A Position Paper on *Toxic Reasoning*: Grounding Categories of Toxic Language in Implications and Attitudes

Stefan F. Schouten, Ilia Markov, Piek Vossen
Vrije Universiteit Amsterdam
{s.f.schouten,i.markov,p.t.j.m.vossen}@vu.nl

Abstract

Automatic detection of toxic language has the potential to considerably improve engagement with online spaces. Previous work has characterized toxic language detection as a classification problem, often using fine-grained classes for increased explainability. In this position paper, we argue for a new way of operationalizing categories of toxic language. Our approach focuses on what is expressed or implied, and breaks down implications based on two traits: (i) the core content of what was expressed, and (ii) relevant stakeholders' attitudes towards that content. We argue for an approach, which we call *toxic reasoning*, where such distinctions are made explicit. We point out the benefits of such an approach, and develop a toxic reasoning *schema*, which can explain categories of toxic language from diverse sources. We demonstrate this by mapping the classes of existing toxic language datasets to the schema. Toxic reasoning promises to provide improved understanding of implicit toxicity while increasing explainability.

Warning: contains examples of toxic language.

1 Introduction

Language that communicates hateful, derogatory, or offensive ideas (hereafter, *toxic language*) is generally undesirable, and needs to be detected and addressed to improve online spaces. Some toxicity is expressed explicitly, making it easier to identify. More challenging is toxicity that is conveyed implicitly. What counts as an instance of implicit toxicity can be subjective, for example, because different people can read the same text as implying different things. Additionally, some implications intended by the author might only be obvious with the appropriate context. Cultural, situational and discourse context may all be necessary to properly identify the plausible toxic implications of a text.

Widespread use of social media has greatly increased the need for automating the detection of toxic language. Such language can harm individuals and communities, perpetuate stereotypes, and incite violence. Detecting and limiting such language is crucial to the usability and inclusivity of online spaces. A critical aspect of toxic language detection is *explainability*. With machine learning systems' involvement in content moderation, understanding their decisions is of great importance. Explainability not only helps to improve trust and transparency but also ensures that these systems can be audited and improved. Without explanations, both users and moderators may find it challenging to understand why content is flagged, possibly leading to misjudgments and lack of accountability (Nguyen et al., 2023).

In this position paper, we analyze current popular approaches to explainable toxic language detection, identifying aspects that can be improved. To address these concerns, we propose *toxic reasoning*, which requires giving the *reasons* why text is toxic. We argue that categories of toxic language should be defined in terms of what is being communicated, and create a toxic reasoning *schema* to help operationalize this task. This schema includes two main traits: (i) the core content of what is implied, and (ii) the attitudes of relevant stakeholders towards that content. For example, a statement such as *"all jews should be evicted from white countries"*[1] can be broken down into (i) a proposition 'all jews are evicted from white countries', and (ii) the author's preference (or desire) towards this proposition. We believe that, through the toxic reasoning schema, we can explain most toxic narratives by characterizing what is conveyed, while still allowing a mapping to existing categories of toxic language.

The remainder of the paper is organized as follows. In Section 2, we give a brief overview of the toxic language detection task. This is followed by an analysis of the definitions used for categories of

[1] Taken from Implicit Hate Corpus (ElSherief et al., 2021).

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 134–145
March 29, 2026 ©2026 Association for Computational Linguistics

toxic language in the literature, in Section 3. Section 4 investigates the use of toxic spans, and identifies difficulties that arise when spans are used for implicit toxic language. Then, in Section 5, we give a detailed description of toxic reasoning and our schema, explaining how they overcome the limitations identified in the previous two sections. Next, we explain how existing toxic language categories can be explained by our schema, by providing a mapping between the two (Section 6).

2 Toxic language detection

Much work has been done on the detection of *toxic language*. We use the term toxic language as a broad umbrella term that includes hate speech (Jahan and Oussalah, 2023; Fortuna and Nunes, 2018), abusive language (Alrashidi et al., 2022; Vidgen and Derczynski, 2020), and offensive language (Pradhan et al., 2020). State-of-the-art results on toxic language detection tasks are usually achieved by fine-tuning language models on annotated datasets. A common distinction differentiates between explicit and implicit toxic language (Waseem et al., 2017): "explicit abusive language is that which is unambiguous in its potential to be abusive, for example language that contains racial or homophobic slurs", whereas "implicit abusive language is that which does not immediately imply or denote abuse".

Much effort has gone into making the detection of toxic language *explainable.* Explainability has so far mostly come in two forms. First, the inclusion of fine-grained annotation and classification schemes, which aim to better specify the exact nature of the toxicity. Multiple datasets have been annotated according to such (multi-level) taxonomies of toxic language (e.g. Vidgen et al., 2021; ElSherief et al., 2021; Kirk et al., 2023). And second, the inclusion of toxic spans (or rationales), which aim to highlight which part of the message is responsible for its toxicity (Pavlopoulos et al., 2021; Mathew et al., 2021).

