the collective understanding of a technology’s capabilities, utilities, and risks. When these capabilities are overestimated, there is an enhanced risk of subjecting the public to dangerous or harmful technology, artificially restricting research and development directions, and enabling misguided or detrimental policy. The dangers of technological hype are particularly relevant in the rapidly evolving space of AI. Centering the research community as a key player in the development and proliferation of hype, we examine the origins and risks of AI hype to the research community and society more broadly and propose a set of measures that researchers, regulators, and the public can take to mitigate these risks and reduce the prevalence of unfounded claims about the technology.

Introduction

Artificial Intelligence (AI) has become a ubiquitous topic in the modern world as it has captivated the popular imagination, spurred rapid technological development, and captured a significant share of the investment market. However, despite being presumably a scientific field, many of the descriptions and promises of AI systems are not empirically proven or conceptually well grounded.

The media, and even some researchers themselves, unfoundedly ascribe language understanding, general reasoning ability, or even sentience to AI systems, and claim that we are close to achieving artificial general intelligence (AGI) (Grace et al. 2017). Through various intentional or accidental mechanisms including failures in testing or system design (Kapoor and Narayanan 2022b), poorly constructed or characterized data sets (Raji et al. 2021, 2022), or outright deception (Raji et al. 2022), researchers and developers can create AI systems that fail to live up to their performance and reliability promises.

This culture of hype has real world consequences. Inaccurate or misrepresented AI systems can cause harm to the public and exacerbate societal biases at an unprecedented scale (Raji et al. 2022; Wang et al. 2022). Unscientific and

misleading claims about AI also negatively impact the research and development ecosystem itself by incentivizing certain research directions over others and affecting how broader society views the validity and utility of the field.

Here, we define AI hype as any non empirically or rigorously theoretically supported performance claims, capability narratives, or system descriptions. We consider empirically supported claims to mean performances or capabilities that demonstrated through properly designed, reproducible scientific research, that generalize outside of the initial experimental context, and are precisely characterized in their descriptive language (i.e. not saying a model exhibits language understanding when what is meant is that it achieved high accuracy on a multiple choice benchmark data set like MMLU (Hendrycks et al. 2020)), while we consider theoretically supported claims to be provably derived directly from statistical or mathematical theory.

Centering the research and development community as a key responsible player in the development and proliferation of hype, this paper will explore origins of AI hype and the dangers that hype poses to both society and the research ecosystem. We will then present actions that the research community can take to address these concerns and ways that those efforts can be scaffolded by political action and community work.

Current Landscape of AI Rhetoric

Recent advancements in AI, though undoubtedly impressive, have been surrounded by unprecedented levels of hype. Particular attention has been paid recently to large language models (LLMs). Some claims like scientific reasoning and understanding of the short-lived model Galactica (Taylor et al. 2022) or the ascribed sentience of more general LLMs like LaDMa (Tiku 2022) received relatively little uptake or have been quickly debunked (Heaven 2022b; Shardlow and Przybyła 2022).

More pernicious and common are claims of general or approximate language understanding (Bender and Koller 2020) and reasoning abilities (Bubeck et al. 2023). This belief enables additional sociological claims that LLMs will soon replace human writers in a variety of domains (Bilton 2022), precipitate a breakdown of traditional education (Marche 2022), and transform search and information retrieval (Shah and Bender 2022), which, while entirely possible due to fi-

nancial power structures, are largely ungrounded in scientifically proven abilities of these models (indeed, we have already seen the threat of AI replacement play a central role in the Hollywood writers strike (Coyle 2023)). While LLMs have demonstrated exceptional performance on certain purported reasoning related benchmark data sets, such as MMLU, they perform substantially worse than humans on others, such as ConceptARC (Mitchell, Palmarini, and Moskvichev 2023) and PlanBench (Valmeekam et al. 2023). As discussed further in Section , these apparent discrepancies in ability are in part attributable to the design and conceptualization of benchmark tasks themselves.

Furthermore, research in NLP has demonstrated that despite these LLMs performing well on some established benchmark tasks, this performance can be attributed to learning to manipulate established linguistic forms (Bender and Koller 2020) or to leverage artifacts from the training data sets (Niven and Kao 2019; Bras et al. 2020), and in fact this performance often does not generalize. For example, Niven and Kao demonstrated that BERT’s performance on the Argument Reasoning Comprehension Task is accounted for by exploitation of spurious statistical cues in the data set (Niven and Kao 2019).

Moreover, these claims obfuscate known limitations and issues with the technology. Notably, they are prone to hallucination and can generate coherent and convincing yet false text, including factually incorrect historical or scientific facts, causally inaccurate reasoning, or fabricated sources and references (Gao et al. 2022; Bickmore et al. 2018; Jack Brewster 2023). Additionally, they often demonstrate bias against certain groups, owing in large part to their training corpus of text scraped from the internet (Bender et al. 2021; Brown et al. 2020; Xu et al. 2020; Abid, Farooqi, and Zou 2021). Unfortunately, some initial exploration demonstrates that these issues persist even in models that are nominally safeguarded, such as ChatGPT (Biddle 2022). These substantial limitations and the risks they pose, as described in Section 4, currently render them largely non-robust and untrustworthy, and perhaps of limited practical utility.

Similar hype and controversy surround other types of generative AI, namely text-to-image and text-to-video generators like DALL-E 2, Stable Diffusion, Make-A-Video, Imagen Video, and others. These models have been described as demonstrating visual reasoning skills or human-like creativity (Heaven 2022a). As with LLMs, news outlets, researchers, and artists themselves have questioned if, or even predicted that, this technology will soon replace (or at least substantially transform) a wide range of creative careers (Heaven 2022a; Bilton 2022).

While the capabilities of these generative AI models are undoubtedly impressive, many of the claims surrounding this technology are, again, not empirically grounded. In particular, recent research has demonstrated that these models are unable to correctly interpret prompts containing specific spatial relation or object multiplicity information (Lin and Srikanth 2023; Gokhale et al. 2022; Cho, Zala, and Bansal 2022). Furthermore, they are not robust to prompt variation, and small adjustments or re-orderings of the prompt can result in substantially different outputs (Witteveen and

Andrews 2022; Gokhale et al. 2022). Additionally, they are susceptible to the same issues of bias present in LLMs; they demonstrate gender and skin-tone biases when given neutral text prompts (Cho, Zala, and Bansal 2022) as well as context bias when provided with gender or race specified input (Heikkilä 2022).

