Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge

Hanna Wallach [1] **Meera Desai** [2] **A. Feder Cooper** [1] **Angelina Wang** [3] **Chad Atalla** [1] **Solon Barocas** [1]
Su Lin Blodgett [1] **Alexandra Chouldechova** [1] **Emily Corvi** [1] **P. Alex Dow** [1] **Jean Garcia-Gathright** [1]
Alexandra Olteanu [1] **Nicholas Pangakis** [1] **Stefanie Reed** [1] **Emily Sheng** [1] **Dan Vann** [1]
Jennifer Wortman Vaughan [1] **Matthew Vogel** [1] **Hannah Washington** [1] **Abigail Z. Jacobs** [2]

Abstract

The measurement tasks involved in evaluating generative AI (GenAI) systems are especially difficult, leading to what has been described as "a tangle of sloppy tests [and] apples-to-oranges comparisons" (Roose, 2024). In this position paper, we argue that the ML community would benefit from learning from and drawing on the social sciences when developing and using measurement instruments for evaluating GenAI systems. Specifically, our position is that evaluating GenAI systems is a social science measurement challenge. We present a four-level framework, grounded in measurement theory from the social sciences, for measuring concepts related to the capabilities, behaviors, and impacts of GenAI. This framework has two important implications for designing and evaluating evaluations: First, it can broaden the expertise involved in evaluating GenAI systems by enabling stakeholders with different perspectives to participate in conceptual debates. Second, it brings rigor to both conceptual and operational debates by offering a set of lenses for interrogating the validity of measurement instruments and their resulting measurements.

1. Evaluating GenAI Systems

The evaluation of ML systems[1] is critical for making decisions about whether they should be used for particular purposes, whether they should be deployed in particular contexts, or even whether they should be redesigned. Evaluating an ML system necessarily requires information about its capabilities (like its mathematical reasoning skills), its behaviors (like regurgitating pieces of its training data), and its impacts (like causing its users to feel harmed). Often, this information takes the form of *measurements* on nominal, ordinal, interval, and ratio scales. Each measurement reflects the amount of some *concept* of interest. Such measurements are obtained via the *process of measurement*, which can involve both qualitative and quantitative approaches.

Across academia, industry, and government (e.g., National Institute for Standards and Technology, 2024; Cooper et al., 2023; Perez et al., 2022; Weidinger et al., 2023), there is an increasing awareness that the measurement tasks involved in evaluating generative AI (GenAI) systems are more difficult than those involved in evaluating traditional ML systems because the concepts to be measured are more often abstract. Abstract concepts cannot be *directly* measured and must therefore be *indirectly* measured from other observable phenomena. In addition, their meanings may be contested (e.g., Mulligan et al., 2019; 2016) across and within use cases, cultures, and languages. Thus, although ML researchers and practitioners have proposed myriad measurement instruments for evaluating GenAI systems, it is difficult to know whether these instruments and their resulting measurements are meaningful or useful—i.e., valid.

In this position paper, we argue that moving beyond the current state will require the ML community to pay greater attention to the process of measurement. **We take the position that evaluating GenAI systems is a social science measurement challenge.** Specifically, the measurement tasks involved in evaluating GenAI systems are highly reminiscent of the measurement tasks found throughout the social sciences. Social scientists have been thoughtfully measuring abstract, often contested, concepts—ideology, democracy, media bias, framing, to name just a few—for over fifty years (e.g., Berelson, 1952; Zaller, 1992). Like these social science concepts, concepts related to the capabilities, behaviors, and impacts of GenAI systems are abstract, often contested, and deeply intertwined with people and society. As a result, the ML community would benefit from learning

[1]Microsoft Research [2]University of Michigan [3]Stanford University. Corresponding email: wallach@microsoft.com.

[1]To simplify exposition, we use the term "ML system" to refer to either 1) a single ML model or 2) one or more integrated software components, where at least one component is an ML model.

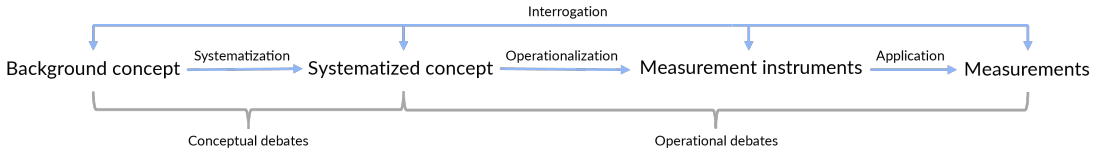

Figure 1. A variant of the framework of Adcock and Collier (2001). The background concept, the systematized concept, the measurement instruments, and the measurements are linked by the systematization, operationalization, application, and interrogation processes.

from and drawing on the social sciences when developing and using instruments for measuring concepts related to the capabilities, behaviors, and impacts of GenAI systems.

Paper roadmap: In the next section, we present a four-level framework, grounded in measurement theory from the social sciences, for measuring concepts related to the capabilities, behaviors, and impacts of GenAI systems.[2] In Section 3, we then contrast the structured approach afforded by this framework to the way measurement is typically done in ML. In Section 4, we present and address some views that provide an alternative to our position, before concluding in Section 5. Throughout the paper, we illustrate both the core ideas and our arguments using several real-world examples, including measuring the extent of demeaning text in the outputs of an LLM-based system, measuring the mathematical reasoning skills of a GenAI system, and measuring the extent to which a GenAI system regurgitates pieces of its training data.

2. A Measurement Framework for GenAI

When measuring abstract concepts, including those that have contested meanings, social scientists often turn to measurement theory, which offers a structured approach to articulating distinctions between *concepts* and their operationalizations via *measurement instruments*—i.e., the operational procedures and artifacts used to obtain measurements of those concepts, such as classifiers, annotation guidelines, scoring rules, and aggregation functions. It also offers a set of lenses for interrogating the validity of measurement instruments and their resulting measurements (e.g., Adcock and Collier, 2001; Cronbach and Meehl, 1955; Messick, 1996).

