Constructing Capabilities: The Politics of Testing Infrastructures for Generative AI

GABRIEL GRILL, School of Information, University of Michigan, United States of America

The advertised and perceived capabilities of generative AI products like ChatGPT have recently stimulated considerable investments and discourse surrounding their potential to aid and replace work. The prominence of these systems, and their promise to be general-purpose, has resulted in an avalanche of tests to discover and certify their capabilities. This new testing regime is concerned with creating ever-more tasks for generative AI products instead of testing a model for one specialized task. Beyond efforts to understand products' capabilities, the construction of tasks and corresponding tests are also performative enactments meant to convince others and thus to gain attention, scientific legitimacy, and investment. The current market concentration of a few big AI companies points to a concerning conflict of interest: those with a vested interest in the success of the technology also have control over globalized testing infrastructures and thereby the exclusive means to create extensive knowledge claims about these systems. In this paper, I theorize capabilities as contested constructions and situated accomplishments shaped by power imbalances. I further unpack the globalized testing infrastructures involved in the construction and stabilization of generative AI products' capabilities. Furthermore, I discuss how the testing of these AI models and products is externalized, extracting value from the unpaid or under-paid labor of researcher and developer communities, content creators, subcontractors, and users. Lastly, I discuss a reflexive and critical approach to testing that challenges depoliticization and seeks to produce lasting critiques that serve more emancipatory goals.

CCS Concepts: • **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; • **Human-centered computing** → **Natural language interfaces**.

Additional Key Words and Phrases: generative AI, testing, ML benchmarks, capabilities, affordances, infrastructure studies

ACM Reference Format:
Gabriel Grill. 2024. Constructing Capabilities: The Politics of Testing Infrastructures for Generative AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3630106.3659009

1 INTRODUCTION

Testing is integral to machine learning (ML) as it promises to produce more understanding of the capabilities and limitations of complex models. It also provides guidance to identify possibilities for improvement and justification for investment and deployment. Traditionally, models were developed and tested for a specialized task, often characterized by a problem description and corresponding datasets. In contrast, recent generative AI products, like ChatGPT, are based on large models, claimed to be general-purpose [115] with evermore "emergent" capabilities as they increase in size. This promise has motivated a new machine learning testing regime concerned with identifying and defining tasks for which generative AI products could be employed. An avalanche of tests and claims around model capabilities has thus ensued, circulating in various media and shaping expectations of AI technology, sometimes in problematic ways.

Author's Contact Information: Gabriel Grill, ggrill@umich.edu, School of Information, University of Michigan, United States of America.

Manuscript submitted to ACM

These large models are complex and opaque as their creation involves extracting and relating patterns from immense amounts of varied data such as text, image, and video. Indeed, as they can also be understood as a compressed and optimized representation of all kinds of content publicly accessible on the internet, they have been described as "blurry [images] of the web" [30]. This scale and the strive to be general-purpose has created a need for direction as it is impossible to manually inspect all data or to test for all situated uses and eventualities. Thus, researcher and developer communities have created great amounts of tests and corresponding tasks to almost approximate a universal benchmark and be able to measure some notion of general progress. Some big meta-benchmarks now encompass hundreds of tasks to cover a great variety of capabilities [50], simultaneously enabling practitioners to report performance via one or several numbers quickly. This kind of large-scale testing promises insight into models' capabilities and possibilities for improvement. Besides efforts at understanding, tests are also performative enactments [57], intended to convince others of the capabilities of large models. For example, tech companies, like Google [109] and OpenAI [105], report performance metrics on academic benchmarks and other tests to advertise their new models. As scholarship in the sociology of expectations has established, performances of capabilities often feature in the promotion of new technologies to gain attention and investments, and attract legitimacy across areas of expertise [18]. However, expectations also co-create new realities. It is thus important to investigate how they are constructed, circulated, and what sustains them.

In this paper, I unpack the construction of so-called "capabilities" of generative AI, and discuss efforts to stabilize extractive globalized testing infrastructures that enable them. I understand capabilities not as static, decontextualized, and simply quantifiable phenomena but as both contested constructions and situated accomplishments of human and non-human actors [136]. When I discuss generative AI capabilities, therefore, I am referring to constructions that arise from different testing performances, demonstrations and experiences. The question of what capabilities can be ascribed to generative AI products is contested, as I will highlight. Nevertheless, powerful actors try to make these capabilities real — as their existence, or a wider belief in their existence, serves these actors' interests and stimulates investment. Given the significance of testing in this process, and the vast amounts of money and expectation now heaped upon generative AI, a more nuanced, critical, research-based understanding of capabilities is sorely needed. In what follows, I unpack the testing infrastructures involved in capabilities' invention and construction, and examine how actors use influence and power to make them "real." In the final section, I reflect upon the testing of generative AI and ML more broadly, and consider how practices of reflexive and critical testing could point these technologies toward more emancipatory goals.

My analysis is guided by sensibilities from situational analysis [33], infrastructure studies, and the field of science and technology studies (STS) more broadly, to unpack discourses on capabilities and contemporary testing infrastructures. I analyze public materials, like news articles and social media interactions, and also engage with literature on human-computer interaction (HCI) and evaluation in ML. My examination is partial and situated, bounded by what information was made public, or reported at the moment of writing. I am aware that the term "AI" is an overstatement, often deployed for marketing purposes [39], but I still use it due to its wide adoption. I also use the terms "generative AI product" to emphasize that these "AI's" are commercial products and "algorithmic system" [58] to refer to socio-technical systems that involve algorithms and models. In dialog with current debates on the politics of testing in machine learning and generative AI, the main contributions of this paper include: (1) Providing an infrastructure studies lens on capabilities in generative AI products, conceptualizing them as contested constructions and situated accomplishments, foregrounding the importance of discourse, and relationality in their stabilization; (2) Unpacking contemporary global testing infrastructures for generative AI products, revealing underlying extractive practices; (3) Presenting approaches and interventions, informed by STS literature, that support critical and reflexive testing.

2 UNPACKING THE CONSTRUCTION OF CAPABILITIES

In the current strive for market dominance, generative AI companies and developers advertise their products and models using grandiose claims about their capabilities. For example, Google introduced Gemini Ultra as their "largest and most capable AI model" [109] in December 2023. The company emphasized this point by presenting benchmark performance metrics on different broad "capabilities" such as "general," "reasoning," "math," "video," and "code." The capability "general" referred to the Massive Multitask Language Understanding (MMLU) benchmark [65] which consists of multiple-choice questions from 57 subject areas. The "image" capability was represented by seven benchmarks, ranging from reasoning to document understanding. Google claimed the performance of its model "exceeds current state-of-the-art results on 30 of the 32 widely-used academic benchmarks" [109].

It is a common practice in industry and the ML community to frame capabilities of models as quantifiable phenomena, which can be determined based on static, decontextualized testing data. For example, Gemini Ultra's performance in the "general" category was reported through a single number (90 %) meant to illustrate it's superiority compared to ChatGPT (86.4%) [109]. Longstanding scholarship in the sociology of testing illustrated how testing of technologies always embodies partial and situated perspectives [62, 111], and can not account for all eventualities. Recent scholarship in ML [24, 41, 115] and critical studies of algorithms and AI [17, 32, 57], is pushing back against an overreliance on a few metrics and pointing out current harms of this practice. Google's presentation of capabilities offers a flawed impression of certainty. This is concerning as such descriptions carry an aura of scientific authority as they are based on academic benchmarks [114], which may result in misplaced trust in the applicability and reliability of the technology and produces hype based on grand promises. The circulation of numbers, demonstrations, and descriptions is also an act to shape how individuals and the broader public understand capabilities. In this section, I will discuss two alternative ways to understand and theorize capabilities with more nuance in this current environment of massive investment and contested claims around generative AI.

