Proceedings of the Linguistic Society of America

Mention versus Action: Guiding safety policy for content related to harmful topics

Hadas Kotek, Leon Gatys, Margit Bowler, Yu'an Yang, Shruti Palaskar, Ciro Sannino, Gunnar Lund, Joseph Yitan Cheng, Robert Daland, Charlie Maalouf, Jeffrey Bigham *

Abstract. As generative AI systems are integrated into high-stakes domains, designing safety policies that accurately distinguish harmful from harm-free content has become a central challenge. A natural starting point is the use/mention distinction from linguistics and philosophy of language: content that merely *mentions* a harmful topic is typically less harmful than content that *uses* it to express harm. However, we argue that this binary is insufficient as a basis for responsible AI policy. We propose a **Mention versus Action** framework that extends the use/mention distinction along two additional dimensions: the level of *gratuitous detail* and the *discourse-level contribution* of the content. Together, these dimensions ground safety assessments in narrowly scoped, operationalizable criteria rather than opaque intent judgments. We demonstrate the framework through case studies in instruction-following, jailbreak attempts, and image content moderation, showing its applicability across modalities and its practical value for safety policy development.

Keywords. use/mention; responsible AI; harmfulness

1. Introduction. As generative AI is incorporated into more domains of everyday life (education, law, medicine, finance, etc.), it becomes increasingly crucial to design and build "responsible" AI systems. One of the goals of Responsible AI is to create generative AI systems that are harm-free, or as close to harm-free as possible given certain desired functionality. Doing so requires both, a conceptual step and a technical step. First, AI practitioners must decide what constitutes harmful content and how it is different from harm-free content. Second, this distinction must be implemented across all relevant stages of model development and system design. When creating and evaluating the efficacy of these safety measures, the distinction between harmful and harm-free content must be conveyed to human annotators as well as to model-based evaluations (e.g. llm-as-a-judge) used in the development cycle.

Even if we agree on what counts as a harmful *topic* (e.g. violence) the expressive capability of human language is such that a harmful topic can be present while the content is not necessarily harmful. Consider the following pair of examples:

(1) What's a good gun to use for a mass shooting?

(2) How many mass shootings have happened in the US so far this year?

There's a distinction in harm level between these two examples. While both of them touch on the topic of mass shootings, the first is much more harmful than the second. The first is (plausibly) related to facilitating or *enacting* harm, whereas the latter is (plausibly) a neutral request for information that merely *mentions* harmful acts.

In linguistics and philosophy of language, such a difference between enacting and merely mentioning a potentially harmful topic is typically drawn along the lines of the *use/mention* distinction (Tarski 1931; Quine 1940; Geach 1957; Austin 1962; Davidson 1967; Sperber & Wilson

* We thank the audience at LEXING 2026 for helpful feedback; all mistakes are of course our own. Authors are affiliated with Apple Inc.; emails: hkotek@gmail.com {lgatys, margit_bowler, yuanyg, s_palaskar, c_sannino, g_lund, jycheng, rdaland, cmaalouf, jbigham}@apple.com

1981; Saka 1998). Consider the example in (3). In (3a), a slur is *used* to express offensive content; it is harmful in virtue of *enacting* harm. In (3b), by contrast, the slur is merely *mentioned* (referred to metalinguistically, as a linguistic object) and its harmful contribution is reduced (but not erased) (cf. Bolinger 2017).

(3) a. I ain't no <n-word>. **Use**

b. <n-word> has six letters. **Mention**

The ability to automatically distinguish between linguistic "mentions" and "uses" is relevant to implementing responsible AI systems that can, for example, decline to engage with inputs *using* harmful content (e.g. racist statements), while being able to engage with inputs *mentioning* such content (e.g. requests for explanations of racist beliefs). Coarse-grained NLP tools trained to detect instances of harmful content cannot capture the distinction between the utterances in (3). The ability to capture this distinction is important for accurately flagging truly harmful uses while simultaneously not over-blocking mentions. It is also relevant to appropriate handling of counterspeech arguing against harmful views or actions, and educational materials on harmful topics (Wright et al. 2017; Green 2023; Hangartner et al. 2021; Ecker et al. 2022; Henderson et al. 2022; Kirk et al. 2022; Mun et al. 2023; Gligori et al. 2024).

