Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor

Alexandra Olteanu[1*] **Su Lin Blodgett**[1] **Agathe Balayn**[1] **Angelina Wang**[2]
Fernando Diaz[3] **Flavio du Pin Calmon**[4] **Margaret Mitchell**[5]
Michael Ekstrand[6] **Reuben Binns**[7] **Solon Barocas**[1]
[1]Microsoft Research [2]Cornell Tech [3]Carnegie Mellon University [4]Harvard University
[5]Hugging Face [6]Drexel University [7]University of Oxford
[*]Corresponding author: `alexandra.olteanu@microsoft.com`

Abstract

In AI research and practice, *rigor* remains largely understood in terms of *methodological rigor*—such as whether mathematical, statistical, or computational methods are correctly applied. We argue that this narrow conception of rigor has contributed to the concerns raised by the responsible AI community, including overblown claims about the capabilities of AI systems. Our position is that a broader conception of what rigorous AI research and practice should entail is needed. We believe such a conception—in addition to a more expansive understanding of (1) *methodological rigor*—should include aspects related to (2) what background knowledge informs what to work on (*epistemic rigor*); (3) how disciplinary, community, or personal norms, standards, or beliefs influence the work (*normative rigor*); (4) how clearly articulated the theoretical constructs under use are (*conceptual rigor*); (5) what is reported and how (*reporting rigor*); and (6) how well-supported the inferences from existing evidence are (*interpretative rigor*). In doing so, we also provide useful language and a framework for much-needed dialogue about the AI community's work by researchers, policymakers, journalists, and other stakeholders.

1 Rigor in AI Research and Practice

Rigor remains a subject of intense debate in science [e.g., 2, 20, 43, 53, 68, 81, 104, 107, 129], a debate we are unlikely to resolve here. We thus keep our goal relatively modest: to help broaden the AI community's perspective of what rigorous AI[1] work should entail. We argue this is critically needed as *relying on impoverished conceptions of rigor can have an undesirable, yet formative impact on the quality of both AI research and practice*—heightening concerns ranging from unsubstantiated claims about AI systems [e.g., 35, 115, 127, 141, 147] to a plethora of unintended consequences [e.g., 80, 110, 126].

Rigor in AI: The debate surrounding rigor in science has by no means evaded the AI community [e.g., 65, 87, 117, 130]. In AI research and practice, rigor remains largely understood in terms of *methodological rigor*, which is typically conceptualized as whether mathematical, statistical, or computational methods are correctly applied; whether new methods, models, or systems are tested on large-scale or complex benchmarks and compared with a sufficient number of competing methods, models, or systems; whether the methods or analyses can scale or generalize; or whether the phenomena under analysis were—in contrast to more qualitative work—mathematically formalized

[1]Throughout this paper, we deliberately leave the term AI under-specified in order to capture the full constellation of meanings the term has been used for, and thus any work that practitioners or researchers would describe as AI—i.e., if they consider themselves as working on AI, our call is for them. We thus mainly use the term as a modifier to help scope the community this call is addressing, and the work and artifacts this community does or uses.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Position Paper Track.

Facet of rigor	What the facet is concerned with	What the facet asks for
Epistemic rigor (§2.1)	What *background knowledge* informs which problems are addressed and how?	Is the *background knowledge* clearly and explicitly communicated? Is the *background knowledge* appropriate, well-justified, and appropriately applied?
Normative rigor (§2.2)	Which disciplinary, community, organizational, or personal *norms, standards, values, or beliefs* influence the work and how?	Are these *norms, standards, values, or beliefs* clearly and explicitly communicated? Are these *norms, standards, values, or beliefs* appropriate & appropriately followed?
Conceptual rigor (§2.3)	Which *theoretical constructs* are under investigation?	Are the *theoretical constructs* clearly and explicitly articulated? Are the *theoretical constructs* appropriate and well-justified?
Methodological rigor (§2.4)	Which *methods* are being used?	Are these *methods* and their use clearly and explicitly described? Are these *methods* appropriate, well-justified and appropriately applied?
Reporting rigor (§2.5)	What are the *research findings*?	Are the *research findings* clearly communicated? Is the presentation of *research findings* appropriate and well-justified?
Interpretative rigor (§2.6)	What *inferences* are being drawn from the research findings?	Are these *inferences* clearly and explicitly communicated? Are these *inferences* appropriate and well-justified?

Table 1: Overview of the six facets of rigor in AI research and practice. For each facet, we highlight what the facet is concerned with (*descriptive* overview)—i.e., what the *objects of concern* is for the facet—and what the facet asks for (*evaluative* overview).

and quantified [e.g., 22, 26, 49, 60, 65, 75, 130, 131, 142, 156]. These conceptualizations are often shaped by implicit assumptions that more complex methods and architectures and larger data samples are better [e.g., 9, 47, 83, 139, 142], by the common practice of relying on benchmark-driven evaluations [e.g., 86, 144, 156], and by a dominant mode of thinking oriented towards algorithmic formalism (centered on ideas of objectivity/neutrality, abstract and mathematical representations, and universalism/generalization) [e.g., 22, 44, 60, 89, 149]. Yet such conceptualizations of rigor may fail to demand, for instance, that benchmarks be fit-for-purpose or proven to measure what they claim to be measuring [e.g., 28, 113, 145], or that the knowledge or assumptions the work relies on be reliable or valid [e.g., 28, 92, 98, 115]. This puts into question the integrity and reliability of any conclusions drawn based on such benchmark-centered evaluations or that depend on questionable assumptions.

While such failures threaten the scientific integrity of AI research, foreseeing and addressing the consequences of putting insufficiently rigorous AI artifacts (e.g., models, systems, applications, outputs, data) into practice are often seen as the purview of responsible AI—which is nevertheless generally seen as separate from rigor, as it is understood as more concerned with ethics, stakeholders, societal impacts and harms, and real-world deployment scenarios. However, it is often only in real-world deployment scenarios or when considering stakeholders that the inadequacies and impact of current approaches to (lack of) rigor in AI research and practice become clear. We thus recast this distinction, and argue that by making visible and demanding attention to such failures, *responsible AI asks researchers and practitioners to uphold principles of scientific integrity in their work*. Although responsible AI is often seen as out of scope for an AI researcher or practitioner not engaged in that space, *any scientist should see rigor as well within scope*. In other words, even though unrigorous AI work implicates what are often seen as responsible AI consequences, we argue that a broader notion of rigor that accounts for what produces these consequences needs to be considered by *all* AI researchers and practitioners.[2]

We take the position that **doing rigorous AI work requires broadening our understanding of what rigor in AI research and practice should entail by drawing on work by the responsible AI community.**[3] To this end, we foreground a wider range of critical considerations from responsible AI literature (broadly construed) for rigorous AI work. Instead of prescribing rigid criteria for rigorous AI work, our goal is to provide useful scaffolding for much-needed dialogue about and scrutiny of the community's work by researchers as well as policymakers, journalists, and other stakeholders.

2 Facets of Rigor in AI Research and Practice

We foreground six key facets of rigor—*epistemic*, *normative*, *conceptual*, *methodological*, *reporting*, and *interpretative*—that we argue AI research and practice should contend with. While it may be

[2]*Positionality statement:* In recent years, we have noticed increasing tensions within and across AI and AI-adjacent communities concerning what research questions should be prioritized, who can even be trusted to conduct certain types of work, and what rigorous research entails. We also observed trends in the types of work submitted and published at both AI and AI-adjacent venues, raising concerns about a pernicious lack of conceptual clarity, growing pockets of research motivated by hypothetical, speculative settings, and an increasing use of ambiguous, non-productive terminology. These experiences inform and motivate this paper. Our perspective is also informed by our diverse disciplinary backgrounds spanning computer systems, theoretical and applied ML, information retrieval, computational social science, natural language processing, computational linguistics, computer vision, human-computer interaction, law and policy, science and technology studies, and philosophy.

[3]We use *responsible AI* to broadly refer to critical work on the impact of AI artifacts on people and society from several communities, including ethical AI, responsible AI, ethics of AI, and science and technology studies.

