

# The Benchmark Trap: Structures of Power and Injustice in AI Evaluations

Jason Branford<sup>\*1</sup>, Angelie Kraft<sup>\*2</sup>

<sup>1</sup>University of Hamburg, Hamburg

<sup>2</sup>Weizenbaum Institute, Berlin

jason.branford@uni-hamburg.de, angelie.kraft@weizenbaum-institut.de

## Abstract

Artificial intelligence (AI) benchmarks are not neutral tools of evaluation but socio-technical artefacts that shape competition, power, and research priorities within AI. Benchmarks standardise the assessment of systems and facilitate the creation of leaderboards that reward state-of-the-art performance with prestige, citations, trust, and institutional influence. As the costs of developing competitive AI systems rise, these rewards increasingly concentrate among powerful, industry-funded labs. This paper situates these concerns within Iris Marion Young’s theories of oppression and structural injustice. It argues that current benchmarking practices may perpetuate systematic harms affecting various actors in AI research, aligning with four of Young’s “faces of oppression”. Benchmarking culture is further framed as a source of structural injustice, as these harms emerge from normalised, individually defensible practices and network effects, even without explicit wrongdoing. By reinforcing existing power structures and narrowing possible research trajectories, benchmarking may in fact prevent the field from advancing in epistemically robust and socially beneficial ways.

## 1 Introduction

Benchmarks do not merely measure the performance of artificial intelligence (AI) systems. They help determine what these systems are taken to be and, therefore, the ends to which they are put, transforming what is inherently a limited technical performance into public evidence of, for example, ‘reasoning’ or ‘safety’. We assume that many, if not most, individual developers and researchers in the field are genuinely interested in improving benchmarks for the good of the discipline and society. However, AI is not only a field of scientific inquiry but a lucrative business, a (geo)political lever, and a growing object of regulatory concern, which incentivises the involvement of players with different sets of interests. Thus, benchmarks are inherently socio-technical artefacts that are shaped by and contribute to existing power dynamics in AI research. This is especially true of prominent benchmarks that are visible beyond specified academic sub-communities, are broadly utilised to communicate the capabilities of commercial and open-source models towards

technical and non-technical audiences, and commonly referenced within AI leaderboards.<sup>1</sup>

Our central argument is that current benchmarking culture in AI may feed into structures of oppression and injustice towards smaller labs and marginalised researcher communities, on the one side, and towards marginalised user groups, on the other. In so doing, we further suggest that they may therefore be hindering the potential of the field. We draw on the pioneering work of Iris Marion Young (1990; 2010) to assess different issues in the economy and epistemology of AI benchmarking and conceptualise the kinds of harms that benchmarking perpetuates. After characterising the current culture of benchmarking in AI (Section 2), we detail how stakeholders involved in and affected by the benchmark race may be subject to injustices that align with four of Young’s five “faces of oppression” (Section 3). We argue that the identified issues with AI benchmarking go beyond the ill-intentioned acts of individual ‘bad apples’ in the AI community but are rather systematic and pose structural harms (Section 4). To this end, we utilise Young’s notion of *structural injustice* (2006; 2010) and argue, following McKeeown (2021; 2024), that the types of injustices we identify can be considered *avoidable*. Finally, we consider the implications of these challenges for benchmarking and the future of AI evaluations (Section 5).

## 2 The Influence and Troubles of Benchmarks

According to the frequently cited definition by Raji et al. (2021), a “benchmark [is] a particular combination of a dataset or sets of datasets [...], and a metric, conceptualized as representing one or more specific tasks or sets of abilities” (p. 2). The dataset comes with so-called ‘ground truth’ labels, which distinguish ‘true’ from ‘false’ or ‘desired’ from ‘undesired’ responses. The metric is then used to compute a performance score, which enables the comparison and ranking of systems (e.g., in leaderboards) to determine which fares ‘best’ at a given point in time; noting that a single model is usually evaluated via a variety of benchmarks to assess the extent of its different capabilities. Critical examinations of benchmarks ought to focus on the datasets and metrics used, as well as the ends to which they are put.

Datasets are a typical focal point, and in the case of

<sup>\*</sup>These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>1</sup>See e.g. <https://llm-stats.com/> (access date: July 6, 2026)

AI benchmarks, those used are comparable to AI training datasets in terms of their diversity of content, form, provenance, etc. In order to understand datasets and their societal impacts, it is necessary, as proposed by Gebru et al. (2021), to consider both their creators and their users. Bordes et al. (2025) identify four scenarios of AI evaluation with different stakeholders: *development-focused* evaluation helps researchers and engineers to analyse model performance and errors quickly and informs design decisions and development approaches; *selection-focused* evaluation guides decision-makers in identifying the right model to deploy for their use case; *deployment-focused* evaluation serves operators and regulators as a measure of “production readiness” and the likelihood of “consequential errors”; *research-focused* evaluation give researchers an insight into what a model is and is not capable of in a more general sense.

These scenarios indicate that benchmarks are not used by a single community for a single purpose, and serve to identify researchers, engineers, decision-makers, operators, and regulators as relevant stakeholders who rely on evaluations in different ways and contexts. Another stakeholder group are journalists reporting on the overall state of AI progress who also draw from (and thereby perpetuate) leaderboards and scores on specific benchmarks.<sup>2</sup> Finally, data authors and annotators are further essential stakeholder groups of AI evaluations, given that the creation of datasets involves significant amounts of human labour, e.g., for writing and annotating test cases (Kraft, Simon, and Schimmler 2025). This work is often carried out under precarious working conditions, characterised by, e.g., low and unstable pay, gig work, isolation, physical and psychological distress (cf. Gebrekidan 2024; Kapania et al. 2026). Moreover, many datasets are based on openly accessible web sources, the scraping of which has been criticised as an extractive practice, often associated with breaches of license agreements and intellectual property rights (cf. Longpre et al. 2025). There are, as such, a variety of stakeholders to be considered when discussing the harms of current benchmarking practices. Benchmarking cannot, therefore, be treated as a narrow methodological concern *internal* to machine learning.

## 2.1 Lack of Construct Validity

A widely discussed concern around AI benchmarks has been their lack of *construct validity*. “[C]onstruct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not ‘operationally defined’”, i.e., directly measurable (Cronbach and Meehl 1955, p. 2). AI benchmarks are different from rulers or scales in that they indirectly measure abstract constructs for which no definite metric exists. To study this, Wallach et al. (2025) provide an analytical lens that draws from a social scientific framework and distinguishes between a *background concept*, a *systematised concept*, the actual *measurement instrument*, and its resulting *measurements*. A background concept comprises different conceptualisations of a term, for dif-

ferent purposes and by different stakeholders. For instance, software engineers might want to measure a model’s mathematical abilities to assess its accuracy in a specific application context. Cognitive scientists, on the other hand, might want to measure mathematical abilities to study similarities and differences to human cognition. Wallach et al. (2025) argue that proper validation of a measurement instrument first requires a systematisation of the concept to be measured. In our example, this entails that a clear-cut definition of ‘mathematical ability’, given a specific purpose and context, must be designated as the systematised concept. The measurement instrument itself can then be designed for and validated against this systematised concept in a structured way. This is to ensure that the final measurement is sufficiently indicative of what it is supposed to be indicative of.

For years, scholars have pointed out that computer scientists mostly fail to provide clear concept definitions and evidence-based construct validation (Jacobs and Wallach 2021; Raji et al. 2021; Liu et al. 2024). A recent study delivers empirical support for this critique through a systematic review of 445 large language model (LLM) benchmarks. More than half of those were found to be designed on the basis of contested concept definitions or no definition at all, and a bit less than half of these benchmarks were published without any reported construct validation in the form of a justification rationale or empirical evidence (Bean et al. 2025). The systematic lack of (evidence for) construct validity renders many benchmarks largely useless, misleading, and guilty of overselling claims. This should alarm us, especially considering the significance of benchmarks in our current practices and debates surrounding AI as a subject of research, a business product, a canvas for projecting our most and least desired futuristic scenarios, and so on.

