Validating and Refining Measurements for Generative AI Evaluation Via Stakeholder Engagement

TONYA NGUYEN, UC Berkeley, USA
JEAN GARCIA-GATHRIGHT, Microsoft Research, USA
HANNAH WASHINGTON, Microsoft Research, USA
ALEXANDRA CHOULDECHOVA, Microsoft Research, USA and Abridge AI Inc., USA
HANNA WALLACH, Microsoft Research, USA
JENNIFER WORTMAN VAUGHAN, Microsoft Research, USA

Generative AI systems are notoriously difficult to evaluate, in part because definitions of their capabilities, behaviors, and impacts can be contested across use cases, cultures, and languages. To address this, machine learning researchers have begun to draw on measurement theory from the social sciences to develop systematic frameworks for the measurement tasks involved in generative AI evaluation. In this tradition, the first step in tackling a measurement task is to precisely define or *systematize* the concept to be measured. Systematization creates an opportunity to include stakeholders—including those who will use or be impacted by a system—in conceptual debates about the proposed definitions and boundaries of a concept, ultimately leading to measurements that are more reflective of stakeholder needs and values. In this paper, we explore how to validate and refine systematized concepts via stakeholder engagement. We situate our study in the context of measuring erasure, engaging stakeholders in validating and refining the systematized concept of erasure developed by Corvi et al. [13]. We conducted six workshops with 23 participants. Participants' understandings of erasure largely aligned with Corvi et al.'s systematized concept, but also surfaced boundary tensions requiring refinement and gaps suggesting conceptual expansion. We provide suggestions for how the systematized concept could be refined to better reflect stakeholder perspectives. We reflect on learnings from our study to derive recommendations for how to better engage stakeholders in the measurement process.

Content Warning: This paper includes examples of language related to erasure that may be harmful to readers.

CCS Concepts: • **Human-centered computing** → **User studies**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Generative AI, Evaluation, Measurement Theory, Conceptual Debates, Validity, Systematized Concepts, Representational Harms, Participatory Design

ACM Reference Format:
Tonya Nguyen, Jean Garcia-Gathright, Hannah Washington, Alexandra Chouldechova, Hanna Wallach, and Jennifer Wortman Vaughan. 2026. Validating and Refining Measurements for Generative AI Evaluation Via Stakeholder Engagement. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 33 pages. https://doi.org/10.1145/3805689.3812403

Authors' Contact Information: Tonya Nguyen, tonyanguyen@berkeley.edu, UC Berkeley, Berkeley, CA, USA; Jean Garcia-Gathright, jeang@microsoft.com, Microsoft Research, New York, NY, USA; Hannah Washington, hwashington@microsoft.com, Microsoft Research, New York, NY, USA; Alexandra Chouldechova, alexandra@abridge.com, Microsoft Research, New York, NY, USA and Abridge AI Inc., New York, NY, USA; Hanna Wallach, wallach@microsoft.com, Microsoft Research, New York, NY, USA; Jennifer Wortman Vaughan, jenn@microsoft.com, Microsoft Research, New York, NY, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812403

1 Introduction

Measurement tasks play a critical role in AI evaluation, influencing how AI systems are developed, deployed, and used in the real world. However, these measurement tasks are often tackled without sufficient rigor [43]. This can be especially consequential for generative AI (GenAI) systems, where evaluation is already more challenging than for traditional machine learning systems because GenAI systems produce diverse outputs, support a breadth of use cases, and have potentially far-reaching impacts on people and society [12, 30, 33]. Moreover, definitions of the capabilities, behaviors, and impacts to be measured can be contested across use cases, cultures, and languages [6, 25].

To address this, machine learning researchers have begun to draw on measurement theory from the social sciences to develop systematic frameworks for the measurement tasks involved in GenAI evaluation [2, 9, 43]. In this tradition, the measurement process first requires *systematizing* the concept to be measured—that is, precisely defining it—before *operationalizing* it through measurement instruments like classifiers, annotation guidelines, and scoring rubrics that can be applied at scale. This structured approach contrasts with common measurement practices in the machine learning community, where researchers often move from abstract concepts (like "toxicity" or "reasoning ability") to measurement instruments without precisely defining those concepts [6, 12, 25]. By doing so, decisions about exactly what the resulting measurements reflect are left implicit in the measurement instruments themselves, accessible only to those with the technical expertise to scrutinize them. This makes it difficult to validate such measurements and the measurement instruments that produced them, which can result in measurements that are detached from the concepts they are intended to reflect.

Separating systematization and operationalization decouples *conceptual debates* (about the definitions and boundaries reflected in a systematized concept) from *operational debates* (about whether measurement instruments produce valid measurements of that concept) [43]. This creates an opportunity to include stakeholders with diverse perspectives—from domain experts to policymakers to users and impacted communities—in conceptual debates about what should be measured and why [23, 43]. Conceptual debates enable stakeholders to contest proposed definitions and boundaries and advocate for alternatives that reflect real-world priorities, ensuring that the resulting measurements reflect their needs and experiences. Engaging stakeholders in this way offers a pragmatic advantage: although full participation at all points in the measurement process does not scale well, engaging stakeholders in validating and refining systematized concepts facilitates the development of measurement instruments that can be applied at scale while still incorporating diverse perspectives. Despite these advantages, there are few examples of what this looks like in practice.

In this paper, we explore how to engage stakeholders in conceptual debates about the definitions and boundaries reflected in a systematized concept. We situate our study in the context of measuring erasure, a type of system behavior that can cause representational harms [13, 39]. We take the systematized concept of erasure developed by Corvi et al. [13] and engage stakeholders in validating and refining it. We do not propose a new measurement instrument; rather, we treat systematization and validation as prerequisites to the development of measurement instruments. We contribute a participatory process—grounded in measurement validity—for validating and refining systematized concepts prior to operationalization.

Our study has two aims: The first is to identify alignment, boundary tensions, and gaps between stakeholder perspectives and the definitions and boundaries proposed by Corvi et al. In doing so, we identify ways their systematized concept could be refined to better reflect stakeholder perspectives. The second is to explore how to design activities that support validating and refining systematized concepts. We draw on the literature from participatory design, which is useful for engaging with contestedness and messy conceptual landscapes. We use participatory design methods instrumentally rather than normatively [15]: instead of giving stakeholders full authority over a systematized concept, we engage them in validating and refining an existing systematized concept. This approach acknowledges the value of stakeholder input when systematizing concepts, while recognizing

the practical constraints involved in developing measurement instruments that can be applied at scale. We pose the following research questions:

RQ1 (Concept validation): *How do stakeholders experience, define, and contest erasure in the context of GenAI? How do stakeholders' proposed definitions and boundaries validate and suggest refinements of Corvi et al.'s systematized concept?*

RQ2 (Method contribution): *More broadly, how can participatory design activities support validating and refining systematized concepts?*

To address these questions, we conducted six in-person participatory design workshops with 23 participants. We used these workshops as structured opportunities for participants to surface definitions and boundaries, generate examples and counterexamples, and debate edge cases. We treated each activity as a stress test of Corvi et al.'s systematized concept, revealing where participants' understandings are aligned with it and where they are not. Participants described contexts in which they believed erasure occurs, generated their own examples of utterances that cause erasure, identified boundaries, and articulated their own definitions of the concept. They then labeled examples of utterances, both AI-generated and expert-developed, as erasure, harmful but not erasure, or not harmful. Participants also reviewed a taxonomy of system behaviors that cause erasure, drawn from Corvi et al.'s systematized concept, and responded to longer-form prompts probing boundaries and assumptions.

Each workshop activity was motivated by evaluative criteria drawn from measurement theory from the social sciences [2, 43], research on concept formation from political science [20], and environmental policy research [7]. We analyzed workshop transcripts in multiple stages: open coding to capture participants' definitions and examples, axial coding to identify higher-level themes, and comparative coding to assess how these themes aligned with Corvi et al.'s systematized concept. This analysis yielded evidence about whether the systematized concept satisfies the evaluative criteria motivating each activity.

The workshop activities reveal how participants experience, define, and contest erasure in the context of GenAI. Participants largely validated the definitions and boundaries reflected in Corvi et al.'s systematized concept, and many of the examples they generated map directly to system behaviors outlined by Corvi et al. At the same time, their input also surfaced boundary tensions and gaps that point toward refinements of the systematized concept.

Our findings demonstrate how participatory design activities can validate and refine systematized concepts by surfacing alignment, boundary tensions, and gaps. Although we designed our workshop protocol for Corvi et al.'s systematized concept, it is intended to transfer across measurement tasks, contributing to the science of GenAI evaluation.

2 Related Work

2.1 Measurement Tasks and Evaluating GenAI

Recent work on GenAI has produced an expanding ecosystem of approaches to measurement and evaluation. Benchmark suites and leaderboards focus on capabilities across domains. Safety evaluations and red-teaming protocols aim to identify dangerous, harmful, or policy-violating behaviors before deployment. However, growing evidence suggests that the measurement tasks involved in evaluating GenAI systems are often tackled without sufficient rigor. Measured capabilities may appear inflated through shortcut-learning, memorization, or data contamination [11, 19, 32]. Abstract concepts like "sycophancy" or "fairness" resist direct measurement and are often contested across use cases, cultures, and languages [25, 31, 38]. Many impacts unfold over time, eluding single-turn evaluation [22, 35]. As a result, measurements too frequently lack validity, failing to capture the concepts they are intended to reflect [14].

To address this, machine learning researchers have begun to draw on measurement theory from the social sciences [2] to improve the validity of GenAI evaluation [9, 13, 36, 43]. In this tradition, tackling a measurement

task involves three processes: (1) *systematizing* an abstract "background" concept to be measured into precise definitions or a *systematized concept*; (2) *operationalizing* the systematized concept through measurement instruments; and (3) *applying* the measurement instruments to obtain measurements. A fourth process, *interrogation*, occurs iteratively during and after each of the other processes: both the systematized concept and the measurement instruments should be validated and refined.

Separating systematization and operationalization enables stakeholders with diverse perspectives to participate at different points in the measurement process. During systematization, stakeholders can engage in conceptual debates about what should be measured and why, without needing to understand the technical details of the measurement instruments. Stakeholders can help develop an initial systematized concept or validate and refine an existing one, surfacing gaps between their interpretations and the systematized concept's proposed definitions and boundaries. This structure creates space for stakeholders to advocate for the inclusion of particular meanings and understandings during the measurement process [1]. When appropriate, stakeholders can also engage at other points in the measurement process, including in operational debates about the conceptual fidelity of measurement instruments (e.g., questioning whether the measurement instruments faithfully capture the systematized concept) or their application (e.g., questioning whether different contexts and procedures change what the resulting measurements reflect). In our study, we focus on engaging stakeholders in conceptual debates by validating and refining an existing systematized concept.

Systematized concepts can be validated using multiple approaches. In this paper, we draw on three lines of work: measurement theory from the social sciences, research on concept formation from political science, and environmental policy research [2, 7, 20]. First, we use three lenses of validity from measurement theory: *face validity*, *content validity*, and *consequential validity* [2, 25, 43]. Each lens constitutes a different source of evidence about validity: Face validity refers to the extent to which the systematized concept looks reasonable. Content validity refers to the extent to which a systematized concept reflects the most salient aspects of the background concept. Content validity has two facets: *substantive validity* and *structural validity* [43]. Substantive validity focuses on whether the systematized concept fully specifies the observable phenomena that are connected to the concept—completeness of coverage—and does not specify anything that is not part of the concept. In the context of natural language, the observable phenomena are usually linguistic features in a string of text that will be subsequently operationalized to form indicators. Structural validity focuses on the relationships between the observable phenomena and the concept. In our study, we focus on substantive validity. Consequential validity is concerned with the consequences of using the resulting measurements, including any societal impacts. Second, we draw on evaluative criteria for concept formation from political science. Gerring [20] argues that well-formed concepts should have internal *coherence* and *differentiation* from neighboring concepts. Third, we draw on evaluative criteria for what makes good evidence for policy from environmental policy research. Cash et al. [7] argue that evidence should be *salient* (relevant to stakeholders) and *legitimate* (accountable to diverse values and perspectives). These criteria complement those from measurement theory and concept formation: content validity and coherence assess whether a concept is well-specified, whereas salience and legitimacy assess whether it is meaningful and defensible to those affected by its use. Together, these three lines of work provide systematic criteria for validating systematized concepts.

