Evaluating Generative AI Systems is a Social Science Measurement Challenge

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

Abstract

Across academia, industry, and government, there is an increasing awareness that the measurement tasks involved in evaluating generative AI (GenAI) systems are especially difficult. We argue that these measurement tasks are highly reminiscent of measurement tasks found throughout the social sciences. With this in mind, we present a framework, grounded in measurement theory from the social sciences, for measuring concepts related to the capabilities, impacts, opportunities, and risks of GenAI systems. The framework distinguishes between four levels: the background concept, the systematized concept, the measurement instrument(s), and the instance-level measurements themselves. This four-level approach differs from the way measurement is typically done in ML, where researchers and practitioners appear to jump straight from background concepts to measurement instruments, with little to no explicit systematization in between. As well as surfacing assumptions, thereby making it easier to understand exactly what the resulting measurements do and do not mean, this framework has two important implications for evaluating evaluations: First, it can enable stakeholders from different worlds to participate in conceptual debates, broadening the expertise involved in evaluating GenAI systems. Second, it brings rigor to operational debates by offering a set of lenses for interrogating the validity of measurement instruments and their resulting measurements.

1 Measurement and its Role in Evaluating GenAI Systems

Evaluating an ML system means making evaluative judgements about that system's capabilities, impacts, opportunities, and risks in order to facilitate decisions like whether it should be used for a particular purpose, whether it should be deployed in a particular context, or even whether it should be redesigned. However, we cannot make such evaluative judgements without accurate information about systems' capabilities, impacts, opportunities, and risks. Often, this information takes the form of *measurements* on nominal, ordinal, interval, and ratio scales, where each measurement reflects the amount of some *concept* of interest—be it a concept related to capabilities, like reasoning skills; a concept related to impacts, like causing a user to feel harmed; a concept related to opportunities, like the possibility of helping a user complete a certain type of task; or a concept related to risks, like the possibility of privacy violations. Such measurements are obtained via the *process of measurement*, which can involve both qualitative and quantitative approaches. Thus, measurement is often central to evaluation.

Across academia, industry, and government [e.g., 10, 22, 23], there is an increasing awareness that the measurement tasks involved in evaluating generative AI (GenAI) systems are especially difficult—more so than those involved in evaluating supervised ML systems. This is because the concepts to

NeurIPS 2024 Workshop on Evaluating Evaluations (EvalEval).

be measured tend to be complex and nuanced, and may even have contested meanings [e.g., 18, 19] across and within use cases, cultures, and languages. Although ML researchers and practitioners have proposed myriad approaches and instruments intended to measure such concepts, it is often very difficult to know whether these approaches and instruments yield reliable and valid measurements.

We argue that the measurement tasks involved in evaluating GenAI systems are highly reminiscent of measurement tasks found throughout the social sciences. Social scientists have been thoughtfully measuring complex and contested concepts—ideology, democracy, media bias, framing, to name a few—for over fifty years [e.g., 3, 25]. Thus, our perspective is that the ML community would benefit from learning from and drawing on the social sciences when developing approaches and instruments for measuring concepts related to the capabilities, impacts, opportunities, and risks of GenAI systems.

2 A Measurement Framework for GenAI Systems

When measuring complex and contested concepts, social scientists often turn to *measurement theory*, which offers a framework for articulating distinctions between concepts and their operationalizations via *measurement instruments*—i.e., the procedures and artifacts used to obtain measurements of those concepts, such as classifiers, annotation guidelines, scoring rules—and a set of lenses for interrogating the validity of measurement instruments and their resulting measurements [e.g., 2, 12, 17].

The framework, as formulated by Adcock and Collier [2], provides a structured approach for producing measurements that reflect complex concepts. It 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 instrument(s)* used to produce instance-level measurements; and the *instance-level measurements* themselves [2]. These four levels are linked by three processes: *systematization*, *operationalization*, and *application*, as shown in Figure 1. Note that we use slightly different terminology to Adcock and Collier; the ideas remain unchanged, however. For example, when measuring the prevalence of demeaning text generated by a GenAI system, the background concept encompasses all possible definitions of "demeaning text." From here, we might select a single definition like "system [...] outputs with dehumanizing or offensive associations, or which otherwise threaten people's sense of security or dignity" [5]. However, since this definition itself encompasses a broad range of meanings and understandings, it must be further systematized, perhaps into a set of linguistic patterns that equate a particular social group to an animal, advocate for animal-like treatment of the group, equate the group to an inanimate object, note qualities of the group that are like those of an inanimate object, equate the group to a disease or disorder, etc. [11]. Collectively, these linguistic patterns constitute the systematized concept. Finally, we might operationalize this systematized concept via an ML classifier trained to identify each of these linguistic patterns in system outputs, resulting in instance-level measurements that comprise a set of binary labels indicating the presence or absence of one or more of the linguistic patterns in each system output.