## 2.2 Bias and Misrepresentation

The next issue is, in a sense, a subset of the problem of *what* a measurement instrument, such as a benchmark, *is actually* measuring. In the context of AI benchmarking, the issue of data bias can be understood as a form of miscalibration of the instrument towards a data distribution that is not representative of the distribution the instrument was (presumably) intended to be calibrated to. Kraft, Simon, and Schimmler (2025) analysed 20 question-answering benchmark datasets and found that several were over-representative of questions related to, e.g., male individuals and Western locations. Even dedicated *bias benchmarks* are not immune to bias (Powers et al. 2024; Demchak et al. 2024). Just as any dataset, benchmark datasets are *situated* (Raji et al. 2021) and are likely to reflect the priorities, concerns, realities, conceptualisations, etc. of their creators. In Kraft, Simon, and Schimmler (2025), popular benchmarks were predominantly created by Western, elite institutions. A benchmark that is marketed as measuring the ability to correctly reproduce facts about the world, but mostly contains questions concerning facts relevant to the Western world, is effectively misleading. What is more is that, when applied, such benchmarks effectively reward biased model outputs and feed into algorithmic bias

<sup>2</sup><https://www.forbes.com/sites/johnkoetsier/2023/03/14/gpt-4-beats-90-of-lawyers-trying-to-pass-the-bar/> (access date: April 26, 2026)

in AI systems (Bowman and Dahl 2021).<sup>3</sup>

This worry should not be understood as only epistemic—in the sense that biases are technical distortions affecting individual measurements—as the score is neither merely academic nor containable to the context in which it was produced. Rather, once adopted, such scores exert an influence that hardens in various ways and domains. Accordingly, these methodological shortcomings cannot be assessed (nor addressed) independently of this social context (nor without attending to various harms discussed in Section 3).

### 2.3 Naturalisation of ‘Ground Truths’

Benchmarks are used by different stakeholders to classify ‘good’ from ‘bad’ or ‘safe’ from ‘unsafe’ models, which shapes their choice of which model to develop further or deploy in real-world applications. Through such choices, the community has witnessed many benchmarks become *de facto* standards in their practice and—as with any standard embedded in practice—this has a “material force in the world” (Bowker and Star 1999, p. 39). The European Union (EU) AI Act, for instance, requires the benchmarking of “accuracy, robustness and cybersecurity” in high-risk systems (Regulation (EU) 2024/1689), which determines if a model will be allowed to be deployed on the market. Another example of how benchmark scores become natural ‘ground truths’ about a phenomenon is the widely adopted use of the *RealToxicityPrompts* benchmark, which utilises the (soon-to-be terminated) Perspective API for the automated output scoring of hateful and toxic language.<sup>4</sup> This has become the presumed and accepted standard for toxicity classification and—despite its many conceptual and implementation limitations and inherent biases (Hartmann et al. 2026)—is used in responsible AI assessments as part of larger LLM evaluation suites (Bommasani, Liang, and Lee 2023).

As such, AI benchmarks give rise to competitive dynamics, with the ‘winners’, i.e., the labs that manage to beat the current state-of-the-art (SOTA), receiving praise, recognition, downloads, citations, and public and institutional trust. Hence, there is a strong pull towards benchmarks, motivating the community to design systems that will be able to compete, which, consequently, structure the community’s innovation efforts (Orr and Kang 2024). In short, benchmarks determine the type of systems that are built and which truths they are calibrated towards. Once benchmarks acquire this status, their flaws do not merely mislead specialists but also enter the broader public and commercial imaginary of AI. Weak measurements, biased representations, and *de facto* standards thus serve to fuel myths about what AI systems are and what they are becoming.

<sup>3</sup>Uzunoglu, Li, and Khashabi (2025) experimentally show that imbalanced representation of ‘subdomains’ within a benchmark dataset, paired with a metric that is based on averaging, yields measurements that obscure potential weaknesses on lesser represented domains. If we conceive of, for instance, different demography-related data points as subdomains, this phenomenon can be expected to generalise also to the type of dataset bias discussed here.

<sup>4</sup><https://perspectiveapi.com/#/home> (access date: May 6, 2026)

### 2.4 Fuelling the Myth

The global ‘AI race’ motivates well-funded labs to outdo their competitors, and the pace, scale, and diversity of improvements have led to a state where established benchmarks become too easily solved or simply outdated. In response, benchmarks are created to pose new challenges to the field or to prove ‘human-like’ capabilities. The authors of *Humanity’s Last Exam* (HLE) (Center for AI Safety, Scale AI, and HLE Contributors Consortium 2026), for instance, claim that “[h]igh accuracy [...] would demonstrate expert-level performance on closed-ended, verifiable questions and cutting-edge scientific knowledge”.

Above, we argued that many established benchmarks are highly misleading as they do not measure what they claim to measure, while systematically concealing concerns related to social bias. This calls for a re-assessment of several beliefs about AI that have been circulating in different areas of society and actively promoted by Big Tech. We argue that claims about human-like or close to human-like ‘reasoning’ or ‘knowledgeability’ levels should be met with scepticism, because those profiting most from these anthropomorphising framings are those who claim to be most familiar with the state of research. Hence, we deem it plausible and necessary to assume that there is another reason to continue overselling benchmark results and that this may be of economic and political nature, as opposed to the mere pursuit of knowledge about AI technology’s real capabilities. As it stands, many highly visible benchmarks achieve nothing more than: (1) mislead the public into thinking that AI systems become more and more ‘human-like’, and (2) facilitate gatekeeping that excludes users, policymakers, journalists, and other relevant stakeholders by making it appear as though AI systems are beyond comprehension. As such, AI is also made to appear as though those stakeholders cannot truly contribute to shaping it. When scores enable not only the ability to define what ‘counts’ but, in the process, confer reputational and commercial advantage, an almost cliché temptation to game the established system emerges.

### 2.5 Gaming of the System

As guides for selection and deployment, AI benchmarks become an important lever for marketing. With regards to Meta’s Llama 4 model family which was released in 2025,<sup>5</sup> the company’s former Chief AI scientist, Yann LeCun, in fact, stated in an interview “that the ‘results were fudged a little bit,’ and the team used different models for different benchmarks to give better results.”<sup>6</sup> When OpenAI announced the release of the o3-model, at the end of 2024, it advertised its capabilities as groundbreaking, among other things, as measured by a particularly difficult mathematics benchmark called *FrontierMath*.<sup>7</sup> This benchmark had

<sup>5</sup><https://ai.meta.com/blog/llama-4-multimodal-intelligence/> (access date: May 6, 2026)

<sup>6</sup><https://arstechnica.com/ai/2026/01/computer-scientist-yann-lecun-intelligence-really-is-about-learning/>, (access date: May 6, 2026)

<sup>7</sup><https://techcrunch.com/2024/12/20/openai-announces-new-o3-model/> (access date: August 15, 2026)

been created by the company EpochAI who contracted expert mathematicians to author difficult mathematics problems with the goal of testing a new generation of ‘reasoning models’. OpenAI had not only funded the creation of FrontierMath, but both companies had also agreed to remain secretive of their partnership until the release of the o3-model.<sup>8</sup> Moreover, they had only verbally agreed that OpenAI would not train their newest model on the benchmark.

It is crucial that a model is not directly trained on any of the examples that occur in a benchmark that it is later evaluated by. This phenomenon—called *data contamination*—must be avoided to ensure that the evaluation score is a true measure of model performance and not only a result of memorisation (cf. Schaeffer et al. 2025; Sainz et al. 2023). The ‘gaming’ of the AI benchmarking system can also happen in more implicit ways. It is, for instance, assumed that to reach high scores on the coding benchmark *SWE-Bench* (Jimenez et al. 2024), developers “craft [...] approaches that are too neatly tailored to the specifics of the benchmark”,<sup>9</sup> meaning that the training examples are selected to closely resemble those presented during testing, which yields effects similar to memorisation.<sup>10</sup>

The deal between OpenAI and EpochAI undermines the necessary conditions for meaningful and trustworthy benchmarking practices. Indeed, such behaviour invalidates the very idea of AI evaluation. While it is hard, if not impossible, to prove widespread intentional misconduct, cases like the FrontierMath scandal, as well as the admitted manipulation of benchmark results to exaggerate Llama 4’s capabilities, are clear warning signals: Benchmarks have become a tool of power which resourceful institutions are not hesitating to use towards their own ends.