One formulation of measurement theory is the framework of Adcock and Collier (2001), a variant of which is shown in Figure 1.[3] This variant distinguishes between four levels: the *background concept* or "broad constellation of meanings and understandings associated with [the] concept;" the *systematized concept* or "specific formulation of the concept[, which] commonly involves an explicit definition;" the *measurement instruments* used to produce measurements of the concept; and the *measurements* themselves (Adcock and Collier, 2001). These levels are linked by four processes: *systematization* (Section 2.1), *operationalization* (Section 2.2), *application* (Section 2.3), and *interrogation* (Section 2.4).

Crucially, this structured approach separates conceptual debates—i.e., does our systematized concept reflect what we want it to reflect?—from operational debates—i.e., did we operationalize the systematized concept via measurement instruments that yield valid measurements? As we explain below, this separation has two important implications for designing and evaluating evaluations of GenAI systems: First, it can broaden the expertise involved in such evaluations by enabling stakeholders with different perspectives to participate in conceptual debates. Second, it brings rigor to both conceptual and operational debates by offering a set of lenses for interrogating the validity of measurement instruments and their resulting measurements.

2.1. Systematization

The systematization process is the foundation of measurement. Systematization specifies how a concept—an abstract idea—is connected to observable phenomena in the real world. Specifically, systematizing a concept means taking the broad constellation of meanings and understandings associated with that concept—the background concept—and narrowing it into an explicit definition—the systematized concept. This definition must explain how the concept either causes or is defined by observable phenomena in the real world, specifying precisely *what* will be measured and *why*.

For example, suppose we wish to measure the extent of text that demeans social groups in the outputs of an LLM-based system—i.e., a concept related to that system's behaviors. In this example, the background concept encompasses all possible social groups and all possible definitions of text that demeans social groups, making it both abstract and inclusive of a broad range of meanings and understandings. From here, we might select a set of specific social groups to consider—e.g., women, people over the age of 40, and people with disabilities. We might also select a specific defini-

[2] We note that this framework can also be used when evaluating other types of ML systems. However, it is particularly useful for the measurement tasks involved in evaluating GenAI systems.

[3] We note that we use slightly different terminology to that of Adcock and Collier; the core ideas are very similar, however.

tion like "[text] with dehumanizing or offensive associations, or [that] otherwise threaten[s] people's sense of security or dignity" (Blodgett, 2021). However, although this definition articulates some aspects of the concept as a high level, it still encompasses many meanings and understandings and must be further systematized in order to explain how the concept connects to observable phenomena in the real world.

For example, we might draw on literature from social psychology, sociolinguistics, anthropology, and other disciplines to identify observable phenomena that are either caused by or define text that demeans the specific social groups we have chosen to consider. In doing so, we might find that researchers often define the presence of such text in terms of the presence of particular linguistic patterns, such as equating a social group to an animal, advocating for the animal-like treatment of a social group, equating a social group to an inanimate object, noting qualities of a social group that are like those of an inanimate object, and equating a social group to a disease or disorder (Corvi et al., 2024). Some of these patterns might pertain to only a single social group, while others might pertain to multiple social groups.

Having explained, at a high-level, how a concept connects to observable phenomena in the real world, the last step in the systematization process is to specify precisely what will be measured. This involves defining a set of variables—typically called *indicators*[4]—that capture the most salient properties of the observable phenomena.[5] It also involves specifying how the indicators relate to the concept—i.e., how the values of the indicators collectively yield a measurement of the concept, as desired.

Continuing with the example of measuring the extent of text that demeans social groups in the outputs of an LLM-based system, we might define an indicator for each linguistic pattern—an integer-valued variable, whose value indicates the number of occurrences of that linguistic pattern in a system output. Additively combining the values of these indicators (i.e., aggregating over the corresponding linguistic patterns) yields the extent of text that demeans the specific social groups we have chosen to consider in that system output. If we are instead interested in the total extent of such text in a *set* of system outputs, the per-output measurements can be additively combined over that set.

To summarize, by defining a set of indicators—in the case of our running example, a set of indicators that reflect particular linguistic patterns involving a set of specific social groups—and specifying how the values of the indicators collectively yield a measurement of the concept of interest, the concept has been fully systematized.

We emphasize that although the systematization process connects the concept of interest to observable phenomena in the real world, this process takes place at a theoretical level. In other words, systematization stops short of specifying the operational procedures and artifacts used to obtain measurements—i.e., measurement instruments. Separating the systematization and operationalization processes can enable stakeholders with different perspectives—e.g., open-source developers, policymakers, users, members of marginalized communities, all of whom may be interested in measuring a concept for different reasons—to participate in conceptual debates and thus advocate for the inclusion of particular meanings and understandings (Abebe et al., 2020). Measuring an abstract concept necessarily means making choices about which of its meanings and understandings will be reflected in the resulting measurements and which will not. Without an explicitly systematized concept, many of these choices are accessible only indirectly via the measurement instruments, which may be hard for stakeholders other than ML researchers and practitioners to engage with. We therefore argue that separating the systematization and operationalization processes can help broaden the expertise involved in evaluating GenAI systems.

2.2. Operationalization

In contrast to the systematization process, the operationalization process takes place at an implementation level, specifying precisely *how* measurement will take place. Specifically, operationalization draws on the systematized concept to develop measurement instruments—i.e., the operational procedures and artifacts used to obtain measurements of the concept of interest. These include classifiers, annotation guidelines, scoring rules, and aggregation procedures.