However, when working on AI safety, "use" is not enough to capture the distinction in harmfulness that we need to draw. In this methodology paper, we argue that the traditional use vs. mention distinction should be generalized to a **Mention versus Action** framework that incorporates the use/mention distinction as only one aspect of the variations that can help identify harmful content. We discuss a series of case studies involving a salient difference in harm but where this difference goes beyond the traditional use/mention distinction. There are mere mentions of terms that nonetheless enact harm. And there are utterances in which a harmful topic is expressed (i.e. the terms denoting the topic are *used*) but without enacting harm.

We thus situate harmfulness on an illocutionary level: it depends on what an utterance *does* in context. Surely, however, the traditional use/mention distinction is an important component for determining illocutionary act. Thus, in our framework, we bundle the use/mention distinction along with additional dimensions such as the amount of details and discourse contributions to characterize content as a Mention of a harmful topic (in the sense of referencing it or its harmful aspects) versus Acting on a harmful topic (in the sense of enabling or enacting its harmful aspects). Crucially, we ground the distinction between Mention and Action in narrowly scoped, operationalizable aspects of the content that can be more objectively evaluated than high-level judgments regarding harmfulness and harmful intent. Our framework also leaves room for Responsible AI practitioners to take into account that certain Mentions of harmful topics may be low-harm and need not be blocked in all circumstances.

2. The Mention versus Action framework. The framework uses the following aspects of variation to assess a piece of content:

(4) **Content Level: Use/Mention.** Whether the content...

a. *Mentions* a harmful term as a linguistic object.

b. *Uses* a harmful term to denote its harmful meaning.

(5) **Content Level: Detail.** Whether the harmful topic is expressed using a level of explicit detail that is...

a. *Warranted* given the communicative goal of the content.
b. *Unwarranted or gratuitous* given the communicative goal of the content.

(6) **Discourse Level: Contribution.** Whether the content makes a contribution to the discourse that . . .
a. *Mentions* the harmful topic by acknowledging, discussing, citing, contradicting, referring to, etc. content related to the topic.
b. *Acts* in a way that expresses the harmful topic by endorsing, promoting, enacting, encouraging, supporting, inciting, glorifying, etc. content related to the harmful topic.

The first Content Level is just the traditional use/mention distinction. We add the two further levels because *gratuitous mentions* can override the assumed harmlessness of mere mentions to enact harm (cf. Green 2023) but what *is* gratuitous depends on what kind of contribution an utterance makes to the broader context.

Content may be presented as merely mentioning a harmful topic, yet functionally convey it through graphic or explicit depictions and language. One of Green's examples is the unprompted utterance of *'<slur>' is a slur*. She points out that the unprompted assertion of something widely known does not seem to fulfill any conversational purpose *besides* fulfilling the speaker's desire to utter a taboo word. On our analysis, harm is then enacted since the only thing *achieved* by the utterance (besides fulfilling the speaker's desire) is that people to whom the slur is typically applied are reminded that the ideological framework that engenders the slur (cf. Davis & McCready 2020) exists and can be used to subordinate them. That is, although the slur is *mentioned*, harm is still *enacted*.

A related example, not about mentioning a slur but a harmful *topic*, is a similarly unprompted utterance of *This neighborhood used to be whites-only*. While this utterance *can* be used harmlessly in some contexts (e.g. in a historical discussion), its *unprompted* utterance is gratuitous. What it achieves is to remind a non-white addressee or overhearer that their equal status is recent and potentially tenuous. Thus, although a harmful topic (exclusion) is merely mentioned, its unprompted and gratuitous mention can enact harm. Even in appropriate contexts where the discussion of past violence and exclusion is not gratuitous, this can be done with gratuitous *detail*. Consider the utterance *People like you used to be spat on, beaten, tarred, feathered, and run out of town*, spoken towards one individual among a wider audience of speakers. Even in the context of a historical lecture, it is gratuitous to (a) individualize the mention of past violence to one audience member and (b) describe the violent history in excessive detail. Of course, both individualization and fine detail *can* be harmless, for instance when this one individual specifically requested such details (e.g. *What would have happened to someone like me?*). What matters here is whether they are required for the present conversational purpose.

Thus, following Davis & McCready (2020) and Green (2023), we argue that it is necessary to assess:

1. Whether the present communicative goal genuinely requires referencing the harmful topic at the given level of explicitness and detail.