Figure 1: Simplified overview of the objects of concern for each **facet of rigor** and of common dependencies among them. For instance, the research findings typically determine what inferences and claims can be made, while normative considerations may influence the choices of theoretical constructs, of what methods to use, of what research findings to report, or of what inferences and claims to make. Dependencies are illustrated through both arrows as well as nested boxes.

difficult to draw clear boundaries between some of these facets, we discuss them separately to provide distinct lenses for reflecting on and interrogating the quality and scientific integrity of AI work (see Table 1). Specifically, for each facet, we provide a *descriptive overview*—what the facet is concerned with, i.e., what the *object of concern* is for that facet—and an *evaluative overview*—what the facet asks for. All facets of rigor are inherently about the *choices* we make about an *object of concern* (e.g., *epistemic rigor* is concerned with *background knowledge*, while *conceptual rigor* is concerned with *theoretical constructs*). Having these distinct facets encourages researchers and practitioners to think carefully about the *choices* they make for each *object of concern*, and disentangles possible debates about these choices—e.g., it separates debates about which *norms* we should follow (*normative rigor*) from what *background knowledge* should inform the work (*epistemic rigor*). For each facet, we discuss examples of mechanisms that help promote rigor along that facet, which can include a mix of processes and desiderata. While we foreground each mechanism for only one of the facets, we note that some mechanisms might help foster rigor along more than one facet (e.g., engaging with construct and internal validity concerns can foster both *methodological* and *interpretative rigor*).

Figure 1 provides a simplified overview of the *objects of concern* for each facet of rigor and of possible common dependencies among them. It illustrates how limiting our conception of rigor to methodological concerns may obfuscate how *our work and the claims we make are shaped by a variety of choices that both precede and succeed any methodological considerations*, even when some of those choices remain tacit or implicit (e.g., as is often the case for common disciplinary norms, standards, or practices [56, 156]). The figure also underscores how it may be difficult to make good choices for "downstream" objects when "upstream" objects are poorly chosen. For instance, making poor construct choices (*conceptual rigor*) may reduce our chances of operationalizing those constructs well (*methodological rigor*). The different facets of rigor may, however, also be tangled in complex relations of mutual interdependency [e.g., 44, 57, 145]—as methodological choices may, for instance, in turn "limit the structure of one's theoretical con[structs]" [44] if the methods determine what is observed and thus theorized about—and present choices may (and usually do) impact future work. For example, a lack of *interpretative rigor*—when ambiguous or baseless claims are being made—can have *epistemic* consequences for future work relying on those claims [e.g., 41, 65]. We further unpack these below:

2.1 Epistemic rigor

What background knowledge informs which problems are addressed and how? Is this background knowledge clearly and explicitly communicated? Is the background knowledge appropriate, well-justified, and appropriately applied?

Before considering any methodological questions, we have to contend with what it is that is being investigated, why it merits consideration, and what knowledge is "not under investigation, but [is] assumed, asserted, or essential" [61]. That is, we need to clarify the facts and background assumptions upon which the work relies in order to establish what the foundation for the work is and whether that foundation is sound. A frequently given example for epistemic failures is work asserting that records of physical appearance (e.g., photographs of faces) can be used to predict latent character traits (e.g., sexuality, political ideology, or criminality) [10]. Such work assumes that actions (what one does) or inner character (what one likes, thinks, or values) can be predicted based on physical attributes (how one looks). Although this assumption has its roots in "physiognomy"—regarded as pseudo-scientific as it relies on a set of epistemically baseless and extensively debunked claims [5, 137]—pockets of AI research recurrently draw upon it [5, 10]. This example also illustrates a broader class of epistemic failures: cases where a failure to scrutinize background assumptions may lead to work whose validity depends on whether existing methods, tools, or systems function—or can be made to function—on given tasks when they do not or cannot [115], including because the tasks are conceptually or practically impossible or because there is no reliable evidence that those methods, tools, or systems are fit-for-purpose or reliable [98, 115, 119, 156].

Epistemic rigor is thus not only about making sure that the background knowledge new work builds on and how that knowledge is acquired are appropriate and appropriately applied, but also that that knowledge is appropriately justified [36]. Rigorously applying statistical or computational methods (*methodological rigor*) will do little to mitigate epistemic concerns if the problems being tackled or the assumptions underpinning the use of some methods are baseless, theoretically implausible, nonsensical, or grounded in pseudo-scientific or scientifically shallow work [3, 10, 80, 115, 133]. Why and what is being studied, built, or deployed [15]; why, what and whose problems are being prioritized [23, 124]; what implicit or tacit assumptions are being made about the stated problems or solutions under consideration [3, 93, 115]; and whether the choices of problems and methods are drawing on valid and well-founded evidence [24] are all examples of the types of key epistemic considerations that we should engage with. By contrast, easily available or common artifacts (such as datasets, methods, tools, or systems) tend to serve as "tools of opportunity, not instruments of epistemic rigor" [153], with researchers and practitioners often prioritizing work based on whether there are some existing artifacts available without appropriately scrutinizing the knowledge and assumptions underpinning them [156].

Epistemic rigor, however, does not necessarily require specific epistemological commitments or choices but rather that those commitments and choices be made explicit. While doing so may not definitively answer whether some problems or assumptions are baseless, nonsensical, or unethical, articulating epistemic commitments and the existing knowledge the work relies on lays down the grounds on which people can have discussions and work out disagreements [32].

Mechanisms to promote *epistemic rigor*: Ensuring work is *appropriately grounded* in past literature [e.g., 10] and that underlying assumptions are made explicit and *appropriately interrogated* [e.g., 93, 115, 156] can help foster epistemic rigor, and in turn all other facets of rigor (§2.2–§2.6).

Appropriate grounding: Common goals in AI research often include producing new insights, theories, or artifacts, providing evidence in support of existing theories, evaluating existing artifacts, or debunking prior work. A failure to review, appropriately situate within, and acknowledge prior literature, and how it influenced the questions being asked or the solutions being considered, can cast doubt on whether any meaningful progress towards those goals was actually made [e.g., 63, 68]. When such failures become systematic, research communities risk "curl[ing] up upon themselves and becom[ing] self-referential systems that orient more [internally]" [37] and developing their separate "terminology, source texts, and knowledge claims" [7]. Indeed, concerns that "[f]indings within [a research] community are often self-referential and lack [quality]" [67] are critical as this can make it difficult to trace the *provenance of central claims* that form the foundation that new work builds on or is motivated by. It can also result in an overall narrowing of what questions AI research tackles and the epistemic marginalization of certain viewpoints [6, 82, 98]. Appropriate grounding requires due diligence when reviewing and acknowledging work in one's subfield and related fields [18, 68, 101].

Interrogating assumptions: Any research work necessarily relies on assumptions about what is important, what is possible, or what represents sufficient or useful evidence [e.g., 118]. Not only does a lack of epistemic rigor make it hard to situate new work and the knowledge it produces in the context of existing knowledge it may draw from, relate to, corroborate, or dispute, but a lack of due diligence about the background knowledge underlying the work further risks bringing to the fore long-debunked claims from other disciplines—e.g., risking the reanimation of physiognomic methods [5, 10, 137]. Interrogating the assumptions underpinning such tasks—e.g., why would outer attributes be useful proxies for someone's character?—can expose them as a category error where observable traits (e.g., skin tone) are incorrectly treated as direct indicators of internal states (e.g., personality). Making background assumptions explicit thus facilitates our ability to scrutinize them [14, 156].

2.2 Normative rigor

Which disciplinary, community, organizational, or personal norms, standards, values, or beliefs influence the work and how? Are these norms, standards, values, or beliefs clearly and explicitly communicated? Are these norms, standards, values, or beliefs appropriate and appropriately followed?

Because of differences in underlying evidentiary standards, background assumptions, goals, ideals, and theoretical frameworks, different disciplines and communities have different ways of determining what methods to use, what types of evidence are needed or sufficient, and why the resulting work or resulting knowledge matters [31, 36]. The assumptions and theories that underpin our work are not derived "out of thin air, but as a function of our experiences with the world in general and specifically with our colleagues, through dialogue, and the literature" [61], all of which can shape the type, direction, and quality of research [44, 57, 69]. Disciplinary and community norms are intertwined with a researcher's personal norms and drivers [45], e.g., what catches their interest or advances their goals. Making explicit how this mix of influences shapes the AI community's work can help others understand them and enable debates on their appropriateness.