2.1 Capabilities as Contested Constructions

Companies, and other actors with a vested interest in the success of generative AI products, typically present new models and technologies in ways that try to convince others of the models' utility, shaping product expectations [18, 57]. For example, similar to Google, OpenAI presented in public announcements the capabilities of it's GPT-4 model by highlighting academic metrics [105], including MMLU, that promise to provide insights into the performance of it's otherwise opaque system mostly inscrutable to outsiders. The public circulation of such figures while maintaining the appearance of a scientific claim, is often more about enticement. Thus, capabilities should be also understood as constructions shaped by established power imbalances and infrastructures. In the examples, Google and OpenAI use the performativity of tests to advance a particular construction—one that presents its new model in a favorable light, influencing wider discourse. Such efforts matter to companies; they are acts of power, and, when successful, attract attention and investment. Correspondingly, uncontested performances of failure can be costly. For example, an error made during a product showcase by Google Bard decreased Google's company valuation by approximately 100 billion dollars in a single day [103, 137].

The presentation of seemingly high performance metrics are based on curated tests that represent a situated, partial perspective of a given capability [57]. The "ground truth" these tests depend upon is based on assumptions about defined tasks and constructs to be measured [71], and the labor of data workers [55]. These workers are instructed to mitigate ambiguity and uncertainty based on guidelines meant to produce consistency [93], which is required to achieve numbers

indicating high performance [57]. Inherent uncertainty is often resolved by either ignoring contentious labeling choices or defining majority opinions as truth, making other possibilities invisible while products are presented as highly accurate [19, 57]. These large meta-benchmarks and datasets usually contain various errors (e.g., MMLU [46]), and reproduce often sexist and racist stereotypes (e.g., ImageNet [15]). Since these problematic ideas are considered true in the tests, they also certify their embedding in AI products [57]. Depending on test implementation, which includes choosing metrics, metric aggregation methods, prompting techniques, and tasks for meta-benchmarks, calculated performance numbers can vary widely, producing differing impressions of performance [35, 57]. Since large models are based on so much data, testing them faces the challenge of memorization. It may inflate their perceived performance as they seem to excel in certain instance while failing in others where no accepted solutions were part of their training data. Various additional issues with current testing practices have also been examined in prior scholarship [19, 41, 50, 57, 83, 115].

Currently, many computer scientists and technology companies have a vested interest in the success of generative AI. When they construct, choose, and present, often inscrutable or unverifiable tests, they are often embroiled in a conflict of interest. For example, various scientists at Microsoft Research released a preprint paper claiming that ChatGPT was an "early (yet still incomplete) version of an artificial general intelligence" [23]. Various scholars have criticized Microsoft's paper, noting the lack of available data for independent verification of its claims, and the simplistic and problematic notions of intelligence that it pushes [91, 99]. Without a wider ability to directly disprove the paper, mystical claims of intelligence capabilities become simply a matter of trust—whether to place one's trust in the authors or their critics.

This example also highlights how public constructions of generative AI capabilities are contested. As these models claim to be general purpose, and possess complex and ill-defined capabilities, this is to be expected. The goals of actors are often not concerned merely with deepening understanding of AI technologies, but rather with exerting power and redefining the meaning of powerful ideas like intelligence and humanity in ways that benefit big technology companies. The stated goals of many of this products is to achieve Artificial General Intelligence (AGI), which OpenAI defines as "highly autonomous systems that outperform humans at most economically valuable work" [104]. This definition highlights a drive toward replacing human work. Tests are seen as one way to certify progress toward this goal as companies illustrate, by comparing the performance of generative AI products to humans via standardized exams [14, 105, 135], and these efforts also reveal how AI tools could aid or replace humans in the workplace. Such territorial claims on concepts are often realized through mapping and redefining certain bodily experiences, ideas, and human practices as operational tasks.

The belief in the emergence of new capabilities until generality is reached is based, in part, on the old promise of big data [134] — that size, complexity, and variety can bestow large statistical models with an almost enchanted level of knowledge and capabilities [26]. A second, and likely more interesting, foundational promise is that these models provide different modalities and operators to recombine and remix what they encountered in training, allowing them to produce outputs not present in the training data. This way of interacting with models makes it possible, for example, to generate images or texts in a certain style or, put differently, to combine and generate based on identifiable and nameable patterns.

I interpret this search for capabilities as a new machine learning testing regime, concerned with identifying and defining tasks for which generative AI products could be used. This includes tasks like completing human exams or playing chess and similar games. In contrast, the previous regime, which also still persists, was more concerned with building and testing models for singular tasks. This new regime is based on the belief in emergent capabilities - the unexpected availability of capabilities as models get larger without being specifically trained for them. Recent work [121] has deconstructed this idea, by investigating several papers that propagate this claim. These scholars

have highlighted how metric choice and test data resolution made it appear as though capabilities emerged suddenly and unpredictably on very large models. This work thus also highlights how emergence is a matter of perspective and expectation, as different test constructions, notions of suddenness, and notions of unpredictability, may result in different interpretations. The structures that incentivize the search for new capabilities remains intact. It is thus likely that new capabilities will continue to be discovered, while other researchers will continue to demonstrate how these tests and associated notions could be constructed differently—dismissing the conclusion that new results are particularly surprising or grand.

Beyond standardized tests, tech demos can also be involved in the public construction of capabilities. For example, when Google published videos on YouTube to demonstrate the capabilities of its new Gemini model, the company was criticized for providing a misleading impression of the model's response speed [92]. Social media plays a role, too, as influencers post test results and discuss their positive or negative experiences with the technology. The initial high adoption of ChatGPT can likely be attributed to the technology going viral online, further highlighting the importance of social media for the introduction of new technologies. Google, therefore, introduced Gemini with a video [53] featuring famous YouTuber Mark Rober. It was meant to illustrate how general and useful the tool can be, even aiding Rober in video creation. This demonstration was not particularly successful, with commentators observing that Rober had to guide the tool, and how some of the tool's suggestions were fairly mediocre. Of course, the broader traditional media landscape is involved in the construction of capabilities, as journalists report on new tools and write about their own experiences using them [28, 76]. While some journalists also seek to curb hype, concerningly, increasingly media depended on big tech infrastructures and funding, and struggle to operate in a highly competitive environment [125]. Consultancy firms and the reports they release also play a major role fueling hype on capabilities [85].

2.2 Capabilities as Situated Accomplishments

Beyond theorizing capabilities as constructions, they should also be understood as situated accomplishments of different human and non-human actors [136]. This lens decenters the model or technology, and instead highlights how capabilities are relational, and emerge in particular contexts of use. To accomplish an action, capabilities require interactions between human and non-human actors — technology use requires ingenuity and creativity from users, who are part of a situated network. In the case of generative AI products — a new technology where users have very little prior experience to draw upon—users must undertake a lot of creative work to make the technology useful to them in a given situation.

This theoretical lens thus focuses on the network that makes an accomplishment possible. The network may include prompt influencers on social media who make a user aware of how to elicit a desired response. For example, prompt influencers have highlighted how informing a chatbot that it would receive a "tip" can lead to better responses [**?**]. In another case, one ChatGPT user was able to correctly identify an illness that doctors were unable to diagnose [**?**]. Such demonstrations also socially construct imagined affordances [40, 97, 98] of generative AI products, which enable certain uses. Imagined affordances emerge as an interaction of user expectations, the materiality of technologies, and the intentions of designers. The concept helps us appreciate how public demonstrations enable users to see the technology differently, and to change their expectations, as they recognize how it might be used in new ways. This is especially important for a new technology, such as generative AI, where people's imagined affordances are unstable. Indeed, this can be understood as a form of articulation work [68, 130], that makes systems 'work' for users. However, this also poses a danger because certain accomplishments may not translate. For example, ChatGPT may provide completely bogus health advice when confronted with a different prompt and symptoms. Such unexpected errors are often called hallucinations, and many potential users may not yet be fully aware of machine learning systems' brittleness . As

imagined affordances stabilize, both good and bad surprises regarding AI responses are likely to become less prevalent, and prompt influencers thus less relevant. Prompt influencers could also become obsolete as interfaces are refined and adjusted, and generative AI products become personalized — prompts may increasingly yield different results depending on the particular user.

This network perspective on capabilities also seeks to foreground infrastructures, materiality, and the largely invisible actors who enable generative AI products. These products should be understood as social technologies, because they are trained via internet content produced by people, and tuned and refined based on people's preferences. Users can only make sense of these tools' responses because tools are built upon extracted, chatbot-optimized, patterns of recognizable sociality. This also means certain social cues can have non-intuitive consequences, as the tipping example above illustrates. In the next section, I unpack the stabilization of testing infrastructures that enables capabilities' discursive and material construction.