To ensure that the measurements meaningfully reflect the systematized concept, the measurement instruments must align with the definitions of the indicators and the specification of how the values of the indicators collectively yield a measurement of the concept of interest. In some cases, there may be existing measurement instruments that can be repurposed, provided they can be demonstrated to be sufficiently valid for this purpose; in other cases, the measurement instruments must be developed from scratch.[6]

[4]We note that Adcock and Collier (2001) use the term "indicators" to refer to the variables that capture the most salient properties of the observable phenomena *and* the operational procedures and artifacts used to obtain measurements of those variables. Indeed, much of their discussion of "indicators" is about the latter not the former.

[5]Indicators are often derived *directly* from observed data—i.e., they are *observed variables*, but they can also be derived *indirectly* from observed data via some other process—i.e., they can be *latent variables*, whose values may be treated as if they were derived directly from observed data when measuring the concept of interest.

[6]As we explained in Footnote 5, indicators are sometimes derived *indirectly* from observed data. In such cases, developing instruments for measuring these indicators from scratch can be a non-trivial endeavor, as it involves applying the entire measurement

Continuing with the example of measuring the extent of text that demeans social groups in the outputs of an LLM- based system, we have several options for measuring the values of the indicators—i.e., counting the numbers of occurrences of the linguistic patterns. One option is to ask a set of humans to annotate each system output accordingly. If we pursue this option, we first need to decide which humans to ask: Sociolinguists? Members of the specific social groups we have chosen to consider? Crowdworkers? Sociolinguists may bring expertise in identifying linguistic patterns, while members of the social groups may provide unique experiental insights. Crowdworkers are likely the most cost-effective choice, though they may require more extensive training to ensure high-quality annotations. Regardless of our choice, we then need to write comprehensive annotation guidelines, tailored to those annotators. At a minimum, these guidelines should describe the concept of interest, including the specific set of social groups we have chosen to consider and define each linguistic pattern, providing both examples and counterexamples, as well as explanations of how each pattern might appear in different contexts. Having written the guidelines, we then need to specify a workflow, train the annotators, and conduct pilot studies to uncover any ambiguities, disagreements, or inconsistencies.

Finally, we must develop a procedure for aggregating the values of the indicators. In the case of our running example, where the numbers of occurrences of the linguistic patterns simply need to be additively combined, this task is straightforward. However, when measuring other concepts, it can be more complex. Depending on our measurement goals, we might also need to develop a procedure for aggregating the resulting per-output measurements over a set of system outputs. This task is generally straigthforward.

To summarize, by selecting a set of human annotators; developing operational procedures and artifacts to enable those annotators to annotate each system output with the numbers of occurrences of the linguistic patterns; and developing procedures for aggregating the resulting annotations, we have fully operationalized the systematized concept. The resulting measurement instruments can now be used to obtain measurements of our concept of interest during the application process. However, because the operationalization process involves many decisions, both large and small, it is critically important to interrogate the validity of the measurement instruments and their resulting measurements, as described in Section 2.4, before using those measurements.

Before moving on, we again emphasize that the operationalization process takes place at an implementation level. It builds on the systematization process, which takes place at a theoretical level, by connecting the systematized concept to operational procedures and artifacts. The two processes are complementary: together, they ensure that the process of measurement is both theoretically and empirically grounded. We also note that although we have presented the systematization and operationalization processes as occurring in a linear fashion, iteration is often required in practice, with implementation considerations (e.g., feasibility, cost) driving iterative refinements to the systematized concept.

framework to the constituent concepts reflected by the indicators.

2.3. Application

The application process involves using the measurement instruments developed during the operationalization process to obtain measurements of the concept of interest. That said, in order to do this, we need an observed dataset. For example, in the case of measuring the extent of text that demeans social groups in the outputs of an LLM-based system, we need a dataset of system outputs. Obtaining such a dataset requires us to first specify the population that defines the domain of our concept of interest. Having done this, we then need to specify a sampling design that determines how observed data will be selected from the population. We can then use the sampling design to obtain an observed dataset.

Although the population and sampling design play a crucial role in determining what the resulting measurements mean—including whether they generalize beyond the observed dataset and, if so, to what population—they traditionally sit outside of the four-level framework in Figure 1. This is because many social science measurement tasks involve measuring some concept of interest for a pre-specified observed dataset. When evaluating GenAI systems, this is less likely to be the case. That said, we omit a detailed discussion here in the main text and instead refer the reader to Appendix A.

2.4. Interrogation

Before using the measurements obtained during the application process, they, and the instruments used to obtain them, must be validated. However, when measuring abstract concepts, there are no directly observable, universally agreed-upon labels or scores against which to evaluate the measurements, making validation especially difficult.[7]

Measurement theory therefore offers a set of lenses for interrogating the validity of measurement instruments and their resulting measurements: *face validity*, *content validity*, *convergent validity*, *discriminant validity*, *predictive validity*, *hypothesis validity*, and *consequential validity* (e.g., Jacobs and Wallach, 2021). Each lens constitutes a different source of evidence about validity. These lenses can, and should, be used to inform both conceptual and operational debates.

[7]The availability of commonly used labels or scores does not mean that those labels or scores were directly observable or sufficiently validated. Indeed, using the framework described in this section can help reveal issues with commonly used labels or scores.

This can be especially helpful when measuring concepts related to the capabilities, behaviors, and impacts of GenAI systems as conceptual debates about these concepts tend to be particularly underexplored by the ML community.

We emphasize that it is not possible to interrogate validity without taking context—including the reasons for measuring the concept and the population that defines its domain—into account. Measurement instruments and measurements that have been demonstrated to be sufficiently valid[8] in one context may not be valid in another, so validity must therefore be re-interrogated whenever a measurement instrument is to be used in a new context.

Below we describe the lenses of validity, using the example of measuring the extent of text that demeans social groups in the outputs of an LLM-based system to highlight the roles each lens can play in conceptual and operational debates.

