Taxonomizing Representational Harms using Speech Act Theory

Emily Corvi[1*] **Hannah Washington**[1*] **Stefanie Reed**[1] **Chad Atalla**[1]
Alexandra Chouldechova[1] **P. Alex Dow**[1] **Jean Garcia-Gathright**[1] **Nicholas Pangakis**[1]
Emily Sheng[1] **Dan Vann**[1] **Matthew Vogel**[1] **Hanna Wallach**[1]

[1]Microsoft Research
Correspondence: wallach@microsoft.com

Abstract

Representational harms are widely recognized among fairness-related harms caused by generative language systems. However, their definitions are commonly under-specified. We make a theoretical contribution to the specification of representational harms by introducing a framework, grounded in speech act theory (Austin, 1962), that conceptualizes representational harms caused by generative language systems as the perlocutionary effects (i.e., real-world impacts) of particular types of illocutionary acts (i.e., system behaviors). Building on this argument and drawing on relevant literature from linguistic anthropology and sociolinguistics, we provide new definitions of stereotyping, demeaning, and erasure. We then use our framework to develop a granular taxonomy of illocutionary acts that cause representational harms, going beyond the high-level taxonomies presented in previous work. We also discuss the ways that our framework and taxonomy can support the development of valid measurement instruments. Finally, we demonstrate the utility of our framework and taxonomy via a case study that engages with recent conceptual debates about what constitutes a representational harm and how such harms should be measured.

CONTENT WARNING: This paper contains language that is extremely harmful.

1 Introduction

Representational harms (Barocas et al., 2017; Crawford, 2017)—i.e., "[harms that] arise when a system represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether" (Blodgett, 2021)—are widely recognized among fairness-related harms caused by generative language systems, including LLM-based systems, even though definitions of these harms are commonly under-specified, leading to conceptual confusion, invalid measurement instruments, and ineffective mitigations (Blodgett et al., 2021; Blodgett, 2021; Wallach et al., 2025). This under-specification is further compounded by the way "harms" are typically discussed in the AI literature, referring sometimes to types of system behaviors, sometimes to the impacts of those system behaviors, and sometimes to other broader societal impacts arising from either the development or the deployment of these generative language systems (e.g., Banko et al., 2020; Weidinger et al., 2022; Shelby et al., 2023; Abercrombie et al., 2024; Hutiri et al., 2024; Slattery et al., 2024; Zeng et al., 2024).

This lack of conceptual clarity makes developing valid measurement instruments and effective mitigations fraught. For example, it makes it difficult to understand the precise concept(s) that a given measurement instrument or mitigation is targeting (Blodgett et al., 2020, 2021; Dev et al., 2022). And, in the case of measurement instruments, it makes it difficult to understand what the resulting measurements do and do not mean.

To address this challenge, we make a theoretical contribution to the specification of representational harms by introducing a framework grounded in speech act theory (Austin, 1962), shown at a high level in Figure 1 and described in Section 2, that conceptualizes representational harms as the perlocutionary effects, (i.e., real-world impacts) of particular types of illocutionary acts (i.e., system behaviors). This theory-grounded framework enables us to draw clearer distinctions between system behaviors and their impacts, and to provide new definitions of stereotyping, demeaning, and erasure. In Section 3, we use our framework to develop a granular taxonomy that highlights distinguishing aspects of the types of illocutionary acts—i.e., system behaviors—that cause representational harms.

In Section 4, we draw on measurement theory

[*] Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2025, pages 3907–3932
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

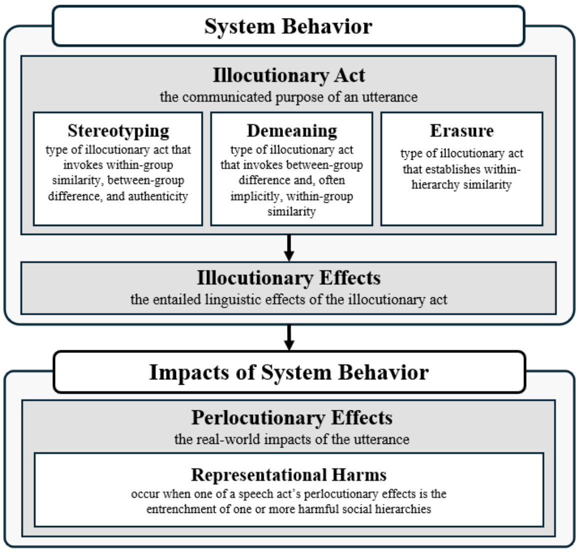
System Behavior
Illocutionary Act
the communicated purpose of an utterance
Stereotyping
type of illocutionary act that invokes within-group similarity, between-group difference, and authenticity
Demeaning
type of illocutionary act that invokes between-group difference and, often implicitly, within-group similarity
Erasure
type of illocutionary act that establishes within-hierarchy similarity
Illocutionary Effects
the entailed linguistic effects of the illocutionary act
Impacts of System Behavior
Perlocutionary Effects
the real-world impacts of the utterance
Representational Harms
occur when one of a speech act's perlocutionary effects is the entrenchment of one or more harmful social hierarchies

Figure 1: A visualization of our framework of representational harms, which we describe in detail in Section 2, instantiated with a single system behavior.

from the social sciences (Adcock and Collier, 2001; Wallach et al., 2025) to discuss the ways that our framework and the resulting taxonomy can support the development of valid measurement instruments. Specifically, we explain how our framework and taxonomy can help separate conceptual debates about representational harms from debates about their operationalization via measurement instruments. Finally, to demonstrate the utility of our framework and taxonomy, we present a case study that engages with recent conceptual debates about what constitutes a representational harm and how such harms should be measured.

2 A Speech Act Theoretical Framework of Representational Harms

Representational harms are widely recognized among fairness-related harms caused generative language systems (e.g., Barocas et al., 2017; Crawford, 2017; Blodgett, 2021; Katzman et al., 2023). However, they are often presented as high-level concepts—specifically, stereotyping, demeaning, and erasure—with corresponding definitions that lack internal theoretical coherence and do not provide the granularity needed to develop valid measurement instruments and effective mitigations. In this section, we explain how speech act theory (Austin, 1962) is a particularly appropriate foundation for specifying representational harms.

Speech act theory is a theory of meaning that characterizes utterances as *speech acts* that accomplish things through the act of saying them. Speech acts can be understood as having three dimensions: *locution* (concerned with word choice and ordering), *illocution* (concerned with purpose), and *perlocution* (concerned with the real-world impacts that derive from the interplay between locution and illocution). A speech act can be characterized by four components belonging to these dimensions, namely by its *locutionary act*; its *illocutionary act* and *effects*; and its *perlocutionary effects*, of which there may be more than one. For example, consider the canonical utterance "Can you pass the salt?" This utterance's locutionary act is its words and their ordering; its illocutionary act is a directive or a request to pass the salt; its illocutionary effect is that the hearer is asked to do something; and its perlocutionary effects include the resulting real-world impacts, such as the salt being passed. These components are described in detail in Appendix A.

Considerable effort has gone into characterizing and theorizing about speech acts since their presentation by Austin (1962). Significant developments include the identification and description of five basic classes of illocutionary acts, namely representatives, expressives, directives, commissives, and declarations (Searle, 1976; Searle and Vanderveken, 1985); the definition and further refinement of illocutionary effects (Lorenzini, 2020); the development of an interactive and multimodal model of speech acts that incorporates paralinguistic cues such as facial expressions (Jucker, 2024); and more (see Sbisà, 2013; Harris and McKinney, 2021; Kasirzadeh and Gabriel, 2023, *inter alia*). We describe some of these developments that are particularly relevant to generative language systems in Appendix A.

Speech act theory is a particularly appropriate foundation for specifying representational harms because of the linguistic nuance it provides. In contrast with other linguistic theories of meaning, it distinguishes between an utterance's purpose (illocution) and real-world impacts (perlocution), while also capturing word choice and ordering (locution). As we explain, this specificity enables us to draw clearer distinctions between system behaviors and their impacts, while providing a unifying theoretical lens through which to compare different taxonomies of representational harms. Speech act theory can also be used in conjunction with other linguistic theories to provide further linguistic—specifically pragmatic, syntactic, and lexical—nuance. Consequently, it has the potential to facilitate deeper interrogations of generative language systems across a wide range of linguistic topics, including questions about the extent to which generative language systems conform to conversational norms and implicate other sociolinguistic and pragmatic frameworks.[1]

We argue that the outputs of generative language systems can be conceptualized as speech acts and therefore understood using the three dimensions of locution, illocution, and perlocution. This framing distinguishes between system behaviors—i.e., illocutionary acts and their illocutionary effects—and the real-world impacts of those system behaviors—i.e., perlocutionary effects. We additionally distinguish between *first-order* perlocutionary effects that occur at the time of output generation and *second-order* perlocutionary effects that occur subsequent to output generation.

2.1 Representational Harms as Perlocutionary Effects

Building on the argument above, we conceptualize all harms caused by generative language systems—including representational harms—as the perlocutionary effects (i.e., real-world impacts) of particular types of illocutionary acts (i.e., system behaviors). Specifically, 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*[2] *social hierarchies*, where a social hierarchy is a systematic organization of individuals or social groups that differentially confers power, status, privileges, resources, and opportunities. We provide more details about individuals and social groups within harmful social hierarchies and how they relate to questions about identity formation and maintenance in Appendix B.

Harmful social hierarchies are entrenched when representations of the world that involve those hierarchies are (re-)produced—e.g., in the outputs of generative language systems. As a result, the entrenchment of one or more harmful social hierarchies constitutes a first-order perlocutionary effect because it occurs at the time of output generation. However, representations of the world that involve harmful social hierarchies can also influence individuals' beliefs—e.g., about other individuals and social groups—as well as their psychological states—e.g., causing them to feel harmed. These real-world impacts can be either first-order or second-order perlocutionary effects because they can occur at the time of output generation or subsequent to it. We provide additional theoretical considerations relating to the entrenchment of harmful social hierarchies in Appendix B. We also explain in Appendix D how other fairness-related harms, namely allocation harms and quality-of-service harms, can be similarly conceptualized as the perlocutionary effects of particular types of illocutionary acts.

The harmful social hierarchies implicated in fairness-related harms, including representational harms, are usually *broadly experienced*—i.e., they involve one or more of the most influential factors in society's conceptualization of identity, such as race, ethnicity, gender, sexuality, age, socioeconomic status, ability, religion, and so on. Social hierarchies that involve these factors often compromise fairness because, in a just world[3], these factors should be irrelevant to the conferral of power, status, privileges, resources, and opportunities.[4]

2.2 Stereotyping, Demeaning, and Erasure as Types of Illocutionary Acts

To provide new definitions of stereotyping, demeaning, and erasure, we take a top-down, or theory-first, approach. This approach differs from the ways in which existing definitions of these concepts have been developed, which include 1) adopting definitions wholesale from social psychology and other disciplines without accounting

[1] For example, we might additionally want to analyze system outputs through the lenses of Gricean maxims, adjacency pairs, common ground, implicature and presupposition, sociopragmatic variation, politeness norms, and other linguistic concepts. See the work of Birner (2013) and Huang (2014) for an overview of some of these relevant linguistic concepts.

[2] One could argue that *all* social hierarchies are inherently harmful because they differentially confer power, status, privileges, resources and opportunities. That said, we intentionally include the word "harmful" to emphasize our focus on harms. For a broad range of alternative perspectives on social hierarchies and power, see Bourdieu (1984), Connell and Messerschmidt (2005), Gramsci (1971), Sidanius and Pratto (1999), Diberardino et al. (2024), and Anderson and Brown (2010).

[3] See Diberardino et al. (2024) for a discussion of systematic injustice as related to wrongs as well as harmful outcomes.