3 STABILIZING EXTRACTIVE TESTING INFRASTRUCTURES

Recent, high-profile advances in generative AI technologies are typically attributed to big technology companies ability to use extensive data and computing power to produce large models. These companies monopolize the possibility of creating these models as only they have the required funds and infrastructure [88, 140]. Some companies and research organizations open-source large models. This allows other actors to fine-tune them using new, context-specific data, resulting often in increased performance on specialized tasks and to extend them with new interfaces to different technologies, including other models [140]. In this section, I unpack how AI model and product testing is externalized, often extracting value from low-paid or unpaid labor supplied by researcher and developer communities, content creators, subcontractors, and users. These actors form part of key globalized testing infrastructures, which stabilize the enactment of ever-improving generative AI products. Testing goes beyond aiding understanding, or advertising the performance of models and systems on certain tasks; testing provides valuable guidance and can aid model improvement. For example, new benchmark data can be used to train, fine-tune, and adapt models and newly-crafted prompting techniques, too, can be used to improve performance.

3.1 Researcher and Developer Communities

Metric-oriented standardized testing, based on corresponding benchmark datasets, has been foundational to ML because it promises to quantify progress in the field and industry [42]. Dedicated platforms exist that enable ML practitioners to release and test models against different benchmarks, with "leaderboards" used to rank and compare them [10, 87]. Some platforms are run by researchers and open-source developers, while others are run by large companies, like HuggingFace. Researchers and other ML practitioners often develop new tasks and benchmarks, resulting in additional leaderboard websites that chart progress over time. These leaderboard and testing platforms benefit from the participation of researchers, companies, and developers, who contribute new or finely-tuned models in order to beat benchmarks and set high scores. This can be considered a form of gamification — whereby participants are individualized and an aura of play obscures how companies extract value from their largely unpaid labor [25, 126]. The practice in ML of relying on few metrics to quantify progress also potentially inflates bubbles fed with expectations that may burst and result in another so-called "AI winter." Such boom-and-bust circles have characterized the AI field for decades [61], and ML practitioners are aware of these issues. Nevertheless, due to various incentives, the practice persists. With funding and recognition regimes deeply tied to industry, academic actors' agency is circumscribed [29], as researchers balance demands for

impartiality and public interest in science on the one hand, and the need to maintain good long-term relations with big tech companies on the other — companies whose influence is sometimes akin to powerful governments [8, 142].

Researchers may develop benchmarks in order to publish papers, further their careers, and attract funding, with the amount of funding increasing as testing becomes a bigger concern for companies. The data work necessary to create benchmarks was historically neglected by academics, as this was perceived to be less prestigious than building and refining models [119]. Recently, however, creating benchmarks has become a more valued task. Big tech companies have created incentives that make the testing of their systems appealing to researchers. First, they framed their systems as general purpose, which invited researchers to investigate this claim and create tests for all kinds of tasks. Second, by making their models only accessible to a few select external researchers and developers, testing became one of the few ways that the top-performing models could be engaged with by third parties. Government agencies, like National Institute of Standards and Technology [141], are also increasingly involved in building benchmarks, also to increase trust in the current landscape. It is likely that more big initiative involving different actors will be created as trust in current leaderbaords and benchmarks may decline over time as performance claims of products stagnate after constant rise and big testing datasets and new standards may be seen as one way forward.

Beyond benchmark data and testing methods, ML practitioners are also concerned with identifying security issues through "red teaming" and new prompting techniques, as these also influence the quality of results. The recognition of the importance of prompts has even resulted in the creation of a new role: the prompt engineer [43]. ML practitioners are incentivized to contribute by making sense of prompting for products and what it enables and identifying valuable prompts that can be resold or published. Companies and researchers also conduct studies with users and potential customers to investigate how far generative AI can support user practices. As user experiences become more central, with greater recognition paid to the limitations of quantified benchmarks, these tests will likely only increase in importance. Some platforms already facilitate leaderbaords based on community assessment of responses of different models [69]. Many startups also take on the risk of testing different monetization models for generative AI applications, and successful startups are typically acquired by big technology companies at a certain stage. Various startups also develop their own models, but recent acquisitions [81] highlight how, upon company purchase, their benchmark data, models, and experience are all internalized by big technology companies.

The promise of AGI also ideologically entices various ML practitioners to contribute to the project of testing generative AI. This encompasses controversial AI safety initiatives that fund and push testing to identify risks deemed existential for humanity [74]. For example, they conduct workshops and studies in order to assess, via expert forecasts, when certain levels of accuracy will likely be reached on prominent benchmarks [2]. When sufficient accuracy is achieved before the predicted time, these studies function as evidence of the urgency regarding the emergence of dangerous AI capabilities. These notions of risk have been critiqued for diverting attention toward unlikely risks that shield big tech companies from critique on more pressing current issues and harms of their products [2]. As a result, such efforts to capture attention and regulatory oversight may further tilt testing infrastructures — and the (unpaid) labor that sustains them — in directions that benefit big tech companies at the expense of marginalized people.

3.2 Subcontractors

Some of the most important testers in this globalized testing infrastructure are also some of the most invisible [55] to many generative AI customers: underpaid gig workers and subcontractors. They are usually part of complex data supply chains [36] that often obfuscate outsourcing from companies in the Global North to a precarious workforce in Global South contexts (also within the Global North), providing no or little means for worker recourse [93]. Their labor

develops the baseline for acceptable results from generative AI, which is then used to fine-tune or align models via methods such as reinforcement learning from human feedback (RLHF). One of these actors' goals is to identify results that could cause controversy and reputational damage. This includes "red teaming" work to identify toxic, violent, blatantly incorrect, and other egregious responses. These actors can be thought of as content moderators for generative AI products, who aim to identify responses that could become controversial [52]. At OpenAI, this internal safety testing and alignment work was seemingly conducted to such a degree that OpenAI managed to avoid, or be less affected by, the sort of controversies that have tanked previous chat programs. For example, Meta's Galactica model was taken offline after three days due to its problematic responses [64], and Microsoft's TayBot learned from interactions in ways which made it blatantly racist [70]. ChatGPT is also plagued by these issues, but seemingly in more subtle ways that did not hamper its early adoption.

The use of these generative AI technologies is also an experience that includes a performance of "artificial" intelligence. Human testing thus not only involves rational concerns, like correctness and accuracy (as may be claimed in tech demos), but also focuses on user experiences — which are intended to leave people in awe. The initial hype around ChatGPT, in part, can likely be attributed to the intentional decision to tune the model toward anthropomorphizing itself, thereby performing human intelligence [34]. Given this, subcontractors also undertake the affective labor of adjusting products' tone to make the tools feel helpful and like an "AI." Subcontractors receive guidance on how to rate AI responses against certain defined preferences meant to capture broad public opinion, such as helpfulness. Subcontractors receive guidelines on how to rate responses according to defined preferences meant to capture broad public opinion. OpenAI employees create these guidelines and also consider external input from experts and the public [104]. Thus, the work of subcontractors is about optimizing these tools in ways to appeal expectations of customer majorities, and to avoid responses that can create controversies. Such policies have been successful for tech companies, for example, some claim to be neutral platforms [51] and others use simple fixes instead of really addressing underlying social issues to curb controversy [54]. OpenAI released example instructions for reviewers such as refusing requests for inappropriate content and "avoid taking a position on controversial topics" [104], but, when explicitly asked, also create arguments for a controversial position. Thus, the goal is seemingly to model the default behaviour after perceived dominant discourses. This, in practice, often means adhering to a hegemonic white individualist male standard recognized as "normal" by many customers [11, 101]. The product is also fine-tuned to create desired responses for different non-dominant positions, as noted above, when explicitly prompted.