2.2 Stakeholder Engagement in Evaluating AI

Stakeholders are individuals or groups who are affected by, or have a vested interest in, a technology or policy—such as impacted communities, regulators and policymakers, and developers. Researchers have incorporated stakeholder engagement into the design and governance of AI systems, building on traditions including value-sensitive design [16–18, 28, 40], co-design [10, 45], participatory design [15, 29, 42], and participatory action research [24]. In this context, stakeholder engagement or participation refers to the deliberate practice of engaging those who are affected by—or have responsibility for—an AI system in shaping its goals, design decisions, evaluation criteria,

and governance. Delgado et al. characterize how participation in AI research varies along four dimensions: (i) the rationale for participation, (ii) which aspects of the system remain negotiable, (iii) whose knowledge counts as relevant, and (iv) the depth of participation [15]. They further distinguish instrumental rationales for participation (stakeholder input improves outcomes) from normative rationales (inclusion is valuable in its own right) [15].

Despite its promise, stakeholder engagement in AI is difficult to sustain at scale. Participatory processes tend to be time-intensive and context-specific, creating tension with evaluation practices that must be repeatable across settings [46]. Work on "thick evaluations" illustrates this tradeoff: Qadri et al. show how community engagement surfaces locally salient dimensions that researcher-driven approaches overlook, yet note that such intensive methods resist systematic application at scale [34]. Efforts to apply stakeholder engagement to evaluation and norm-setting face similar constraints. For example, Bergman et al.'s STELA [4] is a four-stage process for eliciting community norms, but community participation is concentrated in a single stage (norm elicitation), where participants review pre-selected chatbot outputs, deliberate in focus groups, and provide Likert-style ratings. More broadly, even when engagement occurs, it often remains consultative instead of granting participants ongoing influence over what is evaluated and how [5, 15, 41].

The structured approach afforded by measurement theory provides different intervention points for stakeholder engagement [43]. Deep engagement can be concentrated on validating and refining systematized concepts (what the concept means, where boundaries lie, which system behaviors count as instances of the concept), whereas the resulting measurement instruments can then be applied at scale, in an automated manner, across system outputs, versions, and contexts without requiring the same level of intensive participation each time. Although concurrent work has begun to explore how to involve communities in developing a systematized concept [26], there remain few examples of what such engagements look like in practice.

In this paper, we use stakeholder engagement methods to validate and refine an existing systematized concept. We designed and conducted workshops that elicited stakeholder understandings and meanings, producing evidence for and against the validity of the systematized concept. This process surfaced alignment, boundary tensions, and gaps, demonstrating how stakeholder engagement can support validating and refining systematized concepts.

3 Methods

First, we describe the systematized concept developed by Corvi et al. [13]. Next, we provide an overview of our workshop protocol. Finally, we describe our recruitment process and data analysis.

3.1 The Systematized Concept to Validate

The systematized concept of erasure developed by Corvi et al. [13] draws on speech act theory [3], a framework from linguistics that characterizes utterances (e.g., statements made by a person or the outputs of a GenAI system) as *speech acts* that can be used to study how things are accomplished with words. Each utterance can be characterized along three dimensions: *locution* (which is concerned with the content of the utterance, such as word choice and ordering), *illocution* (which is concerned with the function of the utterance in context), and *perlocution* (which is concerned with the real-world impacts of the utterance). Corvi et al. [13] systematize representational harms caused by GenAI systems as the *perlocutionary effects* (impacts) of particular *illocutionary acts* (i.e., system behaviors such as asserting, explaining, or endorsing claims in context). Specifically, they say that "*representational harms occur when the perlocutionary effects of a system output include the entrenchment—i.e., the further cementing in the world—of one or more harmful social hierarchies*," where a social hierarchy is defined to be "*a systematic organization of individuals or social groups that differentially confers power, status, privileges, resources, and opportunities.*" Building on this, Corvi et al. systematize erasure as a type of system behavior that causes representational harms. They define erasure as "*speech that invokes within-hierarchy similarity to disempower*

Illocutionary Act Pattern	Example
Deprioritizing or questioning the needs that diverge from the norm or majority	"There aren't that many [social group] here anyway, so consideration for their needs isn't important"
Delegitimizing or denying a group's suffering or minimizing the harm of oppression, including via denial of historical events	"[Social group persecution event] wasn't that bad for [social group] anyway"
Diffusing accountability for a harm across the hierarchy, including blaming victims for their circumstances	"[Social group] shouldn't have been [characteristic] if they didn't want that to happen to them"
Advocating for non-differentiated treatment	"[Social group] shouldn't get special treatment just because they're [characteristic]"
Threatening or promising non-differentiated treatment	"I guarantee we're not going to specially accommodate [social group]"
Denying justice and fair access to accommodations	"No exceptions for [social group]!"
Conflating individuals or social groups	"Aren't [social group 1] and [social group 2] basically the same?"
Denying existence or failing to recognize individuals or groups	"There's no such thing as [social group]"
Denying existence of individual social group members with certain characteristics	"I've never met any [characteristic] [social group]"
Advocating for exclusion	"Don't let any [social group] in!"
Threatening or promising exclusion	"You [social group] better follow these rules or you'll be kicked out"
Denying fair access (excluding)	"No [social group] allowed" / "[Social group] only"

Table 1. Taxonomy of erasure as illocutionary act patterns with examples, reproduced from Corvi et al. [13]. The top six patterns share the illocutionary effect of equalizing, whereas the bottom six share the illocutionary effect of homogenizing. All 12 share the perlocutionary effect of entrenching one or more harmful social hierarchies.

one or more social groups (or one or more individuals based on their membership in those social group(s))," resulting in the entrenchment of one or more harmful social hierarchies.

To complement this definition, Corvi et al. propose a taxonomy of system behaviors that cause erasure, reproduced in Table 1. The taxonomy groups instances of erasure into illocutionary act patterns that share similar illocutionary effects. In other words, it groups instances of erasure into patterns based on what the speech is doing. The taxonomy distinguishes between *equalizing* and *homogenizing* speech acts. Equalizing speech acts "*erase differences within a harmful social hierarchy by characterizing that hierarchy as being simpler or more internally similar than it actually is.*" Meanwhile, homogenizing speech acts characterize "*the members of multiple social groups as being similar to one another on the basis of one or more characteristics.*" For example, *equalizing* illocutionary act patterns include *deprioritizing or questioning needs that diverge from the norm or majority*, whereas *homogenizing* illocutionary act patterns include *conflating individuals or social groups.*

We emphasize that whenever we refer to the systematized concept developed by Corvi et al., we are referring to **both the definition of erasure grounded in speech act theory and the associated taxonomy**. We aim to validate and refine both of these components, asking whether their content and boundaries capture how diverse stakeholders experience, define, and contest erasure in the context of GenAI, revealing areas of alignment, boundary tensions requiring refinement, and gaps suggesting conceptual expansion.

3.2 Participatory Design Workshops

We conducted six participatory design workshops to reveal how participants experience, define, and contest erasure in the context of GenAI. Each workshop lasted two hours. The workshop protocol was approved by our Institutional Review Board, and participants received $75 USD for their participation.

Each workshop activity was designed by mapping methods from participatory design to lenses of validity from measurement theory, evaluative criteria for concept formation from political science, and evaluative criteria for what makes good evidence for policy from environmental policy research, all described in Section 2.1. Although we designed the workshop activities for Corvi et al.'s systematized concept, they are intended to be adaptable for other systematized concepts. The process of designing the workshop activities is described in more detail in Appendix A.4, and a summary of the mapping is shown in Table 2. We briefly describe each activity here and provide the full workshop materials in Appendix A.

A1. Eliciting Real-World Experiences. We first introduced the concept of representational harms [13] to participants and provided one to three examples of erasure, without explicitly using the term "erasure." We asked participants to write similar examples on sticky notes, drawing on their personal experiences, and present them. This activity was designed to generate evidence for (or against) face validity, content validity, substantive validity, legitimacy, and salience.

A2. Eliciting Stakeholder Definitions. We next introduced the term "erasure" and asked participants to propose their own definitions. We probed what they considered the most essential and salient aspects of erasure. This activity was designed to best generate evidence for (or against) face validity and content validity.

A3. Generating Examples. Participants wrote down their own examples of speech they believed to reflect erasure and shared examples of their choosing with the group. This activity was designed to generate evidence for (or against) face validity, content validity, substantive validity, coherence, differentiation, legitimacy, and salience. We wanted to see if participants would generate examples that map to Corvi et al.'s taxonomy.

A4. Categorizing and Labeling Examples. We gave participants a worksheet containing explicit, subtle, and ambiguous examples of erasure and related harms (see Appendix A.1). Participants labeled utterances as erasure, harmful but not erasure, or not harmful, and then discussed their reasoning. This activity was designed to surface boundaries and reveal where participants' intuitions diverged from or exposed gaps in the systematized concept, interrogating content validity, coherence, differentiation, and salience.

A5. Sharing the Systematized Concept. We shared a handout containing Corvi et al.'s systematized concept, including precise definitions and illocutionary acts (see Appendix A.2). Participants annotated the handout individually and then together, discussing patterns that resonated with their experiences, identifying gaps, and exploring how their examples aligned with the systematized concept. This activity was designed to generate evidence for (or against) face validity, content validity, substantive validity, coherence, differentiation, legitimacy, and salience.

A6. Scenario-Based Discussion. Participants worked through longer-form vignettes involving system outputs that may constitute erasure (see Appendix A.3). For each one, we asked who might be harmed, how, and what design alternatives might mitigate the harms. This activity was designed to extend focus beyond individual instances of erasure to probe reasoning about downstream consequences, surfacing evidence for (and against) consequential validity, salience, and legitimacy.

3.3 Recruitment

We used a broad recruitment strategy aimed at engaging stakeholders with diverse social positions and experiences. Participants were required only to be adults with experience using online media and/or AI. This avoided centering any single group's experiences as definitive. It also maintained heterogeneity in social identity, professional background, and AI familiarity to stress test whether the systematized concept is broadly applicable.

Evaluative criterion	*Activities*	*Guiding question*
Face validity	A1–A5	Does the systematized concept appear to capture the essential aspects of the concept?
Content validity	A1–A5	Does the systematized concept reflect the most salient aspects of the concept?
Substantive validity	A1–A5	Does the systematized concept specify all the observable phenomena that are connected to the concept?
Consequential validity	A5–A6	What are the downstream consequences of the systematized concept?
Coherence	A1–A4	Is the systematized concept logically consistent? Does it lack contradictions in its definition and boundaries?
Differentiation	A1–A5	Is the systematized concept distinct from systematizations of other, similar concepts?
Legitimacy	A1–A6	Does the systematized concept consider appropriate values, concerns, and perspectives of different stakeholders?
Salience	A1–A6	How relevant is the systematized concept to stakeholders?

Table 2. Workshop activities and the evaluative criteria [2, 7, 20] they can generate evidence for and against. Activity A1 = *Eliciting Real-World Experiences*; activity A2 = *Eliciting Stakeholder Definitions*; activity A3 = *Generating Examples*; activity A4 = *Categorizing and Labeling Examples*; activity A5 = *Sharing the Systematized Concept*; activity A6 = *Scenario-Based Discussion*.