Together, these five interrelated concerns highlight why the trouble with benchmarks cannot be reduced to a list of technical defects. Benchmarks transform uncertain and situated measurement practices into public claims about quality, safety, and progress. Their flaws matter because benchmark results circulate far beyond the conditions under which they are produced. Growing recognition of the these limitations does not appear to have diminished the importance of benchmarks and, in particular, the intensity with which they are used to promote LLMs. Benchmarks continue to be used because they are an important structural element of the AI development culture and, so far, there is a lack of a better alternative. This suggests that their authority cannot be explained by measurement quality alone, and that the problems—highlighted in Sections 2.3, 2.4 and 2.5—extend beyond technical fixes and pertain rather to the choices and actions of those seeking to market or steer community efforts towards their own goals. Closer scrutiny of the nature

of these operations and development practices—of the kinds of relations and power structures they give rise to, who is being pushed out as a result and at what cost—is, therefore, warranted in order to appreciate various harms, we argue, that follow from the just described dynamics. The harms outlined below are, therefore, not offered as a second, independent, set of objections, but rather specify what is normatively at stake as complicating factors for tackling the issues above as well as demanding redress in their own right (as *injustices*).

### 3 Benchmarking and the Faces of Oppression

Due to how cases like FrontierMath or the beautified evaluation of Llama 4 leverage an unfair advantage over other competitors, most would likely agree that some wrong has been committed. Yet, a claim that this also amounts to an ‘injustice’ or is ‘oppressive’ would likely be contentious. Nevertheless, we argue that features of existing benchmarking practices are not merely regrettable and warrant remedial efforts but also perpetuate forms of oppression.

For Young (1990), injustices denote a specific kind of pervasive, systemic, and instituted wrong that occurs when social processes routinely privilege some groups while disadvantaging others. She explains that injustice primarily refers to two forms of disabling constraints: *oppression* and *domination*. Oppression concerns the myriad ways in which socially instituted mechanisms undermine or limit self-development and expression, while domination is reserved by Young to capture instances of political exclusion that bar groups from decision-making procedures or formal institutional control mechanisms, thereby undermining self-determination. Accordingly, social justice is about promoting and securing both self-development and self-determination (see Naudts 2024, p. 702).

While we believe that an argument could be made that benchmarking practices can be dominating in Young’s sense, defending this would require detailing and unpacking the many still unfolding instances of political influence and lobbying perpetrated by the AI industry, and propped up by benchmarks, which space forbids doing meaningfully here. As such, in this paper, we will focus only on the notion of oppression. Famously, Young (1990) identifies “five faces of oppression”—*exploitation, marginalisation, powerlessness, cultural imperialism, and violence*—the first four of which, we argue, arise in the case of AI benchmarking.<sup>11</sup> Importantly, the faces are not necessarily exclusive, as a single benchmarking practice (or facet thereof) may instantiate several, which nevertheless capture distinctive kinds of wrongs that ought to inform any attempted reforms. Ex-

<sup>8</sup><https://techcrunch.com/2025/01/19/ai-benchmarking-organization-criticized-for-waiting-to-disclose-funding-from-openai/> (access date: May 6, 2026)

<sup>9</sup><https://www.technologyreview.com/2025/05/08/1116192/how-to-build-a-better-ai-benchmark/> (access date: May 6, 2026)

<sup>10</sup>In the meantime, OpenAI has published its own analyses illustrating data contamination of flagship LLMs regarding this benchmark; <https://openai.com/index/why-we-no-longer-evaluate-swe-bench-verified/> (access date: May 6, 2026)

<sup>11</sup>It is worth briefly noting that benchmarks do also, at least indirectly, serve a function in perpetuating salient forms of violence. Naudts (2024) notes that “[o]nce data-driven systems have normalized a particular set of norms, values and beliefs in a specific setting, deviancy can be identified and made subject to interference” (p. 6) and AI benchmarks act as a shared blueprint and certifier for detectors of conformism and deviancy. Benchmarks thus indirectly shape and grant legitimacy to violence-promoting systems.

exploitation concerns the transfer and capture of labour and benefits; marginalisation, exclusion from recognised participation; powerlessness, participation without authority over its terms; and cultural imperialism, the universalisation of one evaluative standpoint.

We are not, of course, the first to apply Young's notions in the context of AI and algorithmic decision-making. Yet, the existing literature has primarily focused on two aspects, namely, data- and model-specific practices and features (e.g., giving rise to concern around bias, transparency, opacity, explainability, consent, etc.), on the one hand, and the usages those technologies are put to (e.g., negatively impacting individual lives and livelihoods), on the other (cf. Browne 2023; Herzog 2021; Herzog and Branford 2025; Naudts 2024). Our focus on AI benchmarking, however, highlights underexplored elements of the broader AI ecosystem. AI development proceeds iteratively through conceptualisation, data acquisition and preparation, model training and evaluation, and deployment, usage, and monitoring, with feedback loops from model evaluation informing which data are added to the model training and how the model design and training are further adjusted (Bordes et al. (2025) call this development-focused evaluation). As the SWE-Bench case demonstrates, to outperform earlier models, researchers and engineers repeatedly adjust systems through result-guided experimentation. Evaluation results directly influence which models are put to real-world use and serve the AI research community as indicators of progress, constituting an enclosing steering mechanism that, we will argue, is normatively consequential. In particular, we consider three instances of related harms: (1) ethical harms perpetuated against those involved in AI development due to inequitable competition; (2) ethical and epistemic harms stemming from AI systems with inferior performance, robustness and with problematic biases, all of which are concealed by invalid or miscalibrated benchmarks; and (3) epistemic harms that pertain to the scientific and technological field as a whole due to flawed and misleading indicators of progress.

Note that, in presenting the below harms as widespread structural concerns, we do not claim that all researchers and developers who create and apply benchmarks are blameworthy complicit. These issues result from collective activities and standards formed over time through many small, local, and often well-intentioned decisions (see Section 4). Still, we distinguish between 'powerful players'—well-funded, highly visible, and influential industry labs—and less influential researcher and developer groups. The former are not accused of having imposed this system on everyone, but of leveraging and shaping it. They have a greater capacity to ameliorate the issues identified and, therefore, bear a greater share of the responsibility for doing so. Still, collective action will certainly be needed to motivate this, and so responsibility is also dispersed across the field.

### 3.1 Exploitation

Talk of exploitation in AI is likely to evoke images of precarious, harmful, low-paid data labour, asymmetrical and abusive power relations, and large-scale appropriation of data

and intellectual property (cf. Hao 2025; Crawford 2021; Muldoon, Graham, and Cant 2024). While these deplorable practices are also evident in AI evaluation datasets, affecting the stakeholder group of data authors and annotators (including those whose data were extracted without their knowledge and consent), they do not exhaust the kinds of exploitation evident in this context. According to Young, exploitation refers to structural relations and "social processes that bring about a transfer of energies from one group to another to produce unequal distributions, and in the way in which social institutions enable a few to accumulate while they constrain many more" (1990, p. 53). Exploitation is thus a patterned conversion of the burdens, risks, and productive energies borne by (or extracted from) some actors into benefits that accrue elsewhere.

AI benchmarks are coordination devices that organise labour, attention, prestige, and investment among involved stakeholder groups, i.e., researchers and engineers (Raji et al. 2021; Orr and Kang 2024). They create an environment in which it is rational, and often professionally necessary, for researchers to direct their labour toward improving a narrow set of public scores. That labour may not, in itself, be objectionable. What is objectionable is that the terms on which it is solicited, conducted, and rewarded are routinely set by actors positioned to convert benchmark outcomes into proprietary advantage, reputational capital, and agenda-setting authority that primarily serves their own narrow interests, all the while giving the impression that such efforts are necessary for advancing the field as a whole (cf. Hao 2025).