Although there is substantial hype around recent AI systems, this phenomena is by no means new. Autonomous driving systems are perhaps the most over-hyped AI technology. Various executives and researchers have claimed for years that fully autonomous or self-driving vehicles are on the verge of entering the market; for example, in 2017 Jensen Huang, the CEO of Nvidia, proclaimed that “we can realize this vision [of self-driving cars] right now”, clarifying that “AI is the solution to self-driving” (Stilgoe 2018). Two years previously, Elon Musk, the CEO of Tesla, promised that self-driving is “almost...a solved problem”; Musk has continued to tout Tesla’s self-driving capabilities since, despite having to repeatedly walk back, temper, or qualify these claims (Stilgoe 2018). Regardless of the continued hype and large-scale financial investments, only a handful of prototype vehicles have been tested on public roads, and safe and reliable autonomous vehicles remain beyond the immediate horizon of technological capabilities (Stilgoe 2019). In fact, backlash to the sustained hype around this technology and its lack of realization has led some to conclude that “self-driving cars are going nowhere” (Chafkin 2022).

AI hype is not limited to over-estimations of the ability of AI systems to perform on general reasoning tasks. Many focused or narrow applications of AI are subject to misguided or misleading narratives. AI systems have been purported to reliably perform on a diverse array of tasks, ranging from predicting someone’s appearance, political affiliation, criminality, and more from their voice or physical attributes (Oh et al. 2019; Stark and Hutson 2021) to triaging patients (Obermeyer et al. 2019), allocating welfare benefits (Lecher 2018), and selecting employees for layoffs (Rashida Richardson 2010). Unfortunately, the performance claims surrounding many of these systems are not empirically proven or are not replicated when the systems are deployed in the real world.

As described in the following section, some of these unsubstantiated claims can be attributed to engineering or design mistakes that were not noticed during development. However, a substantial portion of the inaccurate rhetoric surrounding this type of AI is due to attempts to apply AI to what Raji et al term “conceptually impossible tasks” where no causal connection exists between the observable data and the proposed task (Raji et al. 2022), despite developers’ or marketers’ claims that AI can learn information that simply is not perceived by the human brain.

The culture of hype surrounding different types of AI systems, as well as factors described in the next section, contribute to perhaps the most ubiquitous non-scientifically grounded belief about AI: the expectation of near future AGI. Although there is no agreed upon scientific definition of AGI (Morris et al. 2024) and AI researchers hold extremely varied opinions about the near-term likelihood of AGI (Grace et al. 2024) (in fact, many have stated that AGI

is far from being realized (Prunkl and Whittlestone 2020; Atkinson 2016)), this rhetoric has captured substantial public and media attention, garnered significant financial investment, and enabled the proliferation of AGI focused start-ups.

Origins of AI Hype

In order to fully characterize the dangers associated with AI hype and identify potential mitigating strategies, we must first understand the factors that allow hype to develop and spread. While many analyses of technological hype focus on popular media representations or financial incentives as origins of hype, here we will focus largely on issues inherent to the current AI research and development (R&D) ecosystem and explore how these concerns enable and enhance broader hype cycles.

AI researchers and developers exercise nearly complete control over the design, and thereby applicability and reliability, of a given AI system. Nonetheless, they typically perceive both the underlying data and eventual impacts of these systems as outside of their control (Slota et al. 2020). Along with the issues described in the following section, this fosters a culture of distance between R&D and the societal context of technology. This can lead to researchers and developers prioritizing certain technical considerations like efficiency over more human-centered concerns like preventing harm (Dobbe, Gilbert, and Mintz 2019). This perceived distance becomes particularly problematic in the current fast-paced technological R&D ecosystem that tends to prioritize the race to innovation and the capture of market share.

By focusing primarily on the R&D ecosystem, we hope to re-center researchers and developers as critical actors in the construction and propagation of misleading or dangerous rhetoric around AI and shift some of the responsibility for counteracting these narratives back to these groups.

Research and Development Environments

AI hype originates in part from the culture of R&D environments and how researchers and developers mathematically and scientifically characterize current AI technology. Many of the issues outlined below are fundamentally intertwined; responsible and honest R&D must reckon with these considerations collectively.

A primary origin of over-hyped AI capabilities is the fact that many AI systems are developed in sterile R&D environments and then deployed in more complex real-world settings without appropriate testing or oversight. In some cases, a mismatch between performance in R&D and deployment settings is attributable to sampling bias in the test set distribution (Kapoor and Narayanan 2022b). Choices made during data collection and pre-processing can render a data set non-representative of the population the AI system is designed to be used on; for example, a disease prediction tool trained on medical images may not generalize from one hospital environment to another (Zech et al. 2018) or may not be robust to small perturbations caused by malfunctions of the imaging device (Antun et al. 2020).

In other cases, performance disparities are attributable to a combination of data set and evaluation metric design

choices. Namely, discrepancies between evaluations in an R&D settings and a meaningful characterization of how the AI system will actually perform in the real world (Raji et al. 2021). A lack of so-called ‘construct validity’ can lead to performance issues by allowing researchers and developers to neglect to include distribution of societal harms and benefits in their evaluation metrics (Raji et al. 2022) and by focusing R&D efforts on improving performance on artificially constructed benchmark data sets that do not meaningfully relate to the general reasoning tasks they purport to measure (Raji et al. 2021).

Designing effective and robust benchmark data sets for AI models is a challenging task, and mischaracterizing what a benchmark is designed to do is a significant contributor to AI hype. For example, ImageNet is widely accepted as a benchmark for visual object recognition, yet it seemingly arbitrarily relies on the WorldNet ontology to derive class labels, and thus is highly limited in the classes it actually represents. It has therefore been critiqued as actually representing a very specific, rather than general, vision benchmark (Raji et al. 2021).

Some R&D environments and applications of AI also eschew critical scientific and software development practices. For example, as demonstrated by surveys of AI researchers, the predictions made by AI systems are often taken at face value without a meaningful understanding of the associated uncertainty (Slota et al. 2020; Fleischmann and Wallace 2009). In other scientific fields, the uncertainty of a measurement or model is an essential component of understanding its significance and utility. Neglecting to provide accounts of the uncertainty associated with AI systems allows developers to gloss over potential concerns with the underlying training data, possible mismatch between the predicted labels and the real world decision being modeled, and many other critical design issues (Raji et al. 2022; Kapoor and Narayanan 2022b; Wang et al. 2022). Additionally, studies have found that many AI developers often disregard traditional software development practices like testing and quality assurance (Wan et al. 2021) and that ML and AI development occurs substantial ‘technical debt’ in terms of documentation, reproducibility, and reliability (Sculley et al. 2015). These issues coalesce to create the present so-called ‘reproducibility crisis’ in AI and ML R&D.