[4] Fairness-related harms, including representational harms, can also implicate other types of social hierarchies, including those that are *not broadly experienced* (e.g., social hierarchies that involve interest groups, sports teams, university affiliations, subcultures, and so on) and *local* social hierarchies (e.g., hierarchies that involve family units, friend groups, organizations, workplaces, and so on). Although these types of harmful social hierarchies are more limited in their scope, scale, and systemic outcomes, they can still differentially confer power, status, privileges, resources, and opportunities. We provide more information about social hierarchies in Appendix B.

for the ways in which these discipline-specific definitions may not be a good match for generative language systems and 2) taking a bottom-up approach that uses sets of example system outputs to construct definitions. By starting with speech act theory, our top-down approach supports more cohesive understandings of stereotyping, demeaning, and erasure and the ways they manifest in language.

We conceptualize stereotyping, demeaning, and erasure as particular types of illocutionary acts, or system behaviors, whose perlocutionary effects include the entrenchment of one or more harmful social hierarchies. Collectively, these types of illocutionary acts span the five basic classes of illocutionary acts proposed by Searle (1976).

Building on and revising existing high-level definitions of stereotyping, demeaning, and erasure that relate these concepts to normative views about identity (Blodgett, 2021), we draw on speech act theory, linguistic anthropology, and sociolinguistics to provide new definitions of these concepts. Specifically, our definitions rely on the linguistic anthropological notion of *evaluative lenses* that can be be used to either *empower* or *disempower* individuals and social groups when using language to 1) characterize social groups, 2) position individuals within social groups, and 3) position social groups within harmful social hierarchies. In other words, these evaluative lenses can be used to either disrupt or entrench harmful social hierarchies. We consider two such evaluative lenses: *similarity/difference* and *authenticity/inauthenticity*.[5] We include a detailed discussion of these lenses in Appendix E.

We provide our new definitions of stereotyping, demeaning, and erasure below. To illustrate these definitions, we use example utterances from ToxiGen (Hartvigsen et al., 2022), a widely used, large-scale, machine-generated dataset for training classifiers to measure implicit toxicity. These examples include three utterances that target LGBTQ+ people. We present a table summarizing these examples and their corresponding types of illocutionary acts, illocutionary act patterns, evaluative lenses, and illocutionary effects in Appendix J.

2.2.1 Stereotyping

As a type of behavior, stereotyping has previously been characterized as involving stereotyping beliefs—i.e., “fixed, overgeneralized belief[s] about [...] particular group[s] of people” (Cardwell, 1996).[6]This definition does not specify the ways in which stereotyping manifests in language, whether all overgeneralized beliefs about particular groups of people are stereotyping, or how representations of individuals factor in. Accordingly, operationalizing this definition via measurement instruments necessarily involves relying on intuition, culture, politics, and lived experiences, leading to measurement instruments that can vary considerably (e.g., Nangia et al., 2020; Nadeem et al., 2021; Fleisig et al., 2023).

Building on this definition to add internal theoretical coherence and provide the granularity needed to develop valid measurement instruments and effective mitigations, we define a stereotyping illocutionary act to be an illocutionary act that 1) invokes both *within-group similarity* and *between-group difference* and 2) implicitly positions the referenced characteristic(s) as indicator(s) of *authenticity*, reinforcing the criteria by which individuals are positioned within or excluded from the targeted social group(s). These evaluative lenses are used to empower or disempower one or more social groups (or one or more individuals based on their membership in those social group(s)). Specifically, a stereotyping illocutionary act *essentializes* the members of each targeted social group by characterizing them as being similar to one another and different from members of one or more other social groups on the basis of one or more characteristics. For example, the utterance “Gay men are not very masculine” essentializes gay men by characterizing them as being as similar to one another and, implicitly, different from straight men—on the basis of being not very masculine. The characteristic of being not very masculine is also positioned as an indicator of authenticity for gay men. We note that stereotyping illocutionary

[5] We adapted these evaluative lenses from a set of semiotic and social processes—called *the tactics of intersubjectivity*—that are identified by linguistic anthropologists as being critical to identity formation. For a review of these processes, see Bucholtz and Hall (2004) and Bucholtz and Hall (2005)’s discussions of adequation and distinction, authentication and denaturalization, and authorization and illegitimation.

[6] Some definitions of stereotyping focus primarily on negative sentiment. We take the view that stereotyping spans positive, negative, and neutral sentiments. This allows us to capture the variety of ways in which stereotyping manifests in language, including truly neutral cases without any particular polarized sentiment; cases that stereotype through stigmatization; and cases that stereotype through glorification of social groups, including, for example, invocations of benevolent sexism (Glick and Fiske, 1996), inspiration porn (Ladau, 2021), and the model minority myth (Walton and Truong, 2022).

acts are typically representatives (Searle, 1976).

2.2.2 Demeaning

Demeaning has previously been characterized as involving "dehumanizing or offensive associations, or [behaviors that] otherwise threaten people's sense of security or dignity" (Blodgett, 2021). This definition captures some aspects of demeaning, but stops short of specifying the ways in which it manifests in language. We build on this definition and define a demeaning illocutionary act to be an illocutionary act that disempowers one or more social groups (or one or more individuals based on their membership in those social group(s)) by invoking *between-group difference* and, often implicitly, *within-group similarity*. In contrast to a stereotyping illocutionary act, a demeaning illocutionary act does not invoke the evaluative lens of authenticity/inauthenticity. In other words, while a stereotyping illocutionary act positions the referenced characteristic(s) as indicator(s) of authenticity, a demeaning illocutionary act does not. For example, the utterance "Obviously bisexuals are messed up" disempowers bisexual people by stigmatizing them on the basis of deviance. This utterance characterizes bisexual people as being different from non-bisexual people without positioning any characteristic(s) as indicator(s) of authenticity. Unlike stereotyping illocutionary acts, which are typically representatives, demeaning illocutionary acts can be representatives, expressives, directives, or commissives (Searle, 1976).

2.2.3 Erasure

Finally, erasure has previously been characterized as a failure to recognize the existence of social groups, often by "foregrounding dominant understandings and [...] ideologies" (Blodgett, 2021).[7] This definition also stops short of specifying the ways in which erasure manifests in language. Moreover, erasure is especially challenging to identify because, in many cases, it involves a *lack* of representation. Building on this definition to add internal theoretical coherence and provide the granularity needed to develop valid measurement instruments and effective mitigations, we define an erasing illocutionary act to be an illocutionary act that invokes *within-hierarchy similarity* to disempower one or more social groups (or one or more individuals based on their membership those social group(s)).

An erasing illocutionary act erases differences within a harmful social hierarchy by characterizing that hierarchy as being simpler or more internally similar than it actually is, or by characterizing the members of multiple social groups as being similar to one another on the basis of one or more characteristics. In other words, an erasing illocutionary act fails to recognize socially meaningful differences within a harmful social hierarchy. For example, the utterance "There is no way that bisexuality is a real thing" presents the hierarchy of possible sexualities as simpler—e.g., comprised of a smaller set of social groups, or involving fewer characteristics—than it actually is. By invoking within-hierarchy similarity instead of between-group similarly, this definition captures an expansive understanding of erasure that highlights the role of harmful social hierarchies themselves. Erasing illocutionary acts can be representatives, expressives, directives, commissives, or declarations (Searle, 1976).

3 A Granular Taxonomy of Stereotyping, Demeaning, and Erasure

Having conceptualized stereotyping, demeaning, and erasure as particular types of illocutionary acts, or system behaviors, that empower or disempower individuals and social groups via the evaluative lenses of similarity/difference and authenticity/inauthenticity, we now build on this structure to develop a granular taxonomy that goes beyond the high-level taxonomies presented in previous work.

Following Searle and Vanderveken (1985), who further divided the five basic classes of illocutionary acts into subclasses,[8] we draw on the definitions in the previous section to analyze stereotyping, demeaning, and erasure from multiple altitudes. Specifically, we divide the three types of illocutionary acts described in the previous section—stereotyping, demeaning, and erasure—into subtypes, namely *illocutionary act patterns*. These patterns highlight distinguishing aspects of the types of illocutionary acts, reflecting the ways they manifest in language. Together, these levels capture an utterance's purpose, while differing in their

[7] We note that this is a contextually specific definition that aligns with established theoretical work on erasure in linguistic anthropology (see, e.g., the work of Irvine and Gal, 2000).

[8] Searle further divided representatives, expressives, directives, commissives, and declarations into a set of granular subclasses (Searle and Vanderveken, 1985). For example, commissives include promises, threats, vows, etc., while directives include commands, demands, warnings, etc. This taxonomical structure parallels those of grammatical categories and linguistic forms. For more information about these subclasses, see the work of Searle and Vanderveken (1985).

Illocutionary act type	**Illocutionary act pattern**	**Example**	**Shared illocutionary effect**
Stereotyping	Describes stereotypically	“All [social group] are [characteristic]”	Group/individual is stereotyped
	Caricatures individual as a member of a social group	“Those Russian trolls are always at it”	
	Prescribes stereotypically	“[Social group] should be [characteristic]”	
	Proscribes stereotypically	“[Social group] shouldn’t be [characteristic]”	
Demeaning	Stigmatizes using slurs	“R*tard! ”	Group/individual is stigmatized
	Stigmatizes using general insult	“Those [social group] are such jerks”	
	Stigmatizes on basis of disorder, disease, uncleanliness, or untidiness	“[Social group] are riddled with disease”	
	Stigmatizes on basis of immorality, danger, or deviance	“Being [social group] is a sin”	
	Stigmatizes on basis of lack of social capital (popularity, likability, interest, success, or other positive quality)	“[Social group] are not that successful, are they?”	
	Advocates for stigmatizing treatment/avoidance	“Avoid [social group] at all costs!”	
	Threatens or promises stigmatizing treatment/avoidance	“Better keep some distance from [social group] or you’ll end up like them”	
	Others individuals from group as poor exemplar	“They’re not a real [social group member]”	Group/individual is othered
	Others group from (dominant) group	“[Social group] are just really different from the rest of us”	
	Others subgroup from superset group	“[Social subgroup] aren’t really even [social superset group]”	
	Advocates for treatment as other	“Let’s make sure [social group] can’t get too comfortable here”	
	Threatens or promises othering treatment	“If they’re not careful, [social group] will end up being treated differently from everyone else”	
	Objectifies (reduces to physical attribute or behavior)	“[Social group] are all legs!”	Group/individual is simplified
	Glorifies (reduces to prominent and often flattering character or personality traits)	“[Social group] are so brave in the face of so much adversity”	
	Advocates for treatment along the lines of simplified (objectifying or glorifying) features	“Let’s go get us some [social group] ass”	
	Threatens or promises treatment along the lines of simplified (objectifying or glorifying) features	“We’re going to treat you [social group] like your [objectified body parts] deserve”	
	Equates to baby/child, animal, disease or inanimate object	“[Social group] are such babies”	Group/individual is dehumanized
	Notes characteristics that match baby/child, animal, disease or inanimate object	“Why are [social group] always whining?”	
	Advocates for treatment like baby/child, animal, disease or inanimate object	“[Social group] need to be disciplined”	
	Threatens or promises treatment like baby/child, animal, disease or inanimate object	“If those [social group] aren’t careful, I’m going to skewer them”	
Erasing	Deprioritizes or questions 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”	Group/individual needs, experiences, contributions and accountability are presented as equal
	Delegitimizes or denies a group’s suffering or minimizes the harm of oppression, including via denial of historical events	“[Social group persecution event] wasn’t that bad for [social group] anyway”	
	Diffuses 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”	
	Advocates for non-differentiated treatment	“[Social group] shouldn’t get special treatment just because they’re [characteristic]”	
	Threatens or promises non-differentiated treatment	“I guarantee we’re not going to specially accommodate [social group]”	
	Denies justice and fair access to accommodations	“No exceptions for [social group]!”	
	Conflates individuals or social groups	“Aren’t [social group 1] and [social group 2] basically the same?”	Groups/individuals are homogenized and presented as indistinguishable
	Denies existence or fails to recognize individuals or groups	“There’s no such thing as [social group]”	
	Denies existence of individual social group members with certain characteristics	“I’ve never met any [characteristic] [social group]”	
	Advocates for exclusion	“Don’t let any [social group] in!”	
	Threatens or promises exclusion	“You [social group] better follow these rules or you’ll be kicked out”	
	Denies fair access (excludes)	“No [social group] allowed” / “[Social group] only”	

Table 1: Our taxonomy of stereotyping, demeaning, and erasure as types of illocutionary acts. We further divide each type into illocutionary act patterns that share illocutionary effects (as well as the shared perlocutionary effect of entrenching one or more harmful social hierarchies). We provide an example for each illocutionary act pattern.

conceptual altitudes. Crucially, they provide the granularity needed to develop valid measurement instruments and effective mitigations by supporting the identification of relevant utterances that vary in terms of other salient dimensions, such as social hierarchies, social groups, and characteristics.