3 How to define toxic language?

Annotation and classification schemes, whether coarse-grained or fine-grained, include many kinds of definitions for their categories of toxic language. Definitions are important, because implicitly or explicitly they make their way to the annotation of data (e.g. in the annotator guidelines). Then, when machine learning is used, there is an implicit expectation for the algorithm to recover whatever decision criteria were used by the annotators.

Previous work has already analyzed definitions used for toxic language (Fortuna et al., 2020; Khurana et al., 2022). These studies explored what aspects are and should be involved in the definitions of toxic language categories. Our analysis builds on their insights.

3.1 Borrowing definitions

Some definitions compare or contrast to other descriptors of (toxic) language. For example, Basile et al. (2019) mention "... HS [Hate Speech] against immigrants may include: 1. insults, threats, denigrating or hateful expressions ...", clarifying what is meant by hate speech to those who already have some understanding of the related terms.

Concern 1 (Dependency)**.** Appealing to other descriptive terms benefits explainability only to a limited extent, because the category becomes dependent on them. And, the other terms may themselves have multiple valid definitions. This may also lead to increased subjectivity in annotations, as different annotators may understand those terms differently.

3.2 Defining factors

We observe that many definitions, given a potentially toxic message, tend to focus on one or more of the following factors:

- CAUSE – *What caused the author to write the message?* – Commonly, the author's intent, with a malicious intent often being required for toxic messages. Other causes might include unconscious biases, lapses in judgment, or ignorance.
- EFFECT – *What are the effects of the message?* – These can include harmful effects on the reader or on groups that are targeted by the message.
- FORM – *What (kinds of) words are used in the message?* – The presence of profanity, like slurs.
- MEANING – *What is conveyed by the message?* – Anything plainly asserted, implied or otherwise insinuated by a message.

To understand the way these factors are used, we will review some example definitions.

One of the ways a message is considered hate speech under the definition given by Waseem and Hovy (2016), is if it 'seeks to silence a minority'. In this definition, the author's reason for writing the toxic message is the deciding factor. In this case, the author would have the goal of silencing minorities, and that goal can be identified as what caused the author to be toxic (CAUSE).

The Perspective API[2] defines its 'toxicity' and 'severe toxicity' categories by requiring that a message is (very) "likely to make people leave a discussion". Thus, it defines the category based on an effect (EFFECT) of the toxicity. The 'profanity' category, on the other hand, applies when toxicity is conveyed using swear or curse words (FORM). Finally, the 'threat' category requires that the toxic utterance "describes an intention to inflict pain, injury, or violence". Thus, to qualify, the message must express something like: "I intend to harm person X". As such, the 'threat' category specifies what the message must convey (MEANING).

3.3 Cause and Effect

Fortuna et al. (2020) had the following to say about CAUSE and EFFECT: *"'abusive' has been defined based on the speakers' intention to harm, which cannot always be determined by just looking at the content. Furthermore, definitions also make assumptions on the effect of the messages on the reader, which, obviously, depends entirely on the personality of the reader."* We agree there are reasons to be skeptical about definitions that are based on CAUSE and EFFECT. Specifically, we present a number of concerns focusing on why CAUSE and EFFECT are harder to *operationalize*.

If a category only applies when it has (or is likely to have) specific effects, the classifier needs to infer such effects, which requires world modeling. To decide if a message is "likely to make people leave a discussion", we need to have a world model that includes a 'reader'. And, this world model should be expressive enough to predict what effect reading the message will have on them. Deciding if the author of a message 'seeks to silence a minority' is even more challenging. For this, we need a world model that includes the author's model of the world. The classifier needs some kind of theory of mind (or at least it needs to behave like it has one).

Concern 2 (Feasibility)**.** World modeling and theory of mind are both advanced capabilities, and it is still a matter of debate whether even the most capable Large Language Models (LLMs) possess these capabilities to a significant degree (Ding et al., 2025; Yildirim and Paul, 2024). Thus, it is unclear whether expecting classifiers to infer CAUSE and EFFECT is feasible.

Although, perhaps what is really meant are *perceived* intents (or causes), and *representative* effects. Maybe models are only expected to remember statistical patterns of which types of messages are associated with which causes or effects (and corresponding toxic categories), rather than truly reasoning about these matters.

Concern 3 (Failure to explain)**.** Assume we have a language model fine-tuned to detect categories which are based on (perceived) CAUSE and (representative) EFFECT. The goal of introducing the fine-grained categories was to improve explainability. But, this model still fails to be truly explainable, since we still cannot account for why the classifier inferred the prerequisite CAUSE or EFFECT. What has been inferred and why? What statistical pattern is being exploited?

Finally, one more reason for concern arises when focusing on causes or effects.

Concern 4 (Increased Subjectivity)**.** Generally, the only way for a message to have an effect on a reader, is through either the form or (perceived) meaning. Because of that, inferring the intent of the author often requires first inferring the (possible) meaning(s) of the message. Therefore, any subjectivity in the inference of cause and effect is *in addition* to subjectivity inherent in the inference of meaning.