There are also issues inherent to the culture of technology R&D including how technologists view other fields and how we cognitively and linguistically frame the problems that we work on. Many AI systems seek to make predictions about complex social behaviors and future outcomes and many researchers and developers hold a belief that such predictions are possible with currently available data. However, as Campolo and Crawford describe, this belief requires an “epistemological flattening” of complex societal contexts into acute signals that neglect the probabilistic worldviews traditionally employed in social science research (Campolo and Crawford 2020). As demonstrated in a study by Sambasivan and Veeraraghavan, AI researchers and developers also tend to neglect the importance of domain expertise in understanding the appropriate development and limitations of AI systems and frequently reduce researchers and practitioners

in other fields to data collectors rather than equal collaborators (Sambasivan and Veeraraghavan 2022). This domain hubris can result in a host of technical issues including feature illegitimacy (Chiavegatto Filho, Batista, and Dos Santos 2021), lack of appropriate data segregation (Oner et al. 2020), and insufficient pre-deployment testing (Nagendran et al. 2020).

Additionally, despite some advances in explainable and interpretable AI, there remains a profound mismatch between the high-dimensionality of many AI systems and the interpretation capacity of human developers and users (Burrell 2016). There is a profound disconnect between technological progress and a fundamental scientific understanding of how AI systems function (Campolo and Crawford 2020). This leads researchers and developers to describe their work in less precise or exact terminology and creates a perceived lack of control over the technological narrative (Slota et al. 2020). As discussed further in Section 3.3, the language used by researchers and developers to describe their work can substantially contribute to AI hype.

Financial Incentives

The R&D ecosystem and its outputs cannot be divorced from the financial mechanisms that support it. Technological R&D is often viewed as a primary driver of economic prosperity and national security (West 2011) and thus attracts substantial public and private investment. AI R&D in particular currently dominates global scientific output (UNESCO 2021) and continues to attract increased funding; the private investment in AI was 18 times higher in 2022 than in 2013, totaling around \$93.5 billion (Daniel Zhang and Perrault 2023). Unlike some other areas of scientific R&D, the AI ecosystem is dominated by this private money; in fact, many ‘top’ academic researchers receive substantial financial support from private tech companies (Abdalla and Abdalla 2020). This funding landscape can incentivize hype and skew publishing norms (Ebell et al. 2021).

In particular, the relative abundance of funding but expected short timelines for innovation and return on investment lead to what Rayner calls the ‘novelty trap’. Researchers and developers are incentivized to exaggerate the speed and inventiveness of their work while de-emphasizing potential safety and reliability concerns (Rayner 2004). Insidiously, this means that there is often substantial profit to be earned by marketing technological abilities that may not exist (Elish and danah boyd 2018). For example, Chinese technology company Baidu was discovered to have cheated on the popular benchmark data set ImageNet in an effort to include exaggerated performance claims in marketing materials (Simonite 2015); similarly, Amazon misrepresented the ways in which their facial recognition software, Rekognition, could be used by law enforcement in sales and promotional material (CLU 2018). This type of behavior dangerously shifts the responsibility for critically evaluating system performance claims to consumers, regulators, and external researchers, who are often less equipped to undertake this work or are not given necessary access to the systems in question. For example, the landmark Gender Shades study, conducted by external researchers Buolamwini and Gebru

uncovered gender and racial bias in commercial facial recognition technology (Buolamwini and Gebru 2018) and was essential for incentivizing the involved tech companies to make their systems safer (Raji and Buolamwini 2019).

Media and Society

Additionally, R&D and the surrounding financial landscape are impacted by public and media discourse and representations of AI. AI has been part of collective human fantasy and science fiction narratives long before technological advances made any of these imagined capabilities remotely possible. These depictions and the resulting cultural norms formed around AI influence not only how the public and reporters perceive this technology, but also how researchers and developers imagine the capabilities and limitations of the very technology they are creating.

These normative representations, as well as a real lack of deep mathematical understanding of how many modern AI systems (particularly deep learning systems) function (Raghu et al. 2016; Lu et al. 2017), impact how the individuals closest to AI describe their work. Both developers (Campolo and Crawford 2020) and critics (Vinsel 2021) of technology employ magical, anthropomorphic, or superhuman language to describe AI systems. For example, in the official paper for AlphaZero, a Go-playing AI system, researchers described the technology as “superhuman” and operating “without human knowledge” despite the fact that the entire system was, in fact, developed by humans with an understanding of the mechanisms and goals of the game (Silver et al. 2017). Similarly, Demis Hassabis, the CEO of DeepMind, the company behind AlphaZero, likened the system to a chess-playing alien (Knight 2017).

It is unsurprising, then, that journalists reporting on AI face difficulties in accurately discussing the technology. This is further complicated by the reality that journalists may lack the technological and scientific foundations to fully understand and characterize AI even in cases when it is precisely and truthfully described by researchers and developers (Jones, Jones, and Luger 2022). Some researchers and developers have described concerns that outcomes and nuances of AI research are distorted as they leave the immediate R&D ecosystem and cite concern over a lack of scientific precision in media coverage (Slota et al. 2020).

Together, these factors can foster a cycle of misrepresentation where R&D environments, funding mechanisms, and social and media representations influence each other to create a collective understanding of AI that may be distant from the scientific reality.

Technologies as Imagined Futures

When interrogating AI hype, it is crucial to emphasize that rhetoric around technology, regardless of the source, is discursive world building. Technology not only represents a solution to a problem, but through its development and presentation asserts what problems exist and should be considered. Stilgoe discusses this phenomena in the context of autonomous vehicles: in media coverage and public policy, vehicular accidents have traditionally been blamed on driver error, and by taking up this rhetoric, autonomous vehicle

companies are able to present self-driving cars and AI as a solution to the “driver problem”. By doing so, they construct an imagined future that entrenches consumer reliance on individual vehicles and detracts financial and intellectual resources from other potential solutions like improved public transit (Stilgoe and Mladenović 2022).