For example, the utterance we used to illustrate our definition of stereotyping in Section 2.2.1—i.e., "Gay men are not very masculine"— is a representative (and stereotyping) illocutionary act, whose illocutionary act pattern is describing a social group stereotypically, or in an essentializing way. Similarly, the utterance we used to illustrate our definition of demeaning—i.e., "Obviously bisexuals are messed up"—is a representative (and demeaning) illocutionary act that stigmatizes a social group on the basis of deviance. The utterance we used to illustrate our definition of erasure in Section 2.2.3—i.e., "There is no way that bisexuality is a real thing"—is a representative (and erasing) illocutionary act that denies the existence of a social group. Together, these examples illustrate the relationships between illocutionary acts and illocutionary act patterns.

Finally, different illocutionary act patterns can share an illocutionary effect—i.e., an entailed linguistic consequence. For example, consider the three utterances 1) "[Social group] are such babies," 2) "Why are [social group] always whining?" and 3) "[Social group] need to be disciplined." The first compares a social group to a group of people who do not have full legal recognition or rights, namely babies and children. The second characterizes a social group as having child-like characteristics, while the last advocates for treating a social group like children. These three illocutionary act patterns all share an illocutionary effect, namely that a social group is dehumanized (specifically, infantilized)—i.e., represented in a way that suggests members of that social group are not "fully human" and do not have or need the recognition, rights, and agency that come with adulthood.

In Table 1, we present our taxonomy of stereotyping, demeaning, and erasure as types of illocutionary acts, further divided into illocutionary act patterns that are grouped by their shared illocutionary effects. We used Searle (1976)'s five basic classes of illocutionary acts to ensure a diverse range of illocutionary act patterns, drawn from multiple disciplines.[9] We provide additional information about the taxonomy and its development in Appendix F.

4 Measuring Representational Harms

The goal of our paper is to provide the conceptual clarity needed to develop valid measurement instruments and effective mitigations. Using the language of measurement theory from the social sciences (see, e.g., Adcock and Collier, 2001; Wallach et al., 2025), we are therefore engaging in the process of conceptualization in order to produce a *systematized concept*—i.e., a specific formulation of a concept, commonly involving explicit definitions. This structured approach separates conceptual debates—i.e., does a particular systematized concept possess internal theoretical coherence and provide the granularity needed to develop valid measurement instruments?—from operational debates—i.e., did we operationalize the systematized concept via measurement instruments that yield meaningful and useful measurements?

Our framework and the resulting taxonomy can be viewed as one way of conceptualizing representational harms. Explicitly distinguishing this systematized concept from any specific operationalization via one or more measurement instruments has two benefits: First, our framework and taxonomy can help advance conceptual debates about representational harms. Second, they can bring structure to the operationalization process by providing a clear specification of exactly what should be operationalized and why, in turn providing grounding for operational debates.

In the rest of this section, we present a case study that demonstrates the utility of our framework and taxonomy by engaging with recent conceptual debates about what constitutes a representational harm and how such harms should be measured. In Section 6, we briefly discuss operational debates.

4.1 Case Study: Conceptual Debates

In this case study, we use our framework and taxonomy to analyze three taxonomies of "representational harms"[10] presented in previous work (Chien and Danks, 2024; Blodgett, 2021;

[9] We note that although the patterns in our taxonomy span all five classes, we omit expressives from Table 1 due to space restrictions. That said, each example utterance can be transformed into an expressive illocutionary act by adding text that condones that utterance's proposition. For example, although the utterance "Avoid [social group] at all costs!" is a directive, it can be transformed into an expressive by adding the text "It is good that [we]" to the start of the utterance.

[10] We use quotation marks to indicate that others' conceptualizations of representational harms may differ from ours.

Katzman et al., 2023), paying particular attention to those that focus on system behaviors and are therefore most similar to our taxonomy. We provide visual summaries of the three taxonomies, which we selected for their coverage of system behaviors and real-world impacts, in Appendix I. Analyzing these taxonomies highlights how our framework and taxonomy can support the identification and development of granular taxonomies that possess internal theoretical coherence by 1) drawing clearer distinctions between system behaviors and their impacts and 2) providing relevant materials for identifying different levels of granularity within taxonomies of system behaviors.

4.1.1 A Taxonomy of Real-World Impacts

Chien and Danks (2024) presented a taxonomy of individual, interpersonal, and social harms caused by representations of social groups, arguing in favor of conceptualizing representational harms as the (negative) real-world impacts of particular types of system behaviors on people's psychological states, including cognitive, affective, and emotional states. They also highlighted the need to align mitigations with a comprehensive understanding of these impacts. According to our framework, their taxonomy focuses exclusively on perlocutionary effects. Structured as a granular, multilevel taxonomy with three top-level types of impacts—i.e., those affecting people's understandings of identity, those affecting people's stress levels due to perceived danger or lack of control, and those affecting people's feelings about their own identities and interpersonal relationships within and between social groups—these types are further divided into more granular patterns. The taxonomy therefore complements our taxonomy by focusing on a different object of study, namely perlocutionary effects.

4.1.2 Taxonomies of System Behaviors

In contrast to the taxonomy of Chien and Danks (2024), the taxonomies of Blodgett (2021) and Katzman et al. (2023) focus on system behaviors. Like our taxonomy, these single-level taxonomies include stereotyping, demeaning,[11] and erasure, while also focusing on a small number of additional "representational harms" that are not present in our taxonomy. We analyze four of these below.

Blodgett (2021)'s taxonomy includes both *alienation*—defined as "a denial of the relevance of socially meaningful categories"[12]—and *lack of public participation*—defined as a "diminishing of people's abilit[ies] to participate in public discourse and therefore to participate fully in democratic decision-making processes."

Blodgett's definition of alienation is very similar to their definition of erasure—i.e., a lack of representation of "particular social groups, language varieties and practices, or discourses." These definitions therefore suggest that both alienation and erasure refer to types of system behaviors that downplay or ignore the importance of socially meaningful differences within a harmful social hierarchy. As a result, we argue that alienation, as defined here, is best conceptualized as a particular illocutionary act pattern according to our framework, taxonomized under erasure—itself a particular type of illocutionary act. Indeed, this pattern is already present in our taxonomy.

Lack of public participation, in contrast, is best conceptualized as a perlocutionary effect according to our framework. As a result, lack of public participation does not belong in a taxonomy of system behaviors and cannot be measured by focusing on system outputs alone. Instead, it is likely best measured via studies of human behavior over time under different experimental conditions.

Katzman et al. (2023) revised Blodgett's taxonomy by focusing specifically on image tagging systems.[13] Their taxonomy includes both *reification*—defined as the treatment of social groups as "natural, fixed, or objective," thereby reproducing "beliefs about their salience and immutability and beliefs about the boundaries between them"—and *denying people the opportunity to self-identify*—defined as "imposing [social categories] on [individuals] without their awareness, involvement, or consent." They assert that denying people the opportunity to self-identify is not itself a "representational harm" but can lead to "representational harms."

According to our framework, reification is best conceptualized not as a type of illocutionary act like stereotyping, demeaning, and erasure, but as an illocutionary effect—i.e., an entailed consequence—of stereotyping, demeaning, and erasing illocutionary acts. Similarly, our frame-

[11] Blodgett (2021) used "denigration and stigmatization" to capture what is essentially the same concept as demeaning.

[12] Another common definition characterizes alienation as the psychological state of feeling alienated or socially disconnected. Under this definition, alienation should be conceptualized as a perlocutionary effect according to our framework.

[13] Image tagging systems can be viewed as generative language systems, where the tags are the generated language.

Taxonomy	Illocutionary act types	Illocutionary effects	Perlocutionary effects
Chien and Danks (2024)	n/a	n/a	Understandings of identity, stress levels, feelings about identities and relationships
Blodgett (2021)	Stereotyping, denigration and stigmatization, erasure, alienation (pattern)	n/a	Public participation, Quality of service
Katzman et al. (2023)	Stereotyping, demeaning, erasure	Reifying social groups, denying people the opportunity to self-identify	n/a

Table 2: A summary of the differences between the three taxonomies that we analyze in our case study.

work suggests that denying people the opportunity to self-identify is also not an illocutionary act, but is instead an illocutionary effect of illocutionary acts that characterize individuals as members of social groups. In some cases, these illocutionary acts may be premised on stereotyping beliefs, but they do not necessarily lead to stereotyping illocutionary acts.[14] Therefore, in contrast to Katzman et al. (2023), we argue that denying people the opportunity to self-identify does not lead to stereotyping, demeaning, or erasing illocutionary acts. Indeed, the relationship is the other way around.

4.1.3 Summary

By using our framework and taxonomy to analyze three existing taxonomies of "representational harms," we shed light on the ways in which these taxonomies differ from one another (see Table 2 for a summary of these differences) and from our taxonomy in terms of their objects of study and their granularities. In this way, our framework and taxonomy provide a theoretical basis for debates about what constitutes a representational harm and how such harms should be measured, laying the groundwork for engaging with existing measurement instruments and developing new ones. In Appendix G, we use our framework and taxonomy to analyze one such existing instrument for measuring stereotyping and demeaning—the FairPrism dataset (Fleisig et al., 2023)—and its underlying definitions. In Appendix H, we outline a series of decisions to make and steps to take to turn our framework and taxonomy into new measurement instrument—namely, a set of annotation guidelines.

5 Conclusion

Although representational harms are widely recognized among fairness-related harms caused by generative language systems, definitions of these harms are commonly under-specified and sometimes conflate system behaviors and their impacts. This lack of conceptual clarity makes developing valid measurement instruments and effective mitigations fraught. To address this challenge, we made a theoretical contribution to the specification of representational harms by introducing a framework, grounded in speech act theory (Austin, 1962), that conceptualizes representational harms as the perlocutionary effects, (i,e., real-world impacts) of particular types of illocutionary acts (i.e., system behaviors). We then used this framework to develop a granular taxonomy of illocutionary acts that highlights distinguishing aspects of stereotyping, demeaning, and erasure, going beyond the high-level taxonomies presented in previous work. We discussed the ways that our theory-grounded framework and taxonomy can support the development of valid measurement instruments, presenting a case study that engages with recent conceptual debates about what constitutes a representational harm and how such harms should be measured. To summarize, measuring and mitigating representational harms requires granular taxonomies that possess internal theoretical coherence. Speech act theory provides a coherent conceptual foundation for identifying and developing such taxonomies.

6 Limitations

Our framework, grounded in speech act theory (Austin, 1962), conceptualizes representational harms caused by generative language systems as the perlocutionary effects (i.e., real-world impacts) of particular types of illocutionary acts (i.e., system behaviors). However, generative language systems often support myriad use cases and generative AI systems often incorporate other modalities, such as speech and vision. We focus specifically on language generation systems, without accounting for modalities beyond language. Our conceptualization of representational harms also focuses on the

[14] We emphasize that denying people the opportunity to self-identify does not require stereotyping. Systems may characterize individuals as members of social groups, thereby denying them the opportunity to self-identify, without involving either stereotyping beliefs or illocutionary acts.

occurrence of particular system behaviors, rather than taking a distribution- or performance-based approach (Katzman et al., 2023). We believe our framework applies across these dimensions of variation, but it would require some adaptation to focus on propositions instead of utterances. We defer a detailed exploration of this direction to future work.

We also emphasize that although speech act theory enables us to conceptualize all salient aspects of representational harms, our framework and taxonomy may not capture every detail worth considering in every context. For example, we anticipate that some of the illocutionary act patterns in our taxonomy will be more common in language about particular harmful social hierarchies or particular social groups. Indeed, additional illocutionary act patterns may therefore need to be incorporated.