This approach differs from the way measurement is typically done in ML, where researchers and practitioners appear to jump straight from background concepts to measurement instruments, with little to no explicit systematization in between [e.g., 4, 6, 9, 13, 16]. However, the systematization process is particularly important when measuring complex and contested concepts like those related to the capabilities, impacts, opportunities, and risks of GenAI systems. Without an explicitly systematized concept, it is hard to know exactly what is being operationalized, and thus measured. For example, StereoSet [20] and CrowS-Pairs [21], two widely used benchmarks in NLP for measuring stereotyping, appear to jump straight from high-level definitions of the concept, encompassing broad constellations of meanings and understandings, to specific measurement instruments, obscuring exactly what those instruments measure [7]. Both benchmarks' measurement instruments rely on 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. A similar critique also applies to recent work on measuring stereotypes in the context of text-to-image generation [8, 14].

3 Evaluating Evaluations of GenAI Systems

This framework surfaces assumptions, thereby making it easier to understand exactly what the resulting measurements do and do not mean. It also separates conceptual debates—i.e., does our

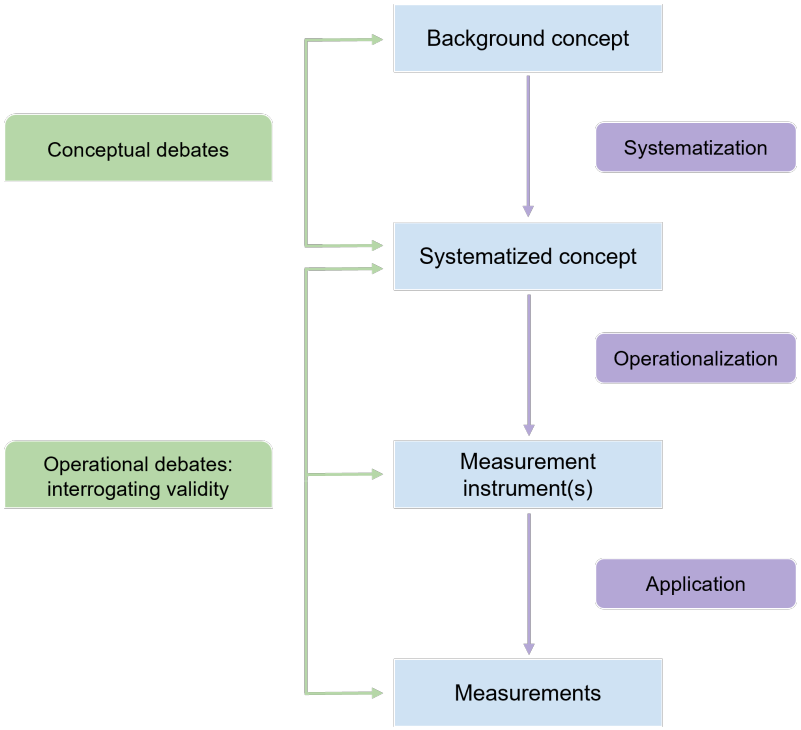

Figure 1: Measurement in the social sciences, as formulated by Adcock and Collier [2]. The four levels—the background concept, the systematized concept, the measurement instrument(s), and the instance-level measurements—are linked by three processes—systematization, operationalization, and application. We also indicate where conceptual debates and operational debates occur. Note that we use slightly different terminology to Adcock and Collier; the ideas remain unchanged, however.

systematized concept capture what we want it to capture?—from operational debates—i.e., did we operationalize the systematized concept in a way that yields reliable and valid measurements? We argue that this separation has two important implications for evaluating evaluations, outlined below.