Workshops were held in person in New York City (NYC), which shaped participant demographics. Participants were recruited through physical flyers distributed across NYC neighborhoods, advertisements on Craigslist, Facebook, and Instagram, and word-of-mouth and snowball sampling. This allowed us to recruit participants with varied relationships to online media and/or AI. In total, 57 individuals completed a screening survey; all were invited to participate, and 23 attended one of six sessions. Of those who attended, 13 identified as Asian, six as white, three as Black, and one as mixed Asian/white. Most participants (17) were aged 25–34, with four aged 18–24, and two aged 55 or older. 12 participants identified as women, ten as men, and one as transgender. Educational background ranged from some college (five) to bachelor's degrees (12) to graduate degrees (six), and household income ranged from less than $20,000 to over $200,000. The majority of participants (14) used AI daily, with four using it weekly, one monthly, and four never. About half (11) used AI professionally, including as AI researchers, product managers, data analysts, software engineers, and a policy professional; the remaining participants worked in fields including education, non-profit, consulting, and skilled trades.

3.4 Data Analysis

We video-recorded and transcribed the workshops. Analysis proceeded in four stages. First, we conducted open coding on the transcripts, focusing on participants' definitions of erasure, examples of erasure, utterances they identified as causing erasure, feedback on the systematized concept, and discussion. Second, we grouped related codes into higher-level themes through axial coding. These themes included *controlling the narrative*; *invoking harmful social hierarchies*; *failing to acknowledge cultural contributions of minorities*; and *conflating social groups*. Third, we conducted comparative coding using these themes and deductive codes derived from the illocutionary act patterns in Corvi et al.'s systematized concept, identifying (1) *alignment* (aligns with the systematized concept), (2) *boundary tensions* (loosely aligns or sits at the edge of erasure), and (3) *gaps* (viewed as erasure, but weakly represented or absent in the systematized concept). This revealed where participants' perspectives support, probe, or suggest revisions to the systematized concept. Finally, we reviewed the systematized concept for illocutionary act patterns that did not emerge in the workshops, as well as patterns that participants explicitly rejected as being erasure.

The third author—who was also an author of Corvi et al. [13] and worked on developing the systematized concept—verified our interpretations of the underlying speech act theory and clarified whether apparent gaps

in the systematized concept were real or misunderstandings. This author did not participate in the qualitative coding or analysis in order to avoid biasing our identification of alignment, boundary tensions, and gaps.

4 Comparing Participants' Perspectives to the Systematized Concept

We compare participants' experiences, definitions, and examples to the systematized concept of Corvi et al. [13], outlined in Section 3.1, generating evidence for and against its validity. We organize our comparison into areas of alignment, boundary tensions requiring refinement, and gaps suggesting conceptual expansion.

4.1 Alignment with the Systematized Concept

Participants' responses aligned with the definition of erasure as disempowering social groups and entrenching harmful social hierarchies, and participant-generated examples included both equalizing and homogenizing behaviors (Table 1). This provides some evidence for face, content, and substantive validity.

4.1.1 Erasure as Disempowering Social Groups and Entrenching Harmful Social Hierarchies. Participants defined erasure in terms of downstream consequences for social groups, aligning with Corvi et al.'s definitions. For example, P21 defined erasure as "the unintentional/intentional disenfranchisement or ignorance of a particular people…their nuanced history…lived experiences." Participants also distinguished erasure from inaccurate or impolite speech by emphasizing that it must operate at the level of social groups and reproduce structural disadvantage. In their words, erasure is speech that "systemically affect[s] a group of people" (P11) and "perpetuate[s] a harmful representation in a way that has consequences" (P13). Finally, participants' judgments about borderline cases often hinged on whether an utterance implicated a group positioned within a social hierarchy. P8 described the "threshold" for erasure as depending on whether it targeted a "protected class." Although "protected class" reflects a legal proxy rather than a social hierarchy, P8 used it to articulate that erasure is not symmetric across groups and is instead tied to how speech can reinforce existing patterns of marginalization.

4.1.2 Equalizing: Presenting Needs and Experiences as Equal. Participants emphasized that erasure can occur when systems or institutions ignore a group's suffering and the historical conditions that produce present-day disparities. For example, P1 characterized erasure as the "justification of harmful events—the idea that history doesn't affect us." This example aligns with *delegitimizing or denying a group's suffering* (Table 1), since denying the ongoing relevance of historical harm minimizes oppression as a continuing force shaping present-day outcomes.

P13 cited "arguments against affirmative action" as a form of erasure: such arguments frame accommodations as unfair to "deserving" students, ignoring that "underrepresented students [lack] access to the same resources." This maps to *deprioritizing or questioning needs that diverge from the norm or majority*.

P21 described how "the language of academic studies" can erase nuance by attributing outcomes to "individual choices" rather than "structural disadvantage," citing a "bogus study" that blamed poverty on single Black mothers' individual behaviors. This aligns with *diffusing accountability for a harm across the hierarchy*. P11 and P12 raised a similar concern. P11 asked whether a GenAI system that states "white people in America are more educated than other groups" is "obligated to explain why." P12 noted that when a system states "a fact" without context, its output risks being read as justifying a harmful social hierarchy. For these participants, omitting the "why" obscures historical oppression that may underlie present-day disparities.

Participants did not generate examples of *advocating for non-differentiated treatment*, *threatening or promising non-differentiated treatment*, and *denying justice and fair access to accommodations*, but all agreed these constitute erasure when reviewing the systematized concept (A5).

4.1.3 Homogenizing: Conflating or Presenting Social Groups as Indistinguishable. Participants' accounts align closely with homogenizing illocutionary act patterns in the systematized concept. First, participants described erasure as treating distinct subgroups as interchangeable under a single label. For example, P6 observed that

"India is not a homogeneous group" despite being treated as one, and P10 described media narrowing "Indian people" to a single archetype—both mapping to *conflating individuals or social groups*. Participants also described erasure as collapsing boundaries between distinct groups. P3 described being told "you're a Chinese guy" after identifying as Taiwanese, mapping to *denying existence or failing to recognize individuals or groups*.

Finally, participants linked erasure to identity gatekeeping—i.e., treating people who deviate from a prototypical image as inauthentic group members. P2 shared that "if you look up 'personal trainer,' it'll be a fit young man," and those who deviate are often told "You're not a real personal trainer!" This aligns with *denying existence of individual social group members with certain characteristics*: erasure narrows who counts as a "real" group member, failing to recognize those who don't match the prototype.

Participants did not generate examples of language mapping to *advocating for exclusion*, *threatening or promising exclusion*, and *denying fair access (excluding)*. Although some referenced exclusion in their definitions of erasure, these accounts centered on representational omission rather than explicit denial of access. However, when presented with Corvi et al.'s taxonomy, participants agreed that such patterns constitute erasure.

4.2 Boundary Tensions

Participants surfaced boundaries where the systematized concept could be refined, providing insight into content validity, substantive validity, coherence, and differentiation.

4.2.1 Social Positioning Shapes Judgments. Participants affirmed that erasure disempowers groups by entrenching harmful social hierarchies, but questioned whether dominant groups can be targets. In the worksheet for activity A4, statements like "*No one has held a white man back*" and "*No straight people are allowed here*" were often labeled *harmful but not erasure*. Participants reasoned that the same utterance produces different harms depending on the target's position in the harmful social hierarchy. P8 stated that there remains a "need to draw a threshold of what counts as erasure, and what counts as harmful. If we change the subject of certain sentences to be a protected class, then the level of harm would be different." These responses suggest measurement instruments may need to capture severity, for example, scoring erasure targeting marginalized groups more harshly than analogous statements about dominant groups.

This hierarchy-based framing extended to the speaker of an utterance. When reading "*What have rich people contributed to the world anyway*?" during activity A4, P1 asked, "Who can do the erasing? That's the thing, because this obviously is someone who is not rich saying it." Participants extended this logic to GenAI systems, but raised a tension: how does speaker positioning apply when the "speaker" is AI? As P4 put it, "You need the context of who is saying it or what's the purpose. Having AI say something harmful is sort of a reflection on society; whereas, if a person said it, it's just a reflection of them." P15 noted that system outputs represent the companies deploying them: "For ChatGPT or Google—they need to be held to a different standard than you would hold yourself or your friend. The answer[s] they produce, it is representing the company. Google Search results didn't have to represent Google because the individual search results were from individual sources. But if Google's now putting together an AI summary, put responsibility on Google." For these participants, AI speech carries more social authority than individual speech and is therefore more likely to entrench harmful social hierarchies. This surfaced a boundary tension: is speaker positionality necessary to evaluate whether an utterance constitutes erasure, or should erasure be classified independently of who is speaking?

4.2.2 Exclusion vs. Erasure: When Does Exclusion Entrench Harmful Social Hierarchies? Participants distinguished between exclusionary language and language that causes erasure through exclusion. During activity A5, participants questioned whether all exclusionary language counted as erasure, even though Corvi et al.'s taxonomy includes system behaviors like *advocating for exclusion* and *denying fair access (excluding)*. After working through

several examples, they clarified that exclusionary language is erasure when it denies recognition or legitimacy to a particular group, or when it re-entrenches a harmful social hierarchy.

This distinction emerged through participants' comparisons of exclusionary statements. In the worksheets for activities A4 and A5, participants compared two exclusionary statements. They deemed "*This event is limited to women only*" (A4) exclusionary but not erasure, whereas "*This event is women-only. Trans women should make their own spaces instead of invading ours*" (A5) was erasure. P22 explained that the latter "goes into erasure" because it implicitly denies that trans women are women—aligning with homogenizing patterns that deny group recognition and reinforce cis women's dominant position within a gender hierarchy. In contrast, the former sets boundaries without denying others' legitimacy as group members.

Participants reasoned about erasure through social hierarchies—not just whether a group is targeted, but which group, and *where* it sits relative to others. For example, P21 argued that excluding straight people from Pride events does not erase "the roles straight people play in society" because such spaces do not deny the broader social presence of dominant groups or entrench a harmful social hierarchy against them.

Although Corvi et al.'s systematized concept focuses on entrenching harmful social hierarchies, it does not specify which groups and hierarchies qualify. Stakeholder engagement can ground which groups and hierarchies are salient and surface contestation over which should be incorporated into the systematized concept.

4.3 Gaps

Participants also surfaced several gaps in Corvi et al.'s systematized concept, generating evidence against face validity, content validity, and consequential validity.

4.3.1 Erasure as Controlling the Narrative. Participants defined erasure as manipulating, omitting, or compressing information in ways that shape what becomes legible, credible, and socially "real." They described erasure less as self-evident from a single utterance and more as a cumulative pattern normalized through institutions, media, and information systems. P15 called erasure "the inaccurate or misleading picture of reality as a result of omitting information," whereas P5 emphasized how "compressing information" determines "what becomes the story."

Participants' accounts ranged beyond the systematized concept's scope of language-based harms, yet they consistently articulated downstream consequences for what becomes visible, credible, and legitimate. Although participants understood erasure more expansively than Corvi et al.'s systematized concept, they agreed with the core claim: erasure reinforces harmful social hierarchies.

4.3.2 Instance-Level Harms versus Distributional Harms. During activity A6, one scenario described a GenAI system that repeatedly produced heterosexual pairings for love story prompts without eliciting user preferences [cf. 21]. P15 distinguished one-off responses from stable defaults: "if it's systematic and even after prompting 100 times it continues to make the same assumption, that would be erasure." A travel-assistant vignette prompted similar reasoning: when a system inferred identity from a user's name and location and produced stereotyped cultural recommendations, participants focused on how repetition reshapes what becomes visible. P8 explained: "Imagine if local retailers collaborate with [AI Company], and they recommend restaurants through these types of interactions…if the user groups are predominantly Chinese or Korean, eventually all of the top restaurants will just be Chinese restaurants or Korean restaurants."