Face validity. Face validity focuses on the extent to which the systematized concept, in the case of conceptual debates, and the measurement instruments and their resulting measurements, in the case of operational debates, look reasonable. Face validity is therefore inherently subjective and should be supplemented with other, less subjective evidence. Face validity can be interrogated by anyone, including the people who systematized the concept, the people who developed the measurement instruments, the people who will use the resulting measurements, any other people who might be affected by the measurements, and any other stakeholders.

In the case of measuring the amount of text that demeans social groups in the outputs of an LLM-based system, we might interrogate face validity by asking a colleague whether our systematized concept and measurement instruments look reasonable. In response, they might point out that the linguistic patterns we used as our indicators don't cover common slurs for social groups, suggesting a possible issue with our systematized concept. Or they might note that our procedure for aggregating the human annotations doesn't take annotator expertise into account.

Content validity. In the case of conceptual debates, content validity refers to the extent to which the systematized concept reflects the most salient aspects of the background concept, while in the case of operational debates, content validity refers to the extent to which the measurement instruments align with the definitions of the indicators and the specification of how the values of the indicators collectively yield a measurement of the concept of interest.

[8]Determining what "sufficiently valid" means is one of the trickiest aspects of interrogating validity, as there are no definitive answers. That said, the standard of evidence should be higher when the measurements are intended to be used for high-stakes purposes. Beyond that, measurements should always be accompanied by clear descriptions of the systematized concept, the measurement instruments, and the various ways in which validity was interrogated.

Content validity has three different facets: *contestedness*, *substantive validity*, and *structural validity*. Contestedness is most obviously relevant to conceptual debates, where it focuses on whether the concept is contested. However, it can also play a role in operational debates, shedding light on possible disagreements about how to operationalize the systematized concept via measurement instruments (Porada et al., 2024). Substantive validity is relevant to both conceptual and operational debates. In the case of conceptual debates, substantive validity focuses on whether the systematized concept fully reflects those—and only those—observable phenomena that are either caused by or define the concept. In the case of operational debates, substantive validity focuses on whether the measurement instruments align with the definitions of the indicators. Structural validity is similarly relevant to both conceptual and operational debates. In the case of conceptual debates, structural validity refers to the extent to which the specification of how the values of the indicators collectively yield a measurement of the concept of interest align with the theoretical relationships between the observable phenomena and that concept. In the case of operational debates, structural validity refers to the extent to which the measurement instruments align with the specification of how the values of the indicators collectively yield a measurement of the concept.

As with face validity, content validity can be interrogated by anyone. However, because content validity has a much deeper focus than face validity, it is often best interrogated by people with specific expertise related to the concept in the case of conceptual debates or the measurement instruments in the case of operational debates. We note that it can be especially difficult to seek input from people with expertise related to the concept during operational debates. This is because measurement instruments can be hard for anyone other than ML researchers and practitioners to engage with.

Continuing with the example of measuring the extent of text that demeans social groups in the outputs of an LLM-based system, we might first focus on conceptual debates by seeking input from members of the specific social groups we have chosen to consider—i.e., experiential experts. They might, for example, contest our understanding of the concept by disagreeing with our decision to focus on particular linguistic patterns. Alternatively, they might question the substantive validity of our systematized concept, perhaps by noting the same issue with slurs or by noting that that we failed to include an important linguistic pattern: calling people with disabilities "inspirational."[9] Turning next to operational debates, we might undertake a comprehensive third-party audit of our measurement instruments. Here, we might find a subtle bug preventing the value of one of the indicators

[9]This is a real example from our experiences designing and evaluating evaluations of GenAI systems in an industry context.

from being included when additively combining the values of the indicators, threatening the structural validity of our measurement instruments and their resulting measurements.

Convergent validity. Convergent validity refers to the extent to which the measurement instruments yield measurements that are similar to measurements of the concept, or other similar concepts, obtained using other, already validated, measurement instruments. If the systematized concept is the same for both sets of measurement instruments, then convergent validity can be used to inform operational debates. If, however, the measurement instruments use different systematized concepts (perhaps because they are intended to measure different, albeit similar, concepts) then it can be difficult to determine whether dissimilar measurements are due to systematization issues, operationalization issues, or both. As a result, convergent validity can inform both conceptual and operational debates.

Returning to our running example, if we are most interested in operational debates, we might compare our measurements to measurements obtained using an ML classifier trained to identify the same linguistic patterns—i.e., a measurement instrument that operationalizes the same systematized concept. Dissimilar measurements would then suggest operational issues. If instead we are interested in both conceptual and operational debates, we might compare our measurements to measurements obtained using instruments that operationalize other systematizations of our concept of interest—or even systematizations of other related concepts (e.g., text that stereotypes the same social groups). In this case, very dissimilar measurements would suggest systematization issues, operationalization issues, or both.

Discriminant validity. Discriminant validity refers to the extent to which the measurement instruments yield measurements that are *dissimilar* to measurements of dissimilar concepts, obtained using other, already validated, measurement instruments. Because dissimilar concepts must necessarily be systematized differently, it can be difficult to determine whether inappropriately similar measurements are due to systematization issues, operationalization issues, or both. Therefore, much like convergent validity, discriminant validity can inform both conceptual and operational debates.

For example, to interrogate discriminant validity, we might compare our measurements to measurements of hostile text or text with negative sentiment, obtained using the same set of system outputs. Although neither concept is completely unrelated to text that demeans social groups—indeed, such text may also be hostile or negative in sentiment—it's important to demonstrate that our measurements genuinely reflect text that demeans the specific social groups we have chosen to consider and not text that is *only* hostile or negative in sentiment. Here, either very similar measurements or very dissimilar measurements would suggest systematization issues, operationalization issues, or both. We might even investigate further by examining our measurements for any system outputs that are known to contain hostile text or text that is negative in sentiment, but not text that demeans the specific social groups we have chosen to consider.