3.3 Users of Generative AI Products & Content Creators

Generative AI companies make their products available to wider audiences by providing specifically crafted chat interfaces. There are different models for managing access; for example, OpenAI's ChatGPT offers premium features and models to paying customers, while also providing freely accessible alternatives. The company was critiqued during its initial release of ChatGPT for being irresponsible, for not having enough safeguards, and for the immaturity of the technology [117]. However, the company followed the old Silicon Valley motto "move fast and break things," and focused on speed [102], which meant that a lot of testing the viability and safety of the technology was externalized to the market [108]. This configuration benefits these companies by providing ample real-world testers, whose interactions and feedback can be integrated into the models through fine-tuning via human preferences. This allows companies like OpenAI to scale the refinement of their products and business models using the unpaid labor of users, an old practice of internet companies [133]. Users produce interactions, which are then turned into data to potentially enable companies to identify patterns and common prompts. These could enable companies to optimize their products based

on actual majority preferences of their users and to align themselves with their perspective. This may become an issue for minority communities, as the failures and harms they experience are often not as visible in the test data. However, this approach greatly benefits companies with large user bases, as such data may cement their strong market position. Thus, various companies in an effort to catch up with OpenAI have begun to integrate generative AI interfaces in all kinds of products [20].

Many generative AI products also have built-in feedback mechanisms, which can be used to rate responses or warn of toxic or copyright-infringing results. These are also likely fed into models for improvement and may be supplied to subcontractors so they may identify better responses. Companies also learn about certain issues, and what matters to users, through monitoring social media posts about users' prompting accomplishments and controversial failures. For example, when users complained about the laziness of ChatGPT, they promptly got a response [89]. Similar fixes were introduced after complaints were made regarding the generation of copyright-infringing content. This kind of system poses a challenge for marginalized voices, which may not be perceived as sufficiently important, powerful, or possessing enough reach to justify adaptations. It may also be the case that generated content available online is also identified to learn about preferences. It may also concerningly be used to remove frequent critiques highlighting important issues of tech company products. Technology journalists also write extensively about their and users' experiences with these products [110], and their writings are likely closely monitored by companies. All these unpaid inputs provide technology companies with information concerning how to improve the experience and usability of their products.

Internet content is also used in benchmark datasets for model training, testing, and fine-tuning [107]. Content creators and providers, therefore, are often unknowingly enrolled in the testing infrastructures of generative AI products. For example, benchmark datasets for cross-lingual summarization draw on content produced by Wikipedia volunteers and Global Voices translators [78, 100, 139]. Due to big technology companies' lack of transparency, it is unclear whether or how this particular datasets are used. Nevertheless, the drive to include all kinds of available online content into valuable data makes inclusion very likely. The new content produced and disseminated by creators, as it fuels generative AI, also enables all kinds of downstream uses that original creators may disagree with, such as problematic surveillance and classification uses [60].

4 RETHINKING TESTING PRACTICES AND INFRASTRUCTURES

The idea of general-purpose AI comes with a drive to measure, and to turn practices into quantifiable tasks. This goal can never fully be achieved thus (testing) performances are enacted to stabilize certain (limited) notions of universality. The perceived brittleness of generative AI results from the interplay of, assumed capabilities — due to surprisingly successful prompts, and big promises of generality and intelligence — and the surprise when other prompts fail. As these systems are trained on ever greater amounts of internet data and interactions, it may be possible for them to produce better outputs and more successful user experiences in many situations. This could be fueled by memorization while not addressing the issue of the long tail. Indeed, as certain less common errors become rarer, they may become harder to detect, and thus less expected, potentially increasing their negative impact. The current general-purpose paradigm is resource-intensive and slow for many tasks [86] compared to specialized, native implementations and will likely continue to surface also nefarious and dual-use capabilities. It may not be possible at this time to significantly change these current trends stabilized by large investment, but efforts like ongoing litigation may still influence its direction, e.g., copyright lawsuits may create a situation more favorable to creators. It will be also essential to improve knowledge making and communication about these systems, which is why a move away from a focus on a few metrics toward a critical and reflexive testing paradigm is needed.

Such a testing paradigm should foster justice-oriented, independent testing, with the goal of developing sensibilities and infrastructures to include marginalized perspectives [63], as, for instance, illustrated in the literature on participatory evaluation [21, 127, 128]. It also entails comparing systems also via qualitative assessments, for instance, through extensive reports [50] or ethnographic user studies [5, 79, 122, 123]. This also requires the creation of interdisciplinary teams, translational work across academic boundaries, and working against disciplinary hierarchies to welcome insights from fields such as STS [57, 124] and HCI [82] in ML. Such critical testing should also prioritize labor concerns across supply chains and aim to develop sensibilities and guiding lenses that acknowledge inequalities, and necessary assumptions. Since more independent, critical testing is resource-intensive and requires infrastructure and access, it may be against established interests and therefore its implementation will require public pressure, changes to incentive structures, and decisive oversight. In the remainder of this section, I will discuss suggestions for more critical, reflexive, and humble testing.

4.1 Reflexive Communication of Tests

The opacity, scale, and fluidity of generative AI products poses difficulties for claiming reliable capabilities. As higher metrics make testing results appear more certain, and therefore often more difficult to challenge [48], it is important to point to how tests and metrics always enact a partial perspective. Tests need to be handled with care in public communication [75], which requires that meaningful descriptions of results need to be constructed for different audiences. These should include possible conflicts of interest in test construction, epistemic values and assumptions [17], marginalized and critical perspectives that may have been sidelined, and possible and current harms and uncertainties that result from the system and the test. Finally, the supply chains of both the system and test need to be communicated, entailing transparency regarding human labor and possible exploitation, documentation of data and modeling decisions, training and testing data, and information on task and metric construction. The current practice to simply claim capabilities based on often contested benchmarks should be abandoned.

It is important also to rethink current conventions around naming and descriptions in the field. For example, naming a metric “accuracy” can be problematic, since this naturalizes a problem-framing via the presentation of a single, oversimplified number that supposedly captures multifaceted capabilities and limitations. Instead, following the principle of epistemic humility [27, 72], metrics that measure agreement between the results of an algorithmic system and selected test data could be renamed “test correlation” or "test agreement" in public communication to curb expectations. Overstating via sensationalist naming is a systemic problem in industry and academia. Scholars have challenged terms routinely used in the field to ascribe human-like capabilities to algorithmic systems, such as “artificial intelligence,” “learning,” “training,” or “understanding” [22, 39, 45, 132]. Other important examples include the term “ground truth,” which implies that human-labeled data represent some absolute truth about the world [73], “data-driven,” which suggests that human judgment does not matter, or the term “prediction,” which insinuates that systems are able to reliably forecast the future [32]. This issue also extends to how tasks are named, which, in public communication, are then often turned into capabilities. For example, “emotion recognition” wrongly implies that actual emotions can be recognized by algorithmic systems [58]. There is thus a need for the careful renaming of concepts and terms to improve descriptions of capabilities and limitations for different audiences. For example, a humble description of “emotion recognition” could be “expression classification,” based on correlated human interpretations of staged images. This description highlights how algorithmic results are based on human judgment, correlations, and partial observations. Academia, in particular, is well-positioned to come up with new, improved terminologies.

4.2 Engaging Metrics Reflexively

Since quantified metrics will endure in some form due to incentives and their importance to ML, it is important for practitioners and critical scholars to theorize, possibly challenge and rethink them. The field of STS offers methods and ideas that can enable qualitative analysis of metrics, yielding potentially more complex and useful interpretations and descriptions of generative AI products, other algorithmic systems, and their results. Below, I discuss several useful guiding questions, aimed at enabling reflexive engagement with metrics in ways that seek to center social concerns.

First, why does an algorithmic system perform well on a chosen metric? That is, what makes a phenomenon predictable and stable, or, more generally, what socio-technical mechanisms make a metric appear (un)successful? An STS-informed analysis could examine which human and non-human actor-networks, power structures, and historical conditions stabilize a phenomenon, situation, fact, or structure and thereby make it more predictable [12, 80, 94, 112, 138]. Generative AI technologies are based on embeddings, which seek to capture similarities and relations in a constructed space. In turn, other important questions may be posed regarding what produces similarities in collected data. For example, social inequalities embedded in texts partly explain why concepts such as "nurse" are more closely aligned with women, while "doctor" is aligned with men. Subsequently, the goal is to examine the politics of predictability and the structures that enable models to become impressive representations of social patterns. Such an analysis denaturalizes performance claims by showing how social practices, discursive power, and material structures — which need to be constantly remade — enable the functioning of an algorithmic system. The analysis shows, too, how things could be different. This kind of analysis constructs deep, situated, and critical causal relations, employing a perspective that questions established assumptions and power structures.