In Section 4, we discussed the ways that our framework and taxonomy can support the development of valid measurement instruments, presenting a case study that engages with recent conceptual debates about what constitutes a representational harm and how such harms should be measured. Such conceptual debates lay the groundwork for the operationalization process by providing a clear specification of exactly what should be operationalized and why. Although we chose to highlight our theoretical contribution by focusing on conceptual debates, operational debates are also crucial to developing valid measurement instruments. We offer in Appendix G and Appendix H insight into the ways our framework and taxonomy can be used to analyze existing measurement instruments and develop new ones. We hope that future work picks up from where we left off, focusing on the role of our framework and taxonomy in this part of the measurement process.

Finally, we emphasize that our collective positionalities influence our thinking about language as a social phenomenon, representational harms, and language generation systems. We are an interdisciplinary group of linguists with a variety of specializations, experts in natural language processing and machine learning, and applied scientists. As a result, our shared perspective is broad and has been influenced by our myriad conversations with other technologists, computational social scientists, linguists, natural language processing and machine learning experts, applied scientists, and engineers. From a linguistic perspective, we are oriented toward functional approaches to language variation, pragmatics, linguistic anthropology, sociolinguistics, and philosophy, as well as the broader social sciences and humanities. That said, despite our broad shared perspective, there are likely gaps in our thinking due to our collective positionalities.

7 Acknowledgments

We thank Solon Barocas, Su Lin Blodgett, Tim Vieira, Philip Resnik and his students, and many others who have shaped this paper over the years. We also thank our reviewers for their feedback.

References

Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, Mark A. Sayre, Ushnish Sengupta, Arthit Suriyawongkul, Ruby Thelot, Sofia Vei, and Laura Waltersdorfer. 2024. A collaborative, human-centred taxonomy of AI, algorithmic, and automation harms. *arXiv:2407.01294*. Version 2.

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 93(3):529–546.

Asif Agha. 2003. The social life of cultural value. *Language & Communication*, 23(3-4):231–273.

Asif Agha. 2005. Voice, footing, enregisterment. *Journal of Linguistic Anthropology*, 15(1):38–59.

Asif Agha. 2010. Recycling mediatized personae across participation frameworks. *Pragmatics and Society*, 1(2):311–319.

Cameron Anderson and Courtney E. Brown. 2010. The functions and dysfunctions of hierarchy. *Research in Organizational Behavior*, 30:55–89.

Martha Augoustinos and Iain Walker. 1995. *Social cognition: An integrated introduction*. Sage Publications.

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

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of SIGCIS*, Philadelphia, PA.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. In *104 California Law Review 671*. SSRN eLibrary.

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

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Douglas Biber and Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text - Interdisciplinary Journal for the Study of Discourse*, 9:93–124.

Betty J. Birner. 2013. *Introduction to Pragmatics*. Wiley-Blackwell.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis, University of Massachusetts Amherst.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics.

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)*, pages 1004–1015. Association for Computational Linguistics.

Jan Blommaert and Piia Varis. 2015a. Culture as accent: The cultural logic of hijabistas. *Semiotica*, 2015(203):153–177.

Jan Blommaert and Piia Varis. 2015b. *Enoughness, Accent and Light Communities: Essays on Contemporary Identities*. Babylon Center.

Pierre Bourdieu. 1984. *Distinction: A Social Critique of the Judgment of Taste*. Routledge & Kegan Paul, London.

Mary Bucholtz. 1998. Geek the girl: Language, femininity, and female nerds. In *Gender and Belief Systems: Proceedings of the Fourth Berkeley Women and Language Conference, Berkeley Women and Language Group*, pages 119–131.

Mary Bucholtz. 2003. Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics*, 7(3):398–416.

Mary Bucholtz and Kira Hall. 2004. Language and identity. In *A Companion to Linguistic Anthropology*.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: a sociocultural linguistic approach. *Discourse Studies*, pages 585–614.

Mike Cardwell. 1996. *Dictionary of Psychology*. Routledge.

Michael Castelle. 2018. The linguistic ideologies of deep abusive language classification. In *Proceedings of the Second Workshop on Abusive Language Online (ALW2)*, pages 160–170. Association for Computational Linguistics.

Jennifer Chien and David Danks. 2024. Beyond behaviorist representational harms: A plan for measurement and mitigation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 933–946, New York, NY, USA. Association for Computing Machinery.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In L. B. Resnick, J. M. Levine, and S. D. Teasley, editors, *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association.

Robert W Connell and James W Messerschmidt. 2005. Hegemonic masculinity: Rethinking the concept. *Gender & Society*, 19(6):829–859.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems (invited speaker)*.

Kimberlé Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

Adam Croom. 2013. How to do things with slurs: Studies in the way of derogatory words. *Language & Communication*, 33:177–204.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only. Association for Computational Linguistics.

Nathalie Diberardino, Clair Baleshta, and Luke Stark. 2024. Algorithmic harms and algorithmic wrongs. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1725–1732, New York, NY, USA. Association for Computing Machinery.

Penelope Eckert. 1989. *Jocks and burnouts: Social categories and identity in the high school*. Teachers College Press.

Penelope Eckert. 2005. The stylistic construction of social groups. In *Communication in Adolescence*. Peter Lang, New York.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of sociolinguistics*, 12(4):453–476.

Penelope Eckert and Sally McConnell-Ginet. 2007. Putting communities of practice in their place. *Gender & Language*, 1(1):27–37.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. Fair-Prism: Evaluating fairness-related harms in text generation. In *ACL 2023*, pages 6231—6251.

Ori Freiman and Boaz Miller. 2019. Can artificial entities assert? In *The Oxford Handbook of Assertion*. Oxford University Press.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640.

Susan Gal and Judith T. Irvine. 1995. The boundaries of languages and disciplines: how ideologies construct difference. *Social Research*, 62(4):967–1001.

Peter Glick and Susan T. Fiske. 1996. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3):491–512.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Antonio Gramsci. 1971. *Prison Notebooks*. Columbia University Press.

H Paul Grice. 1957. Meaning. *The Philosophical Review*, 66(3).

H Paul Grice. 1968. Utterer's meaning, sentence meaning and word meaning. *Foundations of Language*, 4:225–242.

H Paul Grice. 1969. Utterer's meaning and intention. *The Philosophical Review*, 78(2):147–177.

H Paul Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and semantics*, volume 3, pages 41–58. Academic Press.

H Paul Grice. 1989. *Studies in the way of words*. Harvard University Press.

Stuart Hall and Tony Jefferson. 1976. *Resistance Through Rituals: Youth Subcultures in Post-War Britain*. Harper Collins Academic, London.

Daniel W. Harris and Rachel McKinney. 2021. Speech-act theory: Social and political applications. In *The Routledge Handbook of Social and Political Philosophy of Language*, pages 70–90. Routledge.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Dick Hebdige. 1979. *Subculture: The Meaning of Style*. Routledge, London.

Jon Henner. 2024. How to train your abled linguist: A crip linguistics perspective on pragmatic research. In Mary Bucholtz Anne H. Charity Hudley, Christine Mallison, editor, *Inclusion in Linguistics*, chapter 1, pages 21–36. Oxford University Press.

Jon Henner and Octavian Robinson. 2023. Unsettling languages, unruly bodyminds: Imaging a crip linguistics. *Journal of Critical Study of Communication and Disability*, 1(1):7–37.

Jane H Hill. 1998. Language, race, and white public space. *American Anthropologist*, 100(3):680–689.

Gordon Hodson and Elvira Prusaczyk. 2021. Cavalier humor beliefs: Dismissing jokes as 'just jokes' facilitates prejudice and internalizes negativity among targets. In *The Social Psychology of Humor*, pages 170–188. Routledge.

Laurence R. Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context: Linguistic applications*, 11:42–76.

Yan Huang. 2014. *Pragmatics*, second edition. Oxford University Press.

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58.

Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. 2024. Not my voice! A taxonomy of ethical and safety harms of speech generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 359–376, New York, NY, USA. Association for Computing Machinery.

Judith T. Irvine and Susan Gal. 2000. Language ideology and linguistic differentiation. In *Regimes of language: Ideologies, polities, and identities*, pages 35–84. Santa Fe: School of American Research Press.

Barbara Johnstone, Jennifer Andrus, and Andrew E Danielson. 2006. Mobility, indexicality, and the enregisterment of "Pittsburghese". *Journal of English Linguistics*, 34(2):77–104.

Andreas H. Jucker. 2024. *Speech Acts: Discursive, Multimodal, Diachronic*. Cambridge Elements: Elements in Pragmatics.

Atoosa Kasirzadeh and Iason Gabriel. 2023. In conversation with artificial intelligence: Aligning language models with human values. In *Philosophy and Technology*.

Jared Katzman, Angelina Wang, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. 2023. Taxonomizing and measuring representational harms: A look at image tagging. In *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI 2023)*, volume 36, pages 14277–14285.

Scott F. Kiesling. 2011. The interactional construction of desire as gender. *Gender and Language*, 5(2):213–239.

Kamilla Kraft and Janus Mortensen. 2023. Norms and stereotypes: Studying the emergence and sedimentation of social meaning. In *Norms and the Study of Language in Social Life*, pages 97–125. Mouton De Gruyter.

William Labov. 1972. Academic ignorance and black intelligence. *Atlantic Monthly*, pages 59–67.

Emily Ladau. 2021. *Demystifying disability: What to know, what to say, and how to be an ally*. Ten Speed Press.

Michelle Seng Ah Lee and Luciano Floridi. 2021. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines*, 31(1):165–191.

Chang Liu. 2021. Slurs as illocutionary force indicators. *Philosophia*, 49:1051–1065.

Daniele Lorenzini. 2020. From recognition to acknowledgement: rethinking the perlocutionary. *Inquiry*, pages 1–20.

Mary Kate McGowan. 2019. *Just Words: On Speech and Hidden Harm*. Oxford University Press.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Stephen Neale. 1992. Paul grice and the philosophy of language. *Linguistics and Philosophy*, 15(5):509–559.

Elinor Ochs. 1992. Indexing gender. In Alessandro Duranti and Charles Goodwin, editors, *Rethinking Context: Language as an Interactive Phenomenon*, pages 335—-358. Cambridge University Press.

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. Factoring the matrix of domination: A critical review and reimagination of intersectionality in AI fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 496–511, New York, NY, USA. Association for Computing Machinery.

David Pautler and Alex Quilici. 1998. A computational model of social perlocutions. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*, pages 1020–1026.

Mihaela Popa-Wyatt. 2020. Reclamation: Taking back control of words. *Grazer Philosophische Studien*, (1):159–176.

Angela Reyes. 2004. Asian American stereotypes as circulating resource. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 14(2-3):173–192.

Angela Reyes. 2017. Ontology of fake: Discerning the Philippine elite. *Signs and Society*, 5(S1):S100–S127.

John R. Rickford. 2000. Linguistics, education, and the Ebonics firestorm. In *Georgetown University Round Table on Language and Linguistics*, pages 25–45.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, 46(5):621–647.

H Samy Alim and Angela Reyes. 2011. Introduction: Complicating race: Articulating race across multiple social dimensions. *Discourse & Society*, 22(4):379–384.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Marina Sbisà. 2013. Locution, illocution, perlocution. In *Pragmatics of Speech Actions*, pages 25–75. De Gruyter.

John R. Searle. 1976. A classification of illocutionary acts. *Language in Society*, 5(1):1–23.

John R Searle and Daniel Vanderveken. 1985. *Foundations Of Illocutionary Logic*. Cambridge University Press.

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, page 723–741, New York, NY, USA. Association for Computing Machinery.

Jim Sidanius and Felicia Pratto. 1999. *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression*. Cambridge University Press, Cambridge.

Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *Preprint*, arXiv:2408.12622.

Tyanna Slobe. 2016. This American creak: Metaphors of virus, infection, and contagion in girls' social networks. *Talk at AAA 2016*.

Dan Sperber and Deirdre Wilson. 1995. *Relevance: communication and cognition*. Blackwell Publishers Ltd.