Consider the emerging research on developing AI personas as tools to simulate users and human study participants [e.g., 8, 54, 84, 111]. While this research reflects common norms and beliefs in AI communities around scalability and efficiency [e.g., 4, 22], by aiming to replace human participants, such tools may fail to align or may even conflict with foundational values and norms around representation, participation, inclusion, and understanding [4, 148] that underpin many types of work this research seeks to support, such as user experience or social science studies with human participants. Such value conflicts "cannot be alleviated with better training or improved model performance alone" [4].

Common beliefs and expectations in the AI community—such as the goal of producing generalizable findings that are often abstracted away from any use context [22, 70]—have led to a reduction of problems and use scenarios "to a common set of representations or affordances" [22], and thus (even if implicitly) to de-valuing how context shapes datasets and tools. Reliance on undisclosed normative assumptions about problem statements, artifacts, or constructs that are contested or value-laden, yet treated as if they are neutral or generally applicable, can however be particularly misleading when a "field employs the language of procedural adherence to project a sense of certainty, objectivity, and stability" [59]. While dominant values in AI work like performance, generalization, and efficiency [22] may encourage extensive evaluations across as many metrics, benchmarks, and baselines as possible [22, 156], these evaluations tell us little about when, whether, or which improvements are necessary, desirable, or translate to meaningful benefits when deployed. Such expectations and norms—along resource-related considerations and constraints such as about costs and strict timeliness [64, 154, 156]—nonetheless influence what type of work gets prioritized or rewarded [50].

Mechanisms to promote *normative rigor*: All research is shaped by normative considerations. Normative rigor asks us to make these considerations explicit—such as via *ethical* and *positionality* statements [110] or other disclosure practices—and can include aspects related to expectations around the *significance and impact* of research [e.g., 61, 107] or what *ethical norms* to follow [e.g., 109].

Research significance: Central to normative considerations are expectations about how our work and its outcomes should intervene in the world, or why the work matters and is appropriate. The goal of research is often to "generate knowledge that will have positive practical impact" [12] or that advances the field [107], with norms around what work is worthwhile driven by desires to reward such work [34, 150]. There is however also an increasing recognition that the appraisal of the work's quality and significance should also include the work's potential for harm and not only for benefits [12, 29, 61]. Such appraisal typically rests on claims about the possible impacts from either the work's process or its outcomes. This echoes the concept of *consequential validity* from social science measurement scholarship [99, 145], which asks us when determining a measurement instrument's (in)validity to consider the consequential basis both of its use and of possible inferences (along with actions those inferences may entail) [72, 99]. Reckoning with the impact of our work thus promotes not only discussions about whether current norms around what constitutes good work are appropriate, but also promotes methodological (§2.4) and interpretative rigor (§2.6).

Positionality and ethical statements: Researchers' personal, disciplinary, and institutional backgrounds, their lived experiences, and their goals motivate and shape how they approach their work. *Positionality statements* are a mechanism meant to make such considerations explicit in order to help others contextualize the research and research outcomes [85, 110]. Positionality, however, does not only encompass aspects related to researchers' beliefs and values, but also those related to what knowledge they draw on, how they know what they know, and how they make methodological choices [19]. Thus, it can also aid or compromise epistemic (§2.1) and methodological rigor (§2.4). *Ethical statements* can further complement positionality statements by foregrounding the ethical concerns researchers grappled with or mitigated before or while conducting the work [13, 110]. Yet this practice of disclosing whether such concerns were considered and how they shaped any methodological choices (if at all) remains inconsistently adopted across AI communities.

2.3 Conceptual rigor

Which theoretical constructs are under investigation? Are these theoretical constructs clearly and explicitly articulated? Are these theoretical constructs appropriate and well-justified?

Assume a researcher wishes to evaluate whether a model "hallucinates." In AI research, the construct of "hallucination" has, however, been used to refer to several distinct types of system behaviors [95], including cases of generating content which is nonsensical, which contains factual errors, which is not in the input data, which is not in the training data, or a mix of these behaviors. Further, all these different understandings of what it means for a model to "hallucinate" are markedly different from the more common use of the term that requires an ability to perceive, feel, or have sensory experiences [e.g., 40]; and the term can thus carry meanings incompatible with AI systems. To understand what a researcher's evaluation of whether the model "hallucinates" means, we thus need to know which conceptualization they use, and if that conceptualization is sensible.

While contested constructs—those that have competing or even conflicting definitions, like "hallucination," "value alignment," "AGI," or "human-likeness"—are increasingly common in AI research and practice, their definitions often remain elusive, ambiguous, or poorly specified [e.g. 25, 26, 28, 145]. *Without clarity about what specifically we are analyzing, measuring, or striving for, it can be hard to assess progress or make any useful or reliable claims.* Work can also rely on constructs that are inconsistent with any theoretical tradition, such as treating identity categories as fixed and objective rather than continuous and constructed [92] or more generally failing to recognize that some constructs are innately fluid, non-deterministic, and fuzzy [21]. Such a lack of clarity can hinder replicability and reproducibility or may facilitate speculative post-factum interpretations, yielding possibly unsound and unfounded claims, or a conflation of proxy measurements (which may or may not measure any version of the underlying construct) with the construct under analysis. Thus, a lack of conceptual clarity can also undermining epistemic (§2.1), methodological (§2.4), and interpretative rigor (§2.6).

Mechanisms to promote *conceptual rigor*: Conceptual rigor requires attention to *conceptual clarity*—which construct we are after and how it is defined; appropriate *conceptual systematization*—the process by which the definition is made specific; and *terminological rigor*—that the terms used to refer to a construct do not harbor meanings that can lead to a misinterpretation of what the construct is.

Conceptual clarity: A growing number of objects in AI research are ambiguous or poorly specified, or else are objects for which we lack consensus about what they are or what they are for. The examination by Saphra and Wiegreffe [125] of what is meant by "mechanistic interpretability" is instructive: not only does the term have multiple competing meanings, but those meanings also reflect distinct disciplinary orientations and epistemic origins; a lack of clarity about which meaning is under use can obfuscate not only what the work does, but also why and how the work is done. Similar critiques have been made about the lack of conceptual clarity about what unlearning [42], bias [27], model collapse [128], interpretability [88], or generalization [89] mean. While we see growing concerns about the lack of conceptual clarity surrounding many aspects of AI research, from how desired capabilities are described to what metrics to optimize for [28, 65, 73, 125, 145], these remain largely overlooked in discussions about research integrity and quality in AI.

Conceptual systematization: In practice, many constructs involve a "broad constellation of meanings and understandings" [1], and working with them requires making choices about which meanings to use, "narrowing [them] into an explicit definition" [145]. The process of conceptual systematization asks researchers to engage not only with a high-level construct in the abstract, but to grapple more concretely with what it means in the context of the work they are conducting and how it relates to empirical observations or other constructs. Conceptual systematization is a prerequisite for rigorous measurement, (computational) specification, empirical analysis, and theory development [e.g., 1, 65, 113, 145], and is thus a prerequisite for methodological rigor (§2.4).

Terminological rigor: Conceptual rigor depends on terminological choices and what those choices communicate. Many terms in AI often carry over meanings from the human realm or other disciplinary contexts that are incompatible with AI systems, and can mislead [26, 38, 89, 118], suggest "unproven connotations," or lead to "collisions with other definitions, or conflation with other related but distinct concepts" [89]. Blili-Hamelin et al. [26] note that "when researchers equate human faculties with model proxies [...] [t]his rhetorical move is enabled by using colloquial terms like 'imagination' without considering whether it corresponds to the human faculty," leading to inflated claims. Clear, precise language helps "dispel speculative, scientifically unsupported portrayals of [AI] systems, and support more factual descriptions of them" [39], and thus clear scientific communication.

2.4 Methodological rigor

What methods are being used? Are these methods and their use clearly and explicitly described? Are these methods appropriate, well-justified, and appropriately applied?

