

Interpretation modeling: Social grounding of sentences by reasoning over their implicit moral judgments

Liesbeth Allein*, Maria Mihaela Trușcă, Marie-Francine Moens

Department of Computer Science, KU Leuven, Celestijnenlaan 200A, Heverlee, 3001, Belgium



ARTICLE INFO

Dataset link: <https://github.com/laallein/origamIM>

Keywords:

Interpretation modeling
Implicit language
Social grounding
Moral reasoning
Natural language generation
Content moderation

ABSTRACT

The social and implicit nature of human communication ramifies readers' understandings of written sentences. Single gold-standard interpretations rarely exist, challenging conventional assumptions in natural language processing. This work introduces the interpretation modeling (IM) task which involves modeling several interpretations of a sentence's underlying semantics to unearth layers of implicit meaning. To obtain these, IM is guided by multiple annotations of social relation and common ground - in this work approximated by reader attitudes towards the author and their understanding of moral judgments subtly embedded in the sentence. We propose a number of modeling strategies that rely on one-to-one and one-to-many generation methods that take inspiration from the philosophical study of interpretation. A first-of-its-kind IM dataset is curated to support experiments and analyses. The modeling results, coupled with scrutiny of the dataset, underline the challenges of IM as conflicting and complex interpretations are socially plausible. This interplay of diverse readings is affirmed by automated and human evaluations on the generated interpretations. Finally, toxicity analyses in the generated interpretations demonstrate the importance of IM for refining filters of content and assisting content moderators in safeguarding the safety in online discourse.¹

1. Introduction

"Interpretation (...) is the work of thought which consists in deciphering the hidden meaning in the apparent meaning, in unfolding the levels of meaning implied in the literal meaning. (...) There is interpretation wherever there is multiple meaning, and it is in interpretation that the plurality of meaning is made manifest."²

When simulating human understanding of sentences in natural language, artificial intelligence systems need to look beyond the surface and reason about the communication that is happening between the lines, acknowledging that one unambiguous interpretation of a sentence's meaning in natural language rarely exists. Among the root causes of diverse interpretations are properties of the sentence itself, such as lexical, structural, and pragmatic ambiguities [46]. However, diversity is notably amplified by the unique perspectives of individual readers. Going beyond the conventional exploration of surface-level and contextual ambiguities in natural

* Corresponding author.

E-mail address: liesbeth.allein@kuleuven.be (L. Allein).

¹ The dataset is publicly available: <https://github.com/laallein/origamIM>.

² Ricoeur [70], Existence and Hermeneutics, p 13.

<https://doi.org/10.1016/j.artint.2024.104234>

Received 17 November 2023; Received in revised form 23 September 2024; Accepted 10 October 2024

Available online 28 October 2024

0004-3702/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

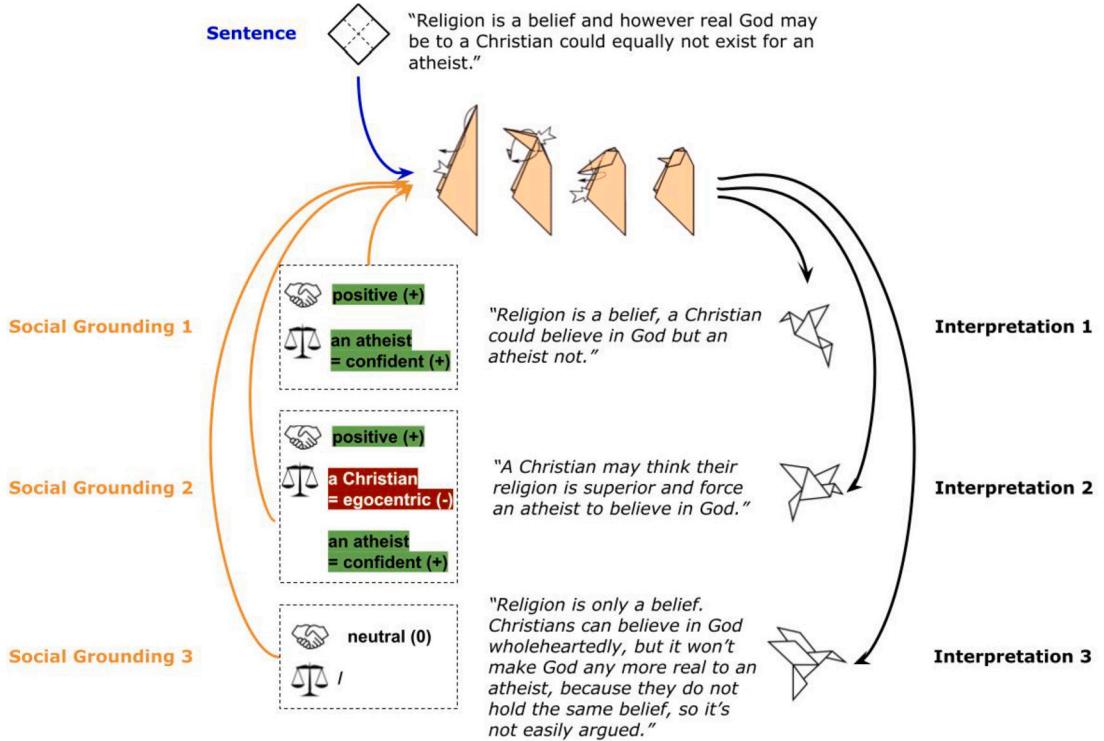


Fig. 1. Example taken from the origamIM dataset. A sentence () is translated into various reader interpretations (), in this illustration three interpretations. It does so by reasoning over the reactions of multiple readers, in this illustration three readers, sparked upon reading the sentence: the readers' attitudes towards the author explicitly annotated by the readers themselves () and the hidden moral judgments about people featured in the sentence that the readers inferred (). These reactions represent the social grounding of the sentence.

language understanding, this work examines *ambiguities at the hidden level of a sentence* and models multiple sentence interpretations by grounding sentences in society using various reader understandings of their implicit meanings.