Shah and Bender similarly position claims that LLMs will replace search systems. These claims, often perpetuated by the developers of the LLMs in question, neglect to consider the societal context and purpose of search. In many cases, it is more useful and desirable for users to explore and make sense of information on their own, rather than be provided with a singular answer as is common with LLMs (Shah and Bender 2022). By promoting a certain narrative about the abilities and utility of technology, we necessarily impact societal behavior and views about technology.

Dangers of Misrepresentation

The pervasiveness of AI hype can prevent researchers and developers, policy makers, journalists, and the public from proactively interrogating these imagined futures. Hype can foster the impression that AI is universally good or bad and can preclude the nuance necessary to use this technology effectively (Slota et al. 2020). Similarly, claims of novelty and innovation can hinder social learning and proper contextualization of AI and create a perceived distance between the technology and its risks (Rayner 2004; Stilgoe 2018).

Campolo and Crawford describe AI hype as ‘enchanted determinism’, allowing AI systems to be seen as powerfully predictive and deterministic, yet also as outside the realm of full human understanding and ability (Campolo and Crawford 2020). The unique qualities of this type of hype make it particularly persuasive to a broad range of stakeholders. Its power and ubiquity pose substantial risks not only to society at large but also, critically, to the very R&D ecosystem in which AI technology is constructed.

Unscientific, Inaccurate, and Dangerous Sociotechnical Systems

When AI capabilities are mischaracterized, there is heightened risk of causing harm to individuals who rely on or are subjected to the predictions of these systems. There are already numerous examples of tangible financial, emotional, and corporeal harm affected by misrepresented or misused AI systems. For example, voice assistants such as Alexa, Siri, and Google Assistant, have been portrayed having the potential to serve as portals for health information (Sezgin et al. 2020). However, a study found that when given medical related queries, a substantial portion of the information these voice assistants returned could result in user harm or death (Bickmore et al. 2018). In another case, an external audit of a widely deployed sepsis prediction tool developed by the healthcare software company Epic found that despite performance claims to the contrary, the tool failed to identify a majority of sepsis cases and generated a considerable number of false alerts, creating extraordinary risk for treatment mistakes (Wong et al. 2021). In yet another example, Allegheny county in Pennsylvania adopted a screening tool de-

signed to identify children at risk of mistreatment and used it to select families for investigations that could involve removing the child from their families. A full two years after the tool was deployed it was discovered that reported performance measurements were over estimated and that the tool exhibited substantially higher error rates on Black families compared to White families (Chouldechova et al. 2018).

The widespread utilization of this type of unreliable predictive AI is directly enabled by AI hype. The specific type of hype surrounding AI, as described in Section 3.3, is used to justify a movement away from explainable decision systems and reduced focus on clear causal mechanisms in high-stakes AI enabled predictions (Campolo and Crawford 2020). The deployment of uninterpretable AI denies the people subjected to these systems opportunities for recourse.

Furthermore, AI hype fosters a culture of R&D and AI system utilization that values ostensible predictive accuracy over deeper contextual or scientific knowledge of the underlying mechanisms or potential impacts of those predictions (Campolo and Crawford 2020). This allows researchers and developers to ignore validity issues intrinsic to the structure of predictive optimization tasks which Wang et al demonstrate carry inherent risks of harm to individuals (Wang et al. 2022).

The way these systems are represented and marketed often results in a displacement of responsibility for negative outcomes onto individual users or subjects who, in practice, lack control over the outcomes (Slota et al. 2020; El-ish 2019). For example, despite emphasizing the predictive power, accuracy, and automation capabilities of their systems, technology companies often attempt to distance themselves from harmful impacts of their technology by claiming that the systems were never intended to be used as decision tools and require extensive human oversight to be used safely (Kapoor and Narayanan 2022a). This framing and misallocation of responsibility poses a threat to legal due process and social agency (Citron and Pasquale 2014).

Critically, the nature of AI allows these risks to proliferate at an unprecedented scale. For instance, the coding question-and-answer site Stack Overflow banned users from posting answers and code examples generated by ChatGPT due to the vast frequency of generated responses that appear convincing yet are subtly incorrect (VINCENT 2022). Similarly, researchers cite increased concern over the risk of AI fueled mis- and dis-information campaigns with vast political and social impacts (Ben Buchanan and Sedova 2021). Of particular concern is the risk of inextricably embedding different types of social bias in our technological infrastructure (Noble 2018; Eubanks 2018).

Artificially Limited Research Landscapes

Though perhaps of less existential importance, AI hype also carries risks to the R&D ecosystem itself. Many of the societal risks described in the previous section are direct consequences of R&D issues. A failure to interrogate these issues precipitates the risk of permanently entrenching these behaviors into the fabric of AI research.

Dishonest rhetoric around the capabilities of AI systems can lead researchers to believe that we have solved certain

problems that we in fact have not; as discussed in the previous section, this can lead to drastic societal consequences. However, it can also profoundly affect the types of research questions that are asked and what areas of development are funded. Wang et al surveyed 387 academic, industry, non-profit, and governmental publications and found widespread critical flaws in predictive optimization systems that use ML to make predictions and decisions about future individual outcomes (Wang et al. 2022). Despite research demonstrating that many life outcomes remain inherently unpredictable with currently available data and computing techniques (Salganik et al. 2020) many of the predictive systems studied by Wang et al were publicly deployed or commercially available. This sets a dangerous precedent that incentivizes more developers to produce similar systems and discourages research into their limitations and shortcomings. It also fosters a culture where younger researchers and students are more likely to erroneously believe that this type of predictive system is inherently trustworthy, further entrenching the idea that alternatives or methods of improvement are not fruitful areas of research.

This belief in the functionality of outcome predictive AI impacts research questions in individual subfields of AI as well, even some areas typically considered to be under the umbrella of ‘AI Ethics’. For example, research in AI fairness often presupposes a trade-off between model accuracy and some notion of fairness (Fish, Kun, and Lelkes 2016; Friedler et al. 2018) and focuses on ways of ensuring AI systems respect existing equal protection laws such as the United States’ ‘4\5ths’ rule for treatment of protected classes in hiring decisions (Wilson et al. 2021; Raghavan et al. 2019). However, as Raji et al demonstrates (Raji et al. 2022), the utility of this research relies on the assumption that the system in question is accurate and reliable, which, as discussed extensively in previous sections, hyped AI systems often are not. Research on interpretable and explainable AI requires similar assumptions about the fidelity of the system being explained. Furthermore, the current belief in the AI R&D ecosystem that large-scale models provide the best accuracy and performance (Bengio and LeCun 2007) pushes researchers to focus on black-box explainability techniques to the detriment of other approaches such as developing improved directly interpretable models (Rudin 2018).