Rigor is often "conceptualized as the appropriate execution of [methods]" [107], with discussions about rigor in AI centering around methodological considerations, from data collection and analysis, to model training and tuning, to experimental practices [e.g., 65, 130, 132], and notions of *theoretical rigor* (of algorithmic and mathematical analysis) and *empirical rigor* (of statistical and experimental approaches). Theoretical rigor typically seeks precise problem formulation using well-defined mathematical notation, accompanied by results (e.g., propositions, theorems, lemmas) with correct proofs—e.g., a clear and correct sequence of mathematically derived steps that support the stated result. Empirical rigor, in turn, seeks comparison of a proposed algorithm with a sufficient number of alternative—often competing—approaches, ablation studies, and some form of statistical analysis (e.g., power analysis, significance tests, or simply error bars). Renewed calls for methodological rigor have often been motivated by reproducibility concerns [52, 76, 94], with checklists and documentation practices proposed as a way to support reproducibility and replicability by standardizing methodological choices, recording them, and making them explicit.

Mechanisms to promote *methodological rigor*: Methodological choices are shaped by considerations about how to operationalize what we know—e.g., the background knowledge—to substantiate existing knowledge or produce new knowledge or artifacts. As methodological concerns have been central to discussions about rigor in AI, here we only focus on foregrounding mechanisms for aspects of methodological rigor we believe deserve added attention, including *construct validity* [28, 73, 145] and the need for *methodological standards*, particularly for high-risk domains [151]. For more comprehensive discussions of methodological rigor we direct the reader to [e.g., 65, 76, 89, 130, 132, 145].

Construct validity: Many problems in AI research, such as assessments of systems and phenomena, are concerned with measurement [73, 103, 145]. Even when there is conceptual clarity, ensuring construct validity—i.e., that measurement instruments (e.g., benchmark metrics) appropriately capture the construct of interest (e.g., reasoning, understanding, values)—is foundational to meaningful measurement and thus methodological rigor. A growing body of work has proposed frameworks and best practices for assessing the validity of measurements [e.g., 73, 91, 109, 143] and illustrated that existing measurement instruments exhibit a range of concerns that threaten their ability to measure what they purport to measure [28, 62, 105]. For example, Northcutt et al. [105] show widespread label errors in benchmark test datasets which can "destabilize ML benchmarks," thereby "lead[ing] practitioners to incorrect conclusions about which models actually perform best in the real world."

Compliance with methodological standards: Establishing methodological standards often involves extensive, community-wide debates about what methods are appropriate and when. Petzschner [112] notes how the failure of ML models intended for medical settings "to generalize to data from new, unseen clinical trials [...] highlight[s] the necessity for more stringent methodological standards," particularly for high-risk settings [151]. This has precipitated calls for developing standards in health datasets in AI applications [11]. The standardization of information retrieval evaluation practices via NIST's Text Retrieval Conference (TREC) was fundamental in revitalizing the research community and impacted the development of web search engines [121]. However, even though established standards can help a research community promote more rigorous debates about methodological choices, they may not by themselves ensure that those choices are explicitly reported or reflected on; Geiger et al. [56] hypothesize "that in fields with widely-established and shared methodological standards, researchers could have far higher rates of adherence to methodological best practices [...] but have lower rates of reporting that they actually followed those practices." By the same token, not making a choice of methods and presenting a kitchen sink (of metrics, methods) [156] undermines critical engagement with why the methods are appropriate. Compliance with standards should include explicit reflections on methodological choices and their application, including aspects related to constraints that researchers and practitioners had to navigate, such as access to participants, computing, or other resources [110].

2.5 Reporting rigor

What research findings are being reported? Are these research findings clearly communicated? Is the presentation of research findings appropriate and well-justified?

The understanding of research findings depends on what is communicated about these findings and how. Reporting rigor is concerned with making sure research findings are clearly and appropriately communicated and justified. For instance, assume a researcher wishes to compare the performance of different recommender systems. Even when reporting only aggregated results, multiple options are possible, including averaging across ratings (treating each rating as equally important regardless of which user provided it or what item it was provided for), users (treating each user equally by first computing performance at user level), or items (treating each item equally by first computing performance at item level).[4] Depending on the data distribution (e.g., number of ratings per item/user), these differences may lead someone to draw different or even contradictory conclusions about which model performs best [e.g., 108]. Some ratings may also be more difficult to predict than others [e.g., 120], and any aggregation can obfuscate where exactly the model fails or succeeds [33, 89, 120]. Further, even such simple aggregations can introduce tacit assumptions about what should be optimized for e.g., to ensure a good predictive performance across all users versus all items [33]. Making these choices explicit can facilitate others' understanding of what specifically is being reported and why.

Such choices of what findings to report, and how—much as with choices of research questions, theoretical constructs, or methods—are shaped by our beliefs, values, and preferences, as well as disciplinary norms and incentives. Negative results are, for instance, less likely to be reported or published [136, 152], potentially "lead[ing] others to develop overly optimistic ideas about scientific progress on a particular topic" [12], and what is reported may be cherry-picked, such as "uncommonly compelling examples to illustrate the output of generative models" [12]. And as the example above also illustrates, even for the same findings, how the findings are reported or what about the findings is reported matters—such as choosing to report inferential uncertainty ("how precisely we have estimated the average for each group," when interested in estimating aggregate outcomes) versus outcome variability ("how much individual outcomes vary around averages for each group") [155]. Communicating the former risks leading people to "overestimate the size and importance of scientific findings" [155].

Poor choices of how findings are reported can thus also undermine interpretative rigor (§2.6)—what inferences or claims are made—as any statement of findings inherently embeds some interpretations while possibly hindering others. While this makes it difficult to fully separate concerns about reporting rigor from those about interpretative rigor, we foreground them separately to help draw attention to the different choices that often can be made about which findings to report and how.

Mechanisms to promote *reporting rigor*: Pre-study practices around disclosing how a study would be run and what about it would be reported like *pre-registration* [e.g., 66, 106, 140], and around reporting more granular, *disaggregated results* [e.g., 16, 33, 102] can promote reporting rigor.

Pre-registration and reporting practices: Reflecting on what to report about a study before conducting it can mitigate concerns about making such choices post-factum by hypothesizing after the results are known or reporting only from positive results [78, 106, 122]. Pre-registration is considered the "practice of specifying what you are going to do, and what you expect to find in your study, before carrying out the study" [140]. Pre-registration is seen as a mechanism for promoting more reliable research findings by differentiating between *confirmatory*—where hypotheses are tested and pre-registration is required—and *exploratory research*—where hypotheses are generated and pre-registration may not be required [106, 135]. While pre-registration often relies on narrow notions of what constitutes exploratory research and might inappropriately situate confirmatory research as more reliable [51], as a mechanism it can be used not only to encourage reflection on what measures of success will be used and reported on but also on how the study is designed [66], including in exploratory settings.

Disaggregated evaluations: As illustrated by the earlier example, aggregate measurements and metrics can obscure rare phenomena and information about where systems or models tend to fail or succeed, and can mask important effects [16, 33, 120, 130]. To mitigate such concerns, many have called for reporting disaggregated evaluations [33, 70, 89, 102], which "have proven to be remarkably effective at uncovering the ways in which AI systems perform differently for different groups of people" [16] and deemed a "critical piece of full empirical analysis" [130]. While conceptually simple, "their results, conclusions, and impacts depend on a variety of choices" [16], and they require careful consideration and justification—including by engaging with domain experts [138]—of how different choices of why, when, what, and how to conduct and report on such evaluations shape what inferences can be drawn and their impact. This, however, can in turn help make such choices explicit and encourage debates about their appropriateness.

[4] While this may often be left implicit due to a failure to explicitly systematize what *performance* entails (§2.3), each option likely reflects a somewhat different underlying conceptualization of performance—e.g., average user, item, rating-level performance—with "[d]ifferent 'findings' created by different conceptualizations" [57].

2.6 Interpretative rigor

What inferences are being drawn from the research findings? Are these inferences clearly and explicitly communicated? Are these inferences appropriate and well-justified?