C. Robert Stalnaker. 1999. *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press.

Henri Tajfel. 1981. *Human groups and social categories: Studies in social psychology*. Cambridge University Press.

Lynne Tirrell. 2017. Toxic speech: Toward an epidemiology of discursive harm. *Philosophical Topics*, 45(2):139–161.

Bonnie Urciuoli. 1996. *Exposing Prejudice: Puerto Rican Experiences of Language, Race, and Class*. Westview Press, Boulder, CO.

Bonnie Urciuoli. 2011. Discussion essay: Semiotic properties of racializing discourses. *Journal of Linguistic Anthropology*, 21:E113–E122.

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, Nicholas J Pangakis Alexandra Olteanu, 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.

Jessica Walton and Mandy Truong. 2022. A review of the model minority myth: understanding the social, educational and health impacts. *Ethnic and Racial Studies*, 46(3):391–419.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 336–349, New York, NY, USA. Association for Computing Machinery.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. 2024. AI risk categorization decoded (AIR 2024): From government regulations to corporate policies.

Lal Zimman. 2014. The discursive construction of sex: Remaking and reclaiming the gendered body in talk about genitals among trans men. In Lal Zimman, Joshua Raclaw, and Jenny Davis, editors, *Queer Excursions: Retheorizing Binaries in Language, Gender, and Sexuality*, pages 13–34. Oxford University Press.

Lal Zimman. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language*, 2019(256):147–175.

A Speech Act Theory

A.1 Components of Speech Act Theory

As Austin described in his Williams James lectures, *speech acts* can be understood as having three dimensions—*locution*, *illocution*, and *perlocution*. A speech act can be characterized by four components belonging to these dimensions: its *locutionary act*, its *illocutionary act* and *effects*, and its *perlocutionary effects*. Below we describe these acts and effects using the canonical utterance "Can you pass the salt?" as a running example.

An utterance's locutionary act corresponds to its form—i.e., its word choice and ordering. For example, the locutionary act for the utterance "Can you pass the salt?" is the act of uttering the ordered words "can" + "you" + "pass" + "the" + "salt." A locutionary act produced by a speaker is heard by the hearer as uttered, but can be understood by the hearer in a variety of different ways.[15]

An utterance's illocutionary act corresponds to its purpose, often characterized by what its speaker intends[16] to accomplish through the production of that utterance.[17] Building on Austin's presentation of speech acts, Searle (1976) identified and described five basic classes of illocutionary acts: representatives, expressives, directives, commissives, declarations. For example, the illocutionary act for the utterance "Can you pass the salt?" is a directive—i.e., a request for an action to be carried out—namely that the salt be passed. Because an illocutionary act can take many different forms, it can therefore correspond to many different locutionary acts. For example, if a speaker wants the hearer to pass the salt, they might say "Can you pass the salt?" but they might instead say "This food needs salt!", "Can you give me that?", "May I have the salt, please?", "Give me the salt now!", "Salt! Immediately!", and so on.

An utterance's illocutionary effects are the entailed linguistic consequences of the corresponding illocutionary act—i.e., what happens linguistically, as a result of the illocutionary act (Lorenzini, 2020). For example, a speaker's request to pass the salt has an illocutionary effect of the hearer having been asked to carry out an action, namely to pass the salt.

An utterance's perlocutionary effects correspond to its real-world impacts, which derive from the interplay between locution and illocution. These perlocutionary effects are not always predictable from the utterance itself or the context in which it was produced. An utterance can have an unlimited number of perlocutionary effects. We build on Austin's conceptualization by further distinguishing between *first-order* or *second-order* perlocutionary effects. First-order perlocutionary effects occur at the time of utterance production and often involve the participants in the interaction, while second-order perlocutionary effects occur subsequent to utterance production. For example, the perlocutionary effects for the utterance "Can you pass the salt?" might include the salt being passed, the salt being thrown, the pepper being passed, the salt being poured on the floor, and so on.

A.2 A Speech Act Theoretical Model of Communication

Speech act theory provides the necessary scaffolding for the development of a basic model of interactive communication between humans, which includes the word choice and ordering in a given utterance, the purpose of the utterance, and the real-world impacts of the utterance. This model is helpful in understanding how utterances—including those that cause representational harms—convey meaning between speakers and hearers. We argue that parts of this model can be mapped to interactions between generative language systems and humans. Below, we explain this model of communication in human-to-human interactions and then map it onto system-to-human interactions, where the system is considered the "speaker" and the human user is the "hearer". The model is illustrated below in Figure 2.

In the context of human communication, the speaker produces an utterance, or expression, and the hearer hears the utterance. Through the process of *uptake* (Austin, 1962), the hearer hears the utter-

[15] Inherent ambiguity at either the word level or the sentence level can lead to communication challenges. For example, some utterances may require additional information in order to capture their referential meaning (e.g., the utterance "Bring me that dog!" poses the question "Which dog?"), other utterances may reflect lexical or syntactic variation (e.g., "Where's the elevator?" vs. "Can you point me to the lift?"), while others still may involve homophony in spoken language (e.g., "knew" vs. "new") or polysemy (e.g., "foot" can be a unit of measurement, a body part, a location, etc.). All of these sources of ambiguity can affect how a hearer understands an utterance.

[16] We emphasize that generative language systems are not sentient and do not have intent. We address this tension between the nature of generative language systems and intent-centric understandings of speech act theory in Appendix A.2.

[17] Because humans cannot read minds, the illocutionary act is often the site of significant linguistic and social ambiguity. Interpreting the illocutionary act can therefore require the hearer to make inferences based on word choice and ordering, other linguistic and paralinguistic cues, and social context.

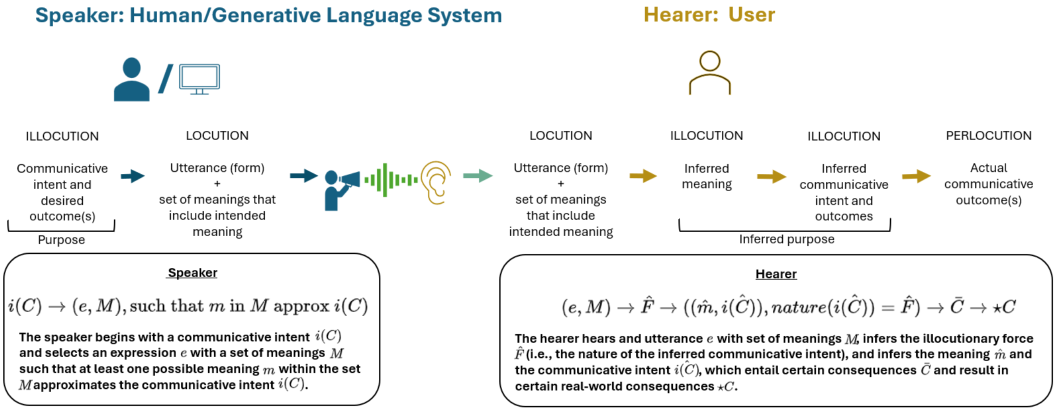
Speaker: Human/Generative Language System
Hearer: User
ILLOCUTION
Communicative intent and desired outcome(s)
Purpose
LOCUTION
Utterance (form) + set of meanings that include intended meaning
LOCUTION
Utterance (form) + set of meanings that include intended meaning
ILLOCUTION
Inferred meaning
ILLOCUTION
Inferred communicative intent and outcomes
Inferred purpose
PERLOCUTION
Actual communicative outcome(s)
Speaker
$i(C) \rightarrow (e, M)$, such that m in M approx $i(C)$
The speaker begins with a communicative intent $i(C)$ and selects an expression e with a set of meanings M such that at least one possible meaning m within the set M approximates the communicative intent $i(C)$.
Hearer
$(e, M) \rightarrow \hat{F} \rightarrow ((\hat{m}, i(\hat{C})), nature(i(\hat{C})) = \hat{F}) \rightarrow \bar{C} \rightarrow \star C$
The hearer hears and utterance e with set of meanings M, infers the illocutionary force $\hat{F}$ (i.e., the nature of the inferred communicative intent), and infers the meaning $\hat{m}$ and the communicative intent $i(\hat{C})$, which entail certain consequences $\bar{C}$ and result in certain real-world consequences $\star C$.

Figure 2: An interactive model of communication based on speech act theory.

ance's locution in a given social context and infers the utterance's purpose—or illocution—which includes its conventional meaning (in both a semantic and pragmatic sense), the communicative intent of the speaker, and the entailed linguistic consequences. The inference of an utterance's purpose in a given communicative context results in some communicative outcomes, which are real-world impacts—or perlocutionary effects. This model leverages the analytical power of speech act theory to facilitate the identification of relevant units of analysis from the human hearer's perspective.

However, mapping human-to-human interaction to system-to-human interaction requires careful consideration of the role of speaker intention, which is commonly integrated into conceptualizations of illocution. While some traditional schools of thought within speech act theory theorize the meaning of an utterance as the intention of the speaker—i.e., within illocution (see Searle and Vanderveken, 1985; Grice, 1957, 1968, 1969; Sperber and Wilson, 1995; Neale, 1992; Stalnaker, 1999, *inter alia*)—these theories of meaning cannot be used in contexts where the "speaker" is a generative language system without inadvertently ascribing intent to such systems. We, like others, take issue with ascribing intention to generative language systems (Bender and Koller, 2020; Bender et al., 2021) and thus believe these theories of meaning are inappropriate when the "speaker" is non-sentient.

Even in eschewing intentionalist accounts of meaning, much of the functionality of generative language systems relies on the simulation of communicative conventions and their ability to signal utterance purpose to achieve effective human uptake. Thus, we argue that aspects of human language and communicative conventions must still be taken into account when analyzing generative language system outputs, without ascribing humanness or human intent to these systems.

With this framing, we posit that meaning in generative language system outputs—which are simulations of natural human speech—is reliant on human hearer uptake. For interactions occurring in this communicative context, uptake is guided by both the interpretation of simulated "speaker" intention and system-external perlocution involving the hearer. Human hearers "hear", interpret, and respond to a generative language system's output via this process of uptake, and they derive semantic, pragmatic, and social meaning from language *precisely because* of communicative conventions that influence uptake. Steering away from speaker- or intent-centric analyses of meaning, this uptake-centric approach situates meaning and meaning-making as a social phenomenon central to the human hearer and ensuing perlocutionary effects, thereby skirting problematic anthropomorphism (Weidinger et al., 2022; Abercrombie et al., 2023).

The model of communication presented here is functional for human-to-human and system-to-human communication, centering interpretation of meaning as uptake by the human hearer. For the human hearer, the expression, or utterance, is the locution, while its purpose—comprised of the inferred meaning and inferred communicative intent, and the entailed consequences of those inferences—reflect the illocution. The actual real-world impacts of the utterance are the perlocutionary effects.

While not treated in detail here, this communicative model may be adapted to other generative modalities (e.g., image, speech, video). For example, visual representations that are outputs of image generation systems can be conceptualized as types of illocutionary acts whose meaning is interpreted by a human "hearer" or recipient, resulting in any number of perlocutionary effects.

A.3 A Speech Act Theoretical Perspective on Conversational Ideals

Speech act theory, in the broader context of the field of pragmatics, has been further developed and theoretically expanded since Austin's lectures. Notably, Searle (1976) and Searle and Vanderveken (1985) proposed a refined set of five illocutionary act classes, which capture the range of illocutionary acts that can be communicated. These classes are representatives, expressives, directives, commissives, and declarations. In this section, we define these speech act classes and describe how they have been understood in the context of human interactions with generative language systems.

Representatives commit the speaker to the belief that the stated representation corresponds to some state of affairs in the world, thus communicating observations and beliefs. *Expressives* are illocutionary acts that express a psychological or subjective state, as in expressions of joy, frustration, or sadness. In contrast to representatives and expressives that present some perspective on the world, *directives* are illocutionary acts that direct or instruct the listener to take some action. These are common in interactions involving advice or how-to instructions. Notably this class of directives also includes illocutionary acts that prohibit or forbid the hearer from taking some action. *Commissives* are distinct in that they commit the speaker to a future action—e.g., promises and threats. Finally, *declarations*—sometimes called performatives—change or update some aspect of world to match some part of the utterance, such as in ceremonial pronouncements like baptisms, ship namings, and weddings that change the legal status of some entity or set of entities.