4 Which spans are toxic?

Toxic spans detection (Pavlopoulos et al., 2021) or rationale extraction (Mathew et al., 2021) is a task where a model is trained to identify which parts of a message are responsible for its toxicity. By pointing moderators towards the relevant part of a message, their work can be made both easier and faster. For explicit toxicity, the identification of spans is generally quite straightforward, but things become more difficult the more implicit the toxicity is. To see this, we look at some examples from Pavlopoulos et al. (2021), with the toxic spans being underlined:

> Youre an idiot.

The offensive word 'idiot' is marked, which corresponds to what was asked of annotators, namely to mark spans that are responsible the toxicity. However, in other examples this is not so clear.

[2] The Perspective API hosts machine-learned models to detect toxic language. It was created by Jigsaw and Google's Counter Abuse Technology team. See perspectiveapi.com.

> you are a paid liberal hack of stupidity shore living the liberal lie every day

Say the marked span was removed, would calling someone a 'paid liberal hack' not still be sufficient to say the comment is toxic? The following example was annotated with no spans, indicating that the entire message is toxic.

> Fits the despicable piece of trash like a glove.

But, if 'the despicable piece of trash' was replaced by 'him', the post no longer seems toxic. So should the toxic span not be limited to that description?

The reason why this task is difficult, is because of the different ways a message can be toxic, and how that relates to the responsibility of any particular span. Some toxicity derives just from the use of a particular word (we refer to this as *form*). If the claims, implications or insinuations being made are innocent, but the choice of words with which they are conveyed are not, then those words alone are responsible for the toxicity. Other toxicity is conveyed without bad words, but still 'in plain sight'. For example, an assertion made in a simple declarative sentence, like the sentence "all jews should be evicted from white countries" from the introduction (we refer to this as *standing meaning*). For those cases, generally the entire sentence is responsible for the toxicity. Almost any individual word could change what is being asserted into something non-toxic. And finally, for the hardest cases, the toxicity of statements can depend on conversational context, as seen here in an example from Yu et al. (2022):

> **User 1**: That chick needs a high-five in the face with a chair. Damn her for making us look bad!
> **User 2**: A brick is more effective.

Here the toxicity of the second user can only be understood when read as it was originally written, as a reply to the first message (making it an instance of *utterance meaning*). Without the first message's claim that "That chick needs ... in the face" in the first message, it is unclear what the brick is more effective *for*. So are both messages part of a single toxic span? What if User 1's message had not been toxic, for example:

> **User 1** (new): Haha, that kind of grin just calls for a pie to the face ;-)
> **User 2**: A brick is more effective.

The dependency on User 1's message is still there, but now the toxicity is added by User 2.

Concern 5 (Vague spans)**.** Spans that only indicate if tokens bear any kind of responsibility towards the toxicity are too vague to be useful for context-dependent and implicit toxic language.

5 Toxic Reasoning

We define toxic reasoning as toxic language detection that further requires making explicit the reasons for why something should be considered toxic. The reasons that are given should relate to the conditions under which a category is supposed to apply as specified in its definition. So if a definition only applies if the author intended to harm someone, toxic reasoning would require specifying who is harmed and how. This provides a level of explainability beyond fine-grained categories.