Assume an AI system achieves high accuracy on a benchmark designed to assess mathematical reasoning [e.g., 58, 123]. As Salaudeen et al. [123] note, based on the system's performance on the benchmark, both of these two different alternative claims could be considered: the system can "solve linear algebra questions from a textbook accurately" or the system "has reached human-expert-level mathematical reasoning." Reliably moving from performance on a benchmark to either of the two claims requires clarity about any background assumptions concerning the feasibility of an AI system reaching human-level reasoning abilities [e.g., 141] (epistemic and normative rigor, §2.1–2.2), about how both "mathematical reasoning" and "human-expert-level" are conceptualized (conceptual rigor, §2.3), about whether the benchmark actually measures mathematical reasoning ability (methodological rigor, §2.4), and about how findings were reported—e.g., do we know what the performance on linear algebra questions is? (reporting rigor, §2.5). Echoing Markham [97]'s observation, a pernicious trap is to believe that methods and evaluation practices "bestow a natural interpretive clarity and self-reflexive awareness on the researcher," and thus "scientists must acknowledge the social and interpretative character of scientific discovery" [24].

Both making and understanding knowledge claims require interpretation. There are often multiple perspectives through which empirical, experimental, or theoretical evidence could be interpreted, and these different perspectives may not only lead to different aspects being emphasized but may also lead to different or even contradictory conclusions being drawn. While *reporting rigor* (§2.5) is concerned with choices about what findings to report and how, *interpretative rigor* is concerned with the conclusions we draw from these findings, and thus with the choices we make when we move from findings to either some *descriptive* claims—e.g., the system solves a given task—or to some *prescriptive* or *normative* claims—e.g., the system should be used to replace humans. Such conclusions or claims rarely directly follow from findings, but require some interpretation and are shaped by choices and considerations related to all other facets of rigor (§2.1–2.4)—they are influenced and made in the context of background knowledge, the relationship with the theoretical constructs under use, and the methods used to produce the findings and their limitations. While critical to how any work ultimately intervenes in the world, how claims are arrived at is often overlooked in discussions about rigor in AI, with the interpretation of results—i.e., what they mean, what people should do next—being treated as *self-evident*. The criteria for interpretative rigor is also "not whether the same interpretation would be independently arrived upon by different" people, but rather that "based on the evidence provided, is a given interpretation credible" or if "given all the same source information, would the interpretation stand up to scrutiny as being a justified, empirically grounded, exposition of the phenomenon" [107].

Mechanisms to promote *interpretative rigor*: Documenting and justifying evidence informing or situating possible claims can promote interpretative rigor, like via *documenting AI artifacts* [e.g., 17, 55, 102] or by engaging with possible threats to *internal/external validity* [e.g., 87, 109, 134].

Documenting AI artifacts, their limitations, and their impacts: Transparency about any AI artifacts (e.g., datasets, models, systems) under use can facilitate others' understanding of what claims may or may not be possible by providing added context about their characteristics and intended uses. To help scaffold and promote more transparent reporting on AI artifacts, researchers have developed tools, resources, or what Boyd [30] terms "context documents" (e.g., for datasets [17, 55, 114], models [102], services [116]). Complementing these efforts, others have argued for and put forward practices for recognizing and disclosing the *limitations*—i.e., "drawbacks in the design or execution of research that may impact the resulting findings and claims" [134]—and *impacts*—i.e., actual or possible consequences from the research, development, deployment, or use—of AI artifacts [13, 29, 48, 74, 90, 96, 100, 110, 134]. Limitations, in particular, directly impact how research findings can be interpreted—a failure to recognize them can lead to unsubstantiated claims, while a failure to disclose them can further lead to misinterpretations or misuse of claims. Impacts, in turn, can further affect prescriptive and normative claims and their implications, such about how one should act given the research findings.

Internal and external validity: Interpretative rigor requires not only careful deliberation on what the claims are about, but also whether there is appropriate evidence to support the claims. Establishing whether the research findings constitute appropriate evidence requires engaging with both *internal validity*—whether there are unaddressed issues with study design or execution that may compromise the findings such as data leakage [e.g., 77] or improper baseline comparisons [e.g., 87]—and *external* *validity*—whether the findings generalize to different settings such as from one dataset [e.g., 113] or construct [e.g., 146] to another. For broader discussions, see Olteanu et al. [109] and Liao et al. [87].

Concluding Reflections

By making the case for better documentation [55, 102, 116], better evaluation practices [70, 145, 156], better development and deployment practices [14, 74, 115], and a better understanding of impacts and limitations [13, 29, 134], *responsible AI research asks for greater scientific rigor.* The AI community has too often cast responsible AI considerations as out of scope, but holds up *research rigor as a virtue*. In AI research and practice, however, rigor still remains largely understood in terms of *methodological rigor*. Nevertheless, this can have unintended consequences as, for instance, "if the methodology is considered to be the *sine qua non* of scientificity, as it usually is, then there will be enormous pressures for the structure of *all* theories to accommodate to the theoretical structure embedded in the methodology [... with each] embedded theory involv[ing] its own value hierarchy" (emphasis original) [44]. In our reconception of rigor in AI—which expands beyond merely *methodological rigor*—we reframe many calls of the responsible AI community as within scope of all AI researchers and practitioners.

We argue that rigor in AI research and practice means more than just *methodological rigor*, and in so doing we bring doing AI work *responsibly*—as it pertains to *epistemic, normative, conceptual, methodological, reporting,* and *interpretative* rigor—under the umbrella of an AI researcher and practitioner's responsibility. By calling attention to these different facets of rigor, we also hope to provide the AI community with useful language that can help researchers and practitioners raise, clarify, and examine a wider range of concerns about existing practices in AI work. Nevertheless, while a broader conception of rigor can improve research integrity and quality, rigor is not a panacea for all problems in AI research and practice. While the facets of rigor we foreground are foundational to good science and apply broadly, we grounded our discussion about these facets in critical work from the responsible AI community because that work discusses these aspects in the context of AI research and practice—which is what our call is concerned with. We also do not claim that these facets are all-inclusive, but rather that they help demonstrate how expanding our conception of rigor beyond methodological considerations can help contribute evidence that AI work is rigorous.

Alternative Views on Rigor in AI Research and Practice

AI work is already rigorous or rigorous enough: Some may view addressing methodological rigor concerns as sufficient for ensuring rigorous AI work; under this view, rigor equates to and is achieved by focusing on methodological concerns. As underscored throughout this paper, this view leaves unaddressed a range of concerns which have produced undesirable outcomes, including a reliance on pseudo-scientific assumptions [e.g., 137], treatment of social phenomena inconsistent with broader scholarly understanding [e.g., gender 46, 79], systems that are not fit-for-purpose or cause harm [e.g., 71, 115], and claims or use of language that impedes public understanding of AI [e.g., 39].

Non-methodological concerns are outside the purview of AI work: This view may arise because AI researchers and practitioners may see such concerns as outside the scope of core AI work [156]; they may also not see themselves as well-suited to addressing such concerns, either because it would be difficult to acquire the expertise needed, or because some concerns ought instead to be addressed by subject matter experts (with whom they may not have the time, desire, or resources to engage). We agree that engaging with subject matter experts can be valuable, and believe that AI work has been strengthened when it has done so [145]. Nevertheless, since all work involves *choices* about what problems are important, why they are important, and what precise objects are under investigation, and since AI work often claims real-world impact in or relevance to particular domains, we argue that it is impossible to do any AI work that does not implicate epistemic, normative, or conceptual questions—and thus researchers and practitioners must grapple with these concerns explicitly rather than implicitly.

All rigor concerns are validity concerns: Under this view, this presentation of rigor concerns simply reframes work already addressing well-understood validity threats in AI research [e.g., 73, 87, 109, 123, 145]. While some of the concerns and mechanisms we describe (e.g., construct clarity/validity, internal/external validity) appear in the literature on validity, many considerations, particularly those related to epistemic, normative, conceptual, and interpretative rigor, precede or are a prerequisite to questioning and establishing validity, and facilitate reflection about issues beyond validity. For example, a failure to interrogate one's epistemological and normative commitments may result in systems that operationalize a construct as defined, but whose definition has been contested.

Acknowledgments

We thank Michael Veale and Hanna Wallach for early conversations that have motivated this paper. We are also grateful to the members of the STAC team at Microsoft Research NYC for their feedback.

References

[1] Robert Adcock and David Collier. Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, 95(3):529–546, 2001.