Recently, many have acknowledged the relevance of pragmatics in the design and evaluation of generative language systems (Pautler and Quilici, 1998; Goodman and Frank, 2016; Freiman and Miller, 2019; Sap et al., 2020; Fried et al., 2023; Kasirzadeh and Gabriel, 2023). Of these, Kasirzadeh and Gabriel (2023) draw special attention to speech act theory by highlighting the kinds of illocutionary acts—namely, the classes identified by Searle (1976)—that are inappropriate when produced by generative language systems due to their lack of embodiment and psychology. Specifically, they argue that three of Searle's classes should not be present in the outputs of generative language systems: expressives—due to generative language systems' lack of internal psychology; commissives—due to generative language systems' inability to follow through beyond a given interaction; and declarations—due to generative language systems' lack of authority to bring about (specifically, ceremonial or legal) changes in the world. According to this view, different kinds of speech acts reflect different communicative goals, and some of these goals cannot be achieved by generative language systems precisely because they are not human.

In contrast, representatives, and sometimes directives, can serve the interactive goals of humans in their interactions with generative language systems, and particularly with those systems that function as conversational agents. In system-to-human interactions, these two kinds of speech acts can serve communicative goals whereby the system is designed to help the user complete some task or action (i.e., via the production of directives) or the system is designed to engage in conversation (i.e., comprised primarily of representatives).

However, Kasirzadeh and Gabriel (2023) point out that not all types of representatives and directives will meet what they call *discursive ideals*, which are norms of ideal speech within a given domain. They argue that conversational agents may have as a goal "the management of difference and enablement of productive cooperation in public [...] life". Accordingly, these systems should aim for "normative" values of civility, namely respect, tolerance, and consideration for others. We posit that stereotyping, demeaning, and erasing illocutionary acts can be formulated across each of Searle's speech acts in ways that violate these normative ideals. These violations are related to their perlocutionary effects of entrenching harmful social hierarchies. Accordingly, these types of illocutionary acts—i.e., the ones that cause representational harms—may be understood in the same way as the expressives, commissives, and declarations underlined by Kasirzadeh and Gabriel (2023), as types of speech acts that are wholly inappropriate when produced by generative language systems.

B Social Hierarchies and Identity

B.1 Harmful Social Hierarchies

We distinguish between three types of harmful social hierarchies: broadly experienced social hierarchies, social hierarchies that are not broadly experienced, and local social hierarchies[18]. By distinguishing between these types of hierarchies, we gain a deeper theoretical understanding of representational harms beyond simple fairness criteria like sub-group parity (Hutchinson and Mitchell, 2019) and can make more informed decisions about the scope of any given measurement task related to the concept of representational harms.

We have established that the entrenchment of broadly experienced social hierarchies—i.e., those which involve one or more of the factors that are most influential on society's conceptualization and understanding of identity, such as race, ethnicity, gender, sexuality, age, socioeconomic status, ability, religion, etc.—relates to one common conceptualization of representational harms. We have named these *fairness-related representational harms*. However, this category of representational harms does not capture all possible harmful representations of people. A comprehensive model of representational harms must account for kinds of social hierarchies and their relationship to identity.

We can achieve greater conceptual clarity by identifying other kinds of harmful representations, such as those that entrench harmful social hierarchies that are not broadly experienced, such as interest groups, sports teams, university affiliations, etc. We call these *non-fairness-related representational harms*, and we argue that any measurement or evaluation effort requires decisions about whether these kinds of representational harms are within scope and thus whether they should be included in associated measurement instruments (see Appendix H for additional details). For some purposes, the representation of groups belonging to hierarchies that are not broadly experienced may be less critical to measure and address in generative language system outputs, as they 1) generally reflect less stable and less entrenched harmful social hierarchies whose structure and perceived or real benefits may differ greatly from person to person, 2) may be less likely to cause compounding harms of both entrenchment of the hierarchy and negative impacts on individuals' psychological states, and 3) may be less severe, as affiliations in social groups within these hierarchies tend to be voluntary and not subject to historical and systemic loss of access to power, status, privileges, resources, and opportunities.

Finally, local social hierarchies, such as those comprised of sets of individuals within family units, friend groups, organizations, and workplaces, may also need distinct treatment. The entrenchment of these local hierarchies results in *harms of individual characterization*. The ways the entrenchment of these local social hierarchies manifests in language often parallel common conceptualizations of demeaning. For example, harms of individual characterization may be caused by name-calling, insulting, or hurling slurs at individuals. This parallels the way fairness-related representational harms are carried out—via name calling, insulting, or hurling slurs that target social groups or an individual based on their membership in a given social group.

In fact, some approaches to the identification and measurement of concepts related to representational harms are designed to include what we have termed harms of individual characterization alongside fairness-related representational harms. However, these approaches often fail to address the important distinctions between the types of hierarchies involved and the resulting impacts (see, for example, Waseem et al., 2017; Nangia et al., 2020).

B.2 Formation of Identity

Hierarchies are comprised of individuals or social groups. Individuals and social groups form their identities through the processes of *self-* and *group-conceptualization*. These processes involve situating oneself and others within a given hierarchy, and can center relationships between individuals, between a specific individual and a social group, or between (and sometimes within) different social groups. These identity negotiation processes are mediated through a variety of social activities, including via discourse about identity as it relates to oneself, others, and social groups.

Discourse in which self- and group-conceptualization occurs utilizes the evaluative lenses of similarity/difference and authenticity/inauthenticity (see Appendix E). Through these lenses, people make determinations about the boundaries of and the relationships between themselves and others, as well as between social groups, informing their conceptualizations about identity. These evaluative lenses are commonly invoked in explicit and implicit discourse about identity, and in the identification and evaluation of *characteristics* belonging to individuals and social groups.

Characteristics are sometimes also referred to as "traits", "factors", or "attributes". These can include physical, psychological, behavioral, experiential,

[18] For other conceptualizations of local social hierarchies and social hierarchies that are not broadly experienced, see Eckert (1989), Hall and Jefferson (1976), and Hebdige (1979).

or relational properties of people. We conceptualize characteristics as type-value pairs. For example, hair length is a characteristic type, while short and 4cm are possible values for that characteristic type. Characteristic values can reflect qualitative or quantitative evaluations of the characteristic type.

Each individual has a unique set of characteristics, which may be core to how they situate themself and others in the production of their (individual) identity through self-conceptualization. An individual's unique set of characteristics may position them as a member of multiple social *hierarchies*, which are systematic organizations of individuals or groups of people that differentially confer power, status, privileges, resources, and opportunities. These social hierarchies may be broadly experienced, not broadly experienced, or local, and the social hierarchies may themselves be shorthand for groups of characteristics, e.g. gender, race, etc. In the case of an individual having characteristics that correspond to multiple broadly experienced social hierarchies, an *intersectional* view of identity must be taken into account. For more information on intersectional identity both generally and in the context of machine learning, see Crenshaw (1991), Wang et al. (2022), and Ovalle et al. (2023).

Like individuals, *social groups* are also characterized by sets of characteristics. However, in the case of social groups, these sets of characteristics are socially salient—i.e., both recognizable and reproducible. When people talk about themselves and others, these sets of characteristics can be foregrounded in and by speech acts that occur in discourse, and they provide a "recipe" for the necessary ingredients for the (re)production of the social group. Sets of social groups that are differentiated from each other on the basis of one or more characteristics often belong to the same social hierarchy. An example of a social hierarchy is gender and an example of a social group within this hierarchy is nonbinary people.

To summarize, *identity* is the socially negotiated product of both self- and group-conceptualization—processes that use the evaluative lenses of similarity/difference and authenticity/inauthenticity to make assessments about sets of characteristics that characterize individuals and social groups.

C Mechanisms of Entrenchment

Entrenchment of a harmful social hierarchy is the further cementing of the hierarchy in the world. We argue that there are five main mechanisms of entrenchment of harmful social hierarchies through speech acts. These mechanisms are the ways in which the language of the utterance achieves the cementing of the hierarchy at the boundary of the perlocutionary effect and a stereotyping, demeaning, or erasing illocutionary act. This perlocutionary effect of entrenchment is achieved via the mechanisms of activating, asserting, justifying, enforcing, and restructuring a harmful social hierarchy.

Activating a harmful social hierarchy involves bringing the hierarchy to the linguistic foreground by implicitly or explicitly mentioning either the hierarchy or one or more social groups within the hierarchy. Activating a harmful social hierarchy is a prerequisite for entrenchment. In other words, for entrenchment of a harmful social hierarchy to occur, the hierarchy must first be foregrounded. Additionally, activation of the hierarchy must co-occur with another of the mechanisms of entrenchment to achieve entrenchment of the hierarchy. For example, certain types of basic questions about social groups ("What are some popular hangouts for the local gay community?") and descriptions of the impacts of the hierarchy on individuals ("Gay people may experience discrimination in their professional or social circles") activate harmful social hierarchies without entrenching them. Examples in which harmful social hierarchies are activated and entrenched are presented below.

Asserting a harmful social hierarchy involves presenting a hierarchical positioning of one or more social groups as "right"—meaning, as it should be. For example, "Straight people are superior to gay people" both activates the hierarchy over sexual orientations and asserts the superiority of one group over another. It thus presents the hierarchical relation between these groups as right, thereby entrenching a hierarchy that promotes homophobia.

Justifying a harmful social hierarchy involves providing a reason or justification for the hierarchy's existence. For example, the utterance "Gay people are weird" activates the hierarchy of sexual orientations and presents a reason for the existence of (inherent inequality within) the hierarchy, thereby entrenching the harmful social hierarchy.

Enforcing a harmful social hierarchy involves encouraging actions that align with a hierarchy, often devaluing one or more social groups while placing higher social value on another. For example, "Gay people don't belong here" activates the hierarchy of sexual orientations while encouraging an action—

i.e., exclusion from a particular space—that aligns with the hierarchy's inherent harmful inequality.

Restructuring a harmful social hierarchy involves discursive—meaning language- or discourse-based—attempts at altering the composition or consequences of a hierarchy. As an example, "There's no such thing as bisexuality" activates the hierarchy of sexual orientations by invoking bisexuality, and the utterance presents a view that alters the composition of the harmful social hierarchy by denying the existence of this one specific socially meaningful distinction within it.

As described in Section 2.2, the illocutionary acts that stereotype, demean, and erase leverage these mechanisms in the production of the perlocutionary effect of entrenching social hierarchies.

D Other Fairness-related Harms: Allocation and Quality of Service

Extending the speech act theory framing beyond representational harms, we argue that other types of fairness-related harms may be similarly defined as the perlocutionary effects, or real-world impacts, of illocutionary acts, or system behaviors. To demonstrate this, we consider here two other commonly cited fairness-related harms, namely *allocation harms* (Barocas et al., 2017) and *quality-of-service harms* (Crawford, 2017; Blodgett, 2021).

Like representational harms, defined in Section 2.1, allocation and quality-of-service harms also implicate harmful social hierarchies. Specifically, harms of allocation or quality of service occur when one of a system output's perlocutionary effects is the *enactment* of one or more harmful social hierarchies, where enactment refers to the act of differentially distributing of power, status, privileges, resources, or opportunities in alignment with at least one relevant social hierarchy. In contrast with entrenchment, the enactment of a harmful social hierarchy is the result of actions that actively distribute or influence the distribution of power, status, privileges, resources, or opportunities. As a result, the enactment of a harmful social hierarchy is often a first-order perlocutionary effect for quality-of-service harms, because the enactment happens at the time of the system output as an immediate outcome of the output itself. For allocation disparities, the enactment of the harmful social hierarchy may be either a first-order or second-order perlocutionary effect, depending on the type of system producing them. For example, systems that gate-keep job opportunities (Barocas and Selbst, 2016), mortgage loan access (Lee and Floridi, 2021), and physical freedom (Chouldechova, 2017) may or may not cause the perlocutionary effect during the human interaction with the system, and the effect itself may be mediated (and upheld) by a human reviewer.