Another related question is: how can correlations that enable high performance metrics be interpreted or constructed differently, such as by grounding analysis in prior, domain-specific literature? ML and generative AI focus on correlation instead of causation, due to the current contested, data-driven paradigm. However, controversies have motivated research into interpretability, seeking to uncover correlative and causal relations based on different needs and priorities [113]. The technological solutionism [84, 96, 118] embedded in the field has motivated the development of new tools — ones that aid in the analysis of such relations — but these cannot replace human judgment. Moving beyond mostly quantitative approaches, such relations can also be uncovered and classified by theorizing, based on information regarding the algorithms, data, and how they are situated. It requires effort to collect and archive such information, which is currently devalued in ML since this analysis does not yield a new algorithm or model, but a new qualitative interpretation instead. Documenting expected causal relations in publications, and other contextual information relevant to understanding algorithmic systems, can also aid researchers in judging the quality of claims made by, and about, these systems [49, 77]. This kind of correlation documentation can also aid in deciding which relations are spurious and problematic, by interpreting and grounding them in empirical studies and pre-existing, domain-specific literature. For example, climate science illustrates how combining evidence from multiple research efforts, perspectives, simulations, theories, measurements, and experiments can produce very strong and trustworthy facts [44].

The qualitative analysis of metrics should also ask: what do they make visible and invisible? The sensibilities that situational analysis [33] provide are useful, as they encourage a search for absences — using maps that can be constructed to encompass an algorithmic system, its supply chain, testing practices, and how these are embedded in different situations. Another important question is: what performance is "good enough," in what context, and for whom? Generative AI applications and other algorithmic systems are not meant to be completely error-free, but instead built to be "good enough" following an engineering logic [7, 37, 106]. Because these systems and models impact people, decisions

about acceptable "working" levels of harm and error are not simply technical matters [129]. Indeed, since modeling and testing are contingent and require human judgment, decisions concerning acceptable levels of performance should also form part of evaluation and interpretation efforts. Investigating what is considered acceptable requires engaging with experts, decision-makers, and affected people. However, simply focusing on thresholds is not enough, as model development and evaluation usually involves multiple, judgement-related steps that also highlight a need for democratic deliberation.

Another central question is: from what partial perspective is an algorithmic system considered performant? That is, what standards are used to evaluate an algorithm? For example, according to available test datasets at the time, facial recognition systems were claimed to be highly accurate. However, the reported metrics were only an indicator of the degree to which the system was consistent with the worldview embedded in the test data. Critical researchers [24] pointed out that dark-skinned people were underrepresented in these datasets, and thereby revealed how the facial recognition systems only appeared accurate according to a standard traceable to white ignorance [95, 131]. In a policing context, facial recognition systems used for identification could also be understood as inherently inaccurate. They are surveillance technologies that enable very narrow ways of seeing the world, focused on producing lists of potential suspects, reproducing logics of criminalization [9] while not providing a deeper understanding of humans and their circumstances or making alternative, structural measures visible, such as social support [47]. This kind of investigation reveals the politics behind a framing that appears at first glance merely technological and instead shows how a ban in this context can be considered be appropriate. Critical testing should involve questioning also naturalized standards, especially when they may involve harm. This entails also questioning generative AI's capabilities or the technology altogether. Furthermore, it means also pushing back against restrictive dichotomies in testing, such as the so-called fairness-accuracy trade-off well-known in the FAccT community [4, 56], which poses that increasing fairness decreases accuracy, but can be also understood as a discursive negotiation over standards and whose perspectives matter. Engaging such situations with a participatory approach [128] is beneficial as marginalized evaluation standards may need to be purposefully co-created. As generative AI products proliferate and become multi-modal, they may be also released with facial recognition capabilities. Due to privacy concerns, ChatGPT is currently configured to only identify public figures, and does not assess faces' gender or emotional states [66]. It is unclear how well this guardrails work and whether OpenAI and others will change their stance. Established critical scholarship on algorithmic systems remains highly relevant, also as their capabilities may be integrated into generative AI products.

4.3 User-Agency and Interactivity

Problematic ascriptions of certainty to singular results of algorithmic systems have resulted in calls for reopening their multiplicity and interpretative flexibility [7, 32]. This requires shifting agency to marginalized and affected people, and away from the powerful organizations that control these systems. Interactivity promises to shift agency by privileging options and making alternatives visible. Thus, fields such as HCI have been exploring different ways that interactivity can be added to automated systems for a long time [6, 67, 90]. There may be a trend towards more interactivity and user-agency, because various controversies have pointed out shortcomings of algorithmic systems that provide singular, seemingly optimal, results, or, responses in the case of generative AI chatbots. For example, instead of providing one translation calculated to be the most likely, contemporary translation services increasingly provide multiple translations when confronted with ambiguous, contested, or sensitive content [116]. Another example is the replacement of Twitter's fully automated cropping algorithm [16] with an interactive interface that allows users to manually crop images as needed [29]. The change was made after users critiqued the algorithm for reproducing sexist and racist stereotypes [31].

Twitter's algorithmic system was repaired by enabling users to do the cropping appropriately, according to their own contexts and needs, and as soon as algorithmic failures became visible. This also partially shifted the responsibility of making the system "work" onto Twitter's users. Similarly, DALL-E, an image generator, externalizes judgment of results to it's users, who tweak parameters, select results that look interesting among many subpar choices, and circulate them, creating publicity for the system. Any failures, meanwhile, like bad queries and unbecoming results, are often individualized and not made public, as tweaking is expected to be the responsibility of users.

The so-called "AI winters" in recent decades have corresponded with increased attention and funding for HCI research, and vice versa [61]. I understand this back and forth between an interaction/HCI and optimization/AI paradigm as an instantiation of the structure-agency dialectic in computing. On the one hand, automation, inference, and optimization introduce standards and calculated presumptions, which may enable certain practices and perspectives while restricting others [129]. In contrast, more user-control, interactivity, and humility can support agency, enabling more people to co-shape results according to their needs by reducing certain default settings. This may introduce friction, as fewer actions are made automatically by algorithmic systems. This relation is dialectical, which means the degree to which agencies and structures are present is negotiated until algorithmic systems are stabilized as infrastructure [112].

Algorithmic systems can be made more interactive, for example, by providing more people with options to change parameters and other inputs to adapt results according to contextual needs. Various generative AI products allow users to generate different responses from the same prompt via chat interfaces. The products also suggest different prompts that may aid in reaching more desirable results, and, by providing these options, also introduce some potentially helpful structure to prompt selection, given the endless possibilities available. Such designs can enable the exploration of different scenarios and results, thereby turning generative AI systems that otherwise may close down debates and reflexivity — by providing seemingly optimal and certain results — into tools with more humility that provide a range of alternatives. Another example is climate modeling, which illustrates how simulation can make predictive reasoning about the future more emancipatory as it foregrounds human agency — an agency that can produce many different futures instead of merely following established path dependencies and power structures. The current optimization paradigm in computer science only provides insights into a view of the world calculated to be the "most likely," according to a limited perspective embedded in data and modeling decisions. Such insights can be valuable, but other, more desirable, shared worldviews that may require more work to achieve are also possible. Such deliberations should be settled through democratic processes instead of the current approach, which frames these matters either as merely technical decisions for engineers or as something to be settled by markets and powerful companies.

The aim of this section is not to principally argue against algorithmic systems that provide singular, optimized results with certainty. Depending on what possibilities they enable or foreclose, how they engage with power structures, and how they are specialized for, and embedded within, contexts that include input from those potentially affected or relevant experts, they can be preferable. Although interactivity can increase the agency of marginalized people in meaningful ways, interfaces also always encompass scripts and structures that co-shape what kinds of interactions are possible, encouraged, or made seamless [3, 90]. For example, simulations provide parameters that can be adjusted to realize different scenarios, but available options are still based on assumptions and circumstances that determine what is made visible. Also, generative AI applications that allow a great variety of inputs are dependent upon lots of hidden organizational decisions and parameters. Agency is also co-determined by how the outputs and systems are presented or made available within a societal context. For example, generative AI chat products like ChatGPT are presented as rational and intelligent. Their tone and responses are purposefully designed to create the appearance of a convincing, human-like intelligence, thereby making it difficult to see how the products embed values and could be different.