Those who define a benchmark's make-up—especially when they can also significantly influence community uptake—can, thereby, exert control over the purposes to which communal labour is put (i.e., the research trajectory), the criteria under which that labour is valued (i.e., metrics of success), and the conversion of resulting gains into institutional and economic advantages (cf. Ott et al. 2022; Thomas and Uminsky 2022). The extent to which benchmarks come to gain scientific authority often obscures this. Put differently, this exploitative power is often exercised through ostensibly technical design decisions that present as merely methodological, but are actually highly normative.<sup>12</sup>

As a result, such decisions structure the field's incentives in precisely the way Young's analysis anticipates: they render certain forms of effort intelligible as 'progress' and others as peripheral, thereby channelling labour toward aims that disproportionately serve already-powerful institutions. One might object that this overstates the agency of benchmark creators. After all, benchmarks are in principle open. However, while less resourced academic labs may produce incremental gains that become part of the field's shared 'improvement curve', the capacity to consolidate these gains through scale, proprietary deployment, marketing, or integration into widely used services largely remains con-

<sup>12</sup>Uberti-Bona Marin et al. (2026) emphasise that several *techno-normative choices* are made when defining and implementing evaluation measures, and the mere choice of metric (e.g., accuracy versus precision or recall) can cause rare cases of a disease to be ignored or social media statements to be incorrectly flagged as problematic.

centrated in well-capitalised institutions. Exploitation, on this reading, consists in the field's collective energies being channelled through evaluative rules that many must comply with but few can meaningfully contest, thereby entrenching dependence on standards others control, and limiting contestation over what and who AI development is for.

### 3.2 Marginalisation

While exploitation concerns how energies are transferred and benefits captured, marginalisation concerns who is kept out of the practices and discussions in which purposes are set, criteria are stabilised, and 'progress' is publicly ratified.

Several popular LLM benchmarks are demographically and geographically biased (biases towards Western, Christian, and male subjects have been identified) and their reporting is opaque, e.g., regarding the identities of annotators (Kraft, Simon, and Schimmler 2025). All the while, the most influential benchmarks are created by a few elite institutions (see also Koch et al. 2021). Benchmarks calibrated to these select perspectives and interests ultimately incentivise the development of models that are biased accordingly. This is because LLM training is always a cost-benefit calculation and data are expensive to obtain, annotate, filter, and process. From such a vantage (objectionable as it is), collecting data representing knowledge from Eastern or Southern regions is disincentivised, given that it would not enhance performance as measured by a benchmark that is calibrated to Western knowledge. Community endorsement of biased benchmarks, thus, solidifies the use and reuse of biased training corpora and, finally, the development of biased models. In these cases, benchmarks are complicit in widely discussed forms of marginalisation, e.g., when AI systems discriminate against particular groups who are, as a result, systematically denied economic opportunities (cf. Bommasani et al. 2022; Jain et al. 2024; Naudts 2024) or harmfully misrepresented (Blodgett et al. 2020). Biased benchmarks certify the 'accuracy' of systems offering discriminatory output and form part of the impetus to adopt such systems that go on to cause harms of this kind.

Young explains that "[m]arginals are people the system of labour cannot or will not use" (1990, p. 53), which accurately describes the situation of researchers and engineers who actively seek to participate in the community efforts to innovate and understand AI but are either entirely precluded or systematically pushed to the periphery. In contemporary AI research contexts, this marginalisation manifests not only as material deprivation but through the withholding of certain kinds of standing, e.g., by having one's work overlooked. Consequently, a "whole category of people is expelled from useful participation" (Ibid), who are "rendered invisible, voiceless, unrecognised and isolated" (Naudts 2024, p. 705). Rather than direct their labour more freely toward ends they deem morally valuable or scientifically promising, these researchers are (at least indirectly) compelled either to chase established benchmarks or to build 'competitors' that nevertheless respond to those benchmarks and so cannot fully shed their terms.

Also marginalised by benchmark-centric research are those who lack the resources to engage in the benchmark

race at all. As cutting-edge benchmarks increasingly presuppose access to large-scale compute, proprietary datasets, and extensive engineering labour (Strubell, Ganesh, and McCallum 2019; Bender et al. 2021), benchmark leaderboards often track access to resources as much as (or more than) methodological ingenuity, scientific value, or social utility (cf. Orr and Kang 2024; Ott et al. 2022). The consequence is that entire research communities (e.g., those based in less-resourced institutions, regions, or disciplines) are effectively marginalised. It may be contested that this is unavoidable, as some kinds of research are simply costly and the price of advancing the field. We reject this as a self-fulfilling prophecy resulting from the current benchmark system, which perpetuates a narrow conception of what constitutes advancement.

There are AI researchers innovating the field whose work is not captured by sitting atop prominent leaderboards. For instance, most popular LLM benchmarks measure English-specific performance only (cf. Kraft, Simon, and Schimmler 2025). While the natural language processing (NLP) community has seen an increase in efforts to build systems attuned to 'low-resource' languages, yet visibility and praise remain rather limited. Such dedicated benchmarks do not play a significant role in the public and more established benchmarks that do cover 'lower-resource' languages are usually those designed to measure multilingual capabilities as an aggregate concept and are based on translations or adaptations from originally English benchmarks, which come with their own quality issues (cf. Umutlu et al. 2025). If model performance on these less represented languages was considered central to AI innovation, respective benchmarks would receive more attention, directing collective efforts accordingly. Not only would this benefit these language communities but also those who have been researching and developing in these areas all along. Our point is not primarily about such work getting its due in terms of visibility, but rather that the benchmark regime structures what counts as intelligible achievement in the first place. Certain kinds of work (e.g., locally grounded evaluation, qualitative assessments of harms, low-resource language work, or alternative task framings) are institutionally disincentivised and rendered epistemically secondary, signalling that such work is less important for the field. What is needed are expanded possibilities for marginalised researchers and labs to do the work they believe will enhance the field without being relegated as a result.

It is tempting to treat marginalisation here as primarily representational. Diversification in the current benchmarking arena is largely lip-service as self-determination (i.e., the ability to set the values and directions of the field) remains severely limited (cf. Birhane et al. 2022). Critical work on data colonialism and decolonial approaches to AI has repeatedly illustrated how 'inclusion' can function as a continuation of colonial patterns of appropriation liable to being repurposed for institutional or commercial ends (Couldry and Mejias 2019; Mohamed, Png, and Isaac 2020; Ricaurte 2019, 2022). The worry is that, without restructuring existing benchmarking practices and their influence on the field, inclusion of marginalised communities may be pursued primarily to improve dominant metrics, expand market reach,

or strengthen claims of generality. Accordingly, ‘difference’ is incorporated only insofar as it can be translated into the prevailing evaluative norms, while the communities’ own standards of success (e.g., what counts as a good model, harmful output, or legitimate use case) remain excluded.

### 3.3 Powerlessness

If marginalisation concerns exclusion from meaningful participation, powerlessness—as just hinted at—concerns participation without authority. It is about whether an individual’s judgement counts as judgement that will be heeded to improve the conditions of their life, whether they can speak in a given setting without being marked as out of place, and whether their participation includes the ability to contest and reshape the terms under which they are evaluated. As such, an individual can be quite involved in the practice—and, as in the case of AI researchers, be recognised as skilled and competent—yet still occupy a subordinate role in entrenched practices that limit their action and inhibit meaningful authority over how and to what ends they labour.

A fine-grained appreciation of the powerless entailed in AI benchmarking requires disentangling some elements of Young’s account, since her analysis is fundamentally concerned with ‘class’, tracking a division between ‘professionals’—who exercise authority and discretion, and who are granted recognition—and ‘non-professionals’—who are supervised, expected to follow orders, and routinely disrespected. Further, Young extends such workplace practices into society more broadly where such standing or lack thereof continues, through feedback loops, to exert influence and shape social expectations and interactions. Accordingly, the powerless are diminished across the board, severely inhibited from developing skills, exercising creativity or judgment, expressing themselves in a manner that is heard, or sharing their experiences (and the positional knowledge entailed therein); in short, means to improve their situation are effectively immobilised.