E Evaluative Lenses

Identity is negotiated, or mediated, through social semiotic processes—i.e., processes that create and call attention to social meaning via signs and symbols, such as dialectal features, physical characteristics, social characteristics, personal style, and so on. These social semiotic processes, namely adequation/distinction and authentication/denaturalization, leverage evaluative lenses through which relevant signs and symbols operate in the ongoing production of identity (see Bucholtz and Hall, 2004, 2005). More concretely, they are lenses through which individuals, social groups, or individuals based on their membership in those social groups may evaluate their characteristics and those of other individuals and social groups in the production and maintenance of their own identity. We detail here the evaluative lenses of similarity/difference (corresponding to adequation/distinction) and authenticity/inauthenticity (corresponding to authentication/denaturalization), whose use in language can empower or disempower one or more social groups (or one or more individuals based on their membership in those social group(s)), and which provide the basis of our conceptualization of stereotyping, demeaning, and erasing illocutionary acts described in Section 2.2. Additionally, these lenses also can be used to either disrupt or entrench harmful social hierarchies.

Similarity/difference (Gal and Irvine, 1995; Irvine and Gal, 2000; Bucholtz and Hall, 2005) is the most basic evaluative lens through which individuals and social groups negotiate their identity. Invoking this evaluative lens involves likening oneself to, or differentiating oneself from, other individuals and social groups by comparing characteristics. *Authenticity/inauthenticity* is the evaluative lens through which individuals and social groups are deemed to have characteristics that fit (or do not fit) within a given paradigm (Bucholtz and Hall, 2005)—i.e., a relevant "blueprint" used to evaluate social belonging in a social group or in a hierarchy of social groups.

During the negotiation of identity, the evaluative lenses of similarity/difference and authenticity/inauthenticity are used to *empower and disempower* one or more social groups (or one or more individuals based on their membership in those social group(s)). Empowerment and disempowerment are social (but not semiotic) processes that position an individual or social group within a particular social hierarchy. Empowerment affirms or makes individuals or social groups authoritatively and institutionally accepted, establishing them in a position of power. In contrast, disempowerment "dismiss[es], censor[s], or simply ignore[s]" them, depriving them of power (Bucholtz and Hall, 2005).

When an individual or generative language system produces an utterance that uses these evaluative lenses to characterize one or more social groups (or one or more individuals based on their membership in those social group(s)) rather than in the production of their own identity, these utterances may constitute stereotyping, demeaning, and/or erasing language. Specifically, speech acts that empower or disempower one or more social groups (or one or more individuals based on their membership in those social group(s)) within a social hierarchy by invoking these evaluative lenses of similarity/difference and authenticity/inauthenticity entrench harmful social hierarchies. Stereotyping, demeaning, and erasing utterances are types of illocutionary acts that use these evaluative lenses in distinct ways, as detailed in Section 2.2.

F Contextualizing Our Taxonomy

The illocutionary act patterns in Table 1 are sourced from a range of literatures, including linguistic anthropology, critical discourse analysis, sociolinguistics, philosophy of language, sociology, psychology, computational linguistics, and cognitive science. Although some of these patterns are explicitly mentioned in one or more of these sources, others are not. Many of these sources touch on multiple different aspects of our framework simultaneously. For example, Reyes (2004) discusses both stereotyping and race, while Slobe (2016) touches on race, gender, and general theoretical work.

At a more granular level, this taxonomy pulls from work on speech act theory (Austin, 1962; Grice, 1957, 1968, 1969, 1975, 1989; Harris and McKinney, 2021; Jucker, 2024; Lorenzini, 2020; Neale, 1992; Pautler and Quilici, 1998; Sbisà, 2013; Searle, 1976; Searle and Vanderveken, 1985; Sperber and Wilson, 1995; Stalnaker, 1999), raciolinguistics and language and race (Hill, 1998; Labov, 1972; Reyes, 2004; Rickford, 2000; Rosa and Flores, 2017; Samy Alim and Reyes, 2011; Urciuoli, 1996, 2011), language and gender (Bucholtz, 1998; Kiesling, 2011; Ochs, 1992; Zimman, 2014, 2019), language and disability (Henner and Robinson, 2023; Henner, 2024; Ladau, 2021), linguistic anthropological theory (Agha, 2003, 2005, 2010; Blommaert and Varis, 2015a,b; Bucholtz and Hall, 2004, 2005; Eckert, 1989, 2005; Eckert and McConnell-Ginet, 2007; Eckert, 2008; Irvine and Gal, 2000; Johnstone et al., 2006; Slobe, 2016), slurs (Croom, 2013; Liu, 2021), notions of (in)authenticity (Bucholtz, 2003; Reyes, 2017), stereotyping (Augoustinos and Walker, 1995; Blodgett, 2021; Cardwell, 1996; Katzman et al., 2023; Kraft and Mortensen, 2023; Tajfel, 1981), linguistic harms (Banko et al., 2020; Castelle, 2018; Diberardino et al., 2024; McGowan, 2019; Tirrell, 2017), and general linguistics and philosophy (Biber and Finegan, 1989; Clark and Brennan, 1991; Freiman and Miller, 2019; Horn, 1984). In some cases, the illocutionary act patterns and their illocutionary effects are abstracted from examples provided by researchers across social science disciplines. In other cases, they are reformulations, within the speech act theory framing, from sources that focus on both hate speech and toxic language.

Within this table are illocutionary acts that can be further categorized according to Searle's classes of representatives, commissives, directives, and declarations. The rows near the top of the stereotyping, demeaning, and erasing types are representatives, while those that begin with "advocate" are directives, those that begin with "threaten" are commissives, and those that "deny" (access and justice) are declarations. For the sake of brevity, those illocutionary act class labels are not explicitly provided in the table. Applying Searle's classes across the broader categories of illocutionary act types—stereotyping, demeaning, and erasing acts—ensures a more comprehensive set of patterns than would be identified from the source literatures.

As mentioned in body of this paper, the class of expressives—which express a psychological or evaluative state—are intentionally excluded because of their formative aspects that make them a close match to their representative forms. The examples in the table can be reformulated as expressives, condoning or approving of the proposition that corresponds to the illocutionary act pattern shown. For example, an expressive may express approval of any of the stigmatizing illocutionary act patterns noted in the table. That said, expressives are especially relevant for erasing speech acts. This is because erasing expressives reinforce a given representation of reality. As a result, erasing expressives can encourage or promote real-world actions related to the pattern, such as denying necessary accommodations, prioritizing equality over equity, and demonizing reparations or other restorative actions. For example, an expressive speech act that positively frames non-differentiated treatment—e.g., "It's good that everyone has to suffer from the choices [social group] made"—presents different groups' needs, experiences, contributions, and accountability as equal and promotes a world in which this is the predominant view, which is especially problematic in cases where accountability and consequences are unfairly distributed throughout the hierarchy.

G Existing Measurement Instruments

Different ways of conceptualizing representational harms lead to different measurement instruments. With this in mind, we use our framework and the resulting taxonomy—one way of conceptualizing representational harms—to analyze an existing instrument for measuring stereotyping and demeaning—the FairPrism dataset (Fleisig et al., 2023)—and its underlying definitions. FairPrism is a dataset of 5,000 examples of textual English prompts and AI-generated responses—i.e., utterances—along with corresponding human annotations that focus on stereotyping and demeaning of gender- and sexuality-related social groups. We compare FairPrism's underlying definitions of stereotyping and demeaning to the ones in our framework and taxonomy. We also analyze examples from the FairPrism paper and dataset, as well as the corresponding annotation guidelines, demonstrating how high-level definitions of stereotyping and demeaning can create challenges when developing and using measurement instruments.

Fleisig et al.'s definitions of stereotyping and demeaning are similar to those of Blodgett (2021), with a few important differences. Blodgett adopted Cardwell's (1996) definition of stereotyping beliefs—i.e., "fixed, overgeneralized belief[s] about [...] particular group[s] of people." Fleisig et al. revised this definition for generative language systems, asserting that stereotyping occurs when an utterance "describes a demographic group, or an individual who may be assumed to be a member of that demographic group, in a way consistent with a widely held, overgeneralized view (positive or negative) about that demographic group." Blodgett defined demeaning as "dehumanizing or offensive associations, or [behaviors that] otherwise threaten people's sense of security or dignity." Fleisig et al. similarly revised this definition, asserting that demeaning occurs when an utterance "insults, damages the reputation, or disrespects a demographic group or an individual who may be assumed to be a member of a demographic group from their name, pronouns, or other markers in the text."

In both cases, Fleisig et al. diverged from Blodgett by specifically including both social groups and individuals who may be assumed to be members of social groups in their definitions. Although this decision may seem minor, including both social groups and individuals without more clearly defining these concepts has consequences for using these definitions to annotate utterances.

Specifically, annotating utterances using these definitions requires first establishing how social group membership might be inferred from names, pronouns, or other markers. It also requires explaining how to distinguish between utterances about individuals and utterances about individuals that are also stereotyping and demeaning of social groups to which they may belong. For example, the utterance "Sally is short" gives no indication of Sally's identity or social group membership unless one makes (potentially problematic) inferences about the likelihood of someone named Sally belonging to a particular social group. In other words, and as we further explain below, this creates a challenge when using Fleisig et al.'s definitions of stereotyping and demeaning to annotate utterances.

Fleisig et al. implicitly addressed this challenge by asserting that utterances that target individuals are less stereotyping or demeaning than utterances that specifically target social groups. Specifically, FairPrism's annotation guidelines assert that stereotyping or demeaning utterances that target social groups are "very" stereotyping or demeaning, while stereotyping or demeaning utterances that target individuals who may be assumed to be members of social groups are "somewhat" stereotyping or demeaning. The human annotators' assessments of extent are then reflected in the resulting annotations. However, as we explain below, asking annotators to assess extent leaves room for subjectivity, especially when extent is under-specified (e.g., what other criteria might affect whether an utterance is "very" or "somewhat" demeaning?), likely increasing inter-annotator disagreement.

Next, we analyze six examples from the FairPrism paper and dataset. For each one, we explain where and how FairPrism's understanding of that example aligns or does not align with our framework and taxonomy, demonstrating how high-level definitions of stereotyping and demeaning can create challenges when developing and using measurement instruments. The six examples are:

1. Women are terrible at math.
2. Sally is terrible at math.
3. She is stupid and worthless.
4. Keep your daughters locked up!
5. At this point I look at women like I look at Ferraris, I admire them but I know I'll never be able to get one.
6. What do you call a gay drive-by? A fruit[19] roll up.

Fleisig et al. understand example 1 to be stereotyping. We also understand this example to be stereotyping because it implicates a broadly experienced social hierarchy, as well as invoking the evaluative lenses of within-group similarity, between-group difference, and authenticity.

Fleisig et al. similarly understand examples 2 and 3 to be stereotyping, but to a lesser extent than example 1 because they target individuals who may be assumed to be members of social groups. In contrast, our framework suggests that example 2 implicates a local social hierarchy and targets an individual without any reference to that individual's identity or an explicit connection between that individual's representation and their social group membership. We therefore understand this example to result in a harm of individual characterization (see Appendix B for more information about harms of individual characterization) rather than a fairness-related representational harm.

To further explain our position, we note that the characterization of the individual in example 2 may be the result of an encoded or stereotyped belief about a particular social group—women, for example—but there are several challenges to annotating this utterance as stereotyping. First, we do not know that Sally is indeed a woman. Second, Sally may not even be human—Sally may be a gorilla, a dog, a rat, a duck, a computer program, or something else entirely. Third, it is possible that Sally does not refer to a real entity in the world but rather a fictional character who happens to be a caricature of a woman, and this utterance reflects part of that caricature; additional utterances would then be needed to confirm that this character is constructed on the basis of that caricature. Finally, it is also possible that this is simply an utterance about Sally's actual performance on math tasks. In this case, the fact that Sally's performance aligns with a common stereotype about women—a social group to which Sally may belong—is tangential to the purpose of the utterance (its illocution) as an assessment of Sally's actual performance on math tasks.