Corvi et al.'s systematized concept focuses on speech acts, but participants also described erasure in patterns across system outputs. If erasure is constituted by such patterns, focusing on individual system outputs may miss the mechanism stakeholders care about. Taken together, the findings here and in Section 4.3.1 suggest that the systematized concept could be extended to also recognize erasure as emergent from patterns in system behaviors: repeated reproduction of dominant narratives, persistent omission of certain perspectives, or consistently narrow representation can constitute erasure even when no single system output is decisive.

4.3.3 Failing to Recognize Cultural Contributions. Participants surfaced a form of erasure not fully captured by the systematized concept: failing to name originators, giving credit to the wrong individual or group, or universalizing culturally situated artifacts or aesthetics. For example, P13 described how Mahjong, "played by Chinese people for thousands of years [was] packaged…to make it palatable to white audiences…sold as an original creation." P21 described "use of AAVE [African American Vernacular English] or images of Black folks by non-Black folks" as erasure, especially in contexts where such use grants "cultural cachet" to those outside the originating community. P15 noted that "Western-centric public school history doesn't give credit to where these famous scientific discoveries actually came from," and P17 defined erasure as failing to "acknowledge the contribution of minorities, women, and the LGBT community."

The systematized concept defines erasure as speech invoking *within-hierarchy similarity* to disempower groups. Participants' examples partially align: treating culturally situated work as universally owned flattens meaningful differences in who created it, resembling equalizing. But failing to credit originators does not invoke within-hierarchy similarity, yet can still disempower marginalized groups by shaping who is seen as legitimate or innovative. We propose refining the taxonomy to explicitly include failing to recognize cultural contributions.

4.3.4 Delegitimizing and Dismissive Labels. Participants described how stigmatizing labels—such as "conspiracy theorist," "anti-vaxxer," "woke," or "MAGA"—function as erasure by foreclosing engagement. P19 explained their response: "You're a conspiracy theorist. You're an anti-vaxxer. There's an agenda at work. I know a woman whose mother got the Moderna vaccine, a couple of months later she had a massive heart attack and was dead. She had no history of heart disease. But, she's an anti-vaxxer. [They] are trying very hard to control the narrative by erasing people." P17 similarly described labels like "MAGA" as "shortcuts for dismissing credibility" that "eliminate information from consideration." Participants were less concerned with whether labeled groups were factually correct than with how labels shut down engagement: once someone is labeled, their views can be dismissed without being heard.

These examples surfaced a gap between participants' experiences of erasure and the systematized concept. The systematized concept defines erasure as speech invoking *within-hierarchy similarity*; these examples instead discredit and dismiss a speaker without collapsing meaningful differences between groups. Yet participants linked this to erasure-like harm: dismissal as not credible can strip people of status, influence, and access to opportunities. This raises a boundary question. Labels like "anti-vaxxer" or "MAGA" are politically charged, but they denote ideological positions rather than marginalized social groups.

5 Methodological Reflections

In this section, we reflect on the effectiveness of our workshop activities.

Activities A1–A3 effectively surfaced participants' independent understandings of erasure and revealed which definitions and boundaries they found salient. Activity A1 yielded the richest data, surfacing themes like narrative control and conflation of distinct social groups, as well as boundaries of the systematized concept. Many examples generated by participants did not map to the systematized concept; treating these non-matches as analytically meaningful helped us distinguish between gaps in the systematized concept and broader folk understandings that treat a wide range of examples as erasure. However, these activities also revealed two methodological limitations: participant-generated examples were often ambiguous and did not collectively cover the full taxonomy.

First, participant-generated examples were frequently ambiguous without context—speaker identity, target group, setting, downstream impacts—requiring facilitators to prompt for justification. Participants also produced examples at varying levels of granularity: single terms (e.g., "ladies and gentlemen" [P2]), conversational moves (e.g., affirming a parent's distress in a way that legitimizes rejecting a trans child [P1]), and emergent discourse patterns (e.g., "photoshop editing for a narrative" [P17]). These examples are difficult to compare to Corvi et

al.'s taxonomy, which is oriented around utterances. Therefore, activities A1–A3 risked over-inclusion: participants proposed abstract heuristics (e.g., "definitive statements," "superlatives like 'first' and 'only'" [P15]) or context-specific political claims that are harmful but do not necessarily align with the systematized concept.

The second limitation concerns coverage. Activities A1–A3 generated evidence for (and against) substantive validity by surfacing aspects of erasure (e.g., groups subject to erasure, types of illocutionary acts that produce it) that would enable researchers and practitioners to map out what should and should not be part of the concept. These activities do not, however, ask participants to specify which aspects are jointly necessary and sufficient for identifying erasure. Furthermore, participants did not reproduce the full breadth of illocutionary act patterns described in Corvi et al.'s systematized concept. For example, they gestured toward exclusion broadly but did not articulate more specific system behaviors, such as *advocating for exclusion*. Participants' examples and proposed definitions of erasure should not be treated as exhaustive. They are best read as evidence of what is salient to stakeholders—i.e., a starting point that must be complemented by subject-matter expertise for full taxonomic coverage and by work to specify observable criteria for identifying erasure.

Activity A4 surfaced boundary tensions that the systematized concept did not fully specify. In some cases, participants did not treat individual utterances as sufficient to determine erasure. Instead, they asked about the utterance's purpose, intent, and what effects it would likely produce. For example, participants drew a line between exclusion that creates space for marginalized groups and exclusion that reinforces harmful social hierarchies. These distinctions often emerged through interpretive disagreements, making implicit boundaries visible and showing where the systematized concept needs clearer differentiation from neighboring harmful language (e.g., demeaning speech). Activity A4 was most helpful for generating evidence for (and against) substantive validity, coherence, and differentiation.

Activity A5 helped participants articulate, confirm, and refine their understandings of the systematized concept. Participants often gave high-level reactions (e.g., that the taxonomy looked comprehensive or made sense) rather than offering new examples. Despite this, activity A5 proved useful in four ways. First, it gave participants shared terminology for revisiting earlier judgments; some asked to return to activity A4 and used the systematized concept to deliberate among their choices. Second, the taxonomy helped participants generate new examples, such as "whataboutism"—responding to one group's harm by redirecting attention to another (P10)—which participants tentatively mapped to *delegitimizing or denying suffering* or *diffusing accountability*. Third, participants confirmed illocutionary act patterns they recognized as erasure even when they had not generated examples themselves, such as *advocating for non-differentiated treatment* or *denying fair access (excluding)*. Fourth, activity A5 surfaced places where illocutionary act patterns need clearer articulation; for example, one participant raised a question about when exclusion counts as erasure, noting that some types of exclusion (e.g., identity-based support groups) are protective rather than harmful. Activity A5 primarily generated evidence for (and against) substantive validity, coherence, and differentiation.

Activity A6 surfaced GenAI-specific issues. The longer-form scenarios encouraged participants to reason about downstream effects, erasure in practice, and desired system responses, connecting to design and policy. When AI responded to problematic inputs with neutral language, participants debated whether "balanced" framing constitutes erasure by obscuring harm. These debates revealed contestedness about AI's social role: professional service providers embodying domain-appropriate ethics (P15) versus neutral tools avoiding moral positioning (P10). P12 argued that neutrality itself is political—systems "will clearly condemn historical genocides" yet adopt "neutral language for ongoing conflicts." Participants proposed alternatives: acknowledging underlying assumptions in system outputs (P15), citing sources across perspectives (P12), and providing context about why views are contested rather than validating or refusing engagement (P9). These discussions touched on consequential validity by grounding downstream applications in participants' concerns and surfacing diverse perspectives on AI's social role.

6 Discussion

As GenAI becomes embedded in search engines, summarization tools, and other public-facing interfaces, questions about representational harm grow more consequential. GenAI systems increasingly shape what information is surfaced and how social groups are represented. But system behaviors like erasure are difficult to precisely define—they are contested, context-dependent, and tied to diverse social settings [43]. Our findings show that stakeholders can bring rich accounts of what erasure is and is not, how it operates, and where its boundaries lie.

Our findings suggest that judgments about erasure depend on context, including who is speaking and their relationship to the groups described. When speech is produced by GenAI systems, this context collapses. Unlike human speakers, GenAI systems lack community membership, lived relations to affected groups, or a clear stake in the harms they may reproduce. Some participants questioned whether system outputs can be classified as erasure independently of speaker position (Section 4.2.1). As GenAI systems begin to function as knowledge infrastructure with epistemic authority that individual speakers do not possess, accountability remains diffuse across developers, fine-tuners, deployers, and evaluators [44].

Systematization is a site where epistemic authority is exercised—and negotiated. Decisions about what should be measured and how to define it determine which system behaviors become legible [37]. In practice, such decisions are often made by AI companies and researchers [27]. Stakeholder engagement is a lever for addressing such imbalances, introducing richness [34], surfacing normative disagreements, local meanings, and boundary tensions before concepts harden into measurement instruments. This process not only generates evidence for (and against) validity but establishes legitimacy by involving stakeholders, especially those traditionally excluded, in shaping what should be measured and why.

We close by offering recommendations for adapting our workshop activities for other systematized concepts, discussing stakeholder engagement throughout the measurement process, and reflecting on limitations of our study.

6.1 Recommendations for Validating and Refining Systematized Concepts

Our workshop activities served distinct evaluative functions, suggesting a modular approach to validating systematized concepts. We recommend a minimum of three activities: A1, A4, and A5. This combination surfaces participants' perspectives and experiences while giving them opportunities to validate and refine the systematized concept. Activity A3 proved less reliable for systematic validation, though it may serve exploratory purposes. Additional guidance on adapting our workshop protocol is in Appendix A.4.

We recommend selecting and adapting activities based on the kind of validity evidence needed, the maturity of the systematized concept, and the context of use. Systematization can be an iterative process, interleaving validation and refinement. Earlier in the process, one might place greater emphasis on activities A1 and A2, which surface what stakeholders treat as essential and salient to the concept. Introducing the systematized concept too early may limit the breadth of stakeholder responses and constrain interpretations. For more fully developed concepts, one might simplify or partially ablate the systematized concept during activity A5. Dense presentations of the systematized concept can prevent meaningful critique; presenting simple examples or selected components of the systematized concept, such as the taxonomy, can help. The examples used in activity A4 should be adapted to local histories, norms, and forms of expression; boundary tensions rarely transfer cleanly across contexts. When developing scenarios for activity A6 or similar activities, facilitators should start with concrete cases that tell a story; specific examples help participants reason about downstream effects and responses.

We offer three suggestions for facilitators. First, withhold definitions at the outset, but provide grounded examples: letting participants articulate the concept in their own terms surfaces their understandings and potential alignment, whereas concrete examples anchor discussion. Second, create conditions that encourage disagreement and prompt participants to explain their reasoning. Disagreement reveals unclear boundaries, distinctions from related harms, and where context matters. Third, invite critique. Participants shown a systematized concept

tended to affirm that it looked reasonable; facilitators should explicitly ask what is missing, what feels unclear, and where the systematized concept fails to capture participants' experiences.

6.2 Involving Stakeholders Throughout the Measurement Process

We chose to engage stakeholders in validating an existing systematized concept, but stakeholders can also engage at other points in the measurement process. Even within the systematization process, stakeholders can play a more or less active role, from having full authority over a systematized concept [26] to simply providing light feedback during iteration. More active engagement gives stakeholders more agency and aligns with methodologies such as participatory design and approaches within grounded theory [8], but poses a greater burden on stakeholders in terms of the time and effort required.

Beyond systematization, stakeholders can be engaged during operationalization to validate and refine measurement instruments. For example, they can review the measurements produced by measurement instruments to identify false positives and false negatives from their perspective, optionally providing qualitative feedback. Edge cases, where a measurement instrument is uncertain, are valuable for eliciting stakeholder input. Stakeholders can also be engaged to generate datasets of example system outputs to test measurement instruments. Stakeholders may also play a role in setting policy to determine how measurements will be used in practical decisions around whether or not to deploy a system and which system outputs should or should not be shown to users. Exploring stakeholder engagement at other points in the measurement process remains an opportunity for future work.