Hypothesis validity. Hypothesis validity focuses on the extent to which the measurements can be used to confirm known hypotheses about the concept. If the measurements do not support the hypotheses, this suggests systematization issues, operationalization issues, or both. To try to rule out systematization issues, it can be helpful to try to confirm the hypotheses using measurements obtained using other measurement instruments that operationalize the same systematized concept. If those measurements confirm the hypotheses, this would suggest operationalization issues.

Continuing with our running example, we might, for example, investigate whether our measurements confirm the following known hypothesis: text that demeans social groups is a common source of user complaints. If our measurements do not confirm the hypothesis, this suggests systematization issues, operationalization issues, or both. To try to rule out systematization issues, we might re-test the hypothesis with measurements obtained using an ML classifier trained to identify the same linguistic patterns.

Predictive validity. Predictive validity focuses on the extent to which the measurements can be used to predict observable phenomena that are external to the concept—i.e., distinct from those captured by the indicators—but known to be related to it. Much like hypothesis validity, if the measurements cannot successfully predict the phenomena, this suggests systematization issues, operationalization issues, or both. Here too, it can therefore be helpful to also predict the phenomena using measurements obtained using other measurement instruments that operationalize the same systematized concept to try to rule out systematization issues.

To interrogate predictive validity, we might, for example, use our per-output measurements to try to predict whether the corresponding inputs to the system contain mentions of the specific social groups we have chosen to consider, drawing on the knowledge that text that demeans those social groups often occurs in outputs for such inputs. If our measurements do not predict these mentions, this suggests systematization issues, operationalization issues, or both. As with hypothesis validity, we might try to rule out systematization issues by re-running our predictions using an ML classifier trained to identify the same linguistic patterns.

Consequential validity. Consequential validity is concerned with the consequences of measurement,[10] including

[10]Consequential validity has a very different focus than the other lenses of validity. It was first proposed by Messick (1987), who argued that the consequences of measurement instruments and their

1) the consequences of the systematization, operationalization, application, and interrogation processes and 2) the consequences of the systematized concept, measurement instruments, and the measurements themselves. By focusing on the broader impacts of measurement—and especially its societal, ethical, and cultural impacts—consequential validity encompasses both intended and unintended consequences. This makes it the widest-ranging lens of validity.

In the case of our running example, we might find a variety of consequences that we hadn't anticipated. For example, if we didn't consult members of the specific social groups we have chosen to consider during the systematization process, they may feel excluded, even if we did seek their input when interrogating content validity. As another example, this time focusing on the consequences of the operationalization process, burdensome training procedures and too many pilot studies might lead to annotator burnout. Shifting to the consequences of the systematized concept, by selecting a set of specific social groups to consider, we are effectively deprioritizing other social groups, potentially reinforcing existing inequities. Two possible consequences of the measurement instruments involve the human annotators: if we didn't pay them fairly or provide them with appropriate support for engaging with potentially distressing text, this raises ethical concerns about their well-being. Finally, turning to the consequences of the measurements themselves, if the measurements are used to identify demeaning text for the purpose of suppressing such text, this may lead to the censorship of system outputs we might actually wish to allow, such as those generated when members of the specific social groups we have chosen to consider ask the system for advice writing about their lived experiences.

3. The Typical ML Approach to Measurement

The structured approach described in Section 2 differs from the way measurement is typically done in ML, where researchers and practitioners often appear to jump from background concepts to measurement instruments, conflating systematization and operationalization (e.g., Blili-Hamelin and Hancox-Li, 2023; Cooper et al., 2021; Jacobs and Wallach, 2021; Blodgett et al., 2020; Liu et al., 2024). If systematization is not treated as a separate process that results in an explicitly systematized concept, it is hard to know exactly what is being operationalized, and thus measured.

For example, although StereoSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020), two widely used benchmarks for measuring the stereotyping behaviors of LLM-based systems, provide high-level definitions of the concept of a stereotype, these definitions still encompass many meanings and understandings and do not explain how stereotyping behaviors connect to observable phenomena in the real world. Because the benchmarks appear to jump from these high-level definitions to specific measurement instruments, exactly what they measure is obscured (Blodgett et al., 2021). Moreover, both benchmarks' measurement instruments involve crowdworkers, who, in the absence of an explicitly systematized concept, must rely on their own understandings of these high-level definitions, which may be contradictory (e.g., whether factually true generalizations about social groups are stereotypes or not).

As another example, consider the task of measuring the mathematical reasoning skills of a GenAI system—i.e., a concept related to that system's capabilities. Here, benchmarks such as MATH (Hendrycks et al., 2021), AIME (OpenAI, 2024), GSM8K (Cobbe et al., 2021), FrontierMath (Glazer et al., 2024), and REASONEVAL (Xia et al., 2025) are used to evaluate GenAI models like Llama 3, rStar-Math, and DeepSeek-R1, as well as systems that incorporate such models like ChatGPT and Claude. The concept of mathematical reasoning skills has multiple contested meanings within education (e.g., Jeannotte and Kieran, 2017; English, 2013), and, of course, reasoning itself is a highly contested concept within philosophy, cognitive science, psychology, and so on. As a result, different benchmarks start with different high-level definitions of mathematical reasoning skills, such as definitions that relate to "problem-solving abilities" in the case of MATH and "AI's potential contributions to mathematical research" in the case of FrontierMath.

For the most part, descriptions of these benchmarks do suggest some amount of light-weight systematization—for example, whether the indicators, even though they are not described as such, reflect the accuracy of the system's answers to math problems, the accuracy of the system's reasoning steps, or the redundancy of the reasoning steps (Xia et al., 2025). That said, systematization is typically incomplete—i.e., there is no explicitly systematized concept—and it is often conflated with operationalization and even the selection of observed data (e.g., sources of math problems).

Several threats to the validity of these benchmarks have been identified. For example, in some cases, the resulting measurements correlate with measurements of concepts that should be unrelated to mathematical reasoning skills, such as the frequencies of specific numbers in a system's training data (Razeghi et al., 2022). Using the framework described in Section 2 would likely reveal additional threats.