[2] Daniel Adler and Randi Zlotnik Shaul. Disciplining bioethics: Towards a standard of methodological rigor in bioethics research. *Accountability in Research*, 19(3):187–207, 2012.

[3] Federico Adolfi, Laura van de Braak, and Marieke Woensdregt. From empirical problem-solving to theoretical problem-finding perspectives on the cognitive sciences. *Computational Brain & Behavior*, 7(4):572–587, 2024.

[4] William Agnew, A Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R McKee. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2024.

[5] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. Physiognomy in the age of AI. *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*, page 208, 2023.

[6] Nur Ahmed, Amit Das, Kirsten Martin, and Kawshik Banerjee. The narrow depth and breadth of corporate responsible AI research. *arXiv preprint arXiv:2405.12193*, 2024.

[7] Shazeda Ahmed, Klaudia Jaźwińska, Archana Ahlawat, Amy Winecoff, and Mona Wang. Field-building and the epistemic culture of AI safety. *First Monday*, 2024.

[8] Danial Amin, Joni Salminen, Farhan Ahmed, Sonja MH Tervola, Sankalp Sethi, and Bernard J Jansen. How is generative AI used for persona development?: A systematic review of 52 research articles. *arXiv preprint arXiv:2504.04927*, 2025.

[9] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.

[10] Mel Andrews, Andrew Smart, and Abeba Birhane. The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns*, 5(9), 2024.

[11] Anmol Arora, Joseph E Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa D Mccradden, Lauren Oakden-Rayner, Stephen R Pfohl, Marzyeh Ghassemi, Francis Mckay, et al. The value of standards for health datasets in artificial intelligence-based applications. *Nature Medicine*, 29(11):2929–2938, 2023.

[12] Carolyn Ashurst, Solon Barocas, Rosie Campbell, and Deborah Raji. Disentangling the components of ethical research in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2057–2068, 2022.

[13] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. AI ethics statements: analysis and lessons learnt from NeurIPS broader impact statements. In *Proceedings of the 2022 ACM conference on Fairness, Accountability, and Transparency*, pages 2047–2056, 2022.

[14] Agathe Balayn, Natasa Rikalo, Jie Yang, and Alessandro Bozzon. Faulty or ready? handling failures in deep-learning computer vision models until deployment: A study of practices, challenges, and needs. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[15] Solon Barocas, Asia J Biega, Benjamin Fish, Jędrzej Niklas, and Luke Stark. When not to design, build, or deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 695–695, 2020.

[16] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 368–378. Association for Computing Machinery, 2021. URL https://doi.org/10.1145/3461702.3462610.

[17] Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[18] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[19] Danielle Berkovic. Researcher positionality. *Qualitative Research–a practical guide for health and social care researchers and practitioners*, 2023.

[20] Frank J Bernieri. Rigor is rigor: But rigor is not necessarily science. *Theory & Psychology*, 1 (3):369–373, 1991.

[21] Abeba Birhane. The impossibility of automating ambiguity. *Artificial Life*, 27(1):44–61, 2021.

[22] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. In *Proceedings of the 2022 ACM conference on Fairness, Accountability, and Transparency*, pages 173–184, 2022.

[23] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. The forgotten margins of AI ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 948–958, 2022.

[24] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280, 2023.

[25] Borhane Blili-Hamelin, Leif Hancox-Li, and Andrew Smart. Unsocial intelligence: An investigation of the assumptions of AGI discourse. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 141–155, 2024.

[26] Borhane Blili-Hamelin, Christopher Graziul, Leif Hancox-Li, Hananel Hazan, El-Mahdi El-Mhamdi, Avijit Ghosh, Katherine A Heller, Jacob Metcalf, Fabricio Murai, Eryk Salvaggio, Andrew J Smart, Todd Snider, Mariame Tighanimine, Talia Ringer, Margaret Mitchell, and Shiri Dori-Hacohen. Position: Stop treating 'AGI' as the north-star goal of AI research. In *Proceedings of the Forty-second International Conference on Machine Learning, Position Paper Track*, 2025. URL https://openreview.net/forum?id=1RlrtH6ydW.

[27] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics, July 2020. URL https://doi.org/10.18653/v1/2020.acl-main.485.

[28] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, 2021.

[29] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. Overcoming failures of imagination in AI infused system development and deployment. In *Navigating the Broader Impacts of AI Research Workshop at NeurIPS 2020*, December 2020. URL https://www.microsoft.com/en-us/research/publication/overcoming-failures-of-imagination-in-ai-infused-system-development-and-deployment/.

[30] Karen L Boyd. Datasheets for datasets help ML engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27, 2021.

[31] E Brister. Epistemological obstacles to interdisciplinary research. *Integration and Implementation Insights*, 2017. URL https://i2insights.org/2017/10/31/epistemology-andinterdisciplinarity.

[32] Evelyn Brister. Disciplinary capture and epistemological obstacles to interdisciplinary research: Lessons from central african conservation disputes. *Studies in history and philosophy of science part C: studies in history and philosophy of biological and biomedical sciences*, 56:82–91, 2016.

[33] Ryan Burnell, Wout Schellaert, John Burden, Tomer D Ullman, Fernando Martinez-Plumed, Joshua B Tenenbaum, Danaja Rutar, Lucy G Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, 2023.

[34] Andrew Burton-Jones, Peter Gray, and Ann Majchrzak. Editor's comments: Producing significant research. *MIS Quarterly*, 47(1):i–xv, 2022. URL https://www.stouras.com/files/MISQ-Editor-Comments-Producing_Significant_Research.pdf. Editorial.

[35] Maarten Buyl, Hadi Khalaf, Claudio Mayrink Verdun, Lucas Monteiro Paes, Caio Cesar Vieira Machado, and Flavio du Pin Calmon. AI alignment at your discretion. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 3046–3074. Association for Computing Machinery, 2025. URL https://doi.org/10.1145/3715275.3732194.

[36] Stacy M Carter and Miles Little. Justifying knowledge, justifying method, taking action: Epistemologies, methodologies, and methods in qualitative research. *Qualitative health research*, 17(10):1316–1328, 2007.

[37] Karin Knorr Cetina. Culture in global knowledge societies: Knowledge cultures and epistemic cultures. *Interdisciplinary science reviews*, 32(4):361–375, 2007.

[38] Allison Chen, Sunnie SY Kim, Amaya Dharmasiri, Olga Russakovsky, and Judith E Fan. Portraying large language models as machines, tools, or companions affects what mental capacities humans attribute to them. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2025.

[39] Myra Cheng, Alicia DeVrio, Lisa Egede, Su Lin Blodgett, and Alexandra Olteanu. "I am the one and only, your cyber BFF": Understanding the impact of genAI requires understanding the impact of anthropomorphic AI. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL https://openreview.net/forum?id=FL7sRPVqni.

[40] Cleveland Clinic. Hallucinations. clevelandclinic.org, last accessed May-2025. URL https://my.clevelandclinic.org/health/symptoms/23350-hallucinations.

[41] A Feder Cooper, Yucheng Lu, Jessica Forde, and Christopher M De Sa. Hyperparameter optimization is deceiving us, and how to stop it. *Advances in Neural Information Processing Systems*, 34:3081–3095, 2021.

[42] A. Feder Cooper, Christopher A. Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu, Alexandra Chouldechova, Jamie Hayes, Yangsibo Huang, Niloofar Mireshghallah, Ilia Shumailov, Eleni Triantafillou, Peter Kairouz, Nicole Mitchell, Percy Liang, Daniel E. Ho, Yejin Choi, Sanmi Koyejo, Fernando Delgado, James Grimmelmann, Vitaly Shmatikov, Christopher De Sa, Solon Barocas, Amy Cyphert, Mark Lemley, danah boyd, Jennifer Wortman Vaughan, Miles Brundage, David Bau, Seth Neel, Abigail Z. Jacobs, Andreas Terzis, Hanna Wallach, Nicolas Papernot, and Katherine Lee. Machine unlearning doesn't do what you think: Lessons for generative AI policy, research, and practice. *arXiv:2412.06966*, 2024.

[43] Chris LS Coryn. The 'holy trinity' of methodological rigor: A skeptical view. *Journal of MultiDisciplinary Evaluation*, 4(7):26–31, 2007.