Even if it is not presumed that we have fully solved certain problems, rhetoric that we are very close to solving particular problems, like AGI, with certain approaches, like increasingly larger models and data sets (collectively referred to as ‘scale’, or the ‘Scaling Hypothesis’ (Branwen 2020)), also profoundly impacts the type of questions researchers are incentivized to pursue. While large scale models have certainly provided state-of-the-art performance on certain tasks, notably language modeling (Kaplan et al. 2020; Brown et al. 2020) and computer vision (Yu et al. 2022; Ramesh et al. 2022), and demonstrated impressive abilities, it has not been proven that scale is universally the correct approach to developing more generalizable and performant AI models. In fact, some research has shown that certain substantially smaller models can match large model performance (Frankle and

Carbin 2018) and that the benefits of scale vary based on model architecture and learning task (Tay et al. 2022). Furthermore, extremely large data sets are typically uncured by necessity, and are therefore unlikely to lead to AI models that can represent diverse view points or truly generalize to real-world tasks (Bender et al. 2021).

Nonetheless, the field of AI R&D as a whole is largely focused on scale approaches as evidenced by the coining of the term ‘foundation models’ to position these large scale models as the central and most important mode of AI (Bommasani et al. 2021). Recently, vast amounts of funding has been poured into research centers and start-up companies focused exclusively on scale approaches. This focus is further driven in part due to large models’ performance on benchmark data sets and tasks, which as discussed in Section 3.1, are often not meaningfully connected to real-world operational and societal goals (Raji et al. 2021). The confluence of these benchmarks, financial incentives, and hype-fueling research claims creates a feedback loop that serves to trap AI R&D in a specific and limited focus.

Rhetoric describing the proximity of scale-driven AGI also impacts what new areas of research are prioritized. Notably, AI Safety and Alignment, the fields of research concerned with ensuring AGI systems do not pose an existential threat to humanity, have become increasingly popular in the past few years and companies in this space have attracted massive amounts of funding. This stands in stark contrast to the relatively small amount of funding being allocated to research organizations and non-profits focused on alleviating the harms that current AI systems are perpetuating.

In all areas of research, there is a limit to the available funding and person-power; prioritizing certain research questions necessarily de-prioritizes others, and this selection can have long-lasting and far-reaching consequences. As a field, we must critically consider what questions are worth focusing on in terms of both their potential scientific and societal impacts and what economic and sociological incentives influence our choices of research topics.

Methods of Addressing and Reducing Hype

In order to appropriately manage expectations around AI, ensure a robust, responsible, and diverse R&D ecosystem, and limit the deployment of unsafe and unreliable AI systems, it is critical to change how AI is discussed and drastically reduce the prevalence of AI hype. As described in the previous sections, many factors interact to cultivate a hype-driven feedback loop that impacts how everyone from researchers and developers to the general public perceives and discusses AI. Consequently, combating AI hype requires a multi-faceted approach involving industry, government, academia, and civil society.

Diversified Funding and Research

As discussed in Section 4.2, AI hype serves to focus R&D on a limited number of questions, typically those that are priorities for technology companies and investors. One way to address this is by encouraging the diversification of AI funding and research at several levels. Although governmental agencies worldwide are investing in AI R&D, prioritizing

that funding towards certain research areas that complement rather than compete with industry will likely be a more effective strategy. There is substantial opportunity for agencies and other funding mechanisms such as research institutes and foundations to support work that re-frames AI R&D as a fundamentally scientific endeavor. They should concentrate R&D funding on scientific validity topics like uncertainty and predictive outcome reliability, more rigorous characterization and understanding of system behavior, and non-scale approaches to generalizable AI. Similarly, they should craft funding solicitations that incentivize meaningful interdisciplinary collaboration. By supporting academic research in these areas, this funding also allows students and younger researchers to develop interests in a wider range of topics that will have outward ripple effects in future generations of R&D.

As members of the AI R&D ecosystem, we must also take a critical eye towards our field and consider structural changes that can encourage more responsible AI development and more accurate representation of our work. We should make a concerted effort to use more precise language to describe ML and AI; although not all of the underlying mechanisms are precisely understood, rather than resorting to vague or magical language to describe AI systems, we should acknowledge what they are: as Campolo and Crawford put it “mathematical optimization at scales beyond expert human play” (Campolo and Crawford 2020). Similarly, we should require a heavier internal burden of proof of the functionality of AI systems before they are deployed. There are multiple proposed frameworks for researchers and developers to interrogate the functionality of their systems (Wang et al. 2022; Kapoor and Narayanan 2022b) and these can be scaffolded by a wider adoption of core scientific practices like uncertainty estimation and significance testing. Additionally, while emergent behavior in large models may be possible, we should refrain from making unsubstantiated claims about an AI system’s reasoning abilities and instead work towards developing rigorous tests to characterize such behavior. As the people initially making claims about an AI system’s performance, this burden of proof rests squarely on the researchers and developers (Lundberg, Johnson, and Stewart 2021).

Additionally, we should move away from using performance on contrived benchmark data sets as the main metric for success in AI R&D, as this impedes hypothesis-based research (Hooker 1995), encourages a narrow concentration on short-term objectives (Thomas and Uminsky 2020), and increases the likelihood of reporting overblown performance claims. Raji et al propose several evaluation methods to serve as an addition or alternative to traditional benchmarking. They also argue that insofar as benchmark data sets are useful for characterizing model performance, we must ensure that they are carefully scoped and context-specific (Raji et al. 2021). We can also increase the utility and applicability of benchmarks by supporting the development of more curated, collaborative, and sociocultural informed archival data sets (Jo and Gebru 2019).

We must also leave aside technological hubris and embrace meaningful interdisciplinary collaboration that places

other domains of knowledge on a level playing field with our own expertise. This will allow us to be more proactive about understanding the societal context and stakeholder value of our work and reduce the likelihood of avoidable feature and label legitimacy issues. Furthermore, it will perhaps discourage some forms of technosolutionism and allow us to consider more diverse imagined futures.