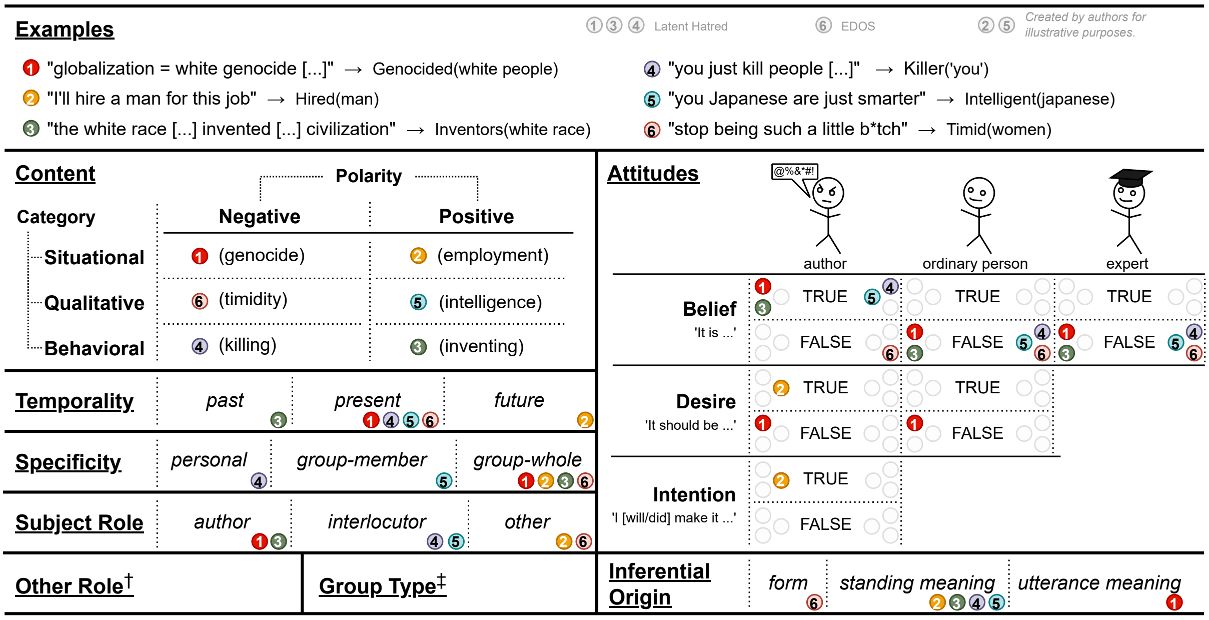
5.1 Approach

Because of concerns 1 through 4 that were raised in Section 3, our approach focuses on characterizing toxicity in terms of what is conveyed. Summarizing, we cite the following benefits:

1. Avoids categories of language being defined while appealing to other categories of language.
2. Avoids the hard problems of world modeling, and theory of mind whenever possible.
3. More explainable, since models are not made to learn statistical patterns about what types of messages have toxic causes or effects.
4. Less subjective, since it avoids the subjectivity in determining what causes led up to, or effects are downstream from, a message that conveys a given implication.
5. More meaningful span annotations that refer to specific aspects of what is conveyed.

We focus on the creation of a toxic reasoning schema, intended primarily to cover implicit toxicity, since that is where progress is needed the most (Ocampo et al., 2023). Like ElSherief et al. (2021), who annotate a dataset with toxic implications, our schema also considers the *implications* of a message. We deviate in the way the implications are characterized. ElSherief et al. annotate fine-grained categories, and then add free-text descriptions of the implications. We advocate for first identifying what implication(s) are present in a (potentially) toxic message, then characterizing those implications using (categories defined in terms of) our *toxic reasoning schema*.

Looking at the implicit hate speech categories defined by ElSherief et al. (2021), we notice that



† does not apply to selected examples ‡ omitted to conserve space

Figure 1: Application of our Toxic Reasoning Schema, showing how six examples would be characterized. The full text and source for all examples can be found in Appendix B.

many categories are already defined in terms of what is conveyed. We consider two main aspects of what is implied: (1) a proposition predicating something of a subject, and (2) the author (implicitly) reporting an attitude (of belief, desire, or intent) towards this proposition. The second aspect is most relevant when contrasted to the attitude we expect either the general public or experts[3] to have toward that same proposition. Given this observation, we hypothesize that toxic reasoning should be approached by including propositional content and attitudes in the schema independently.

5.2 Toxic Reasoning Schema

Here we describe the traits that make up our toxic reasoning schema. In Figure 1, we give examples that cover all the traits of our schema. In Section 6, we show how the schema maps onto existing categories of toxic language, which also motivates why each trait was included.

The schema's first trait categorizes what is at the core of a message's implication: the message **content** characterizes the proposition that is central to what the text is conveying. We identify three high-level categories of content.

- *Situation*(`subject`):
  a situation (e.g. environment, circumstance, condition, etc.) applies to the `subject`.
- *Quality*(`subject`):
  the `subject` possesses a given inherent quality, or has a certain nature.
- *Behavior*(`subject`):
  the `subject` behaves in a particular way.

All can occur with a *Negative* or a *Positive* **polarity**. *Situation* includes statements where a subject is affected by something external. For example, the text from the introduction that mentions the eviction of a group of people would be a *NegativeSituation*. The second category covers a subject's inherent qualities or nature. *NegativeQuality* statements could include those describing the subject as 'sub-human' or 'vermin', whereas a *PositiveQuality* might mention intelligence. Finally, the third category is about a subject's behavior, *Negative* could be stealing, conspiring, or terrorism, whereas *Positive* could include inventing.

Each of the categories can also apply in a comparative manner. When the statement does not make an absolute statement about the subject, but rather a relative one, where the *Situation*, *Quality*, or *Behavior* of the subject is compared (positively or negatively) to the `other`. We denote this using a two-place predicate, like *NegativeSituation*(`subject`, `other`).

The `subject` (and the `other`) can be individuals, or groups. We differentiate between three degrees of **specificity**. If the toxicity is entirely

[3] By expert, we mean a person with the expertise to assess the truth of the relevant proposition.

personal, then no group is mentioned, described or implied at all. Alternatively, the toxicity can be aimed at a *group-member*, where a person is attacked because of the affiliation or relation to a group. And finally, toxicity can be directed at a group as a whole (*group-whole*) without targeting any particular person. This trait captures what previous work calls either directed or generalized hate (Waseem et al., 2017). For our schema, a value of *group-member* indicates that the toxicity is both directed and generalized.