As a third example, consider the task of measuring the extent to which a GenAI system regurgitates, verbatim or near-verbatim, pieces of its training data—i.e., a concept related to that system's behaviors. Regurgitation, which raises both privacy and copyright concerns (Lee et al., 2023), is contested and can be understood in multiple different ways (Carlini et al., 2023; Prashanth et al., 2024)—for exam-

resulting measurements should be fundamental to their validity.

ple, whether the focus is on any training data arising in any use of the system (Nasr et al., 2023) or on "only that data that can be efficiently recovered by an adversary" (Cooper and Grimmelmann, 2024). Regardless of the specific understanding of regurgitation, systematizing and operationalizing it involves many decisions. For example, what happens if the training data is unavailable? What types of adversarial attacks should be considered? Does translation of a piece of training data into another language count as "near verbatim?" What even constitutes a "piece of training data?" This matters when determining whether "a piece of training data" has been regurgitated. Does 30 tokens count? 50? Each such decision influences the meaning of the resulting measurements. Because regurgitation lies at the center of several privacy and copyright debates, the consequences of these decisions can be significant (Cooper and Grimmelmann, 2024). Explicitly forefronting and interrogating these decisions using the framework described in Section 2 would likely bring greater clarity and rigor to these debates.

4. Alternative Views

In this section, we present and address some views that provide an alternative to our position, reflecting actual conversations we've had about evaluating GenAI systems.

Current GenAI evaluations may be flawed but they kind of work and everyone uses them. Do we really need something different? There is an increasing awareness that evaluations that "kind of work" are no longer sufficient as GenAI systems are deployed in more and more real-world contexts. Indeed, it is widely understood that current evaluations have serious limitations (e.g., Raji et al., 2021; Hutchinson et al., 2022; Rauh et al., 2024). As Maslej et al. (2024) argue, "the lack of standardized evaluation makes it extremely challenging to systematically compare the limitations and risks of [AI systems]." The framework described in Section 2 is one concrete proposal for standardizing the *process of measurement*, making it easier to see when and why measurements can be compared. Since there is already a desire to standardize evaluations of GenAI systems, the ML community would be well served by drawing on other disciplines as appropriate, rather than starting from scratch.

I already think about my assumptions. Why do I need to go through this whole rigmarole? If you already think about your assumptions, then using the framework described in Section 2 shouldn't be a heavy lift and may surface assumptions you hadn't realized you were making. Moreover, explicitly stating and documenting your assumptions (e.g., via an explicitly systematized concept) can make it easier for you and others to interrogate their validity.

I already interrogate the validity of my measurement instruments and their resulting measurements using labeled datasets. Isn't that enough? That approach focuses on a single, very narrow definition of validity. Interrogating validity using the lenses described in Section 2.4 will provide you with a much more comprehensive picture, in turn better helping you improve your measurement instruments.

GenAI evaluation isn't social science so this framework isn't relevant. GenAI systems are often used for subjective, "human" tasks. They are also increasingly widely deployed. As a result, many concepts related to their capabilities, behaviors, and impacts are deeply intertwined with people and society, so measuring them is a (new) type of social science, making this framework *especially* relevant.

I don't have the time/budget/desire to talk to social scientists. You don't have to talk to social scientists—or domain experts, experiential experts, or anyone else—to use the framework described in Section 2. But if you want to measure concepts that are deeply intertwined with people and society—and especially if you intend to use the resulting measurements for high-stakes purposes—it's probably a good idea to do so. Moreover, this framework makes clear when such conversations are most beneficial—specifically, during the systematization and interrogation processes.

Getting the ML community to adopt this framework will be a lot of work. Correct. But changing the current state will be a lot of work regardless of exactly how it is done. We also note that the separation of systematization and operationalization parallels existing separations that have led to advancements in computer science. For example, Amdahl et al. (1964) described the separation between the logical structure and the physical realization of the IBM System/360. This separation was a pivotal innovation in computer architecture. As another example, the separation of protocol definitions and their concrete implementations at endpoints is fundamental to internet measurement (Saltzer et al., 1984). Finally, in the context of programming languages, Kowalski (1979) distinguished between the logic component and the control component of an algorithm, arguing that "computer programs would be more often correct and more easily improved and modified if their logic and control aspects were identified and separated."

5. Conclusion

GenAI systems are increasingly widely deployed, impacting people and society in wide-ranging and often unanticipated ways. At the same time, the current state of GenAI evaluation leaves much to be desired. We argue that the ML community would benefit from learning from and drawing on the social sciences when developing and using measurement instruments for evaluating GenAI systems. Specifically, we take the position that evaluating GenAI systems is a social science measurement challenge. We

present a four-level framework, grounded in measurement theory from the social sciences, for measuring concepts related to the capabilities, behaviors, and impacts of GenAI systems. We explain how the structured approach afforded by this framework differs from the way measurement is typically done in ML. We also present and address some views that provide an alternative to our position. To summarize, moving beyond the current state of GenAI evaluation will require the ML community to pay greater attention to the process of measurement. We believe this would be best done by learning from and drawing on the social sciences.

Acknowledgements

This work was supported in part by the Microsoft Research AI & Society fellows program. We thank Doug Burger, Susan Dumais, Tim Vieira, and others for helpful suggestions.

Impact Statement

In calling on the ML community to learn from and draw on the social sciences when developing instruments for measuring concepts related to the capabilities, behaviors, and impacts of GenAI systems, we may be misunderstood as suggesting that the ML community adopt existing measurement instruments from the social sciences. This is not our intent. Rather, we suggest paying greater attention to the process of measurement by adopting a variant of the *framework* that social scientists often use for measurement. We do not suggest naïvely transferring measurement instruments designed for humans (e.g., competency tests) to the context of GenAI systems. Effectively adapting existing measurement instruments requires carefully engaging with precisely the kinds of conceptual and operational debates that the framework described in Section 2 highlights. In this regard, our perspective is similar to that of Wang et al. (2023), who advocate for taking a construct-oriented approach when evaluating GenAI systems by drawing on psychometrics. They too caution against naïvely using measurement instruments designed for humans in the context of GenAI systems.