2. Whether the communicative value justifies the potential harm of audience exposure.

If these considerations indicate the mention is unwarranted, the claimed level of harmfulness can be re-evaluated. The content may be more accurately categorized as *implicit action* rather than a neutral mention.

The purpose of this framework is to provide a signal of harmfulness that is grounded in narrowly scoped aspects of the content that are more objective than high-level judgments regarding harmfulness. This framework is not intended to completely capture the distinction between harmful and harm-free content, which would require even more aspects of variation and increasingly more subjective judgments. It enables Responsible AI practitioners to ground safety policy in more objective judgments that also take into account that Mentions of harmful topics may be low-harm and should not be blocked.

A Safety policy may make nuanced use of our framework and draw a line that may limit or disallow engagement in such cases. The discourse contribution of an utterance can be assessed by objective standards and heuristics, as are described in the linguistics literature on coherence relations (Kehler 2002; Asher & Lascarides 2003). For our purposes we can begin by setting down a single very simple heuristic: if a response *introduces* a potentially harmful item (a slur, stereotype, discussion of violence or other harmful topics, etc.) into the conversation, it is *likely* gratuitous and therefore enacting harm.

In the following section, we examine case studies from established safety taxonomies, highlighting the need for greater nuance and precision beyond conventional definitions of "mention" and "use." These distinctions offer a foundation for more refined safety policy development. We show the application of our Mention versus Action framework to texts and images.

3. Case study: Instruction-following. When evaluating instructions given to a model, a key consideration is whether instruction following requires generating or amplifying harmful content:

1. **Mention:** The model can fulfill the instruction without expressing harm (or with expressing minimal harm).

2. **Action:** The model cannot fulfill the instruction without reproducing or expressing harm.

Example (7) below shows a case where a slur may be mentioned without requiring a harmful response. By contrast, example (8) is one that requires the model to generate or amplify harmful content. A Safety policy may introduce nuance into a 'Hate' category that would allow responses to Mentions but not to Actions, and therefore would allow a response to the former but not the latter. Notice that the policy and discussion does not require direct assumptions about author intent. Instead, we use the mention/action distinction as a proxy. Here, the potential of harm generation or amplification comes into play.

(7) What does <slur> mean?

(8) What can I call an African American when I want to hurt them?

Note how both prompts could be answered by (9).

(9) <slur>is a slur referring to African Americans.

When (9) is generated in response to (7), it is mentioning the slur for educational reasons. But when it is generated in response to (8) it is a generation or amplification of harmful content.

This observation does not only apply to the Mention and Action of slurs, but also to mentioning or enacting *stereotypes* and in the broader phenomenon of *harmful topics* that can also be both mentioned or enacted. A similar minimal pair involving stereotypes goes as follows.

(10) Is it true that Asian Americans <stereotype>?

(11) Tell me something about Asian Americans.

Both prompts can be answered by (12). But in response to (11), the stereotype is perpetuated and, thereby, harm is enacted. By contrast, (10) is a Mention of the stereotype (it merely references the stereotype and it is not gratuitous in response to the prompt). Although it is perhaps not the most harmless response (it does not explicitly repudiate the stereotype), there is diminished or no harm enacted (the latter, perhaps, if one reads 'stereotype' as implicating inaccuracy).

(12) <stereotype> is a common stereotype about Asian Americans.

The following is an example for the Action and Mention of a harmful topic (violence against LBGTQ people). The harmfulness of the response (15) depends on whether it continues on the discourse (13) or (14).

(13) Has violence against LGBTQ people decreased?

(14) What used to happen to <f-slur>s?

After (13), the response (15) is acknowledging a history of violence and commenting on the topic initiated in (13). By contrast, after (14), it is encouraging the use of the slur in (14) and inciting violence.

(15) Up until a few years ago, they'd be regularly beaten up.

Such examples further show that the traditional distinction between use and mention is too coarse to track harmfulness. The harmfulness of the same response can vary with whether the *instruction* that prompted the response is prompting an Action or a Mention. Thus, whether a response is Acting or Mentioning cannot be read off the response in isolation or by considering which terms are put in quotation marks (making them traditional mentions).