The positions of data workers and other subordinated labourers in the AI supply chain closely resemble the non-professional status Young has in mind (cf. Naudts 2024). One might argue that data annotators do own a certain level of power in the sense that they can shape which data are considered ‘right’ or ‘wrong’, ‘desired’ or ‘undesired’. However, due to their material dependency paired with strict annotation criteria and tightly regulated and supervised work environments, their annotations are actually more likely to reflect those of the clients (i.e., companies like OpenAI) (Miceli and Posada 2022).

Focusing on benchmarking as a research-governance practice, two further constituencies matter that complicate Young’s professional/non-professional distinction. First, there are *users* whose powerlessness flows from the extent to which they are drawn into a relation of epistemic deference concerning AI models. Benchmarks are the dominant paradigm to diagnose accuracy, robustness, and security (as reflected in the EU AI Act, Article 15(1)), affecting operator and regulator decisions as well as informing the wider public of the current state of AI (e.g., mediated by journalism). As discussed in Section 2.4, evaluation scores can thus

fuel certain (mythical) conceptions of AI irrespective of numerous concerns related to bias or invalidity. This evaluation infrastructure—reified by legislation—renders the general public powerless. That is, the in-transparency of the practice deprives the public of its means to form a comprehensive position of its own. Benchmark scores and leaderboard rankings function as ready-made ‘evidence’ for what is ‘best’, which invites users to treat the relevant metric as reliable and the use of leading models as responsible. Stretching Young’s language of ‘class’—even though users may be ‘professionals’ who deploy existing models in their own downstream applications—the division becomes between those who are able to set the conditions of evaluation and develop systems able to best meet them, on the one hand, and those who must take this all on face value. This can be conceptualised as a kind of epistemic powerlessness, defined by being governed, in one’s practical decisions, by evaluative machinery one cannot meaningfully contest, audit, or reinterpret.

Second, there are the *researchers* actively designing their own benchmarks, developing models, writing papers, etc. These actors are highly skilled, articulate, and credentialed. They look like ‘professionals’ in Young’s sense, and so—by definition—should be excluded from the powerless. Yet, even though anyone can contribute their own benchmark, not all benchmarks are destined to be considered a widely shared standard. As with any research contribution, relevance to currently shared areas of interest, improvements over existing approaches, and visibility within the community are preconditions. Once performance on a specific benchmark becomes the currency of success, many researchers are pressed into a role that looks less like collective authorship and more like a form of deference and adherence to tasks confined to mere optimisation, comparison, and reporting within the space delineated by the given benchmark paradigm. Orr and Kang’s diagnosis of benchmarking as a kind of ‘sport’ is helpful, as sport is a competitive practice governed by rule-makers, officiating procedures, and legitimacy-conferring institutions that decide what counts as a win (Orr and Kang 2024). This is not to say that professional athletes are powerless in an absolute sense, and neither are AI researchers and engineers. But current benchmarking practices, and in particular their capture through well-funded institutions with political and economic agendas, set significant constraints on the assertion of power. Researchers can act, but only within terms they did not set and which may draw them into activities they may object to. That is because refusing the benchmark race and insisting on alternative criteria of success is not only likely to carry significant professional costs, but their activities may indeed be in some sense unintelligible within the dominant economy of recognition perpetuated by benchmarking and, therefore, likely to be dismissed. This already suggests that the trouble resides in the broader system.

The gravitational force of benchmarks invites myopia, gaming, and the displacement of other substantive goals (Thomas and Uminsky 2022). Even researchers who are personally committed to aims like social utility or responsiveness to harms are likely to find themselves having to translate those commitments into benchmark legibility, or

to treat them as secondary rather than as success conditions in their own right. The result is a domain in which many ‘professionals’ are still, in a decisive sense, highly limited authors of the norms that govern their professional lives.

### 3.4 Cultural Imperialism

For Young, cultural imperialism refers to the social processes under which a dominant group can universalise and establish their experiences and culture as the norm. Conversely, the particular perspectives and lived experiences of less privileged groups become obscured, stereotyped or marked out as ‘other’. The harm is that dominant culture secures the status of common sense, appearing ‘natural’ or authoritative, thereby subsuming its way of seeing (or sense-making) as standing in for reality as such, while alternatives appear partial, parochial, or deficient in some way.

Naudts (2024) transposes this point into the digital ecosystem by arguing that data-driven systems increasingly function as modern “means of interpretation and communication” in that they order, classify, and represent the world in ways that can naturalise particular values as technical facts. AI benchmarking practices exemplify this dynamic. As evaluation encircles the whole AI development and deployment pipeline, it directly shapes future model optimisations. Benchmarks are, as such, both complicit in, and even a source of, those discriminatory system behaviours that “other” (groups of) individuals detailed by Naudts (2024). They also certify automation systems that are then utilised at scale, “ordering of the world” according to dominant views. The worry, thus, is that the current modes of benchmarking have become a dominant—as well as *dominating*—means of interpretation, whereby a particular and exclusionary evaluative culture is normalised and wrongly perceived as neutral, functioning as a shared standard of ‘right’ and ‘wrong’, ‘desirable’ and ‘undesirable’ model behaviour at scale (see Section 2.3). Indeed, their authority depends in part on obscuring that particularity.

As Hartmann et al. (2026) highlight through the example of Perspective API, whole communities of scientific practice came to base large bodies of research on one specific, problematic operationalisation of toxicity. Benchmarks serve an illusion of objectivity, formalising the world into binaries, when—upon closer look—each binary is a reductive formalisation that involves actively choosing which error margins to accept (Uberti-Bona Marin et al. 2026). Raji et al. (2021) illustrate that performance on particular benchmark suites is repeatedly made to stand in for broader claims about flexible or general AI capacities, universalising what is inherently a local, value-laden evaluation and transforming a limited utility into an authority. Benchmarks thereby set the interpretive regime through which activity in the field gains meaning.

This both accentuates the earlier noted difficulties of ‘powerlessness’ while complicating any would-be remedial efforts concerning ‘exclusion’ and ‘marginalisation’ by restricting the conditions of participation in a way that obfuscates that this is in fact a restriction (albeit one that has been internalised by those whose interests this arrangement serves). As a result, benchmarks operate as gatekeeping mechanisms that constrain the scope of self-development

and self-determination, compelling actors in the field to orient their own practices accordingly. Due to this universalising capacity, even attempts to break with establishment (e.g., community-driven benchmarks for low-resource languages, local safety concerns, or domain-specific harms) are more likely to have their value recognised when they assimilate to and render themselves intelligible in the benchmark system. In other words, to show that they outperform on a recognised measure. Important research that cannot ‘travel’ in this way (e.g., participatory assessments of model impacts) may struggle to gain traction or be co-opted to legitimate established benchmarks (Birhane et al. 2022; Sloane et al. 2022).

Benchmarks privilege certain research interests and trajectories while marginalising others. However, the problem goes beyond unfair market capture; it negatively impacts the science as a whole by pre-emptively narrowing what can appear as knowledge, critique, or innovation. This, therefore, adds to those issues related to bias (Section 2.2) and lacking construct validity (Section 2.1) which already underscore how epistemically fraught benchmarks are and significantly challenge their wide acceptance as meaningful indicators of scientific progress. The corporate capture of AI evaluation under the growing global politico-economic appetite for AI breakthroughs amplifies this issue, eroding an epistemic practice that decreasingly provides meaningful epistemic insight and increasingly serves the rationalisation of dominant norms and profitable myths. The harm done here does not only affect researchers and engineers, but also operators and decision-makers, journalists, and consumers.

## 4 Structural Injustice and the Science of AI

The four faces of oppression diagnosed in AI benchmarking are different manifestations of the same underlying structures of power we will now investigate through the notion of *structural injustice* proposed by Young (2010) and further developed by McKeown (2021). This structural view reveals how these forms of oppression yield a further instance of harm affecting the epistemic fabric of AI research itself.