It is important that interactive systems highlight their multiplicity and practice humility through design, and important, too, for critical scholars to call out those who only use interactivity to create an illusion of agency and control whilst stabilizing exploitative structures. The testing of generative AI products should also aim to account for this, with human agency to be further explored beyond focusing merely on simplified metrics, such as task productivity in user testing. Lastly, by increasing the interactivity of algorithmic systems, and giving people more agency, some responsibility is also transferred from technology providers to people. This can also be a problematic tactic to individualize harms and other risks. For example, reporting systems in generative AI products shift the responsibility for removing problematic content onto people. Instead, the company that profits from the platform should be seeking to remove such content proactively. People may not have the time, means, or power to interrogate the systems they are engaging with. Ultimately, such systems need to be designed and maintained in ways that disrupt the unequal and discriminatory status quo, and in ways that make justice-oriented perspectives the seamless default [1, 13, 38]. Interactivity by itself is not a fix.

5 CONCLUSION

This paper unpacked testing infrastructures for generative AI, and suggested improvements for testing more generally. It is not simply a call for more testing, as this would not change the underlying testing culture that prioritizes business interests. Such a call may, in the end, only produce a bigger sea of seemingly conflicting test results, where powerful companies' claims stand against those of critical researchers while companies perpetually suggest improvement is on the horizon. Although the institutionalization of testing is moving forward, it is often captured by technocrats while the matter of testing is political. This undermines the emancipatory promise of testing — to provide understanding and uncover the harms caused by these systems that are often neglected by big tech companies. As this rise in testing unfolds, it will be important that policy makers ensure support and funding for critical, independent testing, like external auditing [120], and not simply establish a universal standard likely capture by corporate interests [59]. Testing depends on perspective and is political. There is a risk that external testing just becomes more unpaid labor, absorbed into company testing infrastructures that move quickly to fix and/or cover up controversial discoveries without addressing underlying deeper, structural challenges. Thus, critical testing should also seek to produce lasting critiques that point to value conflicts and question naturalized standards with the goal of repoliticizing the discourse around technologies like generative AI that affect many people potentially in detrimental ways. Critical testing should also be able to question intended behaviors that produce harm, interrogate uses or development, and consider inequality and power in analysis. It is also important to communicate, and debate the limits of testing, which can never anticipate all harms that may occur. The current incentive structures make it difficult to conduct critical research and communicate with humility, and the ecosystem surrounding ML must support this kind of work more actively. A culture change is needed in ML, which entails building different testing infrastructures that prioritize minoritized perspectives and needs. Policy makers must intervene and create support structures for critical and reflexive research, because, left to itself, the market will likely only further entrench inequalities and path-dependencies. Industry should start by ensuring that data workers receive decent wages, and the broader community should value their essential work.

ACKNOWLEDGMENTS

I would like to thank Silvia Lindtner, Christian Sandvig, the Tech.Culture.Matters. research collective at the University of Michigan, and the Politics of Machine Learning Evaluation workshop participants for their insights and the reviewers and chairs for their helpful feedback.

REFERENCES

[1] Sara Ahmed. 2010. *The Promise of Happiness.* Duke University Press.

[2] Shazeda Ahmed, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. 2024. Field-Building and the Epistemic Culture of AI Safety. *First Monday* (April 2024). https://doi.org/10.5210/fm.v29i4.13626

[3] Madeleine Akrich. 1992. The De-Scription of Technical Objects. (1992).

[4] Doris Allhutter. 2021. Memory Traces in Society-Technology Relations. How to Produce Cracks in Infrastructural Power. In *Edited By*. 426.

[5] Katrin Amelang and Susanne Bauer. 2019. Following the Algorithm: How Epidemiological Risk-Scores Do Accountability. *Social Studies of Science* (July 2019), 0306312719862049. https://doi.org/10.1177/0306312719862049

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300233

[7] Louise Amoore. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others.* Duke University Press, Durham.

[8] Elizabeth Anderson. 2017. *Private Government: How Employers Rule Our Lives (and Why We Don't Talk about It).* Princeton University Press. https://doi.org/10.1515/9781400887781

[9] Kirstie Ball. 2009. Exposure. *Information, Communication & Society* 12, 5 (Aug. 2009), 639–657. https://doi.org/10.1080/13691180802270386

[10] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open LLM Leaderboard.

[11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big. *Proceedings of FAccT* (2021).

[12] Ruha Benjamin. 2016. Catching Our Breath: Critical Race STS and the Carceral Imagination. *Engaging Science, Technology, and Society* 2, 0 (July 2016), 145–156. https://doi.org/10.17351/ests2016.70

[13] Ruha Benjamin. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code.* John Wiley & Sons.

[14] Celeste Biever. 2023. ChatGPT Broke the Turing Test — the Race Is on for New Ways to Assess AI. *Nature* 619, 7971 (July 2023), 686–689. https://doi.org/10.1038/d41586-023-02361-7

[15] Abeba Birhane and Vinay Uday Prabhu. 2021. Large Image Datasets: A Pyrrhic Win for Computer Vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158

[16] Abeba Birhane, Vinay Uday Prabhu, and John Whaley. 2022. Auditing Saliency Cropping Algorithms. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Waikoloa, HI, USA, 1515–1523. https://doi.org/10.1109/WACV51458.2022.00158

[17] Borhane Blili-Hamelin and Leif Hancox-Li. 2023. Making Intelligence: Ethical Values in IQ and ML Benchmarks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 271–284. https://doi.org/10.1145/3593013.3593996

[18] Mads Borup, Nik Brown, Kornelia Konrad, and Harro Van Lente. 2006. The Sociology of Expectations in Science and Technology. *Technology Analysis & Strategic Management* 18, 3-4 (July 2006), 285–298. https://doi.org/10.1080/09537320600777002

[19] Samuel R. Bowman and George Dahl. 2021. What Will It Take to Fix Benchmarking in Natural Language Understanding?. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4843–4855. https://doi.org/10.18653/v1/2021.naacl-main.385

[20] Brady Snyder. 2024. Meta AI Is Taking over WhatsApp, Facebook, Instagram, and Messenger. *Yahoo Tech* (April 2024).

[21] Sharon Brisolara. 1998. The History of Participatory Evaluation and Current Debates in the Field. *New Directions for Evaluation* 1998, 80 (1998), 25–41. https://doi.org/10.1002/ev.1115

[22] Meredith Broussard. 2018. *Artificial Unintelligence: How Computers Misunderstand the World.* The MIT Press, Cambridge, Massachusetts.

[23] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. https://doi.org/10.48550/arXiv.2303.12712 arXiv:2303.12712 [cs]

[24] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[25] Michael Burawoy. 1982. *Manufacturing Consent: Changes in the Labor Process Under Monopoly Capitalism.* The University of Chicago Press, Chicago London.

[26] Alexander Campolo and Kate Crawford. 2020. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society* 6, 0 (Jan. 2020), 1. https://doi.org/10.17351/ests2020.277

[27] John Carson. 2020. Quantification–Affordances and Limits. *Scholarly Assessment Reports* 2, 1 (2020).

[28] Brian X. Chen. 2023. How ChatGPT and Bard Performed as My Executive Assistants. *The New York Times* (March 2023).

[29] Yuchen Chen, Yuling Sun, and Silvia Lindtner. 2023. Maintainers of Stability: The Labor of China's Data-Driven Governance and Dynamic Zero-COVID. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[30] Ted Chiang. 2023. ChatGPT Is a Blurry JPEG of the Web. *The New Yorker* (Feb. 2023).

[31] Rumman Chowdhury. 2021. Sharing Learnings about Our Image Cropping Algorithm. https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.

[32] Wendy Hui Kyong Chun. 2021. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition.* MIT Press.

[33] Adele E. Clarke, Carrie Friese, and Rachel S. Washburn. 2017. *Situational Analysis: Grounded Theory after the Interpretive Turn.* Sage Publications.