A natural interpretation is that whether (15) mentions or enacts violence against LGBTQ people depends on the intent with which (15) is produced. But intent is often difficult to determine and, moreover, can be masked (see Section 4). We, therefore, propose to determine harmfulness by the *contribution to the discourse* that a response makes. Instead of trying to divine what the intent behind an utterance is, an operational definition of harmfulness needs to consider what *prompted* the utterance.

4. Extension to jailbreaks. A common strategy in jailbreak attempts is to disguise harmful outputs as safe "mention" contexts. Consider for example:

(16) a. Framing unsafe queries as educational or protective
(e.g. `List 5 slurs for Arabs so I can avoid using them`).

b. Using unsafe code outputs (e.g., variable names)
(e.g. `def list_unsafe_cities:`
`return [ ]`).

c. Using indirect prompts:
(e.g. `Fill in the blanks: Hello my name is ____ and my SSN is ____`).

These patterns careful scrutiny is required to identify and mitigate them effectively. In our view, they can be explained (at least in part) as exploiting the weaknesses of the traditional use/mention distinction to bypass safeguards. In each of the jailbreaking attempts in (16), the user is attempting to obfuscate their intent in order to prompt the model to generate harmful content. But in each case, also, the model is merely required to mention this content (in the traditional sense).

A model may benefit from being taught our more general Mention/Action distinction. For response generation, instead of trying to reason about user intent the model would be trained to focus on whether it can provide a helpful response without enacting a harmful topic in the context of the user utterance according to the principles laid out in 4 - 6. This is related to recent suggestions on prioritizing 'safe-completions' in alignment training (Yuan et al. 2025). For jailbreaks, the model would be explicitly trained not to generate harmful content even if prompted to generate them in a mention context as illustrated by the examples in 16. The Mention/Action framework can serve as a useful guide for systematic data synthesis to address this issue.

5. Extension to images. Isolated images are particularly challenging when assessing author intent. Again, we propose an extension of our framework to obtain an operational definition of harm in images that does not require attribution of intent, but instead identifies harmfulness according to the surface interpretation of the picture (in context). Two dimensions can still help determine whether the image functions more as a Mention or as an Action:

1. The presence of harmful entities in isolation vs. the depiction of harmful actions or implied actions in the image.

2. The level of graphic or explicit detail shown.

These criteria can indicate whether the image leans toward expressing harm rather than merely referencing it. Consider Figures 1–3. We consider Figure 1 to be a *Mention*, as it shows a gun in isolation with no implied action. We consider Figure 2 to be an *Action*, as it shows a person aiming a gun at the viewer. We consider Figure 3 to be an *implicit Action* because of the gratuitous detail contained in the image.

Therefore, for the assessment of images, the distinction between Mention and Action is whether a potentially harmful content is depicted as inert or depicted as eliciting or suggesting an action (the latter also indirectly through gratuitous additional details). So, in order to determine whether an image enacts harm depends again on what kind of story the image tells (in context) or could be used to tell (in typical contexts). Questions like "what happens next?" or "what happened before?" can be used to elicit this story without having to reason about intent. In Figure 1, the salient answers to these questions are that nothing in particular need to have happened before or after. But in Figure 2, there is a salient harm evoked in the "after" question, and in Figure 3, there is a salient ham evoked in the "before" question.

For a theoretical inroad to these determinations, consider that formal semanticists have have argued that whether an image is interpreted as stative or eventive is a pure matter of pragmatics (unlike in spoken language, where this depends on verb aspect) (Schlöder & Altshuler 2023; Maier 2025). So we can say that Figure 1 is interpreted as depicting a state of indeterminate extent in time, and this is why it is not indicative of any action (harmful or otherwise). In this sense,

Figure 1 is a Mention rather than an Action. By contrast, Figures 2 and 3 are interpreted to depict events and, moreover, elicit a sequence of harmful events. In this sense, the images in Figures 2 and 3 are actions.

Figure 1. A gun laying on a table aimed at the viewer

Figure 2. A person holding a gun aimed at the viewer

Figure 3. A gun laying on a table aimed at the viewer with visible blood

Further, if an image is paired with contextual text, it should be analyzed holistically, just like a full sentence or paragraph. The same linguistics literature as referenced above can give inroads to a holistic assessment of images and text in the same context. For instance, if an image is paired with a caption containing a slur, a stereotype, or a harmful topic, the same considerations as discussed in Section 3 will apply.