Regulatory Action

Changes to the R&D ecosystem are slow, are unlikely to be universally adopted, and are not alone sufficient to tackle the issue of AI hype. While governance and legislation of AI systems is a complex and rapidly evolving field of work, here we provide a few suggestions for external regulation that can reduce the prevalence of unsupported performance claims and increase the reliability of AI discourse.

Many scholars have suggested standardized third-party audits or testing procedures for evaluating AI system performance, particularly in high-stakes application areas (Tutt 2016; Green 2022). If, and only if, these audits do not presuppose system functionality or accept performance claims offered by the developers as a given (Raji et al. 2022), requiring them could significantly diminish AI hype by providing a reliable check on technology companies’ assertions about their products and discouraging them from overstating AI systems’ capabilities in the first place.

Mandated transparency is also a useful regulatory tool. While requiring transparency into the training data and model development and demanding justification of the scientific basis of technical capability claims enables external audits and allows stakeholders to draw informed conclusions about an AI system regardless of claims put forward by the developers. Although transparency is a complex topic, there is a rich literature around transparent documentation practices to draw from when developing regulatory frameworks for technology (Mitchell et al. 2018; McMillan-Major et al. 2021; Crisan et al. 2022; Gebru et al. 2018; Kapoor and Narayanan 2022b). While it is unlikely that regulators could mandate the use of such tools for all AI systems, requiring them for public sector AI systems would help alleviate some of the most dangerous instances of AI hype and perhaps encourage their wider adoption.

To accelerate the regulatory process, regulators can also rely on existing consumer protection, product liability, and fraud law to address harmful AI hype (Raji et al. 2022). Exercising these existing avenues of redress would not only help avoid the dissemination of exaggerated performance and reliability claims, but would also likely impede the willingness of developers to release under-tested AI systems in the first place.

Technical Literacy and Best Practices

Finally, as the consumers and propagators of much of the hype around AI, the media and the general public have a critical role to play in addressing these concerns. Society should not have to rely solely on the developers of a technology to describe its utility (Davis 2020) and thus it is essential that reporters and writers covering AI have the technical knowledge to avoid reproducing and spreading

AI hype. Several excellent frameworks have been proposed to help avoid common pitfalls in AI coverage (Kapor and Narayanan 2022; Shneiderman 2022; Sivadas and Argoub 2021) and it is essential that journalists adopt these techniques to ensure a trustworthy media ecosystem.

Perhaps the most useful tool for combating AI hype is improved technical literacy amongst all facets of the population, as this will allow everyone to participate in interrogating technological promises and holding AI systems and their developers accountable. Developing a broadly shared vocabulary and understanding of AI is necessary for the appropriate socialization of the technology and enabling and effective distribution of responsibilities and enactment of regulation (Slota et al. 2020). Several public facing initiatives like Better Images of AI and AI Myths have made important steps towards improved public understanding and more meaningful conversation around AI. These efforts must be supported by increased public outreach, opportunities for public engagement in the AI design process, and improved technical K-12 education. Empowering all members of society to meaningfully engage with the AI systems affecting their lives is an essential component of responsible and truthful AI R&D. This also fosters a more participatory system where researchers can be informed by the actual technological needs and desires of society.

Conclusion

We have explained how misrepresentations of AI capabilities, or AI hype, originates in many cases from the AI R&D ecosystem. Hype is further propagated through interactions with the media, funding and regulatory institutions, and the general public, which in turn influence the R&D ecosystem creating a feedback loop of false narratives, overstated performance claims, misplaced research incentives, and substantial negative societal impact. While no component of the AI hype cycle can be divorced from the others, AI researchers and developers bear a substantial responsibility for recognizing and addressing issues inherent to the field that encourage misleading and unscientific depictions of AI.

References

Abdalla, M.; and Abdalla, M. 2020. The Grey Hoodie Project: Big Tobacco, Big Tech, and the threat on academic integrity. *CoRR*, abs/2009.13676.

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent Anti-Muslim Bias in Large Language Models. *CoRR*, abs/2101.05783.

Antun, V.; Renna, F.; Poon, C.; Adcock, B.; and Hansen, A. C. 2020. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proc. Natl. Acad. Sci. U. S. A.*, 117(48): 30088–30095.

Atkinson, R. D. 2016. 'It's Going to Kill Us!' And Other Myths About the Future of Artificial Intelligence. *Information Technology Innovation Foundation*.

Ben Buchanan, M. M., Andrew Lohn; and Sedova, K. 2021. Truth, Lies, and Automation: How Language Models Could Change Disinformation. *Center for Security and Emerging Technology*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Online: Association for Computational Linguistics.

Bengio, Y.; and LeCun, Y. 2007. Scaling learning algorithms towards AI.

Bickmore, T. W.; Trinh, H.; Olafsson, S.; O'Leary, T. K.; Asadi, R.; Rickles, N. M.; and Cruz, R. 2018. Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. *J Med Internet Res*, 20(9): e11510.

Biddle, S. 2022. The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques. *The Intercept*.

Bilton, N. 2022. The New Generation of A.I. Apps Could Make Writers and Artists Obsolete. *Vanity Fair*.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R. B.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N. S.; Chen, A. S.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N. D.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M. S.; Krishna, R.; Kuditipudi, R.; and et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR*, abs/2108.07258.

Branwen, G. 2020. The Scaling Hypothesis.

Bras, R. L.; Swayamdipta, S.; Bhagavatula, C.; Zellers, R.; Peters, M. E.; Sabharwal, A.; and Choi, Y. 2020. Adversarial Filters of Dataset Biases.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *FAT*.