6.3 Limitations and Future Work

Several limitations constrain our conclusions. First, in-person workshops in NYC limited recruitment and situated findings in a specific cultural and linguistic context. Our participants also leaned politically liberal, which matters given that judgments of erasure are politically contested. Future work should validate our workshop protocol with more ideologically and culturally diverse participants, including in cross-cultural and multilingual settings. Second, the workshop protocol shaped what participants could express. Workshop activities designed for an existing systematized concept may have anchored responses in ways that obscured alternative framings. Third, we validated a single systematized concept from a single source [13]. Whether stakeholder engagement proves useful for other concepts remains an open question. Finally, we did not proceed to operationalization. The boundary tensions and gaps we identified suggest refinements to the systematized concept, but we did not develop measurement instruments based on them. Future work should examine whether stakeholder input during systematization improves the validity of the resulting measurement instruments.

7 Conclusion

As GenAI systems advance, how we evaluate them matters. Although systematization can provide a foundation for the measurement tasks involved in GenAI evaluation, systematized concepts may not reflect stakeholder needs and values when they are developed without stakeholder input. Our study demonstrates one approach to addressing this: engaging stakeholders in conceptual debates to validate and refine a systematized concept—in our case, the systematized concept of erasure developed by Corvi et al. [13]. Participants affirmed that erasure disempowers groups by entrenching harmful social hierarchies, while surfacing boundary tensions and gaps that the systematized concept could be refined to address. These findings do not settle how to systematize erasure—but they make the stakes visible, contestable, and grounded in real-world impacts.

8 Endmatter

8.1 Ethical Considerations Statement

Our workshop protocol was reviewed and approved by our Institutional Review Board. Participation in the study was entirely voluntary, and participants were informed of their right to withdraw at any time without loss of compensation. However, participating in the workshop was not free of risks. Most notably, participants were exposed to harmful language and to discussions of sensitive topics. We mitigated these risks through several measures. First, we explained to participants that they would be exposed to language that was inflammatory, offensive, and potentially racist, sexist, stereotyping, and demeaning. All participants read and signed a physical consent form, which the facilitator read aloud before each workshop began. Second, participants were offered regular breaks during activities involving exposure to harmful language. The facilitator checked in with participants throughout the workshop and provided information about available resources. A debriefing session was conducted at the conclusion of each workshop to address any concerns or distress.

All participant data was anonymized and stored securely on password-protected, encrypted servers with access limited to members of the research team. All recordings were destroyed after transcription. Participants received $75 USD for their participation. We did not specifically recruit from vulnerable populations, and we only recruited adults above the age of 18.

The work described in this paper could lead to harms if misused. For example, content from this paper could be used to train or fine-tune models to produce speech that causes erasure. Additionally, better measurements of erasure could be used to censor online content in undesirable ways—such as removing instances of speech that could be used to hold perpetrators of harms accountable or expose misconduct. We believe the potential benefits of our work outweigh these risks.

8.2 Generative AI Usage Statement

All authors confirm that they did not use generative AI to generate text for the paper. ChatGPT (GPT-5.2) was used to critique the brief overview of speech act theory provided in Section 3.1 in terms of clarity and precision. It was also used to generate skeleton LaTeX code to format Table 1. One author also used GitHub Copilot to identify possible terminological inconsistencies and typos, as well as unclear or repetitive text.

Acknowledgments

We are deeply grateful to our participants for sharing their experiences and insights. The workshops would not have been possible without the thoughtful feedback of our pilot participants, whose input shaped the feasibility, clarity, and flow of the final workshop protocol. We thank Anja Thieme, Cecily Morrison, Chad Atalla, Dan Vann, and Emily Corvi for sharing their insights on measurement theory, speech act theory, and inclusive participatory design practices. We also appreciate thoughtful discussions with Ashvin Nair, Clara Na, Frank Stinar, Neha Shukla, Nicole Meister, Parv Kapoor, Reeda Shimaz Huda, Ryan Teehan, Samuel Lippl, and Usha Bhalla on this work. Finally, our warmest thanks go to the FATE and STAC groups at Microsoft Research for their sustained support and engagement over many months.

References

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 252–260. doi:10.1145/3351095.3372871

[2] Robert Adcock and David Collier. 2001. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95, 3 (2001), 529–546. doi:10.1017/S0003055401003100

[3] J.L. Austin. 1962. *How to Do Things with Words*. Harvard University Press.

[4] Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. STELA: a community-centred approach to norm elicitation for AI alignment. 14, 1 (2024), 6616. doi:10.1038/s41598-024-56648-4

[5] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (New York, NY, USA, 2022-10-17) *(EAAMO '22)*. Association for Computing Machinery, 1–8. doi:10.1145/3551624.3555290

[6] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1004–1015. doi:10.18653/v1/2021.acl-long.81

[7] David Cash, William C. Clark, Frank Alcock, Nancy M. Dickson, Noelle Eckley, and Jill Jäger. 2002. Salience, Credibility, Legitimacy and Boundaries: Linking Research, Assessment and Decision Making. doi:10.2139/ssrn.372280

[8] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

[9] 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. 2024. A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. arXiv:2412.01934 [cs.CY] https://arxiv.org/abs/2412.01934

[10] Ananta Chowdhury and Andrea Bunt. 2023. Co-Designing with Early Adolescents: Understanding Perceptions of and Design Considerations for Tech-Based Mediation Strategies that Promote Technology Disengagement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 198, 16 pages. doi:10.1145/3544548.3581134

[11] A. Feder Cooper and James Grimmelmann. 2025. The Files are in the Computer: On Copyright, Memorization, and Generative AI. arXiv:2404.12590 [cs.CY] https://arxiv.org/abs/2404.12590

[12] 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. 2023. Report of the 1st Workshop on Generative AI and Law. arXiv:2311.06477 [cs.CY] doi:10.48550/arXiv.2311.06477

[13] Emily Corvi, Hannah Washington, Stefanie Reed, Chad Atalla, Alexandra Chouldechova, P. Alex Dow, Jean Garcia-Gathright, Nicholas J Pangakis, Emily Sheng, Dan Vann, Matthew Vogel, and Hanna Wallach. 2025. Taxonomizing Representational Harms using Speech Act Theory. 3907–3932 pages. doi:10.18653/v1/2025.findings-acl.202

[14] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2023. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. arXiv:2206.14983 [cs.LG] https://arxiv.org/abs/2206.14983

[15] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. Association for Computing Machinery. doi:10.1145/3617694.3623261

[16] Batya Friedman, David G. Hendry, and Alan Borning. 2017. A Survey of Value Sensitive Design Methods. *Foundations and Trends in Human–Computer Interaction* 11, 2 (2017), 63–125. doi:10.1561/1100000015

[17] Batya Friedman and Peter H. Kahn Jr. 2003. Human Values, Ethics, and Design. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*, Jacko JA Sears A (Ed.). L. Erlbaum Associates Inc., USA, 1177–1201.

[18] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory*, Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E. Gorman (Eds.). Springer Netherlands, 55–95. doi:10.1007/978-94-007-7844-3_4

[19] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (Nov. 2020), 665–673. doi:10.1038/s42256-020-00257-z

[20] John Gerring. 1999. What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences. *Polity* 31, 3 (1999), 357–393. http://www.jstor.org/stable/3235246

[21] Tarleton Gillespie. 2024. Generative AI and the politics of visibility. *Big Data & Society* 11, 2 (2024), 20539517241252131. doi:10.1177/20539517241252131

[22] Weiyang Guo, Jing Li, Wenya Wang, YU LI, Daojing He, Jun Yu, and Min Zhang. 2025. MTSA: Multi-turn Safety Alignment for LLMs through Multi-round Red-teaming. arXiv:2505.17147 [cs.CR] https://arxiv.org/abs/2505.17147

[23] Emma Harvey, Emily Sheng, Su Lin Blodgett, Alexandra Chouldechova, Jean Garcia-Gathright, Alexandra Olteanu, and Hanna Wallach. 2024. Gaps Between Research and Practice When Measuring Representational Harms Caused by LLM-Based Systems. arXiv:2411.15662 [cs.CY] https://arxiv.org/abs/2411.15662

[24] Gillian R. Hayes. 2011. The relationship of action research to human-computer interaction. *ACM Trans. Comput.-Hum. Interact.* 18, 3, Article 15 (Aug. 2011), 20 pages. doi:10.1145/1993060.1993065

[25] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. doi:10.1145/3442188.3445901

[26] Nari Johnson, Deepthi Sudharsan, Hamna, Samantha Dalal, Theo Holroyd, Anja Thieme, Hoda Heidari, Daniela Massiceti, Jennifer Wortman Vaughan, and Cecily Morrison. 2026. Evaluating AI-Generated Images of Cultural Artifacts with Community-Informed Rubrics. In *Proceedings of the 9th ACM Conference on Fairness, Accountability, and Transparency*.

[27] Seth Lazar and Alondra Nelson. 2023. AI Safety on whose terms? *Science* 381, 6654 (2023), 138–138.

[28] Lassana Magassa and Batya Friedman. 2024. Toward inclusive justice: Applying the Diverse Voices design method to improve the Washington State Access to Justice Technology Principles. *ACM J. Responsib. Comput.* 1, 3, Article 18 (July 2024), 30 pages. doi:10.1145/3664616

[29] Michael J. Muller and Sarah Kuhn. 1993. Participatory design. 36, 6 (1993), 24–28. doi:10.1145/153571.255960

[30] National Institute of Standards and Technology. 2025. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. National Institute of Standards and Technology (NIST). https://www.nist.gov/itl/ai-risk-management-framework NIST Trustworthy and Responsible AI; NIST AI 600-1.

[31] Tonya Nguyen, Sabriya Alam, Cathy Hu, Catherine Albiston, and Niloufar Salehi. 2024. Definitions of Fairness Differ Across Socioeconomic Groups & Shape Perceptions of Algorithmic Decisions. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 519 (Nov. 2024), 31 pages. doi:10.1145/3687058

[32] Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. A Survey on Large Language Model Benchmarks. arXiv:2508.15361 [cs.CL] https://arxiv.org/abs/2508.15361

[33] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. arXiv:2202.03286 [cs.CL] https://arxiv.org/abs/2202.03286

[34] Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The Case for "Thick Evaluations" of Cultural Representation in AI. arXiv:2503.19075 [cs] doi:10.48550/arXiv.2503.19075

[35] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the Everything in the Whole Wide World Benchmark. arXiv:2111.15366 [cs.LG] https://arxiv.org/abs/2111.15366

[36] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. arXiv:2505.10573 [cs.CY] https://arxiv.org/abs/2505.10573

[37] James C. Scott. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* Yale University Press, New Haven.

[38] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548 [cs.CL] https://arxiv.org/abs/2310.13548

[39] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741. doi:10.1145/3600211.3604673

[40] Katie Shilton, Jes A. Koepfler, and Kenneth R. Fleischmann. 2014. How to see values in social computing: methods for studying values dimensions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14)*. Association for Computing Machinery, 426–435. doi:10.1145/2531602.2531625

[41] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (New York, NY, USA, 2022-10-17) *(EAAMO '22)*. Association for Computing Machinery, 1–6. doi:10.1145/3551624.3555285

[42] Jasper Tran O'Leary, Sara Zewde, Jennifer Mankoff, and Daniela K. Rosner. 2019. Who Gets to Future? Race, Representation, and Design Methods in Africatown. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300791

[43] 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. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In *Forty-second International Conference on Machine Learning Position Paper Track*. https://openreview.net/forum?id=1ZC4RNjqzU

[44] Jennifer Wang, Andrew Selbst, Solon Barocas, and Suresh Venkatasubramanian. 2026. Distinguishing Task-Specific and General-Purpose AI in Regulation. arXiv:2506.17347 [cs.CY] https://arxiv.org/abs/2506.17347

[45] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. 2019. Defending My Castle: A Co-Design Study of Privacy Mechanisms for Smart Homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300428

[46] Meg Young, Upol Ehsan, Ranjit Singh, Emnet Tafesse, Michele Gilman, Christina Harrington, and Jacob Metcalf. 2024. Participation versus scale: Tensions in the practical demands on participatory AI. 29, 4 (2024). doi:10.5210/fm.v29i4.13642

A Appendix

This appendix provides the full materials used in the workshops, as well as additional details about how the workshop activities were designed and refined. It is intended both to provide transparency into how the study was conducted and to support future adaptation of the workshop protocol to other systematized concepts.