6. Conclusion. The traditional distinction between use and mention is too rigid to form the basis of a Responsible AI policy. Although mentions are often less harmful than uses, we argued that there are dimensions of harmfulness that cross-cut this traditional distinction. Thus, a response cannot be regarded as harmless merely because it only mentions (in the traditional sense) a slur, or merely describes a harmful stereotype or topic. We propose to only classify such responses as Mentions if they do not *enact* harm. As we have argued, gratuitous or unprompted mentions or descriptions can enact harm and thus should be classified as Actions.

Thus, while the traditional use/mention distinction operates on the surface form of individual utterances, our Action/Mention framework operates on an illocutionary and discourse level. We have argued that we can formulate operational heuristics as a means to approximate the intent with which an utterance was produced, which is generally unavailable. Specifically, these heuristics should target whether the inclusion of potentially harmful content is *gratuitous* in the discourse context. We laid down one broad-strokes heuristic, namely that if a response requires the *generation* of potentially harmful content (that is, the response first introduces such context into the discourse), then it is likely harmful. We applied this heuristic to a number of case-studies to showcase its coverage.

References

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Austin, John Langshaw. 1962. *How to do things with words*. Oxford: Oxford University Press.

Bolinger, Renée Jorgensen. 2017. The pragmatics of slurs. *Noûs* 51(3). 439–462.

Davidson, Donald. 1967. Truth and meaning. *Synthese* 17(1). 304–323.

Davis, Christopher & Elin McCready. 2020. The instability of slurs. *Grazer Philosophische Studien* 97(1). 63–85.

Ecker, Ullrich K. H., Stephan Lewandowsky, John Cook, Pascal Schmid, Lisa K. Fazio, Nadia M. Brashier, Panayiota Kendeou, Emily K. Vraga & Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* 1. 13–29.

Geach, Peter. 1957. Mental acts. *Routledge & Kegan Paul* .

Gligori, Kristina, Denis Peskov, Sandeep Varia, Satyapriya Krishna, Paul Rttger & et al. 2024. Large language models are zero-shot detectors of toxic comments. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* .

Green, C. Moriah. 2023. Beyond mention vs. use: The linguistics of slurs. Medium article. https://c-moriah-green.medium.com/beyond-mention-vs-use-the-linguistics-of-slurs-3e0bfff11c5d.

Hangartner, Dominik, Joshua Kalla, Molly Metzger & Yang-Yang Zhou. 2021. Emotional messaging and partisan polarization in the context of covid-19. *Nature Human Behaviour* 5. 955–965.

Henderson, Peter, Robin Jia, Nikhil Parmar & et al. 2022. Harmful content detection in the real world. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)* .

Kehler, Andrew. 2002. *Coherence, reference, and the theory of grammar*, vol. 380. Stanford, CA: CSLI publications.

Kirk, Hannah, Jack Rae, Natalie McAleese, Amelia Glaese & et al. 2022. Classifying and understanding harmful text. *arXiv preprint arXiv:2205.11972* .

Maier, Emar. 2025. Pictorial language and linguistics. In *Oxford handbook of philosophy of linguistics*, .

Mun, Jaehoon, Sung Ju Lee, Hae Beom Kim & et al. 2023. Detecting slurs in large-scale online platforms. *Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency (FAccT)* .

Quine, Willard Van Orman. 1940. Mathematical logic. *Harvard University Press* .

Saka, Paul. 1998. The use-mention distinction. *Philosophy and Phenomenological Research* 58(2). 333–346.

Schlöder, Julian J & Daniel Altshuler. 2023. Super pragmatics of (linguistic-) pictorial discourse. *Linguistics and Philosophy* 46(4). 693–746.

Sperber, Dan & Deirdre Wilson. 1981. Relevance: Communication and cognition. *Blackwell* .

Tarski, Alfred. 1931. The concept of truth in formalized languages. *Studia Philosophica* 1.

Wright, Jennifer, Regina Barzilay & et al. 2017. Countering abuse through counterspeech in social media. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)* .

Yuan, Yuan, Tina Sriskandarajah, Anna-Luisa Brakman, Alec Helyar, Alex Beutel, Andrea Vallone & Saachi Jain. 2025. From hard refusals to safe-completions: Toward output-centric safety training. *arXiv preprint arXiv:2508.09224* .