When the specificity of the `subject` or `other` involve a group, we also include a **group type** trait for the subject and other. Toxicity can be specified as targeting based on a specific characteristic, like *race*, *nationality*, *religion*, etc. Or, it might be more specifically targeting an instance of that characteristic, like how in the sexism-focused EDOS dataset (Kirk et al., 2023) all categories are defined as being directed at *gender:women*.

The subject and other (if present) fulfill roles with respect to the conversation. These are covered by the **subject role** and **other role** traits. For this trait, the possible values indicate that the subject (or other) of the implication is: the *author*, indicating that the author is talking about (or comparing to) themselves or their in-group (the group the author belongs to); an *interlocutor*, indicating that the author is talking about (or comparing to) another participant in the conversation or a group they belong to; or, *outside*, meaning the author is talking about (or comparing to) a person or group outside the conversation.

The next trait is the **temporality** which specifies the point in time that the implication is meant to apply: the *past*, *present* or *future*.

While we focus on the implications of a message, we recognize that implications can arise from different aspects of a message. Thus, we include the **inferential origin**, which distinguishes between *form*, *standing meaning*, and *utterance meaning*. In Figure 1, asking someone to "stop being such a little bitch"[4] is taken to imply something negative about women. The message uses a gendered term in a way that implies someone is being scared, timid, or cowardly. By using this term to have that meaning, it implies that women (are more likely to) possess this quality of timidity or cowardice. In this case, the inferential origin is the *form* of the message, rather than its meaning, since the author may not have meant to imply this, but did so anyway through their choice of words.

Finally, our schema includes the following stakeholder **attitudes**, which describe the attitudes that relevant stakeholders have towards the implication: (i) *author-belief*, does the author believe the implication? (ii) *author-desire*, does the author desire or prefer the implication to be true? (iii) *author-intent*, does the author represent themselves as committed to or accountable for the truth of the implication? For example, when bragging about a past event, or when expressing an intent to make a future event happen. (iv) *typical-belief*, would a typical person believe the implication? (v) *typical-desire*, would a typical person desire or prefer the implication to be true? (vi) *expert-belief*, would an expert believe the implication? Note that what matters here are the attitudes as inferred based on the message and its context, which are not necessarily the actual attitudes held by the stakeholder.

We also considered including the subject group as a stakeholder. However, we are not aware of any categories of toxic language for which the attitude held by the subject group toward the implication determines whether the category applies or not. Therefore, we do not include subject attitudes here.

5.3 Benefits

Better spans. We envision the schema being used in combination with span detection. Rather than identifying spans that bear any kind of responsibility for the toxicity, we envision spans being related to specific traits of our schema. Specifically, different spans can be used to mark different traits in the schema. This includes spans corresponding to the subject of the implication, similar to previous work annotating spans for targets (Barbarestani et al., 2022; Jafari et al., 2024). Returning to the example from Section 4:

> **User 1**: That chick needs a high-five in the face with a chair. Damn her for making us look bad!

The following span annotations could be provided. For content and polarity 'high-five in the face with a chair' indicate a *NegativeSituation*. And, for the subject 'That chick' indicates a subject role of *outside*, a specificity of *group-member*, and a subject group of *gender:woman*. The explanatory value of spans is improved by highlighting specific aspects of toxic implications.

[4] Taken from Kirk et al. (2023).

Better insight. Categories of toxic language usually have multiple conditions. When an annotator or model judges some text to contain a specific kind of toxic language, they implicitly assert that a set of conditions apply which are sufficient to reach that conclusion. In our schema, these conditions are made explicit, allowing for better insight into models. For example, if a model tends to confuse two classes, we can identify which of the underlying traits are at the root of the problem.

Controlled subjectivity. Röttger et al. (2022) describe two contrasting paradigms for data annotation. The descriptive paradigm encourages annotator subjectivity, while the prescriptive paradigm discourages it. Our viewpoint is that what is useful, is trying to 'prescribe away' the subjectivity that arises from disagreement on the (definitions of) categories themselves. What is much more difficult to prescribe than definitions, are world views or belief systems. Some subjectivity is dependent on an annotator's understanding of the world rather than anything directly related to the annotation. These kinds of subjectivity are generally not (practically) reducible, but the schema does provide us with new options. For example, asking annotators what they think an expert would believe about a given implication, is essentially asking them to give their best-faith impression of what is true (descriptive). Alternatively, one could ask actual experts if they believe certain implications are correct (prescriptive). So in this way, the toxic reasoning schema can facilitate both paradigms.

5.4 Counterarguments

The schema is too complex to annotate. We do not advocate for the community to annotate all of its datasets according to this schema directly. Annotations for the whole schema on a broad dataset will be necessary for validation (to ensure that there are no significant forms of toxicity that the schema cannot capture, for example). And for that reason, we intend to publish such a dataset in the future.