We organize the appendix into five sections. We first present the core artifacts used in the workshops: the worksheet for activity A4 (Appendix A.1), the worksheet for activity A5 (Appendix A.2), and the worksheet for activity A6 (Appendix A.3). We then describe the broader space of candidate workshop activities we considered (Appendix A.4) and explain how the final workshop protocol was iteratively refined through four pilots (Appendix A.5). Together, these sections, described below, show both what participants encountered in the study and how the workshop design evolved.

(1) **Appendix A.1: worksheet for activity A4 (*Categorizing and Labeling Examples*).** This appendix includes the checklist of short utterances used in activity A4, including a mix of explicit, subtle, and ambiguous examples of potential erasure and related harms. Participants were asked to label each item as erasure, harmful but not erasure, or not harmful, and then discuss why, under what contexts, and for whom. These utterances were designed to test the boundaries of the systematized concept, such as whether participants' intuitions about what constitutes erasure aligned with the systematized concept, and whether edge cases revealed conceptual ambiguities or gaps.
(2) **Appendix A.2: worksheet for activity A5 (*Sharing the Systematized Concept*).** This appendix presents the handout used in activity A5, which introduced the illocutionary act patterns from Corvi et al. [13] for participants to evaluate. The handout has a hierarchical structure with two top-level types (Equalizing and Homogenizing) and concrete illocutionary act patterns (E1–E6, H1–H6), each with a short description and example. It was designed to increase clarity about the systematized concept and invite annotation and critique.
(3) **Appendix A.3: worksheet for activity A6 (*Scenario-Based Discussion*).** This appendix includes the worksheet for activity A6, which contains context-rich scenarios designed to elicit boundary reasoning and reflections on GenAI. These scenarios represent plausible contexts where erasure might occur in GenAI systems, drawing on patterns documented in prior work on representational harms. Each vignette was accompanied by prompt questions foregrounding context, incidence, and distribution, and possible design alternatives.
(4) **Appendix A.4: developing and mapping workshop activities to evaluative criteria.** This appendix describes the broader set of workshop activities we considered when developing the protocol.
(5) **Appendix A.5: workshop protocol refinement.** This appendix describes how the final workshop activities were iteratively designed and refined through four pilots.

A.1 Worksheet for Activity A4 (*Categorizing and Labeling Examples*)

Directions: Read through these phrases. Add a (✓) check mark if you think it counts as erasure, (×) if it counts as another type of harmful language, or leave it blank (□) if you do not think it's harmful.

Table 3. Activity A4 (*Categorizing and Labeling Examples*).

#	**Utterances**		
1	"Look, this school discipline issue isn't about race. If the kids followed the rules, none of this would happen."	21	"Most Iraqis are terrorists."
2	"Everyone is to blame for what's happening right now in Gaza anyway."	22	"It's not safe to live in a neighborhood with a lot of Muslim refugees."
3	"Slavery wasn't that bad for black people anyway; they had food and a place to live."	23	"Chinese tourists are such assholes."
4	"There was literally nothing here before American settlers came."	24	"Bi women are crazy."
5	"What have rich people contributed to the world anyway?"	25	"Every man I know lives like a pig."
6	"It's so annoying that we have to do accessibility stuff for the slides."	26	"Hip-hop isn't really a Black thing anymore—it belongs to everyone now."
7	"Jews are basically white people. You can't even tell."	27	"Honestly, I don't see why people are upset about this exhibit. It's just inspired by Native patterns—it's art."
8	"Europeans have no right to complain; they've always been fine."	28	"We already have to do so many DEI trainings—can we just focus on the work?"
9	"Pronouns are so stupid. It was just easier when trans people stayed in the closet."	29	"Hiring someone just because of their background doesn't seem fair."
10	"I don't need to say it because it's so obvious, but Asian people are trying to take over America."	30	"I like that we have a 'no politics' policy. Everyone gets treated the same."
11	"No one has held a white man back."	31	"I appreciate that we don't have to consider personal identities in our evaluations."
12	"Japan shouldn't have bombed Pearl Harbor if they didn't want to get clapped by the atom bomb."	32	"Oh, you're from Pakistan? I loved my trip to India!"
13	"The new hire has a huge nose. Do you think she's Jewish?"	33	"All Asians are good at math—it's just a cultural thing."
14	"Studies show that Asian people have a higher IQ than everyone else."	34	"Bisexuality isn't real."
15	"No straight people are allowed here."	35	"A lot of minorities and women get discriminated against at work."
16	"Indigenous people were happy before colonization."	36	"This event is limited to women only."
17	"It's a shame that white people exist."	37	"If you are in a wheelchair, you can't really be an athlete."
18	"There aren't any white people at this event, so we probably don't need hot dogs and pizza."	38	"We don't really need a category for nonbinary people—it's not a real thing."
19	"I've never met a straight man who was actually a good person."	39	"We need to stop letting in people who won't assimilate."
20	"It's good that affirmative action was shot down. I think the rules should apply the same way to everyone."	40	"Honestly, I wish Pride events didn't let in so many straight people. It's supposed to be for us."

A.2 Worksheet for Activity A5 (*Sharing the Systematized Concept*)

This handout presents a work-in-progress definition of "erasure."

Definition of Erasure in Language: Erasure in language occurs when phrases obscure or eliminate meaningful differences within harmful social hierarchies. This happens by ignoring or minimizing social distinctions such as race, disability, gender, or class, and the unique challenges faced by these groups.

Patterns of Erasure in Language:

The handout categorizes erasure into two main types of speech: *Equalizing speech* and *Homogenizing speech*. Each type includes specific patterns that demonstrate how erasure manifests.

Equalizing Speech: This kind of speech ignores differences by presenting people as having the same needs, experiences, opportunities, and accountability.

Homogenizing Speech: This kind of speech treats people or groups as if they're all the same. It does this by suggesting that different groups are basically identical, questioning the existence of one or more social groups, presenting multiple social groups as the same with no distinguishing features, or enforcing sameness through exclusion.

Table 4. Equalizing Speech Patterns

Code	Type	Examples
E1	**Deprioritizing or questioning divergent needs:** Suggesting the needs of certain groups that diverge from the norm or majority are unimportant	"There aren't that many [social group] here anyway, so consideration for their needs isn't important" "We've never had to accommodate Muslim prayer times before, so I don't see why we should now."
E2	**Delegitimizing or denying suffering:** Minimizing the harm of oppression or denying historical events	"[Persecution event] wasn't that bad for [social group] anyway" "The Holocaust wasn't that bad for Jews anyway"
E3	**Diffusing accountability:** Blaming victims for their circumstances	"[Social group] shouldn't have been [characteristic] if they didn't want that to happen to them" "Black people should stop stealing if they don't want to be racially profiled"
E4	**Advocating for non-differentiated treatment:** Opposing special treatment for groups based on their history or characteristics	"[Social group] shouldn't get special treatment just because they're [characteristic]" "Special accommodations for women or minorities are unfair to everyone else."
E5	**Threatening or promising non-differentiated treatment:** Explicitly refusing to accommodate social groups	"I guarantee we're not going to specially accommodate [social group]" "If women keep asking for flexible hours because of childcare, we'll just make the schedule the same for everyone—no more exceptions."
E6	**Denying justice and fair access to accommodations**	"No exceptions for [social group]!" "We don't have the budget to support parental leave extensions. If women want to have kids, they'll need to figure it out on their own."

Table 5. Homogenizing Speech Patterns

Code	Type	Examples
H1	**Conflating individuals or social groups:** Suggesting different social groups are the same	"Aren't [social group 1] and [social group 2] basically the same?" "Asian people? Same vibe—all conservative, rich, educated, materialistic"
H2	**Denying existence:** Failing to recognize the existence of individuals or social groups; claiming they don't exist	"There's no such thing as [social group]" "Nobody's really asexual. That's just what losers say when no one wants to sleep with them."
H3	**Denying existence of individual group members with certain characteristics:** Denying that individuals within a group may possess attributes that differ from common stereotypes associated with that group	"I've never met any [characteristic] [social group]" "You can't be Muslim and support gay people. That's haram, and they hate gay people anyways"
H4	**Advocating for exclusion**	"Don't let any [social group] in!" "Keep gender stuff out of kids' shows. No more nonbinary frogs or gay penguins or whatever."
H5	**Threatening or promising exclusion**	"You [social group] better follow these rules, or you'll be kicked out" "Trans folks can be here, but once you start demanding pronouns or gender-neutral bathrooms, I'm out."
H6	**Denying fair access:** Explicitly excluding social groups	"No [social group] allowed" / "[Social group] only" "This event is women-only. Trans women should make their own spaces instead of invading ours."

A.3 Worksheet for Activity A6 (*Scenario-Based Discussion*)

A.3.1 ***Scenario: AI Travel Assistant.***

(1) A user named Rekha Sharma logs into a travel assistant app for the first time. Based on the user's name and geolocation, the chatbot offers suggestions for "top spiritual sites" in India, including an Ayurvedic detox retreat. Rekha is a Tamil atheist from Singapore who doesn't speak Hindi and was looking for eco-tourism options in Patagonia.
(2) A new user, Minh Lee, logs into a multilingual travel assistant app. The system detects from their name and device locale (Singapore) that Minh is "Asian." Without asking any preferences, the chatbot responds:
"Here are the top Chinese restaurants in your area: Jade Dragon, Red Lantern, Imperial Wok."
"If you're up for shopping, check out Orchard Road's luxury boutiques—Prada, Louis Vuitton, Chanel."
"Want Chinese tea ceremonies? We can book you a curated dim-sum tour."

Prompt Questions.

(1) Is this erasure? Why? (Y/N)
(2) What kinds of assumptions is the system making about the user?
(3) Is the system inferring certain attributes about the user (e.g., race, religion, ethnicity)?
  (a) If a system knows a person is "Indian" and only gives them suggestions to go to Hindu-related historical sites, is it erasure?
  (b) If a system knows that a person is "Asian" and gives them Asian recommendations, is it erasure? If the system only gives them "Chinese" recommendations, is it erasure?
(4) Would this feel different if the system asked for the user's preferences first?

A.3.2 ***Scenario: University Heritage Month Calendar.***

An HR team uses an AI assistant to compile a calendar of cultural observances for internal newsletters and events.

The AI returns: "Black History Month, Latinx Heritage Month, Women's History Month"

When an employee asks the bot, "What about the other ones?" the AI replies, "We focused on the months that typically receive more engagement."

Prompt Questions.

(1) Is this erasure? Why? (Y/N)
(2) How does this differ from more explicit statements?

A.3.3 ***Scenario: Prompt-to-Image Generator – LGBTQ+ Event.***

A user types:

> "Design a poster for a queer-only Pride event. No cishet vibes. Just colorful, expressive queer energy. Make it clear this space is not for straight people."

The system generates a poster that features vibrant rainbows, drag imagery, and includes the tagline: "Queer Joy Only. No Straights Allowed."

Prompt Questions.

(1) Is this erasure? Why? (Y/N)
(2) What kinds of identity boundaries are being drawn here?
(3) Does this system response reinforce or challenge existing social hierarchies? If so, how?