### 4.1 An Avoidable Structural Injustice

The instances of oppression laid out in Section 3 are structural in the sense that they arise from a culture of evaluation established over time (Orr and Kang 2024) that is carried and upheld by collective practices. Put differently, they are not merely harms that occur *within* benchmarking but are made possible by the position benchmarking now occupies in the AI ecosystem as a relatively stable arrangement of norms, incentives, resources, institutions, expectations, and interpretive schemes that direct action. We should not, therefore, focus on these harms *solely* as following from the actions of discrete individuals—even though it is typically individuals who often must shoulder their direct costs (cf. Kasirzadeh 2022)—and so seek out or ‘trace’ (Browne 2023) the ‘wrongdoer’, e.g., the team lead who excluded a community driven project, or the business who failed to recognise their lack of diversity, or the company who cheats to get ahead on a benchmark. While in many cases it will be possible to identify some kind of wrongdoing and assign

a portion of responsibility for the injustice in question, this does not *fully* capture what is going on in benchmarking. Rather, as in the case of what McKeown (2021) calls *pure* structural injustice, these also follow from people systematically acting in morally acceptable ways with no ill intent to others or intention to exploit a situation for their gain, and so stem from largely well-intentioned, normalised, and seemingly unobjectionable actions of individuals exhibiting network effects (Young 2010). Concretely, it can be assumed that most researchers and developers are genuinely interested in developing the best possible measures for quality, safety, and progress. This is reflected in increasing collective efforts to improve AI evaluations (see Section 5).

However, McKeown (2021) nuances Young's notion of structural injustice, differentiating two further kinds that better track the aspect of 'power' (pp. 4-5). *Avoidable* structural injustice is where there are powerful agents with the capacity to change unjust structures, but they fail to do so. For example, rich states have the capacity to end homelessness, but they do not. So where Young understood homelessness as a structural injustice simpliciter, McKeown argues that it is avoidable. *Deliberate* structural injustice is where agents are deliberately perpetuating unjust background conditions for their own gain, and they have the power to change them. For example, in the sweatshops case, multi-national corporations deliberately lobby governments under the threat of capital flight, in order to maintain the poorest wages and working conditions for garment workers. McKeown concludes, therefore, that when powerful agents deliberately perpetuate structural injustice or have the capacity to change unjust structures, they bear moral responsibility to do so.

It is difficult to categorise AI benchmarking as a whole, since the landscape of benchmarks, what they are built and used for, and how and whether they become established as a collective standard, is quite diverse. Some benchmarks emerge from shared tasks or are published as test sets in AI publications, others are directly created or commissioned by technology companies. Nevertheless, benchmarking has the hallmarks of *pure* structural injustice given that no single actor creates the benchmark race and the structure is reproduced through the ordinary, often well-intentioned activities of an array of stakeholders adhering to established norms. Yet, the relevant harms are no longer unforeseeable, nor are all participants equally positioned. The benchmarks that come out as the most popular are increasingly those created by well-funded research and industry labs (Kraft, Simon, and Schimmler 2025; Baack, Buschek, and Bohacek 2026). Technology companies, in particular, continue to reap the benefits instead of triggering change in favour of fair and inclusive evaluation. In this sense, we argue, the injustices observed in AI benchmarking fall into the category of *avoidable* structural injustice. The running system was not put into place to deliberately cause harm and, in fact, it is plausible to assume that most stakeholders both intend to avoid harm and seek to advance the field in the best way they know how. Nonetheless, in its current state, influential players hold unequal power in and over this system, and do have a real capacity to alter the current setup, which means that they can choose to either continue to leverage it to their

own benefit or change it in favour of others (or mitigate the concerns raised above).

## 4.2 How Benchmarking Harms AI Research

Acknowledging this structural dynamic permits us to step back from specific instances of injustice and assess how the system they constitute may weaken AI research as a whole. Writing in the context of AI decision-making in medicine, Herzog and Branford (2025) flag a concern directly relevant to ours. They note that AI use and reliance may be "indulging in an inadequate epistemological system that oppresses knowledge production and possession of a particular kind without us having the capability to recognize this" (p.15). By diminishing physician and patient involvement, a particular conception of what counts as 'good' medical practice is being advanced and barriers erected that inhibit the recognition of shortcomings and the means for assessing the value of alternatives. A similar fear underscores the epistemic concerns we have vis-à-vis benchmarking. Specifically, that the current model not only excludes alternative conceptions of what counts as 'good' AI development or 'progress', but that it is itself structured to reward a distinctive perspective concerning this. It is, therefore, necessary to reflect on whether the prominence of benchmarking has produced a similarly "inadequate epistemological system".

Echoing concerns raised by Dotson (2012), we argue that the systems of evaluation in AI are calibrated to exclude certain groups from participating in this significant epistemic process. The interests of less dominant researcher, developer, and consumer groups are inadequately represented in the established frameworks, which has two forms of ethically and epistemically harmful consequences: First, it leads to insights about AI progress and appropriateness that are less relevant to their preferences and experiences. Due to the causality between evaluation, development, and deployment, this yields systems unfit to their needs and wants, and which perpetuate bias and misrepresentation. Second, current frameworks prevent objectivity. For many years, philosophers of science have argued that objectivity can only be approached via temporal and local consensus that is best justified through ongoing engagement with diverse, conflicting, situated, partial, and changing understandings of the matters in view (Longino 2002; Haraway 2016).

To fix the science of AI, the community needs to pursue what Dotson (2014) would categorise as *third-order change*: AI research is trapped in a closed cycle, in which those who determine the mode of evaluation are those who build the systems which are to be evaluated, and those who benefit from their uptake (Kraft 2025). To overcome epistemic oppression upheld by current practices, we must start with the epistemic tools we utilise to judge and tune them. Escaping the evaluation crisis is not only a matter of building more reliable benchmarks. It is also a matter of questioning the concept altogether and the currently accepted norms related to the *how*, *who*, and *for whom* it is conducted. It must, therefore, involve transforming the conditions under which evaluation becomes authoritative, the purposes it serves, the agents empowered to define it, and the institutions that convert their results into material and symbolic power.

## 5 Implications for the Future of AI Research

We have argued that AI benchmarking is not merely a technical practice for measuring progress, but a socially powerful evaluative infrastructure that shapes labour, recognition, authority, and the direction of AI research that helps perpetuate instances of four out of Young's five faces of oppression—exploitation, marginalisation, powerlessness, and cultural imperialism—within the AI ecosystem. Moreover, we have argued that these arise as not only a form of *structural injustice* but an *avoidable* form which is currently epistemically and ethically undermining AI research. But how do we fix an “inadequate epistemological system”? How can we possibly transcend our established frameworks and alter our known ways of doing AI research from within?

Efforts to establish an “evaluation science for AI” have been surging, calling for deeper engagement with the statistical foundations of testing, more mathematically rigorous measurement instrument creation, or more transparent evaluation documentation and sharing of results (cf. Truong and Koyejo 2026, or EvalEvalCoalition<sup>13</sup>). We clearly welcome this collective push towards more valid, reliable, and transparent benchmarking. However, we also warn against circling around individual symptoms (especially in isolation) and technosolutionism. We argue that we must not focus too narrowly on an ‘inside-only’ view on issues and, instead, encourage to zoom out, to conduct power-critical inquiry, and understand the structural entanglements of evaluation. Indeed, it is precisely the need to do so that is highlighted by the accumulative instances of oppression detailed in Section 3 and their relation to the issues noted in Section 2. Again, revising the epistemic system in place calls for *third-order change*, which “involves developing the capacity to recognize and alter elements of operative, instituted social imaginaries that inform and preserve organizational schemata” (Dotson 2014, p. 119). However, because benchmarks constitute the very epistemic schema with which the community is used to judge its practice, assessing and revising the concept as a whole is challenging, in particular, for those who are situated within this community (see also Kraft 2025). Therefore, ‘outside’ perspectives are likely to prove invaluable (cf. Branford, Soulier, and Fichtner 2025).