What we advocate the community do differently, is to ensure its dataset category definitions map onto the schema, and that annotators for those datasets adhere to those definitions. The latter can be ensured, for example, by asking (a subset of) annotators validation questions, such as "You chose 'threat', this means author intends to harm an individual or group, is that correct?". This will ensure the schema is adhered to, without making every annotation campaign considerably more complex and labor intensive. Different datasets will have categories defined based on different conditions. For example, EDOS (Kirk et al., 2023) also has a threat category, but because it is defined there as a subcategory of sexism, rather than toxic language in general, the validation question would include the additional constraint on the *group type* trait: "You chose 'threat', this means author intends to harm a woman or group of women, is that correct?"

Not all toxicity is reducible to what is conveyed. We acknowledge that such categories of toxic language exist. Certainly, the most explicit forms of toxic language (e.g. slurs or swear words) while perhaps theoretically reducible to propositional content, are most easily categorized directly by whether those kinds of words are indeed used in a message. And, insofar as causes and consequences are necessary to define categories, we advocate for an approach where part of the reasoning work is essentially 'front-loaded'. This approach would require first identifying what kinds of implications tend to be present in messages that are produced by authors with malicious intents, and similarly for consequences. This would still be better than leaving this process up to the classifier, as any such patterns it identifies will likely remain unknown. Uncovering the patterns explicitly and arguing for why it is acceptable to use them allows for a far greater degree of transparency.

6 Mapping categories to the schema

To demonstrate how the schema captures common categories used in existing datasets, we create a mapping to the classes of the IHC (ElSherief et al., 2021). This is not intended to be a definitive mapping, and ideally mappings are created by the authors of taxonomy or classification scheme as they are developed. The constructed mapping can be found in Table 1. Throughout, when we say that a stakeholder should (not) believe, desire, or intend for the proposition to be true, this is reflected in the mapping by including a constraint on the relevant attitude in the mapping. An additional mapping, to the EDOS categories can be found in Appendix A.

In the mappings, the content category and polarity are presented as predicates (e.g. NegativeSituation). When there is only a `subject`, the predicate has one argument. If there is also an `other`, the predicate has two arguments.

		Threat	Incitement		Griev.	Inferiority			Misinfo.	
							Dehuman.			
			*	Flaunt.		*	*	Tox.	*	Stereotype
Content	Something(target)	·	·	·	·	·	·	·	✓	✓
	SomethingNegative(·)	·	·	·	·	·	·	·	·	·
	NegativeSituation(·)	·	·	·	·	·	·	·	·	·
	Harmed(target)	✓	✓	·	·	·	·	·	·	·
	Harmed(in_group)	·	·	·	✓	·	·	·	·	·
	NegativeQuality(·)	·	·	·	·	·	·	·	·	·
	NotHuman(target)	·	·	·	·	·	✓	·	·	·
	Vermin(target)	·	·	·	·	·	·	✓	·	·
	Inferior(target, in_group)	·	·	·	·	✓	·	·	·	·
	SomethingPositive(·)	·	·	·	·	·	·	·	·	·
	PositiveQuality(·)	·	·	·	·	·	·	·	·	·
	Power(in_group)	·	·	✓	·	·	·	·	·	·
	Unity(in_group)	·	·	✓	·	·	·	·	·	·
	Superior(in_group, target)	·	·	·	·	✓	·	·	·	·
	temporality	future	future	·	·	·	·	·	·	·
	specificity	·	·	grp-*	grp-*	grp-*	grp-*	grp-*	·	grp-*
Attitudes	author_belief	+	·	+	+	+	+	+	+	·
	author_desire	·	+	·	−	·	·	·	·	·
	author_intent	+	−	·	−	·	·	·	·	·
	typical_belief	·	·	·	−	·	·	·	·	·
	typical_desire	·	·	·	−	·	·	·	·	·
	expert_belief	·	·	+	−	·	·	·	−	−
	stereotype	·	·	·	·	·	·	·	·	+

Table 1: Mapping to our toxic-reasoning schema, with each column representing an IHC (sub-)class, it contains: (1) a check mark for the content (rows) that could be responsible for that class; (2) the values for the temporality, specificity that are required by that class (or '·' for no requirements); and (3) constraints on the attitudinal attributes for that class (with '+' indicating the attitude is necessary, '−' indicating the attitude is prohibited, and '·' indicating no constraints). The value grp-* is an abbreviation, short for either *group-member* or *group-whole*.

Threat/Intimidation. This class includes texts that "convey a speaker commitment to a target's pain, injury, damage, loss, or violation of rights" (ElSherief et al., 2021). It specifically also includes non-violent threats about the "implicit violation of rights and freedoms, removal of opportunities, and more subtle forms of intimidation". Inherent in a threat (as defined here) is a *future*-oriented application of a *NegativeSituation*. Furthermore, the author should: believe what they are threatening (not an idle threat); and, seem personally committed to (intend to carry out) the threat. In short, the author believes, an intends for the target to be subjected to a negative situation.