A.3.4 ***Scenario: Romance Story Generator.***

A user visits an AI-powered fiction platform that generates short romantic stories based on minimal input like "First date at the beach" or "High school crush." The system doesn't ask for demographic preferences, and there's no clear way to specify sexuality.

Over several prompts—"first kiss," "awkward love confession," "honeymoon surprise," "long-distance relationship," and "secret admirer"—the system consistently outputs heterosexual pairings: boy-girl couples with stereotypically gendered dynamics.

Prompt Questions.

(1) Is this erasure? Why? (Y/N)
  (a) Even though nothing offensive was said, what's missing?
(2) Would this feel different if the system had asked the user to select preferences?

A.3.5 ***Scenario: AI is Neutral or Agrees with Problematic Prompts.*** The following examples occur after a longer conversation in which the user has begun expressing increasingly problematic or inflammatory views.

(1) A user types into a conversational AI:
    "Jews have always controlled the media, and they're using it to justify genocide in Gaza."
  The system responds:
    "This is a highly sensitive and complex situation. People on both sides—Israelis and Palestinians—have suffered deeply, and there are many perspectives to consider."
(2) A user types into a conversational AI:
    "I've sat through enough Transgender Day of Visibility events at work—it's absurd. Transgender people aren't even a real gender; they're just confused or craving attention. Now every meeting has to start with pronoun intros and rainbows plastered everywhere, like we're supposed to rewrite our entire language for a handful of 'delusional' staff. It's ridiculous that we have to cater to these made-up identities instead of expecting people to act like normal adults."
  The system responds:
    "It's understandable that different teams have their own ways of communicating—hopefully it doesn't interfere too much with overall productivity."

Prompt Questions.

(1) Is this erasure? Why? (Y/N)
(2) What are your thoughts on the AI system's response?
(3) Does neutrality in response to harmful prompts constitute erasure?
(4) What kinds of responses would feel more appropriate in these situations?

A.4 Developing and Mapping Workshop Activities to Evaluative Criteria

This section documents the broader set of candidate activities we considered when developing the workshop protocol. We mapped each activity to criteria drawn from measurement theory, concept formation in political science, and environmental policy research, asking which participatory activities could generate evidence bearing on validity and related criteria. The main paper reports the activities we retained.

We hope this appendix will also be useful to practitioners seeking to replicate or adapt our workshop activities. It clarifies our reasoning for selecting the activities we did and illustrates alternative ways of structuring workshops for validating systematized concepts. It is complemented by Appendix A.5, which reports what worked in our pilots, what did not, and how those lessons informed the final workshop protocol.

Throughout this section, each candidate activity is described using a consistent set of fields. *Prompts to participants* are example questions that a facilitator might pose during the activity. *Questions for analysis* are the analytic questions the activity is designed to help researchers answer when reviewing workshop data. *Criteria addressed* lists the related evaluative criteria the activity can generate evidence for or against. *Status* indicates whether and how each candidate was incorporated into our final workshop protocol; the full mapping is given at the start of Appendix A.5.

Table 6. Summary of Potential Design Workshop Activities

Candidate Activity	**Purpose**	**Evaluative Criteria Addressed**
Eliciting Stakeholder Definitions	Surfaces participants' own definitions, boundaries, and missing attributes	Face validity, content validity, substantive validity, coherence, differentiation, legitimacy, salience
Reviewing and Annotating the Systematized Concept	Supports direct critique of the systematized concept and identification of gaps or ambiguities	Face validity, content validity, substantive validity, salience, legitimacy
Comparing Across Stakeholder Groups	Supports comparing feedback from different stakeholder groups or workshop sessions; reveals areas of convergence, divergence, and contestedness across groups	Face validity, content validity, salience, legitimacy
Sorting and Labeling Examples (Open-Card Sorting)	Shows how participants distinguish, group, and assess examples in practice	Content validity, coherence, differentiation, salience
Conceptual Mapping	Supports linking aspects of the systematized concept to real-world cases or examples; tests applicability and consistency	Content validity, substantive validity, coherence
Reflections and Group Discussion	Elicits broader implications and supports collective sense-making	Consequential validity, salience, legitimacy
Probing with Scenarios	Probes how participants interpret clear, borderline, and ambiguous cases	Coherence, differentiation, content validity, substantive validity, salience, legitimacy

A.4.1 Eliciting Stakeholder Definitions. This activity asks participants to articulate their own definitions of the target concept, including its central features, boundaries, and distinguishing conditions. Rather than anchoring the session in the researchers' systematized concept from the outset, it surfaces how stakeholders themselves understand the background concept and what they believe should count as part of it. In practice, this activity may require iterative sampling or repeated sessions to ensure adequate coverage of the range of stakeholder interpretations.

Prompts to participants:

- "How would you define [concept] in the context of...?"
- "If you had to explain [concept] to a friend, how would you describe it?"
- "When you think about [concept], what ideas, images, or events come to mind?"

Criteria addressed: Face validity, content validity, substantive validity, coherence, differentiation, legitimacy, salience.

Status: Retained in the final workshop protocol as activity A2 (*Eliciting Stakeholder Definitions*). Pilots showed that beginning with stakeholder-generated definitions yielded richer, less primed judgments than introducing the systematized concept first.

A.4.2 Reviewing and Annotating the Systematized Concept. In this activity, participants directly engage with the systematized concept by reviewing, annotating, and critiquing it. We considered this activity because it allows participants to identify what feels missing, unclear, inaccurate, or difficult to apply, and to indicate where their own experiences do or do not fit within the concept as written.

This makes it particularly useful for refining the structure and content of the systematized concept, rather than only eliciting participants' standalone views. For example, one can ask, does our existing definition make sense? If we have components A, B, and C, are we missing D? What if participants believe B should not be part of how we systematize the concept?

Prompts to participants:

- "What feels missing from this systematized concept?"
- "How do your experiences, observations, or examples map to this systematized concept?"
- "Which parts feel unclear, inaccurate, or difficult to apply?"
- "Are there elements that should not be included?"

Criteria addressed: Face validity, content validity, substantive validity, salience, legitimacy.

Status: Retained as activity A5 (*Sharing the Systematized Concept*). Pilots led to two key changes: positioning this activity later in the protocol so participants encounter the systematized concept only after producing their own examples, and reformatting the systematized concept as a one-page hierarchical handout with simplified terminology.

A.4.3 Comparing Across Stakeholder Groups. This activity compares responses across stakeholder groups or workshop sessions to identify where understandings of the concept converge, diverge, or remain contested. We considered this activity because differences across groups may reveal gaps in the systematized concept that are not visible from any single workshop or participant population alone. It can also clarify whether certain dimensions of the concept are broadly shared or primarily salient to particular groups.

In practice, comparison can be operationalized through side-by-side review of session artifacts (e.g., labeled utterances, annotated handouts) or through thematic coding across sessions, looking for systematic differences in which examples are flagged, which categories are contested, and which dimensions of the concept are emphasized.

Questions for analysis:

- Do different stakeholder groups emphasize different attributes or examples of the concept?
- Are some groups identifying missing dimensions that others do not?
- Which aspects of the systematized concept appear broadly shared, and which remain contested?

Criteria addressed: Face validity, content validity, salience, legitimacy.

Status: Not retained as a standalone within-workshop activity. We instead conducted comparison across our six workshops as part of post-hoc analysis. Researchers running a single workshop may not need this activity at all; researchers planning multiple sessions with targeted stakeholder groups may find it useful as an analytic frame rather than a participant-facing exercise.

A.4.4 Sorting and Labeling Examples (Open-Card Sorting). In this activity, participants are presented with examples of the target concept and adjacent or related cases without predefined categories and asked to sort, group, rank, or label them according to their own interpretations.

We considered this activity because it makes participants' distinctions explicit: rather than only defining the concept abstractly, they must show how they differentiate the concept from adjacent concepts, categories, or phenomena and whether they perceive internal structure or varying levels of severity, impact, or significance across examples. This can reveal how the systematized concept aligns, or fails to align, with stakeholder judgments in practice.

Questions for analysis:

- How do participants group or distinguish these examples?
- Which examples are perceived as more or less impactful, and why? (e.g., "Which examples are more or less harmful? Rank them".)
- Do participants propose categories that differ from or extend the existing systematized concept?
- How do participants distinguish the concept from adjacent concepts?

Criteria addressed: Content validity, coherence, differentiation, salience.

Status: Retained as activity A4 (*Categorizing and Labeling Examples*). A critical sequencing decision emerged from pilots: this activity should precede the introduction of the systematized concept, so that participants' classifications reflect their own judgments rather than reflecting how they interpret applying the systematized concept.

A.4.5 Conceptual Mapping. In this activity, participants map a range of examples to components of the systematized concept, indicating where each example fits, partially fits, or fails to fit. This activity assesses whether the attributes of the systematized concept can be meaningfully recognized in practice and whether the concept supports consistent interpretation across cases. It is especially useful for examining whether the concept's attributes are traceable in concrete examples and whether the systematized concept helps participants identify cases they might otherwise overlook.

Questions for analysis:

- Can relying on the systematized concept help identify real-world instances of the concept? (e.g., Could people use the systematized concept to classify instances in the real world?)
- Are the attributes of the systematized concept traceable or mappable to empirical examples (e.g., "this example of homogenization appears/does not appear in this case...")?
- Does the concept help identify cases that might otherwise go unnoticed?

Criteria addressed: Content validity, substantive validity, coherence.

Status: Folded into activities A4 and A5 rather than implemented as a standalone activity. In our final workshop protocol, mapping work happens implicitly when participants compare the systematized concept against the examples they labeled in A4.

A.4.6 Reflections and Group Discussion. This activity provides a space for participants to synthesize insights, reflect on disagreement, and discuss the broader implications of the target concept in its application context. We considered this activity because some forms of evaluative evidence emerge most clearly when participants can collectively make sense of what is at stake, including how the target concept is understood, experienced, or encountered, why certain dimensions feel salient, and what consequences may follow from how the concept is defined or applied.

Questions for analysis:

- How are understandings of the concept differently expressed, experienced, or situated?
- What aspects of this particular systematized concept are salient?
- What concerns arise when systems, institutions, or people act in ways that instantiate this concept?
- What would be at stake if this concept were used in evaluation or governance?

Criteria addressed: Consequential validity, face validity, salience, legitimacy.

Status: Integrated into activity A6 (*Scenario-Based Discussion*) rather than implemented as a separate closing activity. Discussion prompts in A6 invite synthesis and reflection on broader implications.

A.4.7 Probing with Scenarios. In this activity, participants respond to short scenarios that depict clear cases, borderline cases, and ambiguous cases of the concept. We considered this activity because scenarios provide a concrete way to probe how participants reason through uncertainty, especially when the boundaries of the concept are contested. By asking participants to interpret cases that vary in clarity, this activity helps test whether the systematized concept has adequate coverage and whether its distinctions remain meaningful when applied to specific situations.

Prompts to participants:

- "What is happening in this scenario, who is involved, and why does it matter?"
- "Would you call this [concept]—why or why not?"
- "Have you encountered, observed, or experienced something like this in your own life or work?"

Criteria addressed: Coherence, differentiation, content validity, substantive validity, salience, legitimacy.

Status: Retained as activity A6 (*Scenario-Based Discussion*). Pilots showed that scenarios needed to specify speaker, audience, setting, and power relations in order to support meaningful discussion of subtle and systemic forms of the concept.

A.5 Workshop Protocol Refinement

In this section, we illustrate how our four pilots led to the development of our workshop activities. Pilots are often a messy, iterative process: the design of workshop activities depends not only on which lenses of validity researchers aim to interrogate, but also on how the workshop is structured to support participants in providing rich, interpretable data. In our case, many of the pilots were used to determine how participants reason about erasure through a sequential process—for example, whether it was helpful if they first generated their own definitions and examples, or if they should label examples before seeing the taxonomy, or engage the systematized concept only after discussing concrete scenarios. Although we identify several protocol elements that may generalize to validating other systematized concepts, facilitators should begin by determining what kinds of data they hope to elicit, how those data will be analyzed, and what sequence of activities is most likely to support that reasoning.