Benchmarking is inevitably a reductive mode of analysis and, in important respects, may well remain *per se* irreparably problematic and prone to capture. Yet, this is not an isolated problem, limited to AI, but one long recognised in the philosophy of science, which has also suggested that we can nevertheless structure our practices to best mitigate these deficiencies by accepting the situatedness of knowledge claims and being explicit about their purported utility (understood as a value-laden claim). Specifically, this leads us to advocate for AI research to endeavour towards what Kitcher (2001, 2011) terms a “well-ordered science”, which is intended to “serve the collective good” (2001, p. xiii) by identifying relevant ends and means through the best available approximation of an “ideal conversation, embodying all human points of view, under conditions of mutual en-

agement” (2011, p.106). Through this it might aspire to “answer the *right* questions in the *right* ways, where value judgements and methodological issues are inextricably intertwined in determining what is *right*” (Cartwright 2006, p. 981). Accordingly, the path towards better AI evaluation requires (1) a wider, participatory ecosystem of imagination, debate, and audit, while at the same time (2) accepting and embracing the locality and temporality of what is considered ‘right’ and ‘wrong’, ‘desirable’ and ‘undesirable’.

Regarding (1), the AI community is already trialling more collective forms of evaluation, namely *red-teaming*, i.e. collective adversarial testing of ‘model safety’ (Ganguli et al. 2022), and open voting platforms like Chatbot Arena (Chiang et al. 2024), where anyone can rate models against each other according to one's preferred outputs. However, sarin and Bao (2024) raise the concern that those approaches are falsely marketed as ‘democratic’, and that powerful players are harvesting free labour under this guise. The crowd-sourced performance judgements and preference feedback (on reductive pre-defined dimensions and in strictly standardised formats) are usually absorbed by technology companies towards the improvement of their proprietary models. Furthermore, these platforms do not provide the necessary space for public “deliberation and discourse”, or better even “dissent, disobedience, and difference”, which the authors consider important constituents of a truly democratic process. What is needed are community-led initiatives that ensure that labour and gains rest with the same group of individuals, and that allow for intervention. This directs us to claim (2): Of course, science will always need some methodological consensus and shared standards to function. In striving for rigour and objectivity, such consensus is best identified among and tested against diverse, conflicting, situated, partial, and changing perspectives. Evaluation should be recognised as a bounded practice, never all-encompassing and general-purpose. Only then can we start to create meaningful measures to evaluate evaluations and assure aptness to context.

Finally, for epistemic and ethical reasons, the community has to start holding those in power accountable for capturing, exploiting, and gaming evaluation in their own favour. This is a necessary means for mitigating the extent to which the structural injustices in AI benchmarking are avoidable. This paper is a first step in this direction. Instead of vaguely calling for system monitoring through benchmarks, regulators should require true democratic oversight of AI tools when it comes to assessing their suitability and risks. This can only happen through localised, context- and use-specific, as well as continuously recurring modes of evaluation. To this end, regulators must provide the venues and infrastructures needed for its realisation, where diverse stakeholders can meaningfully deliberate and contest decisions.

## Acknowledgements

This work was supported by the Weizenbaum Institute (grant number 16DIII141), funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) and the State of Berlin.

<sup>13</sup><https://evalvalai.com/research/2025/07/13/eval-science-kickoff/> (access date: May 17, 2026)

## References

- Baack, S.; Buschek, C.; and Bohacek, M. 2026. Unsteady Metrics and Benchmarking Cultures of AI Model Builders. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, 840–872. New York, NY, USA: ACM. ISBN 9798400725968.
- Bean, A. M.; Kearns, R. O.; Romanou, A.; Hafner, F. S.; Mayne, H.; Batzner, J.; Foroutan, N.; Schmitz, C.; Korgul, K.; Batra, H.; Deb, O.; Beharry, E.; Emde, C.; Foster, T.; Gausen, A.; Grandury, M.; Han, S.; Hofmann, V.; Ibrahim, L.; Kim, H.; Kirk, H. R.; Lin, F.; Liu, G. K.-M.; Luettgau, L.; Magomere, J.; Rysström, J.; Sotnikova, A.; Yang, Y.; Zhao, Y.; Bibi, A.; Bosselut, A.; Clark, R.; Cohan, A.; Foerster, J. N.; Gal, Y.; Hale, S. A.; Raji, I. D.; Summerfield, C.; Torr, P.; Ududec, C.; Rocher, L.; and Mahdi, A. 2025. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, NeurIPS '25.
- 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: ACM. ISBN 9781450383097.
- Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22. New York, NY, USA: ACM. ISBN 9781450394772.
- Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, 5454–5476. ACL.
- Bommasani, R.; Creel, K. A.; Kumar, A.; Jurafsky, D.; and Liang, P. S. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 3663–3678. Red Hook, NY, USA: Curran Associates Inc.
- Bommasani, R.; Liang, P.; and Lee, T. 2023. Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, 1525(1): 140–146.
- Bordes, F.; Ross, C.; Kao, J. T.; Spiliopoulou, E.; and Williams, A. 2025. Eval Factsheets: A Structured Framework for Documenting AI Evaluations. arXiv:2512.04062.
- Bowker, G. C.; and Star, S. L. 1999. *Sorting Things Out: Classification and Its Consequences*. The MIT Press. ISBN 9780262269070.
- Bowman, S. R.; and Dahl, G. 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*, NAACL '21, 4843–4855. ACL.
- Branford, J.; Soulier, E.; and Fichtner, L. 2025. Generative AI and Democratic Culture. *Philosophy & Technology*, 38(123).
- Browne, J. 2023. AI and Structural Injustice: A Feminist Perspective. In Browne, J.; Cave, S.; Drage, E.; and McInerney, K., eds., *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*, 328–346. Oxford University Press. ISBN 9780192889898.
- Cartwright, N. 2006. Well-Ordered Science: Evidence for Use. *Philosophy of Science*, 73(5): 981–990.
- Center for AI Safety; Scale AI; and HLE Contributors Consortium. 2026. A Benchmark of Expert-Level Academic Questions to Assess AI Capabilities. *Nature*, 649(8099): 1139–1146.
- Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M. I.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24. JMLR.org.
- Couldry, N.; and Mejias, U. A. 2019. Data Colonialism: Rethinking Big Data’s Relation to the Contemporary Subject. *Television & New Media*, 20(4): 336–349.
- Crawford, K. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press. ISBN 9780300209570.
- Cronbach, L. J.; and Meehl, P. E. 1955. Construct Validity in Psychological Tests. *Psychological Bulletin*, 52(4): 281.
- Demchak, N.; Guan, X.; Wu, Z.; Xu, Z.; Koshiyama, A.; and Kazim, E. 2024. Assessing Bias in Metric Models for LLM Open-Ended Generation Bias Benchmarks. In *Proceedings of the Workshop Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI co-located with NeurIPS '24*.
- Dotson, K. 2012. A Cautionary Tale: On Limiting Epistemic Oppression. *Frontiers: A Journal of Women Studies*, 33(1): 24–47.
- Dotson, K. 2014. Conceptualizing Epistemic Oppression. *Social Epistemology*, 28(2): 115–138.
- European Union. 2024. Regulation (EU) 2024/1689 – Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). *Official Journal of the European Union*.
- Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; Jones, A.; Bowman, S.; Chen, A.; Conerly, T.; Das-Sarma, N.; Drain, D.; Elhage, N.; El-Showk, S.; Fort, S.; Hatfield-Dodds, Z.; Henighan, T.; Hernandez, D.; Hume, T.; Jacobson, J.; Johnston, S.; Kravec, S.; Olsson, C.;