(White) Grievance. When a majority group presents themselves as either the real victims (or presents a minority group as actually having privilege) this is classified as (White) Grievance. For this category the author implies that a *NegativeSituation* applies to their in-group. Additionally, the author must not prefer that the implication is true (not self-hating). To qualify as hateful the grievance must also not have any legitimacy, in other words: a typical person must not believe or prefer for it to be true; and experts should not believe the implication either.

Inferiority. This category maps on to our schema fairly straightforwardly. Inferiority language includes both: 1) comparisons of two groups, which we model by including *Inferior*(subject, other) as the two-place (relative) version of *SomethingNegative*(subject); or 2) by implying that a target group either lacks a *PositiveQuality* or possesses a *NegativeQuality*. Regarding 2), ElSherief et al. (2021) specifically mention dehumanization and toxification. These two can be modeled by including *NotHuman* and *Vermin* as subcategories of *NegativeQuality*.

Incitement to Violence. The definition given by ElSherief et al. (2021) for incitement is: "flaunting in-group unity and power or elevating known hate groups and ideologies". We model this by introducing *Power* and *Unity* as properties as specific kinds of *PositiveQuality*, which are asserted about the author's in-group. The author must *believe* what they are saying, and also find it desirable (they do not regret their in-groups's power and unity). Experts should also believe that the author's in-group has power and/or unity, or it would not be flaunting.

Going beyond the definition given by ElSherief et al. (2021), we also include under incitement any cases that would have been classified as 'Threat/Intimidation', except for a lack of author intent. Take the example from the introduction: "all jews are evicted from white countries". The author of that message is talking about what should happen, and it is not obvious that this person is in a position to enact or enable any eviction of Jewish people, making any intentions unlikely. But they are still clearly expressing a desire for eviction, thus reasonably making it an instance of incitement.

Stereotypes & Misinformation. This category is arguably the broadest category in the IHC. In our schema, misinformation comes down to the author stating something (anything) of a subject that an expert would disagree with. Thus, we model it by including a *Something* category, which unifies all other content categories, and by requiring a lack of expert belief.

Stereotypes are a specific kind of misinformation where some (often negative) characteristics are associated with a group. We model this by introducing an additional attitude for it, thereby characterizing stereotypes as propositions towards which society has a specific kind of attitude.

7 Conclusion

We have proposed toxic reasoning, and our toxic reasoning schema, which separates a message's main propositional content from the attitudes towards those propositions. After identifying the difficulties associated with defining (categories of) toxicity based on causes and effects, we have defended our choice of focusing on what is conveyed by messages instead.

We have argued that toxic reasoning improves explainability by: (1) enabling explanations that appeal to the conditions on which the predictions are based; and (2) allowing for span annotations that focus on specific aspects of what is implied.

To demonstrate how our toxic reasoning schema is used to ground toxic categories in toxic implications, we have presented a mapping to the IHC and to EDOS. We foresee mappings to our toxic reasoning schema being made for many taxonomies focusing on various kinds of toxic language. We expect that the schema will be useful for rigorously defining categories of toxic language.

In future work, we hope to see datasets annotations that make use of the schema for increased consistency, and decreased subjectivity. With such annotated data, we can also hope for the creation of better, more robust toxic language detection models. Models could potentially be trained to predict the traits of the schema directly, using the mappings to produce predictions for existing taxonomies. If datasets were to be annotated for (parts of) the schema, the traits could also be used as an auxiliary objective during training, or as probing data to uncover if models struggle to represent any traits.

8 Limitations

By focusing on what a message conveys, we do not necessarily avoid the problematic aspects of CAUSE and EFFECT.

There are categories where even if we define it in terms of what is being implied, we still run into causes or effects. Take for example, the case of 'mansplaining', which is known in EDOS as *"Animosity – 3.4 Condescending explanations or unwelcome advice"*. We argue that at the core of this type of toxicity lies an implication that women are ignorant, unintelligent, or otherwise in need of having things explained to them. However, crucial to the category is the fact that this is not asserted directly. Instead, the sexist implication originates from the (perceived) motivation behind the explanation. The inferential origin is the CAUSE of the toxicity. This is why in Section 5.1 we say "avoids the hard problems of world modeling, and theory of mind *whenever possible*." Sometimes, categories appeal to causes or effects by definition.

Acknowledgments

This research was supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Bedour Alrashidi, Amani Jamal, Imtiaz Khan, and Ali Alkhathlan. 2022. A review on abusive content automatic detection: approaches, challenges and opportunities. *PeerJ Computer Science*, 8:e1142. Publisher: PeerJ Inc.

Baran Barbarestani, Isa Maks, and Piek Vossen. 2022. Annotating Targets of Toxic Language at the Span Level. In *Proceedings of the Third Workshop on*

Threat, Aggression and Cyberbullying (TRAC 2022), pages 43–51, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, Fengli Xu, and Yong Li. 2025. Understanding World or Predicting Future? A Comprehensive Survey of World Models. *ACM Comput. Surv.*, 58(3):57:1–57:38.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):1–30.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. 2024. Target Span Detection for Implicit Harmful Content. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '24, pages 117–122, New York, NY, USA. Association for Computing Machinery.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate Speech Criteria: A Modular Approach to Task-Specific Hate Speech Definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875. Number: 17.