[44] Kurt Danziger. The methodological imperative in psychology. *Philosophy of the social sciences*, 15(1):1–13, 1985.

[45] Deirdre Davies and Jenny Dodd. Qualitative research and the question of rigor. *Qualitative health research*, 12(2):279–289, 2002.

[46] Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of “gender” in NLP bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, page 2083–2102. Association for Computing Machinery, 2022. doi: 10.1145/3531146.3534627. URL https://doi.org/10.1145/3531146.3534627.

[47] Fernando Diaz and Michael Madaio. Scaling laws do not scale. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 341–357, 2024.

[48] Kimberly Do, Rock Yuren Pang, Jiachen Jiang, and Katharina Reinecke. “That’s important, but...”: How computer science researchers anticipate unintended consequences of their research innovations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[49] Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M Roy. In search of robust measures of generalization. *Advances in Neural Information Processing Systems*, 33:11723–11733, 2020.

[50] Miriam Fahimi, Mayra Russo, Kristen M Scott, Maria-Esther Vidal, Bettina Berendt, and Katharina Kinder-Kurlanda. Articulation work and tinkering for fairness in machine learning. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–23, 2024.

[51] Uljana Feest and Berna Devezer. Toward a more accurate notion of exploratory research (and why it matters). *Theory- and model-building in psychology*, 2025. URL https://www.researchgate.net/publication/388065924_Toward_a_More_Accurate_Notion_of_Exploratory_Research_And_Why_it_Matters. Forthcoming.

[52] Jessica Zosa Forde and Michela Paganini. The scientific method in the science of machine learning. *arXiv preprint arXiv:1904.10922*, 2019.

[53] Roberto Forero, Shizar Nahidi, Josephine De Costa, Mohammed Mohsin, Gerry Fitzgerald, Nick Gibson, Sally McCarthy, and Patrick Aboagye-Sarfo. Application of four-dimension criteria to assess rigour of qualitative research in emergency medicine. *BMC health services research*, 18:1–11, 2018.

[54] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

[55] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[56] R Stuart Geiger, Dominique Cope, Jamie Ip, Marsha Lotosh, Aayush Shah, Jenny Weng, and Rebekah Tang. “garbage in, garbage out” revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*, 2(3):795–827, 2021.

[57] Pulivelil M George. Conceptualization: the central problem of science. *Organon*, 9:23–33, 1973.

[58] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, et al. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

[59] Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the Internation Conference on Machine Learning: the debates workshop*, 2018.

[60] Ben Green and Salomé Viljoen. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 conference on Fairness, Accountability, and Transparency*, pages 19–31, 2020.

[61] Olivia Guest. What makes a good theory, and how do we make a theory good? *Computational Brain & Behavior*, pages 1–15, 2024.

[62] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics, June 2018. URL https://doi.org/10.18653/v1/N18-2017.

[63] Dylan Hadfield-Menell. The need for scientific rigor in AI safety research. Medium.com, 2024. URL https://medium.com/@dhm.csail/the-need-for-scientific-rigor-in-ai-safety-research-3e3c71f29968.

[64] Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, and Hanna Wallach. Understanding and meeting practitioner needs when measuring representational harms caused by LLM-based systems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18423–18440. Association for Computational Linguistics, July 2025. URL https://doi.org/10.18653/v1/2025.findings-acl.947.

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

[66] Jake M Hofman, Angelos Chatzimparmpas, Amit Sharma, Duncan J Watts, and Jessica Hullman. Pre-registration for predictive modeling. *arXiv preprint arXiv:2311.18807*, 2023.

[67] House Science, Space, & Technology Committee. Science committee leaders stress importance of diligence in NIST AI safety research funding. https://science.house.gov/2023/12/science-committee-leaders-stress-importance-of-diligence-in-nist-ai-safety-research-funding, 2023.

[68] Mark B Houston. Four facets of rigor. *Journal of the Academy of Marketing Science*, 47: 570–573, 2019.

[69] Nick Howe. ‘Stick to the science’: When science gets political. *Nature*, 2020.

[70] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM conference on Fairness, Accountability, and Transparency*, pages 1859–1876, 2022.

[71] Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. Not my voice! a taxonomy of ethical and safety harms of speech generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 359–376, 2024.

[72] Dragos Iliescu and Samuel Greiff. On consequential validity. *European Journal of Psychological Assessment*, 2021.

[73] Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency*, pages 375–385, 2021.

[74] Seyyed Ahmad Javadi, Chris Norval, Richard Cloete, and Jatinder Singh. Monitoring AI services for misuse. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 597–607, 2021.

[75] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[76] Sayash Kapoor, Emily M Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A Bail, Odd Erik Gundersen, Jake M Hofman, Jessica Hullman, Michael A Lones, Momin M Malik, et al. Reforms: Consensus-based recommendations for machine-learning-based science. *Science Advances*, 10(18):eadk3452, 2024.

[77] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21, 2012.

[78] Norbert L Kerr. Harking: Hypothesizing after the results are known. *Personality and social psychology review*, 2(3):196–217, 1998.

[79] Os Keyes. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human Computer Interaction*, 2(CSCW), November 2018. URL https://doi.org/10.1145/3274357.

[80] Os Keyes, Jevan Hutson, and Meredith Durbin. A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pages 1–11, 2019.

[81] Halil Kilicoglu. Biomedical text mining for research rigor and integrity: tasks, challenges, directions. *Briefings in bioinformatics*, 19(6):1400–1414, 2018.

[82] Joel Klinger, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. A narrowing of AI research? *arXiv preprint arXiv:2009.10385*, 2020.

[83] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[84] Christopher Klaus Lazik, Christopher Katins, Charlotte Kauter, Jonas Jakob, Caroline Jay, Lars Grunske, and Thomas Kosch. The impostor is among us: Can large language models capture the complexity of human personas? In *Proceedings of the Mensch Und Computer 2025*, MuC '25, page 434–451. Association for Computing Machinery, 2025. URL https://doi.org/10.1145/3743049.3743057.

[85] Calvin Liang. Reflexivity, positionality, and disclosure in HCI. Medium.com, 2021. URL https://medium.com/@caliang/reflexivity-positionality-and-disclosure-in-hci-3d95007e9916.

[86] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

[87] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[88] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[89] Zachary C Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77, 2019.

[90] David Liu, Priyanka Nanayakkara, Sarah Ariyan Sakha, Grace Abuhamad, Su Lin Blodgett, Nicholas Diakopoulos, Jessica R Hullman, and Tina Eliassi-Rad. Examining responsibility and deliberation in AI impact statements and ethics reviews. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 424–435, 2022.

[91] Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. ECBD: Evidence-centered benchmark design for NLP. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365. Association for Computational Linguistics, August 2024. URL https://doi.org/10.18653/v1/2024.acl-long.861.

[92] Christina Lu, Jackie Kay, and Kevin McKee. Subverting machines, fluctuating identities: Re-learning human categorization. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1005–1015, 2022.

[93] Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna Wallach, and Alexandra Olteanu. "one-size-fits-all"? examining expectations around what constitute "fair" or "good" NLG system behaviors. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1054–1089, 2024.

[94] Ian Magnusson, Noah A Smith, and Jesse Dodge. Reproducibility in NLP: What have we learned from the checklist? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12789–12811, 2023.

[95] Negar Maleki, Balaji Padmanabhan, and Kaushik Dutta. AI hallucinations: A misnomer worth clarifying. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 133–138, 2024. URL https://doi.org/10.1109/CAI59869.2024.00033.

[96] Momin M Malik. A hierarchy of limitations in machine learning. *arXiv preprint arXiv:2002.05193*, 2020.

[97] Annette Markham. Response to Nancy Baym. *Internet Inquiry: Conversations about Method. Annette Markham and Nancy Baym, eds*, pages 190–197, 2009.

[98] Lisa Messeri and MJ Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.

[99] Samuel Messick. Validity. *ETS research report series*, 1987(2):i–208, 1987.

[100] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on Fairness, Accountability, and Transparency*, pages 735–746, 2021.

[101] Margaret Mitchell. Oversight of AI: Insiders' perspectives. Testimony before the U.S. Senate Subcommittee on Privacy, Technology, and the Law, September 2024. Available at https://www.judiciary.senate.gov/imo/media/doc/2024-09-17_pm_-_testimony_-_mitchell.pdf.