- Ringer, S.; Tran-Johnson, E.; Amodei, D.; Brown, T.; Joseph, N.; McCandlish, S.; Olah, C.; Kaplan, J.; and Clark, J. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858.
- Gebrekidan, F. B. 2024. Content Moderation: The Harrowing, Traumatizing Job that Left Many African Data Workers with Mental Health Issues and Drug Dependency. In *Data Workers' Inquiry*.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; III, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Commun. ACM*, 64(12): 86–92.
- Hao, K. 2025. *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. New York: Penguin Press. ISBN 9780593657508.
- Haraway, D. 2016. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. In *Space, Gender, Knowledge: Feminist Readings*, 53–72. Routledge.
- Hartmann, D.; Tonneau, M.; Kraft, A.; Seiling, L.; Stauffer, D.; Delobelle, P.; Fillies, J.; Luther, A. R.; Batzner, J.; and Lisker, M. 2026. Bye Bye Perspective API: Lessons for Measurement Infrastructure in NLP, CSS and LLM Evaluation. arXiv:2604.25580.
- Herzog, C.; and Branford, J. 2025. Relational Ethics and Structural Epistemic Injustice of AI in Medicine. *Philosophy & Technology*, 38(4): 160.
- Herzog, L. 2021. Algorithmic Bias and Access to Opportunities. In Véliz, C., ed., *The Oxford Handbook of Digital Ethics*, 413–432. Oxford University Press. ISBN 9780198857815.
- Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 375–385. New York, NY, USA: ACM. ISBN 9781450383097.
- Jain, S.; Suriyakumar, V.; Creel, K.; and Wilson, A. 2024. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 197–206. New York, NY, USA: ACM. ISBN 9798400704505.
- Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues? In Kim, B.; Yue, Y.; Chaudhuri, S.; Fragkiadaki, K.; Khan, M.; and Sun, Y., eds., *International Conference on Learning Representations*, ICLR '24, 54107–54157.
- Kapania, S.; Yang, T.; Abdelkadir, N. A.; Scheuerman, M. K.; Miceli, M.; Taylor, A. S.; and Fox, S. E. 2026. 'The plan is just survival': Data Work in Kenya and the Regime of Entrapment. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: ACM. ISBN 9798400722783.
- Kasirzadeh, A. 2022. Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 349–356. New York, NY, USA: ACM. ISBN 9781450392471.
- Kitcher, P. 2001. *Science, Truth, and Democracy*. New York: Oxford University Press. ISBN 9780195145830.
- Kitcher, P. 2011. *Science in a Democratic Society*. Amherst, NY: Prometheus Books. ISBN 9781616144074.
- Koch, B.; Denton, E.; Hanna, A.; and Foster, J. G. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*, NeurIPS '21.
- Kraft, A. 2025. *On Knowledge in AI: Epistemic and Ethical Limitations of Language Models and Knowledge Graphs*. Ph.D. thesis, University of Hamburg, Hamburg.
- Kraft, A.; Simon, J.; and Schimmler, S. 2025. Social Bias in Popular Question-Answering Benchmarks. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AACL '25*, 1421–1438. AFNLP and ACL. ISBN 979-8-89176-298-5.
- Liu, Y. L.; Blodgett, S. L.; Cheung, J.; Liao, Q. V.; Olteanu, A.; and Xiao, Z. 2024. ECBD: Evidence-Centered Benchmark Design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '24, 16349–16365. ACL.
- Longino, H. 2002. *The Fate of Knowledge*. Princeton University Press. ISBN 9780691187013.
- Longpre, S.; Singh, N.; Cherep, M.; Tiwary, K.; Materzynska, J.; Brannon, W.; Mahari, R.; Obeng-Marnu, N.; Dey, M.; Hamdy, M.; Saxena, N.; Anis, A. M.; Alghamdi, E. A.; Chien, V. M.; Yin, D.; Qian, K.; Li, Y.; Liang, M.; Dinh, A.; Mohanty, S.; and et al. 2025. Bridging the Data Provenance Gap Across Text, Speech, and Video. In *The Thirteenth International Conference on Learning Representations*, ICLR '25. OpenReview.net.
- McKeown, M. 2021. Structural Injustice. *Philosophy Compass*, 16(7): e12757.
- McKeown, M. 2024. *With Power Comes Responsibility: The Politics of Structural Injustice*. Bloomsbury.
- Miceli, M.; and Posada, J. 2022. The Data-Production Dispositif. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).
- Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33: 659–684.
- Muldoon, J.; Graham, M.; and Cant, C. 2024. *Feeding the Machine: The Hidden Human Labour Powering AI*. Edinburgh: Canongate Books. ISBN 9781838859114.
- Naudts, L. 2024. The Digital Faces of Oppression and Domination: A Relational and Egalitarian Perspective on the Data-driven Society and its Regulation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 701–712. New York, NY, USA: ACM. ISBN 9798400704505.

- Orr, W.; and Kang, E. B. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1875–1884. New York, NY, USA: ACM. ISBN 9798400704505.
- Ott, S.; Barbosa-Silva, A.; Blagec, K.; Brauner, J.; and Samwald, M. 2022. Mapping Global Dynamics of Benchmark Creation and Saturation in Artificial Intelligence. *Nature Communications*, 13: 6793.
- Powers, H.; Baldini, I.; Wei, D.; and Bennett, K. P. 2024. Statistical Bias in Bias Benchmark Design. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS '24. Red Hook, NY, USA: Curran Associates Inc.
- Raji, I. D.; Denton, E.; Bender, E. M.; Hanna, A.; and Paullada, A. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, *NeurIPS Datasets and Benchmarks 2021*, NeurIPS '21.
- Ricaurte, P. 2019. Data Epistemologies, the Coloniality of Power, and Resistance. *Television & New Media*, 20(4): 350–365.
- Ricaurte, P. 2022. Ethics for the Majority World: AI and the Question of Violence at Scale. *Media, Culture & Society*, 44(4): 726–745.
- Sainz, O.; Campos, J. A.; García-Ferrero, I.; Etxaniz, J.; de Lacalle, O. L.; and Agirre, E. 2023. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23.
- sarin, p.; and Bao, M. 2024. Democratic Perspectives and Corporate Captures of Crowdsourced Evaluations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS '24. Red Hook, NY, USA: Curran Associates Inc.
- Schaeffer, R.; Miranda, B.; Kazdan, J.; Liu, K.; Ahmed, A. M.; Mireshghallah, N.; and Koyejo, S. 2025. Causally Quantifying the Effect of Test Set Contamination on Generative Benchmarks. In *NeurIPS '25 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, NeurIPS '25.
- Sloane, M.; Moss, E.; Awomolo, O.; and Forlano, L. 2022. Participation Is not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, 1–6. New York, NY, USA: ACM.
- Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, 3645–3650. ACL.
- Thomas, R. L.; and Uminsky, D. 2022. Reliance on Metrics Is a Fundamental Challenge for AI. *Patterns*, 3(5): 100476.
- Truong, S. T.; and Koyejo, S. 2026. *AI Measurement Science: A Science of Knowing Where AI Thrives, Where It Breaks, and How to Respond*. Stanford University.
- Uberti-Bona Marin, L. G.; Rijsbosch, B.; Meding, K.; Spanakis, G.; van Dijk, G.; and Kollnig, K. 2026. Is your AI Model Accurate Enough? The Difficult Choices Behind Rigorous AI Development and the EU AI Act. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, 8184–8201. New York, NY, USA: ACM. ISBN 9798400725968.
- Umutlu, E. E.; Cengiz, A. A.; Sever, A. K.; Erdem, S.; Aytan, B.; Tufan, B.; Topraksoy, A.; Darıcı, E.; and Toraman, C. 2025. Evaluating the Quality of Benchmark Datasets for Low-Resource Languages: A Case Study on Turkish. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics*, GEM<sup>2</sup>, 471–487. ACL. ISBN 979-8-89176-261-9.
- Uzunoglu, A.; Li, T.; and Khashabi, D. 2025. The Flaw of Averages: Quantifying Uniformity of Performance on Benchmarks. arXiv:2509.25671.
- Wallach, H.; Desai, M.; Cooper, A. F.; Wang, A.; Atalla, C.; Barocas, S.; Blodgett, S. L.; Chouldechova, A.; Corvi, E.; Dow, P. A.; Garcia-Gathright, J.; Olteanu, A.; Pangakis, N.; Reed, S.; Sheng, E.; Vann, D.; Vaughan, J. W.; Vogel, M.; Washington, H.; and Jacobs, A. Z. 2025. Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org.
- Young, I. M. 1990. *Justice and the Politics of Difference*. Princeton University Press.
- Young, I. M. 2006. Responsibility and Global Justice: A Social Connection Model. *Social Philosophy and Policy*, 23(1): 102–130.
- Young, I. M. 2010. *Responsibility for Justice*. Oxford University Press.