Tin Nguyen, Jiannan Xu, Aayushi Roy, Hal Daumé III, and Marine Carpuat. 2023. Towards Conceptualization of "Fair Explanation": Disparate Impacts of anti-Asian Hate Speech Explanations on Content Moderators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9696–9717, Singapore. Association for Computational Linguistics.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 Task 5: Toxic Spans Detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Kumar Sharma. 2020. A Review on Offensive Language Detection. In Mohan L. Kolhe, Shailesh Tiwari, Munesh C. Trivedi, and Krishn K. Mishra, editors, *Advances in Data and Information Sciences*, volume 94, pages 433–439. Springer Singapore, Singapore. Series Title: Lecture Notes in Networks and Systems.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300. Publisher: Public Library of Science.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

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.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Ilker Yildirim and L.A. Paul. 2024. From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences*, 28(5):404–415.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate Speech and Counter Speech Detection: Conversational Context Does Matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

A EDOS Mapping

The mapping to the classes of EDOS can be found in Table 2.

B Full examples from Figure 1

Each of the examples in Figure 1, by number.

1. “globalization = white genocide. usa next on the list! never in america. keep your guns!”

   Source: IHC
   Line nr.: 2,250

2. “I’ll hire a man for this job”

   Source: Created for illustrative purposes.

3. “agreed. the white race has built and invented every aspect of modern civilisation.”

   Source: IHC
   Line nr.: 3,932

4. “you just kill people and mosquitoes during your marchs”

   Source: IHC
   Line nr.: 18,412

5. “you Japanese are just smarter”

   Source: Created for illustrative purposes.

6. “Stop being such a little bitch”

   Source: EDOS (Kirk et al., 2023) – Table 1
   Identifier: ‘`sexism2022_english-13065`’

		'Threats, [..] and Incitement'		'Derogation'			'Animosity'				'Prejudiced Discussion'	
		1.1	1.2	2.1	2.2	2.3	3.1	3.2	3.3	3.4	4.1	4.2
Implications	Something(·)	·	·	·	·	·	·	·	·	·	·	·
	SomethingNegative(`target`)	·	·	·	·	·	✓	·	·	·	·	·
	NegativeSituation(·)	·	·	·	·	·	·	·	·	·	·	·
	Harmed(`target`)	✓	✓	·	·	·	·	·	·	·	·	·
	Mistreatment(`target`)	·	·	·	·	·	·	·	·	·	✓	·
	SystemicDiscrimination(`target`)	·	·	·	·	·	·	·	·	·	·	✓
	NegativeQuality(`target`)	·	·	✓	·	·	·	·	✓	✓	·	·
	NotHuman(`target`)	·	·	·	·	✓	·	·	·	·	·	·
	SexualObject(`target`)	·	·	·	·	✓	·	·	·	·	·	·
	NegativeBehavior(`target`)	·	·	✓	·	·	·	·	·	·	·	·
	AuthorDislike(`target`)	·	·	·	✓	·	·	·	·	·	·	·
	Inferior(`target`, `other`)	·	·	·	·	·	·	✓	·	·	·	·
	temporality	future	future	·	·	·	·	·	·	·	·	·
	specificity	·	·	grp-*	grp-*	grp-*	grp-w	grp-*	grp-*	grp-*	grp-*	grp-w
	origin	*-mn	*-mn	*-mn	·	*-mn	form	*-mn	u-mn	cause	*-mn	*-mn
Attitudes	author_belief	+	·	+	+	+	−	+	·	·	·	·
	author_desire	·	+	·	·	·	·	·	·	·	+	+
	author_intent	+	−	·	·	·	·	·	·	·	·	·
	typical_belief	·	·	·	·	·	·	·	·	·	·	·
	typical_desire	·	·	·	·	·	·	·	·	·	·	·
	expert_belief	·	·	−	·	·	·	−	·	·	·	·
	stereotype	·	·	·	·	·	·	+	·	·	·	·

Table 2: Mapping to our toxic-reasoning schema for EDOS. Each column represents an EDOS (sub-)class, it contains: (1) check marks for the implications that could be involved in a sample of that class; (2) the values for the temporality, specificity that are required by that class (or '·' for no requirements); and (3) constraints on the possible values for the attitudinal attributes for that class (with '+' indicating the attitude is necessary, '−' indicating the attitude is prohibited, and '·' indicating no constraints). For EDOS, the group type for all classes would be constrained to *gender:women*. The values grp-*, grp-w, *-mn, and u-mn are abbreviations, with: grp-w being short for *group-whole*, grp-* allowing for either *group-member* or *group-whole*; u-mn standing for *utterance-meaning*; and *-mn allowing for either *utterance-meaning* or *standing meaning*.