[102] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the 2019 conference on Fairness, Accountability, and Transparency*, pages 220–229, 2019.

[103] Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. Measuring data, 2023. URL https://arxiv.org/abs/2212.05129.

[104] National Academies of Sciences and Policy and Global Affairs, Board on Research Data, Information, Division on Engineering, Physical Sciences, Committee on Applied, Theoretical Statistics, Board on Mathematical Sciences, et al. *Reproducibility and replicability in science*. National Academies Press, 2019.

[105] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=XccDXrDNLek.

[106] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

[107] Branda Nowell and Kate Albrecht. A reviewer's guide to qualitative rigor. *Journal of public administration research and theory*, 29(2):348–363, 2019.

[108] Alexandra Olteanu, Anne-Marie Kermarrec, and Karl Aberer. Comparing the predictive capability of social and interest affinity for recommendations. In *15th International Conference on Web Information System Engineering (WISE 2014)*, pages 276–292, 2014.

[109] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13, 2019.

[110] Alexandra Olteanu, Michael Ekstrand, Carlos Castillo, and Jina Suh. Responsible AI research needs impact statements too. *arXiv preprint arXiv:2311.11776*, 2023.

[111] Alexandra Olteanu, Solon Barocas, Su Lin Blodgett, Lisa Egede, Alicia DeVrio, and Myra Cheng. AI automatons: AI systems intended to imitate humans. *arXiv preprint arXiv:2503.02250*, 2025.

[112] Frederike H Petzschner. Practical challenges for precision medicine. *Science*, 383(6679): 149–150, 2024.

[113] Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. Challenges to evaluating the generalization of coreference resolution models: A measurement modeling perspective. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15380–15395, 2024.

[114] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.

[115] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972, 2022.

[116] K Natesan Ramamurthy et al. Factsheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1, 2019.

[117] Ben Recht. The war of symbolic aggression, 2023. URL https://www.argmin.net/p/the-war-of-symbolic-aggression.

[118] Rainer Rehak. The language labyrinth: Constructive critique on the terminology used in the AI discourse. *AI for Everyone*, pages 87–102, 2021.

[119] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3): 199–217, 2021.

[120] Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, 2021.

[121] Brent R. Rowe, Dallas W. Wood, Albert N. Link, and Diglio A. Simoni. Economic impact assessment of NIST's text retrieval conference (trec) program. Technical Report Project Number 0211875, RTI International, Research Triangle Park, NC, July 2010.

[122] Mark Rubin. When does HARKing hurt? identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, 21(4):308–320, 2017.

[123] Olawale Elijah Salaudeen, Anka Reuel, Ahmed M Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Benjamin W. Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL https://openreview.net/forum?id=2Bw6uC49QF.

[124] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "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*, CHI '21. Association for Computing Machinery, 2021. URL https://doi.org/10.1145/3411764.3445518.

[125] Naomi Saphra and Sarah Wiegreffe. Mechanistic? *arXiv preprint arXiv:2410.09087*, 2024.

[126] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman. AI mismatches: Identifying potential algorithmic harms before AI development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2025.

[127] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36:55565–55581, 2023.

[128] Rylan Schaeffer, Joshua Kazdan, Alvan Caleb Arulandu, and Sanmi Koyejo. Position: Model collapse does not mean what you think. *arXiv:2503.03150*, 2025.

[129] Mark Schaller. The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology*, 66:107–115, 2016.

[130] D. Sculley, Jasper Snoek, Alexander B. Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. In *Proceedings of ICLR (Workshop)*, 2018. URL https://openreview.net/forum?id=rJWF0Fywf.

[131] D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Michael Young, Jean-François Crespo, Dan Dennison, Emily Fox, and H. Larochelle. Survey of scientific rigor studied in machine learning. openreview.net, 2023. URL https://api.semanticscholar.org/CorpusID:259300529.

[132] D. Sculley, William Cukierski, Phil Culliton, Sohier Dane, Maggie M Demkin, Ryan Holbrook, Addison Howard, Paul T Mooney, Walter Reade, Meg Risdal, and Nate Keating. Position: AI competitions provide the gold standard for empirical rigor in genAI evaluation. In *Proceedings of the Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=Rxd2TpV6Eg.

[133] Mona Sloane, Emanuel Moss, and Rumman Chowdhury. A silicon valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, 3(2), 2022.

[134] Jessie J Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. Real ML: Recognizing, exploring, and articulating limitations of machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 587–597, 2022.

[135] Anders Søgaard, Daniel Hershcovich, and Miryam de Lhoneux. A two-sided discussion of preregistration of NLP research. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 83–93. Association for Computational Linguistics, 2023.

[136] Fujian Song, Sheetal Parekh, Lee Hooper, Yoon K Loke, Jon Ryder, Alex J Sutton, Caroline Hing, Chun Shing Kwok, Chun Pang, and Ian Harvey. Dissemination and publication of research findings: an updated review of related biases. *Health Technology Assessment*, 14(8): 1–193, 2010.

[137] Luke Stark and Jevan Hutson. Physiognomic artificial intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal*, 32:922, 2021.

[138] Laura M Stevens, Bobak J Mortazavi, Rahul C Deo, Lesley Curtis, and David P Kao. Recommendations for reporting machine learning analyses in clinical research. *Circulation: Cardiovascular Quality and Outcomes*, 13(10), 2020.

[139] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

[140] Emiel Van Miltenburg, Chris van der Lee, and Emiel Krahmer. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, 2021.

[141] Iris Van Rooij, Olivia Guest, Federico Adolfi, Ronald de Haan, Antonina Kolokolova, and Patricia Rich. Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, pages 1–21, 2024.

[142] Gael Varoquaux, Sasha Luccioni, and Meredith Whittaker. Hype, sustainability, and the price of the bigger-is-better paradigm in AI. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 61–75. Association for Computing Machinery, 2025. URL https://doi.org/10.1145/3715275.3732006.

[143] Claudia Wagner, Markus Strohmaier, Alexandra Olteanu, Emre Kıcıman, Noshir Contractor, and Tina Eliassi-Rad. Measuring algorithmically infused societies. *Nature*, 595(7866):197–204, 2021.

[144] Kiri L. Wagstaff. Machine learning that matters. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1851–1856. Omnipress, 2012.

[145] Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the Forty-second International Conference on Machine Learning, Position Paper Track*, 2025. URL https://openreview.net/forum?id=1ZC4RNjqzU.

[146] Angelina Wang. Identities are not interchangeable: The problem of overgeneralization in fair machine learning. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 485–497. Association for Computing Machinery, 2025. URL https://doi.org/10.1145/3715275.3732033.

[147] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Journal of Responsible Computing*, 1(1), March 2024. URL https://doi.org/10.1145/3636509.

[148] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 2025.

[149] Hilde Weerts, Raphaële Xenidis, Fabien Tarissan, Henrik Palmer Olsen, and Mykola Pechenizkiy. The neutrality fallacy: When algorithmic fairness interventions are (not) positive action. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2060–2070, 2024.

[150] Richard E West and Peter J Rich. Rigor, impact and prestige: A proposed framework for evaluating scholarly publications. *Innovative Higher Education*, 37:359–371, 2012.

[151] Christoph Wilhelm, Anke Steckelberg, and Felix G Rebitschek. Benefits and harms associated with the use of AI-related algorithmic decision-making systems by healthcare professionals: a systematic review. *The Lancet Regional Health–Europe*, 48, 2025.

[152] Torsten Wilholt. Bias and values in scientific research. *Studies in History and Philosophy of Science Part A*, 40(1):92–101, 2009.

[153] Eric Winsberg and Ali Mirza. Success and scientific realism: considerations from the philosophy of simulation. In *The Routledge handbook of scientific realism*, pages 250–260. Routledge, 2017.

[154] Annie Zaenen. Last words: Mark-up barking up the wrong tree. *Computational Linguistics*, 32(4):577–580, 2006.

[155] Sam Zhang, Patrick R Heck, Michelle N Meyer, Christopher F Chabris, Daniel G Goldstein, and Jake M Hofman. An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences*, 120(33):e2302491120, 2023.

[156] Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, 2022.