[34] Rick Claypool. 2023. *Chatbots Are Not People.* Technical Report. Public Citizen.

[35] Clémentine Fourrier, Nathan Habib, Julien Launay, and Thomas Wolf. 2023. What's Going on with the Open LLM Leaderboard? https://huggingface.co/blog/open-llm-leaderboard-mmlu.

[36] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding Accountability in Algorithmic Supply Chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).* Association for Computing Machinery, New York, NY, USA, 1186–1197. https://doi.org/10.1145/3593013.3594073

[37] Edward W. Constant. 1984. Communities and Hierarchies: Structure in the Practice of Science and Technology. In *The Nature of Technological Knowledge. Are Models of Scientific Change Relevant?*, Rachel Laudan (Ed.). Springer Netherlands, Dordrecht, 27–46. https://doi.org/10.1007/978-94-015-7699-4_2

[38] Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need.* MIT Press.

[39] Kate Crawford. 2021. *The Atlas of AI.* Yale University Press.

[40] Jenny L Davis. 2020. *How Artifacts Afford: The Power and Politics of Everyday Things.* MIT Press.

[41] Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The Benchmark Lottery. *arXiv:2107.07002 [cs]* (July 2021). arXiv:2107.07002 [cs]

[42] David Donoho. 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics* 26, 4 (2017), 745–766.

[43] Drew Harwell. 2023. Tech's Hottest New Job: AI Whisperer. No Coding Required. *Washington Post* (Feb. 2023).

[44] Paul N. Edwards. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming.* MIT Press, Cambridge, Mass.

[45] Nathan Ensmenger. 2012. Is Chess the Drosophila of Artificial Intelligence? A Social History of an Algorithm. *Social studies of science* 42, 1 (2012), 5–30.

[46] Daniel Erenrich. 2023. Errors in the MMLU: The Deep Learning Benchmark Is Wrong Surprisingly Often. https://derenrich.medium.com/errors-in-the-mmlu-the-deep-learning-benchmark-is-wrong-surprisingly-often-7258bb045859.

[47] Andrew Guthrie Ferguson. 2017. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement.* NYU Press.

[48] Fabian Fischer. 2019. The Accuracy Paradox of Algorithmic Classification. (2019), 16.

[49] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[50] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research* 77 (May 2023), 103–166. https://doi.org/10.1613/jair.1.13715

[51] Tarleton Gillespie. 2010. The Politics of 'Platforms'. *New media & society* 12, 3 (2010), 347–364.

[52] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press.

[53] Google. 2023. Mark Rober Takes Bard with Gemini Pro for a Test Flight. https://www.youtube.com/watch?v=mHZSrtl4zX0.

[54] Nico Grant and Kashmir Hill. 2023. Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's. *The New York Times* (May 2023).

[55] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Houghton Mifflin Harcourt, Boston.

[56] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35, 4 (2022), 90.

[57] Gabriel Grill. 2022. Constructing Certainty in Machine Learning: On the Performativity of Testing and Its Hold on the Future. https://doi.org/10.31219/osf.io/zekqv

[58] Gabriel Grill and Nazanin Andalibi. 2022. Attitudes and Folk Theories of Data Subjects on Transparency and Accuracy in Emotion Recognition. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (April 2022), 78:1–78:35. https://doi.org/10.1145/3512925

[59] Gabriel Grill, Fabian Fischer, and Florian Cech. 2023. Bias as Boundary Object: Unpacking The Politics Of An Austerity Algorithm Using Bias Frameworks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).* Association for Computing Machinery, New York, NY, USA, 1838–1849. https://doi.org/10.1145/3593013.3594120

[60] Gabriel Grill and Christian Sandvig. 2023. Military AI's Next Frontier: Your Work Computer. *Wired* (2023).

[61] Jonathan Grudin. 2009. AI and HCI: Two Fields Divided by a Common Focus. *AI Magazine* 30, 4 (Sept. 2009), 48–48. https://doi.org/10.1609/aimag.v30i4.2271

[62] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist studies* 14, 3 (1988), 575–599.

[63] Sandra Harding. 1992. Rethinking Standpoint Epistemology: What Is" Strong Objectivity?". *The Centennial Review* 36, 3 (1992), 437–470.

[64] Will Douglas Heaven. 2022. Why Meta's Latest Large Language Model Survived Only Three Days Online. *MIT Technology Review* (2022).

[65] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. https://doi.org/10.48550/arXiv.2009.03300 arXiv:2009.03300 [cs]

[66] Kashmir Hill. 2023. OpenAI Worries About What Its Chatbot Will Say About People's Faces. *The New York Times* (July 2023).

[67] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems the CHI Is the Limit - CHI '99.* ACM Press, Pittsburgh, Pennsylvania, United States, 159–166. https://doi.org/10.1145/302979.303030

[68] Linda Huber and Casey Pierce. 2023. Navigating the Empty Shell: The Role of Articulation Work in Platform Structures. *Journal of Computer-Mediated Communication* 28, 4 (June 2023), zmad004. https://doi.org/10.1093/jcmc/zmad004

[69] Hugging Face. 2024. TTS Arena: Benchmarking Text-to-Speech Models in the Wild. https://huggingface.co/blog/arena-tts.

[70] Elle Hunt. 2016. Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter. *The Guardian* (March 2016).

[71] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 375–385. arXiv:1912.05511

[72] Sheila Jasanoff. 2005. Technologies of Humility: Citizen Participation in Governing Science. In *Wozu Experten?* Springer, 370–389.

[73] Florian Jaton. 2017. We Get the Algorithms of Our Ground Truths: Designing Referential Databases in Digital Image Processing. *Social Studies of Science* 47, 6 (Dec. 2017), 811–840. https://doi.org/10.1177/0306312717730428

[74] Amba Kak and Sarah Myers West. 2023. The AI Debate Is Happening in a Cocoon.

[75] Frederike Kaltheuner (Ed.). 2021. *Fake AI.* Meatspace Press.

[76] Alex Kantrowitz. 2022. Finally, an A.I. Chatbot That Reliably Passes "the Nazi Test". *Slate* (Dec. 2022).

[77] Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* 4, 9 (2023).

[78] Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization. https://doi.org/10.48550/arXiv.2010.03093 arXiv:2010.03093 [cs]

[79] Ann-Christina Lange, Marc Lenglet, and Robert Seyfert. 2019. On Studying Algorithms Ethnographically: Making Sense of Objects of Ignorance. *Organization* 26, 4 (July 2019), 598–617. https://doi.org/10.1177/1350508418808230

[80] Bruno Latour and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts.* Princeton University Press, Princeton, N.J.

[81] Laura Bratton. 2024. The Biggest AI Acquisitions by Apple and Other Big Tech Firms. *Quartz* (April 2024).

[82] Q. Vera Liao and Ziang Xiao. 2023. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap. https://doi.org/10.48550/arXiv.2306.03100 arXiv:2306.03100 [cs]

[83] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta-Review of Evaluation Failures Across Machine Learning. (2021), 20.

[84] Silvia Lindtner, Shaowen Bardzell, and Jeffrey Bardzell. 2016. Reconstituting the Utopian Vision of Making: HCI After Technosolutionism. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* Association for Computing Machinery, San Jose, California, USA, 1390–1402. https://doi.org/10.1145/2858036.2858506

[85] Yiwen Lu. 2023. Generative A.I. Can Add $4.4 Trillion in Value to Global Economy, Study Says. *The New York Times* (June 2023).

[86] Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2023. Power Hungry Processing: Watts Driving the Cost of AI Deployment? https://doi.org/10.48550/arXiv.2311.16863 arXiv:2311.16863 [cs]

[87] Dieuwertje Luitse, Tobias Blanke, and Thomas Poell. 2024. AI Competitions as Infrastructures of Power in Medical Imaging. *Information, Communication & Society* (March 2024).

[88] Dieuwertje Luitse and Wiebke Denkena. 2021. The Great Transformer: Examining the Role of Large Language Models in the Political Economy of AI. *Big Data & Society* 8, 2 (July 2021), 20539517211047734. https://doi.org/10.1177/20539517211047734

[89] Arwa Mahdawi. 2024. What Is Going on with ChatGPT? *The Guardian* (Jan. 2024).