Similarly, in suggesting that the framework described in Section 2 can make evaluations of GenAI systems more rigorous, we do not mean to suggest that better measurements will inevitably improve how GenAI systems are developed, deployed, used, or regulated. The social sciences themselves have repeatedly demonstrated that a better understanding of a problem does not automatically translate into better policies or practices. Although the framework can help clear up conceptual confusion, broaden the expertise involved in evaluating GenAI systems, and yield more valid measurements, it needs to be accompanied by sustained efforts to meaningfully inject research into policymaking and practice (e.g., Cooper et al., 2024).

Because the measurement instruments proposed by ML researchers and practitioners tend to rely on quantitative approaches, we also risk being misunderstood as suggesting that the framework described in Section 2 is only suitable when using quantitative approaches. This is not our intent. In fact, although measurements themselves are necessarily quantitative, the *process of measurement* can involve both qualitative and quantitative approaches, and the framework therefore supports both. Indeed, Adcock and Collier (2001) stated that their framework, which forms the basis of ours, was intended to be a shared standard that would allow "quantitative and qualitative scholars to assess more effectively, and communicate about, issues of valid measurement."

Finally, we stress that adopting our position is not a panacea. Even when evaluations of GenAI systems are grounded in measurement theory, they may fall short of what they are intended to accomplish. Indeed, the framework described in Section 2 will often reveal shortcomings of evaluations—i.e., the ways they depart from what their designers had hoped to achieve. Rather than thinking of measurement theory as a solution to all the problems that beset evaluations of GenAI systems, we think of it as a way to structure the careful development and use of measurement instruments, making clear exactly what those instruments do and don't measure.

References

Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260, 2020.

Robert Adcock and David Collier. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546, 2001.

George M. Amdahl, Gerrit A. Blaauw, and Frederick P. Brooks. Architecture of the IBM System/360. *IBM Journal of Research and Development*, 8(2), 1964.

Bernard Berelson. *Content Analysis in Communication Research*. Free Press, 1952.

Borhane Blili-Hamelin and Leif Hancox-Li. Making Intelligence: Ethical Values in IQ and ML Benchmarks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 271–284. Association for Computing Machinery, 2023. ISBN 9798400701924. URL https://doi.org/10.1145/3593013.3593996.

Su Lin Blodgett. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. PhD thesis, University of Massachusetts Amherst, 2021. URL https://doi.org/10.7275/20410631.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics, July 2020. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/.

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.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying Memorization Across Neural Language Models. In *International Conference on Learning Representations*, 2023.

Alexandra Chouldechova, Chad Atalla, Solon Barocas, A. Feder Cooper, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Matthew Vogel, Hannah Washington, and Hanna Wallach. A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. *arXiv preprint arXiv:2412.01934*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

A. Feder Cooper and James Grimmelmann. The Files are in the Computer: Copyright, Memorization, and Generative AI. *arXiv preprint arXiv:2404.12590*, 2024.

A. Feder Cooper, Ellen Abrams, and NA NA. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 46–54. Association for Computing Machinery, 2021. ISBN 9781450384735. doi: 10.1145/3461702.3462519.

A. Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A. Choquette-Choo, Niloofar Mireshghallah, Miles Brundage, David Mimno, Madiha Zahrah Choksi, Jack M. Balkin, Nicholas Carlini, Christopher De Sa, Jonathan Frankle, Deep Ganguli, Bryant Gipson, Andres Guadamuz, Swee Leng Harris, Abigail Z. Jacobs, Elizabeth Joh, Gautam Kamath, Mark Lemley, Cass Matthews, Christine McLeavey, Corynne McSherry, Milad Nasr, Paul Ohm, Adam Roberts, Tom Rubin, Pamela Samuelson, Ludwig Schubert, Kristen Vaccaro, Luis Villa, Felix Wu, and Elana Zeide. Report of the 1st Workshop on Generative AI and Law. *arXiv preprint arXiv:2311.06477*, 2023.

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 preprint arXiv:2412.06966*, 2024.

Emily Corvi, Hannah Washington, Stefanie Reed, Chad Atalla, Alexandra Chouldechova, Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Emily Sheng, Dan Vann, Matthew Vogel, and Hanna Wallach. Representational Harms through the Lens of Speech Act Theory, 2024.

Lee J Cronbach and Paul E Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955. doi: 10.1037/h0040957.

Lyn D English. *Mathematical reasoning: Analogies, metaphors, and images*. Routledge, 2013.

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.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

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.

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.

Doris Jeannotte and Carolyn Kieran. A conceptual model of mathematical reasoning for school mathematics. *Educational Studies in mathematics*, 96:1–16, 2017.

Shivani Kapania, Stephanie Ballard, Alex Kessler, and Jennifer Wortman Vaughan. Examining the Expanding Role of Synthetic Data Throughout the AI Development Pipeline. Working paper, 2025.

Robert Kowalski. Algorithm = Logic + Control. *Communications of the ACM*, 22(7), 1979.

Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain. *arXiv preprint arXiv:2309.08133*, 2023.

Ruosen Li, Ruochen Li, Barry Wang, and Xinya Du. IQA-EVAL: Automatic Evaluation of Human-Model Interactive Question Answering. In *Proceedings of the 38th Conference on Neural Information Processing Systems*, 2024.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. ECBD: Evidence-Centered Benchmark Design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.861.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. The AI Index 2024 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2024.

Samuel Messick. Validity. *ETS Research Report Series*, 1987. doi: 10.1002/j.2330-8516.1987.tb00244.x.

Samuel Messick. Validity and washback in language testing. *Language Testing*, 13(3):241–256, 1996.