- Burrell, J. 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data Soc.*, 3(1): 205395171562251.
- Campolo, A.; and Crawford, K. 2020. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society*, 6: 1.
- Chafkin, M. 2022. Even After \$100 Billion, Self-Driving Cars Are Going Nowhere.
- Chiavegatto Filho, A.; Batista, A. F. D. M.; and Dos Santos, H. G. 2021. Data leakage in health outcomes prediction with machine learning. Comment on “prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning”. *J. Med. Internet Res.*, 23(2): e10969.
- Cho, J.; Zala, A.; and Bansal, M. 2022. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers. *CoRR*, abs/2202.04053.
- Chouldechova, A.; Prado, D. B.; Fialko, O.; and Vaithianathan, R. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *FAT*.
- Citron, D. K.; and Pasquale, F. 2014. The scored society: Due process for automated predictions. *Wash. L. Rev.*, 89: 1.
- CLU. 2018. ACLU Comment on New Amazon Statement Responding to Face Recognition Technology Test.
- Coyle, J. 2023. In Hollywood writers’ battle against AI, humans win (for now).
- Crisan, A.; Drouhard, M.; Vig, J.; and Rajani, N. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.
- Daniel Zhang, E. B. J. E. T. L. J. M. H. N. J. C. N. M. S. E. S. Y. S. J. C., Nestor Maslej; and Perrault, R. 2023. The AI Index 2023 Annual Report. *Stanford Institute for Human-Centered AI*.
- Davis, J. 2020. How Artifacts Afford: The Power and Politics of Everyday Things.
- Dobbe, R.; Gilbert, T. K.; and Mintz, Y. 2019. Hard Choices in Artificial Intelligence: Addressing Normative Uncertainty through Sociotechnical Commitments. *CoRR*, abs/1911.09005.
- Ebell, C.; Baeza-Yates, R.; Benjamins, R.; Cai, H.; Coeckelbergh, M.; Duarte, T.; Hickok, M.; Jacquet, A.; Kim, A.; Krijger, J.; Macintyre, J.; Madhamshettiwar, P.; Maffeo, L.; Matthews, J.; Medsker, L.; Smith, P.; and Thais, S. 2021. Towards intellectual freedom in an AI Ethics Global Community. *AI and Ethics*, 1.
- Elish, M. C. 2019. Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*.
- Elish, M. C.; and danah boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication Monographs*, 85(1): 57–80.
- Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin’s Press.
- Fish, B.; Kun, J.; and Lelkes, D. 2016. A Confidence-Based Approach for Balancing Fairness and Accuracy.
- Fleischmann, K.; and Wallace, W. 2009. Ensuring transparency in computational modeling.
- Frankle, J.; and Carbin, M. 2018. The Lottery Ticket Hypothesis: Training Pruned Neural Networks. *CoRR*, abs/1803.03635.
- Friedler, S. A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E. P.; and Roth, D. 2018. A comparative study of fairness-enhancing interventions in machine learning.
- Gao, C. A.; Howard, F. M.; Markov, N. S.; Dyer, E. C.; Ramesh, S.; Luo, Y.; and Pearson, A. T. 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H. M.; III, H. D.; and Crawford, K. 2018. Datasheets for Datasets. *CoRR*, abs/1803.09010.
- Gokhale, T.; Palangi, H.; Nushi, B.; Vineet, V.; Horvitz, E.; Kamar, E.; Baral, C.; and Yang, Y. 2022. Benchmarking Spatial Relationships in Text-to-Image Generation.
- Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; and Evans, O. 2017. When Will AI Exceed Human Performance? Evidence from AI Experts. *CoRR*, abs/1705.08807.
- Grace, K.; Stewart, H.; Sandkühler, J. F.; Thomas, S.; Weinstein-Raun, B.; and Brauner, J. 2024. Thousands of AI Authors on the Future of AI. arXiv:2401.02843.
- Green, B. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law Security Review*, 45: 105681.
- Heaven, W. D. 2022a. Generative AI is changing everything. But what’s left when the hype is gone? *MIT Technology Review*.
- Heaven, W. D. 2022b. Why Meta’s latest large language model survived only three days online. *MIT Technology Review*.
- Heikkilä, M. 2022. The viral AI avatar app Lensa undressed me—without my consent. *MIT Technology Review*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring Massive Multitask Language Understanding. *CoRR*, abs/2009.03300.
- Hooker, J. N. 1995. Testing heuristics: We have it all wrong. *J. Heuristics*, 1(1): 33–42.
- Jack Brewster, M. S., Lorenzo Arvanitis. 2023. The Next Great Misinformation Superspreader: How ChatGPT Could Spread Toxic Misinformation At Unprecedented Scale. *NewsGuard*.
- Jo, E. S.; and Gebru, T. 2019. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. *CoRR*, abs/1912.10389.
- Jones, B.; Jones, R.; and Luger, E. 2022. AI ‘everywhere and nowhere’: Addressing the AI intelligibility problem in public service journalism. *Digit. Journal.*, 10(10): 1731–1755.

- Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *CoRR*, abs/2001.08361.
- Kapoor, S.; and Narayanan, A. 2022a. The bait and switch behind AI risk prediction tools. *AI Snake Oil*.
- Kapoor, S.; and Narayanan, A. 2022b. Leakage and the Reproducibility Crisis in ML-based Science.
- Kapor, S.; and Narayanan, A. 2022. Eighteen Pitfalls to beware of in AI journalism.
- Knight, W. 2017. Alpha Zero’s “Alien” Chess Shows the Power, and the Peculiarity, of AI. *MIT Technology Review*.
- Lecher, C. 2018. What happens when an algorithm cuts your health care. *The Verge*.
- Lin, J.; and Srikanth, M. 2023. Diffusion Models as Visual Reasoners. In *The AAAI-23 Workshop on Creative AI Across Modalities*.
- Lu, Z.; Pu, H.; Wang, F.; Hu, Z.; and Wang, L. 2017. The Expressive Power of Neural Networks: A View from the Width. *CoRR*, abs/1709.02540.
- Lundberg, I.; Johnson, R.; and Stewart, B. M. 2021. What is your estimand? Defining the target quantity connects statistical evidence to theory. *Am. Sociol. Rev.*, 86(3): 532–565.
- Marche, S. 2022. The College Essay Is Dead. *The Atlantic*.
- McMillan-Major, A.; Osei, S.; Rodriguez, J. D.; Ammanamanchi, P. S.; Gehrmann, S.; and Jernite, Y. 2021. Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards. *CoRR*, abs/2108.07374.
- Mitchell, M.; Palmarini, A. B.; and Moskvichev, A. 2023. Comparing Humans, GPT-4, and GPT-4V On Abstraction and Reasoning Tasks. arXiv:2311.09247.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2018. Model Cards for Model Reporting. *CoRR*, abs/1810.03993.
- Morris, M. R.; Sohl-dickstein, J.; Fiedel, N.; Warkentin, T.; Dafoe, A.; Faust, A.; Farabet, C.; and Legg, S. 2024. Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv:2311.02462.
- Nagendran, M.; Chen, Y.; Lovejoy, C.; Gordon, A.; Komorowski, M.; Harvey, H.; Topol, E.; Ioannidis, J.; Collins, G.; and Maruthappu, M. 2020. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ*, 368: m689.
- Niven, T.; and Kao, H.-Y. 2019. Probing Neural Network Comprehension of Natural Language Arguments.
- Noble, S. U. 2018. Algorithms of oppression. In *Algorithms of oppression*. New York University Press.
- Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.
- Oh, T.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; and Matusik, W. 2019. Speech2Face: Learning the Face Behind a Voice. *CoRR*, abs/1905.09773.
- Oner, M. U.; Cheng, Y.-C.; Lee, H. K.; and Sung, W.-K. 2020. Training machine learning models on patient level data segregation is crucial in practical clinical applications.
- Prunkl, C.; and Whittlestone, J. 2020. Beyond Near- and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society. Association for Computing Machinery. ISBN 9781450371100.
- Raghavan, M.; Barocas, S.; Kleinberg, J. M.; and Levy, K. 2019. Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices. *CoRR*, abs/1906.09208.
- Raghu, M.; Poole, B.; Kleinberg, J.; Ganguli, S.; and Sohl-Dickstein, J. 2016. On the Expressive Power of Deep Neural Networks.
- Raji, D.; Denton, E.; Bender, E. M.; Hanna, A.; and Paullada, A. 2021. AI and the Everything in the Whole Wide World Benchmark. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Raji, I. D.; and Buolamwini, J. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.
- Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents.
- Rashida Richardson, V. M. S., Jason M. Schultz. 2010. Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems.
- Rayner, S. 2004. The Novelty Trap: Why Does Institutional Learning about New Technologies Seem So Difficult? *Industry and Higher Education*, 18: 349–355.
- Rudin, C. 2018. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.
- Salganik, M. J.; Lundberg, I.; Kindel, A. T.; Ahearn, C. E.; Al-Ghoneim, K.; Almaatouq, A.; Altschul, D. M.; Brand, J. E.; Carnegie, N. B.; Compton, R. J.; et al. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15): 8398–8403.
- Sambasivan, N.; and Veeraraghavan, R. 2022. The Deskillling of Domain Expertise in AI Development. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.
- Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; and Dennison, D. 2015. Hidden Technical Debt in Machine Learning Systems. In Cortes, C.; Lawrence, N.; Lee, D.;

- Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Sezgin, E.; Huang, Y.; Ramtekkar, U.; and Lin, S. 2020. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. *NPJ Digital Medicine*, 3(1): 122.
- Shah, C.; and Bender, E. M. 2022. Situating Search. CHIIR '22, 221–232. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391863.
- Shardlow, M.; and Przybyła, P. 2022. Deanthropomorphising NLP: Can a Language Model Be Conscious?
- Shneiderman, B. 2022. Guidelines for journalists and editors about reporting on robots, AI, and computers.
- Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359.
- Simonite, T. 2015. Why and How Baidu Cheated an Artificial Intelligence Test. *MIT Technology Review*.
- Sivadas, L.; and Argoub, S. 2021. How to report effectively on artificial intelligence.
- Slota, S. C.; Fleischmann, K. R.; Greenberg, S.; Verma, N.; Cummings, B.; Li, L.; and Shenefiel, C. 2020. Good systems, bad data?: Interpretations of AI hype and failures. *Proc. Assoc. Inf. Sci. Technol.*, 57(1).
- Stark, L.; and Hutson, J. 2021. Physiognomic artificial intelligence. *SSRN Electron. J.*
- Stilgoe, J. 2018. Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science*, 48(1): 25–56. PMID: 29160165.
- Stilgoe, J. 2019. Self-driving cars will take a while to get right. *Nature Machine Intelligence*, 1.
- Stilgoe, J.; and Mladenović, M. 2022. The politics of autonomous vehicles. *Humanit. Soc. Sci. Commun.*, 9(1).
- Tay, Y.; Deghani, M.; Abnar, S.; Chung, H. W.; Fedus, W.; Rao, J.; Narang, S.; Tran, V. Q.; Yogatama, D.; and Metzler, D. 2022. Scaling Laws vs Model Architectures: How does Inductive Bias Influence Scaling?
- Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; and Stojnic, R. 2022. Galactica: A Large Language Model for Science.
- Thomas, R. L.; and Uminsky, D. 2020. The Problem with Metrics is a Fundamental Problem for AI. *CoRR*, abs/2002.08512.
- Tiku, N. 2022. The Google engineer who thinks the company's AI has come to life. *The Washington Post*, 11.
- Tutt, A. 2016. An FDA for algorithms. *SSRN Electron. J.*
- UNESCO. 2021. UNESCO Science Report: The race against time for smarter development.
- Valmeekam, K.; Marquez, M.; Olmo, A.; Sreedharan, S.; and Kambhampati, S. 2023. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- VINCENT, J. 2022. AI-generated answers temporarily banned on coding QA site Stack Overflow. *The Verge*.
- Vinsel, L. 2021. You're Doing It Wrong: Notes on Criticism and Technology Hype.
- Wan, Z.; Xia, X.; Lo, D.; and Murphy, G. C. 2021. How does Machine Learning Change Software Development Practices? *IEEE Transactions on Software Engineering*, 47(9): 1857–1871.
- Wang, A.; Kapoor, S.; Barocas, S.; and Narayanan, A. 2022. Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. *Available at SSRN*.
- West, D. M. 2011. Technology and the Innovation Economy. *Center for Technology Innovation at the Brookings Institute*.
- Wilson, C.; Ghosh, A.; Jiang, S.; Mislove, A.; Baker, L.; Szary, J.; Trindel, K.; and Polli, F. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Witteveen, S.; and Andrews, M. 2022. Investigating Prompt Engineering in Diffusion Models.
- Wong, A.; Otles, E.; Donnelly, J. P.; Krumm, A.; McCullough, J.; DeTroyer-Cooley, O.; Pestrue, J.; Phillips, M.; Konye, J.; Penzoza, C.; Ghous, M.; and Singh, K. 2021. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern. Med.*, 181(8): 1065–1070.
- Xu, J.; Ju, D.; Li, M.; Boureau, Y.; Weston, J.; and Dinan, E. 2020. Recipes for Safety in Open-domain Chatbots. *CoRR*, abs/2010.07079.
- Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation.
- Zech, J. R.; Badgeley, M. A.; Liu, M.; Costa, A. B.; Titano, J. J.; and Oermann, E. K. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Med.*, 15(11): e1002683.