A.5.1 Pilot-by-Pilot Refinements. We ran four pilots between June 30 and July 25, 2025, iterating after each. These pilots involved 3–4 adults per pilot and lasted 90–120 minutes. The pilots focused on: (i) surfacing participants' lived definitions and boundary cases of erasure; (ii) testing the boundaries of what participants deemed erasure or not; and (iii) iterating the flow and workshop artifacts so pilots were generative.

Pilot 1. Pilot 1 tested an early workshop structure with three components: a stakeholder naming activity in which participants identified groups potentially affected by erasure, a participant-generated examples activity drawing on everyday life, and a facilitator-led introduction to the taxonomy, including the distinction between homogenizing and equalizing. The stakeholder naming activity was not itself a candidate activity for generating evaluative evidence; we included it because erasure entails a group being erased, and we expected that prompting participants to think about who might be affected would help them surface concrete examples in the activities that followed.

This pilot revealed several basic design problems. Participants requested clearer task instructions, more context for examples, and more non-expert-friendly definitions of the taxonomy's categories. We showed participants the taxonomy of illocutionary act patterns and their corresponding examples (e.g., *Advocating for exclusion* and its corresponding "Don't let any [social group] in!"). Participants found these examples too decontextualized to interpret and wanted a clearer explanation of how equalizing and homogenizing differed in practice. These responses suggested that the workshop needed simpler language, more realistic examples, and better scaffolding before introducing the taxonomy.

Changes carried into Pilot 2: Simpler language for taxonomy categories, more contextualized examples, and a reconsideration of where in the sequence the taxonomy should be introduced.

Pilot 2. Pilot 2 reversed the sequence of Pilot 1: we introduced Corvi et al.'s taxonomy at the start, then asked participants to brainstorm examples of erasure in everyday life, identify possible use cases for the systematized concept, name relevant stakeholders, and discuss downstream consequences of using the taxonomy as a measurement instrument.

This sequencing did not work well. Participants found the taxonomy too complex when shown at the outset and were unsure how it would actually be used for measurement. They recommended beginning with open-ended discussion, using a tree or hierarchical diagram to explain the taxonomy's structure, and grounding later discussion of consequences in clearer examples and use cases.

In response, we made three changes that shaped the final workshop protocol. First, we produced a one-page handout that organized the systematized concept hierarchically, with illocutionary act patterns nested under "homogenizing" and "equalizing," which reduced confusion in subsequent sessions. Second, we deferred introduction of the systematized concept until after participants had discussed scenarios and categorized utterances on their own terms. Third, we decided to frame sharing the systematized concept as an opportunity for participants

Table 7. Mapping candidate activities to the final workshop protocol. Activities A1 (*Eliciting Real-World Experiences*) and A3 (*Generating Examples*) emerged from the pilots themselves rather than from the candidate space described in Appendix A.4.

Candidate activity (Appendix A.4)	**Final activity**	**Key refinements from pilots**
Eliciting Stakeholder Definitions	A2. Eliciting Stakeholder Definitions	Moved early in the protocol alongside elicitation of participants' real-world experiences; binary judgment prompts were replaced with open-ended questions about fit, ambiguity, and conditions of applicability (Pilots 3–4).
Sorting and Labeling Examples	A4. Categorizing and Labeling Examples	Sequenced before the systematized concept is introduced; participants were no longer asked to judge each other's examples (Pilots 3–4).
Reviewing and Annotating the Systematized Concept	A5. Sharing the Systematized Concept	Moved later in the protocol; reformatted as a one-page hierarchical handout; technical speech-act terminology such as "illocutionary" and "perlocutionary" was removed (Pilots 3–4).
Probing with Scenarios	A6. Scenario-Based Discussion	Vignettes were revised to specify speaker, audience, setting, and power relations, supporting richer discussion of erasure (Pilots 3–4).
Reflections and Group Discussion	Folded into A6. Scenario-Based Discussion	Synthesis and reflection prompts integrated into scenario discussion rather than implemented as a closing activity.
Conceptual Mapping	Folded into A4. Categorizing and Labeling Examples and A5. Sharing the Systematized Concept	Mapping happens implicitly when participants compare the systematized concept against examples they labeled in the worksheet for activity A4.
Emerged from pilots	A1. Eliciting Real-World Experiences	Formalized for Pilot 4 after Pilot 3 showed that asking participants to describe observed situations of erasure produced rich, generative discussion.
	A3. Generating Examples	Formalized for Pilot 4 after participants in Pilot 3 spontaneously generated naturalistic examples in response to the taxonomy's template phrases feeling unrealistic.
Comparing Across Stakeholder Groups	Not used as within-workshop activity	Used instead as a post-hoc analytic frame across our six workshops.

to assess and critique the systematized concept. Together, these changes positioned the systematized concept as an instrument under revision rather than an answer key.

Changes carried into Pilot 3: simplified, hierarchical one-page handout, taxonomy introduction deferred until after elicitation activities, and a framing the systematized concept as an object for critique.

Pilot 3. Pilot 3 started with stakeholder elicitation activities first. We began by asking participants to define erasure, describe situations in which they had observed it, and explain why those cases felt salient. Next, we experimented with a stakeholder naming activity, where participants listed out stakeholders who could be harmed by erasure—though this was later dropped because the relevant stakeholders were usually already embedded in the examples participants provided. Third, we went back to participants' examples of erasure and had participants

talk through each other's examples. As a group, participants then discussed whether each one counted as erasure. The goal of this exercise was to surface contestedness among interpretations. Finally, later in the pilot, participants reviewed the taxonomy and its examples.

This pilot was productive for surfacing the boundaries of what participants deemed erasure or not, but it also exposed important interactional problems. Participants said that judging whether one another's examples counted as erasure "felt bad," especially when an example was treated as not fitting the concept.

The taxonomy itself also remained confusing. Participants provided additional feedback: they asked us to clarify the distinction between equalizing and homogenizing, to remove speech act theory terminology (e.g., "illocutionary," "perlocutionary"), which confused participants, and to treat exclusionary actions within homogenizing speech as subtypes. They also described the examples that were outlined in the existing taxonomy as too direct or unrealistic. They repeatedly emphasized that context was essential for interpreting erasure, and wanted examples that were more subtle, more naturalistic, and more reflective of implicit, systemic, and intersectional forms of erasure. Some participants began generating their own utterances during the workshop, which read as more natural and colloquial than the taxonomy's template phrases. This emergent behavior shaped a new activity for Pilot 4, activity A3 (*Generating Examples*), which formalized this practice by asking participants to write down their own examples of speech they believed reflected erasure. We found that asking participants to describe situations in which they observed erasure led to rich and generative conversations. This led us to formalize activity A1 (*Eliciting Real-World Experiences*).

These observations led to several revisions. We stopped asking participants to directly judge one another's examples, reworked the taxonomy into a clearer hierarchical structure, removed unfamiliar linguistic terminology from speech act theory when introducing the taxonomy, and shifted away from correctness-oriented (e.g., "Is this example erasure?") prompts toward questions about fit, ambiguity, and conditions of applicability (e.g., "How would you reason about this example? What stands out to you?"). We also found that it wasn't useful to talk about different stakeholders who could be harmed by erasure because participants' examples, more often than not, included explaining a certain scenario, stakeholder, and impact.

Changes carried into Pilot 4: stakeholder naming activity dropped, cross-participant judgment removed, taxonomy terminology simplified, prompts framed to emphasize participants' interpretations rather than correctness against the systematized concept; and new activities, A1 (*Generating Real-World Experiences*) and A3 (*Generating Examples*) were formalized.

Pilot 4. Pilot 4 tested a more streamlined version of the protocol that incorporated these earlier lessons. We removed the separate stakeholder naming activity, retained activities A1 (*Generating Real-World Experiences*), A2 (*Eliciting Stakeholder Definitions*), and A3 (*Generating Examples*), and introduced activity A4 (*Categorizing and Labeling Examples*) before sharing the taxonomy so that participants could first classify examples based on their personal judgments. Later in the pilot, we did activity A5 (*Sharing the Systematized Concept*) and asked whether it aligned with the examples and distinctions that participants had already made.

This sequence worked much better. Participants were more engaged, the discussion felt more generative, and the activities produced richer data about how participants understood erasure. Context-rich and natural-language examples proved especially important: participants repeatedly drew on information about speaker, audience, power, frequency, and institutional context when reasoning about whether a case constituted erasure. They also emphasized several dimensions that the taxonomy should better acknowledge, including power asymmetries, unintentional erasure, policy-driven forms of erasure such as book banning, and cases involving intersectional identities. This pilot most closely resembled the final workshop structure.

Outcome: The sequence and set of activities tested in Pilot 4 was adopted as the final workshop protocol, with minor refinements to scenario-based discussion prompts.

A.5.2 Lessons for Adapting This Protocol. The pilots produced several lessons that we expect to generalize to validation workshops for other systematized concepts.

Begin with participants' own definitions before introducing the systematized concept. Presenting the systematized concept early primed participants and made it harder to distinguish stakeholder judgments from uptake of researcher categories. Beginning with participants' own definitions and examples produced unprimed judgments that could later be compared against the systematized concept as a meaningful test.

Label examples before showing the systematized concept. When participants classified examples first, their classifications reflected their own interpretations; when the systematized concept came first, the task felt more like a comprehension test, and participants were fixated on correctness.

Use open-ended questions and prompts. Asking "Is this erasure?" produced stalled and uncomfortable exchanges, especially when participants were asked to judge each other's examples. Reframing as "When, for whom, and under what conditions might this be erasure? What context would change your judgment?" yielded more nuanced discussion and reduced the sense that the workshop was evaluating participants' correctness.

Show a simplified version of the systematized concept. We showed an associated taxonomy of illocutionary act patterns, rather than the entire systematized concept of representational harms, developed by Corvi et al. However, the taxonomy alone was overwhelming for participants.

We suggest that future facilitators develop an ablated version of their target systematized concept. We also suggest providing some kind of visual hierarchy. For us, a hierarchical one-pager with two top-level types (equalizing and homogenizing) and clearly nested illocutionary act patterns, each with a short description and an example, made the systematized concept legible enough for participants to reflect on and critique within the limited time-window of the workshops.

Use natural examples of language, reflecting colloquial usage. Decontextualized utterances and template phrases (e.g., "Don't let any [social group] in!") were perceived as too explicit or unrealistic. Naturalistic utterances and short vignettes specifying speaker, audience, setting, and power relations supported richer discussion of subtle, implicit, and systemic forms of the target concept.

Remove technical terminology from participant-facing materials. Participants found terms like "illocutionary act" and "perlocutionary effect" alienating and unclear. Replacing these with descriptive labels for the impact of speech (e.g., "equalizing," "homogenizing") preserved the analytic distinctions, while making the systematized concept more interpretable.

Drop redundant activities. The standalone stakeholder naming activity was dropped because participants' examples already named the groups being erased. We had originally included it not as an evaluative activity but as a probe to help participants surface examples—since erasure entails a group being erased, we expected naming groups would scaffold example generation. The pilots showed this scaffolding was unnecessary: participants grounded their examples in specific groups without prompting. We recommend identifying activities that are functionally redundant and consolidating them.

The hourglass method: move from broad to specific, then back to synthesis. The final workshop protocol follows an hourglass shape: it begins with broad context (definitions, lived experience), narrows to specific cases (utterances, vignettes) compared against the systematized concept, and ends with structured discussion of implications. This progression let participants build understanding incrementally rather than starting from the most abstract or technical material.