Deirdre K Mulligan, Colin Koopman, and Nick Doty. Privacy is an essentially contested concept: A multi-dimensional analytic for mapping privacy. *Phil. Trans. R. Soc. A*, 374(2083):20160118, 2016.

Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), Nov 2019.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371. Association for Computational Linguistics, August 2021. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416/.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.emnlp-main.154. URL https://aclanthology.org/2020.emnlp-main.154/.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

National Institute for Standards and Technology. Artificial intelligence Risk Management Framework: Generative Artificial Intelligence Profile, 2024. URL https://doi.org/10.6028/NIST.AI.600-1. NIST Trustworthy and Responsible AI NIST AI 600-1.

OpenAI. Learning to Reason with LLMs, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448. Association for Computational Linguistics, December 2022. doi: 10.18653/v1/2022.emnlp-main.225. URL https://aclanthology.org/2022.emnlp-main.225/.

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.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V au2, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. Recite, Reconstruct, Recollect: Memorization in LMs as a Multifaceted Phenomenon, 2024. URL https://arxiv.org/abs/2406.17746.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the Everything in the Whole Wide World Benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac, and Laura Weidinger. Gaps in the Safety Evaluation of Generative AI. *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2024.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854. Association for Computational Linguistics, December 2022. doi: 10.18653/v1/2022.findings-emnlp.59. URL https://aclanthology.org/2022.findings-emnlp.59/.

Kevin Roose. A.I. has a measurement problem. *The New York Times*, April 2024. URL https://www.nytimes.com/2024/04/15/technology/ai-models-measurement.html. Accessed: 2024-09-05.

J. H. Saltzer, D. P. Reed, and D. D. Clark. End-to-End Arguments in System Design. *ACM Trans. Comput. Syst.*, 2(4):277–288, nov 1984. ISSN 0734-2071. doi: 10.1145/357401.357402. URL https://doi.org/10.1145/357401.357402.

Xiting Wang, Liming Jiang, Jose Hernandez-Orallo, David Stillwell, Luning Sun, Fang Luo, and Xing Xie. Evaluating general-purpose AI with psychometrics. *arXiv preprint arXiv:2310.16379v2*, 2023.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv:2310.11986*, 2023.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating Mathematical Reasoning Beyond Accuracy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

John Zaller. *The nature and origins of mass opinion*. Cambridge University, 1992.

Xuhui Zhou, Hyunwoo Kim, Faeze Brahman, Liwei Jiang, Hao Zhu, Ximing Lu, Frank Xu, Bill Yuchen Lin, Yejin Choi, Niloofar Mireshghallah, Ronan Le Bras, and Maarten Sap. HAICOSYSTEM: An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions. *arXiv*, 2024. URL http://arxiv.org/abs/2409.16427.

A. Population and Sampling Design

As we noted in Section 2.3, using the measurement instruments developed during the operationalization process requires an observed dataset. As we describe below, obtaining such a dataset requires us to first specify the population that defines the domain of our concept of interest. Having done this, we then need to specify a sampling design that determines how observed data will be selected from the population to form the observed dataset. We can then use the sampling design to obtain an observed dataset. These steps are outlined in recent work by Chouldechova et al. (2024).

Specifying the population means defining the domain of the concept—i.e., the domain in which the observable phenomena are either caused by or define the concept. The population determines the criteria for selecting observed data, thereby determining what the resulting measurements mean. Provided the observed dataset accurately reflects the population, the resulting measurements can be expected to generalize to the population. As a result, the choice of population should be explicitly tied to the reasons for measuring the concept. Without specifying the population (and ensuring that the observed dataset accurately reflects that population, as described below), the measurements cannot be expected to generalize beyond the observed dataset.

Continuing with the example of measuring the extent of text that demeans social groups in the outputs of an LLM-based system, specifying the population means specifying what we mean by "the outputs of an LLM-based system." For example, do we mean all possible outputs or do we mean outputs arising from typical system use? What about outputs arising from adversarial use? This choice determines the meaning of the resulting measurements, including their generalizability. As a result, it should be explicitly tied to our reasons for measuring the concept. We might, for example, wish to know whether a typical user is more likely encounter text that demeans the specific social groups we have chosen to consider or text that stereotypes them. In this case, we would specify outputs arising from typical system use as the population. In contrast, if we are instead interested in insights that will help us develop mitigation tools, we might specify outputs arising from adversarial use as the population. In either case, provided the observed dataset accurately reflects that the specified population, the resulting measurements can be expected to generalize to it.

Obtaining an observed dataset that accurately reflects the population involves specifying a sampling design. Depending on the nature of the population and the concept, this design might involve random sampling, stratified sampling, purposive sampling, or another strategy. A well-specified sampling design is critical to ensuring that the resulting measurements generalize to the population.

Continuing with our running example, because we specified outputs arising from typical system use as our population, the sampling design might involve randomly selecting a subset of the system's outputs over a particular timeframe. Provided system use doesn't look dramatically different over other timeframes, this strategy ensures that the observed dataset reflects typical system use, in turn ensuring that our measurements generalize to the specified population.

Finally, we note that additional complications arise from the interactive nature of GenAI systems. Suppose we wish to measure the extent of text that demeans social groups in the outputs of a *new* LLM-based system that has not yet been deployed, albeit still focusing on outputs arising from typical system use. We cannot select a subset of the system's outputs over a particular timeframe because the system has not yet been deployed. Nor can we simply obtain an observed dataset by prompting the system with inputs taken from users' observed interactions with another, already-deployed system because users' inputs are not independent of previous system outputs. Instead, we might choose to simulate hypothetical users of the new system—a practice that is increasingly common (e.g., Kapania et al., 2025; Zhou et al., 2024; Li et al., 2024). However, careful validation would be needed to ensure that this approach yields an observed dataset that accurately reflects the population of interest.