[90] Lev Manovich. 2002. *The Language of New Media.* MIT press.

[91] Gary Marcus. 2023. The Sparks of AGI? Or the End of Science? https://cacm.acm.org/blogs/blog-cacm/271354-the-sparks-of-agi-or-the-end-of-science/fulltext.

[92] Melissa Heikkilä. 2023. Google DeepMind's New Gemini Model Looks Amazing—but Could Signal Peak AI Hype. *MIT Technology Review* (2023).

[93] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[94] Mike Michael. 2017. *Actor-Network Theory: Trials, Trails and Translations.* SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road London EC1Y 1SP. https://doi.org/10.4135/9781473983045

[95] Charles W. Mills. 1997. *The Racial Contract.* Cornell University Press, Ithaca.

[96] Evgeny Morozov. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism* (first edition ed.). PublicAffairs, New York.

[97] Peter Nagy and Gina Neff. 2015. Imagined Affordance: Reconstructing a Keyword for Communication Theory. *Social Media + Society* 1, 2 (July 2015), 2056305115603385. https://doi.org/10.1177/2056305115603385

[98] Peter Nagy and Gina Neff. 2023. *Rethinking Affordances for Human–Machine Communication Research.* SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road London EC1Y 1SP, 273–279. https://doi.org/10.4135/9781529782783.n34

[99] Arvind Narayanan. 2023. GPT-4 and Professional Benchmarks: The Wrong Answer to the Wrong Question. https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks.

[100] Khanh Nguyen and Hal Daumé III. 2020. Global Voices: Crossing Borders in Automatic News Summarization. arXiv:1910.00421 [cs]

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

[102] Beatrice Nolan. 2023. Silicon Valley Has a New Version of Its Beloved 'move Fast and Break Things' Mantra. *Business Insider* (2023).

[103] Emily Olson. 2023. Google Shares Drop $100 Billion after Its New AI Chatbot Makes a Mistake. *NPR* (Feb. 2023).

[104] Open AI. 2023. How Should AI Systems Behave, and Who Should Decide? https://openai.com/blog/how-should-ai-systems-behave.

[105] OpenAI. 2023. GPT-4. https://openai.com/research/gpt-4.

[106] Samir Passi and Phoebe Sengers. 2020. Making Data Science Systems Work. *Big Data & Society* 7, 2 (July 2020), 2053951720939605. https://doi.org/10.1177/2053951720939605

[107] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and Its (Dis) Contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns* 2, 11 (2021).

[108] Sebastian Pfotenhauer, Brice Laurent, Kyriaki Papageorgiou, and and Jack Stilgoe. 2022. The Politics of Scaling. *Social Studies of Science* 52, 1 (Feb. 2022), 3–34. https://doi.org/10.1177/03063127211048945

[109] Sundar Pichai and Demis Hassabis. 2023. Introducing Gemini: Our Largest and Most Capable AI Model. https://blog.google/technology/ai/google-gemini-ai/.

[110] David Pierce. 2023. Google's Bard Chatbot Doesn't Love Me — but It's Still Pretty Weird. https://www.theverge.com/2023/3/21/23650472/google-bard-ai-chatbot-hands-on-test.

[111] Trevor Pinch. 1993. "Testing - One, Two, Three ... Testing!": Toward a Sociology of Testing. *Science, Technology, & Human Values* 18, 1 (Jan. 1993), 25–41. https://doi.org/10.1177/016224399301800103

[112] Trevor J. Pinch and Wiebe E. Bijker. 1984. The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. *Social Studies of Science* 14, 3 (1984), 399–441. jstor:285355

[113] Nikolaus Poechhacker and Severin Kacianka. 2021. Algorithmic Accountability in Context. Socio-Technical Perspectives on Structural Causal Models. *Frontiers in Big Data* 3 (2021). https://doi.org/10.3389/fdata.2020.519957

[114] Theodore M. Porter. 1996. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life.* Princeton University Press.

[115] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2021), 23.

[116] Chelsea Ritschel. 2018. Google Fixes Translate Tool after Accusations of Sexism. https://www.independent.co.uk/life-style/women/google-translate-sexist-masculine-feminine-he-said-she-said-english-spanish-languages-a8672586.html.

[117] Kevin Roose. 2023. How ChatGPT Kicked Off an A.I. Arms Race. *The New York Times* (Feb. 2023).

[118] Cengiz Salman. 2024. COOL IT! The Objective Racism of Carceral Technofixes. *Critical Studies in Media Communication* 0, 0 (2024), 1–15. https://doi.org/10.1080/15295036.2024.2314667

[119] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[120] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014), 1–23.

[121] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage? https://doi.org/10.48550/arXiv.2304.15004 arXiv:2304.15004 [cs]

[122] Nick Seaver. 2017. Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems. *Big Data & Society* 4, 2 (2017), 2053951717738104.

[123] Nick Seaver. 2018. What Should an Anthropology of Algorithms Do? *Cultural Anthropology* 33, 3 (Aug. 2018), 375–385. https://doi.org/10.14506/ca33.3.04

[124] Andrew D. Selbst, Sorelle Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2018. Fairness and Abstraction in Sociotechnical Systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*.

[125] Felix M. Simon. 2022. Uneasy Bedfellows: AI in the News, Platform Companies and the Issue of Journalistic Autonomy. *Digital Journalism* 10, 10 (Nov. 2022), 1832–1854. https://doi.org/10.1080/21670811.2022.2063150

[126] Anubha Singh, Patricia Garcia, and Silvia Lindtner. 2023. Old Logics, New Technologies: Producing a Managed Workforce on on-Demand Service Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[127] Katta Spiel, Christopher Frauenberger, Geraldine Fitzpatrick, and Eva Hornecker. 2019. Effects of Participatory Evaluation - A Critical Actor-Network Analysis. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299049

[128] Katta Spiel, Laura Malinverni, Judith Good, and Christopher Frauenberger. 2017. Participatory Evaluation with Autistic Children. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5755–5766.

[129] Susan Leigh Star. 1990. Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions. *The Sociological Review* 38, 1_suppl (1990), 26–56.

[130] Susan Leigh Star and Anselm Strauss. 1999. Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Computer Supported Cooperative Work (CSCW)* 8, 1 (March 1999), 9–30. https://doi.org/10.1023/A:1008651105359

[131] Nikki Stevens, Anna Lauren Hoffmann, and Sarah Florini. 2021. The Unremarked Optimum: Whiteness, Optimization, and Control in the Database Revolution. *Review of Communication* 21, 2 (April 2021), 113–128. https://doi.org/10.1080/15358593.2021.1934521

[132] Lucy Suchman. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions.* Cambridge University Press, Cambridge, England.

[133] Tiziana Terranova. 2004. *Network Culture: Politics For the Information Age.* Pluto Press, London ; Ann Arbor, MI.

[134] José Van Dijck. 2014. Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology. *Surveillance & Society* 12, 2 (2014), 197–208.

[135] Lakshmi Varanasi. 2023. GPT-4 Can Ace the Bar, but It Only Has a Decent Chance of Passing the CFA Exams. Here's a List of Difficult Exams the ChatGPT and GPT-4 Have Passed. https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1.

[136] Janet Vertesi and D. Ribes. 2019. From Affordances to Accomplishments PowerPoint and Excel at NASA. In *digitalSTS: A Field Guide for Science and Technology Studies*. Princeton Univ. Press, 369–392.

[137] James Vincent. 2023. Google's AI Chatbot Bard Makes Factual Error in First Demo. https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo.

[138] Judy Wajcman. 2004. *TechnoFeminism.* Polity, Cambridge ; Malden, MA.

[139] Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A Survey on Cross-Lingual Summarization. *Transactions of the Association for Computational Linguistics* 10 (Nov. 2022), 1304–1323. https://doi.org/10.1162/tacl_a_00520

[140] David Gray Widder, Sarah West, and Meredith Whittaker. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. https://doi.org/10.2139/ssrn.4543807

[141] Kyle Wiggers. 2024. NIST Launches a New Platform to Assess Generative AI.

[142] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (first edition ed.). PublicAffairs, New York.