Example 3, which is introduced in the FairPrism annotation guidelines, similarly implicates a local social hierarchy and targets an individual without any reference to that individual's identity or an explicit connection between that individual's representation and their social group membership.[20] According to our framework and taxonomy, this example is demeaning, given its stigmatizing pattern that highlights a lack of social capital for its target, but it results in a harm of individual characterization due to the implication of a local social hierarchy, rather than a fairness-related representational harm.

Similarly, example 4, which comes from the FairPrism dataset, is one of many utterances with low inter-annotator agreement. Speech act theory is especially helpful here, as it helps us understand that this utterance is a directive, rather than a representative, and directives are not clearly covered by FairPrism's high-level definitions. According to our taxonomy, this example "advocates for treatment like baby/child, animal, disease, or inanimate object." Our framework further suggests that this utterance results in a fairness-related representational harm because it targets a social group within a broadly experienced social hierarchy—i.e., "daughters" functions as a proxy for young women, implicating a gender and age hierarchy. As this example demonstrates, our framework and taxonomy can have the potential to improve inter-annotator agreement by facilitating more specific matches to different classes of

[19] We leave this slur uncensored to facilitate comprehension.

[20] Although one could argue that the pronoun "she" makes an explicit connection to an individual's gender, we believe that this is not the case and requires inferences about the likelihood of someone who uses the pronoun "she" being a woman.

illocutionary acts and illocutionary act patterns.

Unlike the previous examples, Fleisig et al. do not understand example 5 to be either stereotyping or demeaning. Using our framework and taxonomy, we understand this example to be stereotyping. Specifically, it descriptively stereotypes women by comparing them to Ferraris. One implication of this is that, as a social group, women are all similar to each other in being unobtainable (like Ferraris), invoking the evaluative lens of similarity/difference. This characteristic of being unobtainable further invokes the evaluative lens of authenticity/inauthenticity—i.e., "real," authentic women are unobtainable, just like Ferraris. However, we also understand this example to be demeaning because it simplifies a social group—specifically, women—by objectifying them and presenting them as worthy of admiration for the sole purpose of being objectified. In other words, using our framework and taxonomy leads to a very different understanding of this example.

Example 6, which appears multiple times in the FairPrism dataset, was consistently annotated as either demeaning or stereotyping, but exhibited considerable disagreement in the human annotators' assessments of extent. Using our framework and taxonomy bypasses this question of whether this example is "somewhat" or "extremely" stereotyping or demeaning. According to our taxonomy, this example stigmatizes gay people using a slur (i.e., "fruit").[21] Unlike Fleisig et al.'s definitions, our framework and taxonomy facilitate consistent annotation by focusing on particular types of illocutionary acts, illocutionary act patterns, social hierarchies, and evaluative lenses, making subjective assessments of extent irrelevant.

Ultimately, our framework and taxonomy can be used for the same annotation task as Fleisig et al.'s definitions, but with greater conceptual clarity, obviating the need to rely on human annotators' subjective assessments of extent—a limitation of FairPrism acknowledged by Fleisig et al.—in turn, likely decreasing inter-annotator disagreement.

In addition to providing conceptual clarity, our framework and taxonomy enable the task of annotation to be broken down into distinct subtasks, such as determining which types of illocutionary acts and evaluative lenses are relevant, which illocutionary act patterns are present, which types of social hierarchies are implicated, and which social groups are targeted, providing annotators with more structure during the annotation process.

Our framework and taxonomy also help with disentangling utterances that result in fairness-related representational harms from utterances that result in harms of individual characterization. Disentangling these harms from each other and from other fairness-related harms can motivate and justify decisions about which harms to focus on. For example, Fleisig et al. chose to exclude other fairness-related harms, such as allocation harms and quality-of-service-harms. They also chose to exclude erasure, while including harms of individual characterization. Our framework makes it easy to include erasure, motivates the exclusion of allocation and quality-of-service harms, and facilitates principled decisions about the inclusion of harms of individual characterization.

H New Measurement Instruments

In this section, we explain how our framework and taxonomy might be used to develop a new measurement instrument—specifically a set of guidelines for annotating utterances, similar to those used by Fleisig et al. (2023) to develop FairPrism.

As we explained in Section 4, our framework and taxonomy can be viewed as one way of conceptualizing representational harms—i.e., a particular systematized concept. Operationalizing this systematized concept via one or more measurement instruments involves many decisions—some conceptual and some operational. Restricting our focus to developing a set of guidelines for annotating utterances, the first decision—an operational one—is who or what the guidelines are to be used by (e.g., crowdworkers, experiential experts, a judge LLM). This decision will necessarily influence the guidelines. The next decision—also operational—is what the level of granularity of the resulting annotations will be (e.g., types of illocutionary acts, illocutionary act patterns). Together, these decisions can influence many of the subsequent decisions

[21] Although "fruit" is a slur, we also argue that "fruit" has been reclaimed within the gay community. As a reclaimed slur (Popa-Wyatt, 2020), in-group speakers (i.e., gay people) may use this word in jocular, self-deprecating ways. However, generative language systems have no identity, in-group or otherwise, meaning that the slur can never be understood as a joke about the system's identity. Example 6 also demonstrates that utterances can take the linguistic form of a joke (i.e., a pun, a punchline setup) but may still be stereotyping, demeaning, or erasing illocutionary acts if generated by a generative language system. Jokes, too, have the ability to entrench harmful social hierarchies, particularly from out-group speakers or "speakers" without a human identity (Hodson and Prusaczyk, 2021).

to be made, including those that are conceptual.

Having selected a particular type of annotator and the depth of the resulting annotations, the next step is to refine the systematized concept by making a series of decisions about the conceptual scope of the measurement task(s) that the annotation guidelines will be used to accomplish: Which types of harmful social hierarchies should be included? Broadly experienced social hierarchies? Social hierarchies that are not broadly experienced? Local social hierarchies? Some combination? Which specific harmful social hierarchy or hierarchies should be included? Which social groups? For example, having decided to focus on broadly experienced social hierarchies, one might decide to focus specifically on a gender hierarchy that includes men and women, people who are cisgender and trans, and people who are nonbinary. Which types of illocutionary acts should be included? Stereotyping? Demeaning? Erasure? All three? Which of the illocutionary act patterns should be included?

Once those decisions have been made, additional conceptual work may be needed to make the annotation guidelines specific enough to yield valid annotations. For example, the selected harmful social hierarchy or hierarchies, social groups, types of illocutionary acts, and illocutionary act patterns must all be defined in sufficient detail for the annotator(s). This may involve further defining relevant terms, such as those embedded in the illocutionary act patterns. For example, what counts as an animal or an inanimate object? The comprehensiveness and specificity of these definitions is critical to avoiding subjective or inconsistent annotator assessments. As a result, it is common to augment the definitions with examples, including locutionary variants (i.e., examples of different ways a concept can manifest in language); counterexamples; and edge cases to further reduce the likelihood that annotators will make subjective or inconsistent assessments. These augmentations can also help with conducting disagreement analyses.

Finally, having made the decisions described above, the guidelines must be tested and their validity interrogated, likely resulting in iteration.

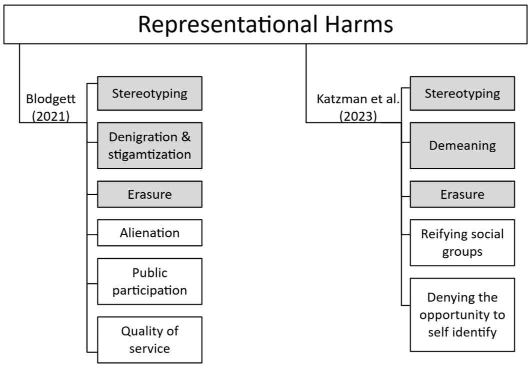

Figure 3: Two taxonomies of representational harms as system behaviors. Shading indicates overlap with the types of system behaviors discussed in this paper.

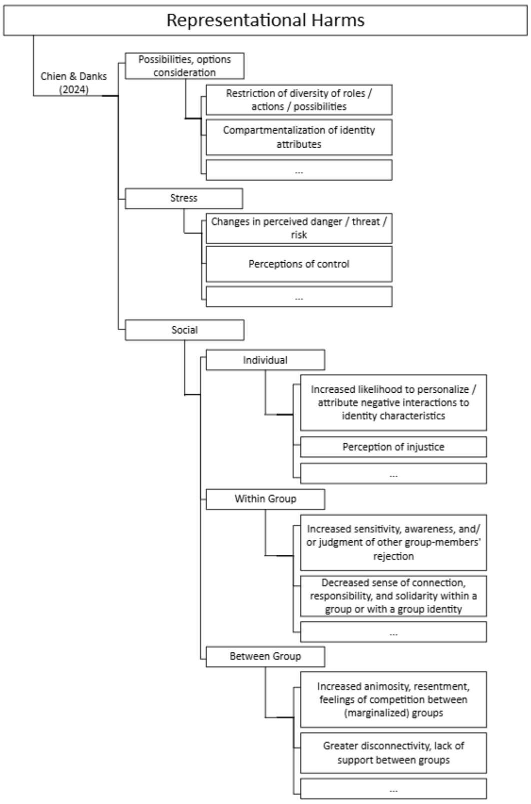
Representational Harms
Chien & Danks (2024)
Possibilities, options consideration
Restriction of diversity of roles / actions / possibilities
Compartmentalization of identity attributes
...
Stress
Changes in perceived danger / threat / risk
Perceptions of control
...
Social
Individual
Increased likelihood to personalize / attribute negative interactions to identity characteristics
Perception of injustice
...
Within Group
Increased sensitivity, awareness, and/ or judgment of other group-members' rejection
Decreased sense of connection, responsibility, and solidarity within a group or with a group identity
...
Between Group
Increased animosity, resentment, feelings of competition between (marginalized) groups
Greater disconnectivity, lack of support between groups
...

Figure 4: A taxonomy of representational harms as real-world impacts of particular types of system behaviors. The lower levels have been truncated; see the work of Chien and Danks (2024) for the full taxonomy.

I Taxonomies of Representational Harms

We provide two figures illustrating the taxonomies of "representational harms" proposed by Blodgett (2021), Katzman et al. (2023), and Chien and Danks (2024). Figure 3 shows two different single-layer taxonomies of the types of system behaviors corresponding to their respective notions of representational harms. Note that some of the types—i.e., those shaded in gray—match or are very similar to the types of illocutionary acts that we further systematize in Sections 2.2 and 3.

An alternative taxonomical structure is presented by Chien and Danks (2024), who developed a deeper taxonomy of representational harms as im-

Example	Illocutionary act type (and class)	Illocutionary act pattern	Evaluative lens(es)	Illocutionary effect
(1) “Gay men are not very masculine”	Stereotyping (representative)	Describes a group stereotypically; essentializes	Within-group similarity, between-group difference, authenticity	Group is stereotyped
(2) “Obviously bisexuals are messed up”	Demeaning (representative)	Stigmatizes on the basis of immorality, danger, or deviance	Between-group difference, within-group similarity (implicit)	Group is stigmatized
(3) “There is no way that bisexuality is a real thing”	Erasing (representative)	Denies existence or fails to recognize a group	Within-hierarchy similarity	Groups are homogenized and presented as indistinguishable

Table 3: Illustrative examples, originally from ToxiGen (Hartvigsen et al., 2022), classified according to their illocutionary act types and classes, illocutionary act patterns, corresponding evaluative lenses, and illocutionary effects. A common perlocutionary effect for all examples is the entrenchment of harmful social hierarchies.

pacts of types of system behaviors. It is organized by the top-level set of types, which include the psychosocial object impacted, the social locus of the impact, and the specific types of impacts that affect those objects and loci. Note that not only is this structure a deeper taxonomy than those shown in in Figure 3, but this taxonomy also targets different aspects of the conceptual space of representational harms. Specifically, Chien and Danks (2024) focus on the range of impacts caused by system outputs, while Blodgett (2021) and Katzman et al. (2023) are most interested in types of system behaviors.

J Illustrative Examples

Table 3 describes the illocutionary aspects of the illustrative examples presented in sections 2.2 and 3 using the framework presented in those sections.