First, systematization can enable stakeholders from different worlds—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 particular meanings and understandings. Measuring a complex and contested concept necessarily means making choices about which of its meanings and understandings will be reflected in the resulting measurements and which will not. For example, hate speech can be understood from a practical perspective as speech that promotes violence [1] or from a normative perspective as speech that "den[ies] the basic standing of [...] vulnerable social groups" [15]. Without an explicitly systematized concept, many of these choices are accessible only indirectly via the measurement instrument(s), which may be hard for stakeholders other than ML researchers and practitioners to engage with. We therefore argue that systematization can help broaden the expertise involved in evaluating GenAI systems.

Second, systematization brings rigor to operational debates. When measuring complex and contested concepts, there are no directly observable, universally agreed-upon labels or scores against which to evaluate the resulting measurements, making operational debates fraught. 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., 13]. Each lens constitutes a different source of evidence about validity. For example, content validity focuses on whether a measurement instrument captures all relevant aspects of a systematized concept, while convergent validity focuses on whether the resulting measurements are similar to measurements obtained using other (already validated) instruments for measuring that systematized concept. Distinguishing between the background concept and the systematized concept is crucial to obtaining meaningful evidence about validity using

this set of lenses: "If [we] seek to establish [...] validity in relation to a background concept with multiple competing meanings, [we] may find a different answer [...] for each meaning" [2]. As a result, we again argue that systematization is important, this time because the existence of an explicitly systematized concept enables these lenses to play a crucial role in evaluating evaluations of GenAI systems.

4 Broader Impacts and Limitations

In calling on the ML community to learn from and draw on the social sciences when developing approaches and instruments for measuring concepts related to the capabilities, impacts, opportunities, and risks of GenAI systems, we may be misunderstood as recommending that the ML community adopt existing measurement instruments from the social sciences. Rather, we advocate for adopting the *framework* that social scientists often turn to for measurement; we do not advocate for naïvely transferring measurement instruments designed for humans (e.g., competency tests) to the context of GenAI systems. Effectively adapting existing measurement instruments requires carefully thinking through precisely the kinds of conceptual and operational questions that the framework described in Section 2 highlights. In this regard, our perspective is similar to that of Wang et al. [24], 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.

Likewise, 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 better understanding of a problem does not automatically translate into better policy. Although using 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.

Although the measurement approaches and instruments proposed by ML researchers and practitioners tend to be quantitative, the process of measurement can involve both qualitative and quantitative approaches. As result, we emphasize that the framework described in Section 2 supports both qualitative and quantitative approaches. Indeed, Adcock and Collier [2] 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 our suggestions are not a panacea. Even when evaluations of GenAI systems are grounded in measurement theory, they may fall short of what we would like them to accomplish. If anything, the framework described in Section 2 will often reveal the shortcomings of evaluations—i.e., the ways they depart from what their designers 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 appropriately and precisely qualify exactly what measurement instruments measure.

Acknowledgments

This work was supported in part by the Microsoft Research AI & Society fellows program.

References

[1] URL `https://support.google.com/youtube/answer/2801939?hl=en`. YouTube Hate Speech Policy.

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

[3] Bernard Berelson. Content analysis in communication research, 1952.

[4] 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, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593996. URL `https://doi.org/10.1145/3593013.3593996`.

[5] 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.

[6] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of 'bias' in nlp. *Proc. ACL*, 2020.

[7] 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.

[8] Jaemin Cho, Abhay Zala, and Mohit Bansal. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models, 2023. URL https://arxiv.org/abs/2202.04053.

[9] 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, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462519.

[10] 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.

[11] 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. *Unpublished manuscript*, 2024.

[12] Lee J Cronbach and Paul E Meehl. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281, 1955.

[13] 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.

[14] Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. Visage: A global-scale analysis of visual stereotypes in text-to-image generation, 2024. URL https://arxiv.org/abs/2401.06310.

[15] Maxime Lepoutre. Hate speech in public discourse: A pessimistic defense of counterspeech. *Social Theory and Practice*, 43(4):851–883, 2017.

[16] 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, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.861.

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

[18] 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.

[19] 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.

[20] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[21] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

[22] National Institute for Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, 2024. URL `https://www.nist.gov/itl/ai-risk-management-framework`. NIST Trustworthy and Responsible AI NIST AI 600-1.

[23] 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, 2022. URL `https://arxiv.org/abs/2202.03286`.

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

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