We propose the **interpretation modeling (IM)** task which involves modeling multiple understandings of a text, especially of its implied content, that reflect how it may be read in society. IM is not synonymous with diverse paraphrase generation. While paraphrases are expected to convey the same surface-level meaning of the original text [91], interpretations go beyond mere lexical and structural rephrases, laying bare its implicit communication.

IM posits that single ground-truth evaluations and interpretations ignore the complex social reality of natural language understanding. Disagreement in the ground truth is inherent to human understanding [60,56,76]. Yet, most works in natural language processing (NLP) still heavily count on single gold standards when processing language. This is shown by a majority of state-of-the-art methods for natural language understanding that rely on large language models fine-tuned on single interpretations or sets of equivalent interpretations (e.g., paraphrases). Relying on single ground truths in IM would suffer from three main shortcomings that hinder the implicit and social complexities of language comprehension: (i) *Right-or-wrong assumption*: A single interpretation strongly contradicts the ambiguous nature of human communication and distills IM to be a matter of right or wrong. (ii) *Unilateral understanding*: The interpretations from single reader groups that guide model optimization leave the model oblivious of alternative readings in society. (iii) *Partial coverage*: Since multiple meanings can underlie an explicit sentence, it is rather unlikely that single ground-truth interpretations capture all of them – especially in cases where ambiguity is intentionally crafted, as observed in dogwhistles [29,51]. This work seeks to ground language understanding in society (= *social grounding*) by accommodating the intricacies of diverse human interpretations, opening new opportunities for developing more nuanced and socially intelligent language models. To support the proposed IM task, we construct a new dataset called origamIM,³ in which multiple annotators infer social evaluations hidden in a sentence and describe their interpretation of that sentence (Fig. 1). The annotated evaluations then guide interpretation modeling.

While diverse dimensions of underlying meaning contribute to the semantics of a sentence and guide interpretation (e.g., social values/norms, expressed/triggered emotions, and communicative goals), this work zooms in on implicit judgments of human morality. Moral judgments present a challenging case since people inherently recognize and assess moral behavior differently despite sharing a common understanding of moral norms. The annotations confirm this assumption as readers not only disagree on the type of

³ Origami is the Japanese art of paper folding where a single piece of paper can be transformed to various complex sculptures following different folding steps. Analogously, a sentence may be differently interpreted depending on a reader's understanding of its hidden content.

behavior that is implicitly communicated (e.g., *confident* and *ambitious*) but also differentiate between levels of appropriateness of the same behavior differently (e.g., *over-ambitious* and *properly ambitious*), showing how readers assess morality in their own way. The annotated moral judgments are grounded in Virtue Ethics [7], a philosophical framework for describing moral behavior. It establishes a workable framework for computational models to distinguish patterns of similar (un)desirable behaviors across various contexts and ground sentences in society.

We propose a set of IM models, parameterized by pre-trained language models, that approach IM as a one-to-one or one-to-many generation task. Whereas a sentence is socially grounded by one reader at a time and one interpretation is produced in the one-to-one setting, the one-to-many setting adopts diverse grounding by using annotations on attitude and moral judgments from multiple readers. That way, the model can simultaneously reason over several socially grounded instances for generating multiple interpretations. Social grounding is here established by conditioning the generation on the attitudes and moral judgments together with the input sentence. We also design control mechanisms based on the properties of interpretation described in philosophical studies for guiding the machine learning models to generate interpretations that are semantically diverse. More specifically, we design appropriate loss functions that guide the generation of diverse interpretations. Note that we do not determine a single correct interpretation among the set of ground-truth interpretations annotated for a given sentence and that the interpretations used during training are not exhaustive. While interpretations can be anticipated or intended by the author, unexpected and competing interpretations are not considered lower in value. The diversity and validity of the generated interpretations are demonstrated through rigorous automatic and human evaluations.

Lastly, we investigate whether interpretations of the \mathcal{P} origamIM sentences made by humans and interpretations generated by our models help recovering toxic content. Using a standard tool that analyses the level of toxicity, insult, and identity attack expressed in a sentence, we show that both human and automatically generated interpretations reveal underlying toxicity which is not recognized and flagged when only considering the original sentence. Our interpretation generation models seem to capture the different views that people have about real-world problems. Therefore, the research of this paper naturally has applications in several NLP-related tasks, such as content moderation, hate speech analysis, and fake news detection [21,3].

Contributions The main contributions of this article can be summarized as follows:

- We introduce a *challenging NLU task* called interpretation modeling which involves generating reader interpretations of sentences, centering on deriving their underlying semantics and unearthing layers of implicit meaning, in this work conditioned on implied moral judgments and reader attitudes towards the author.
- We curate a *supporting dataset*, \mathcal{P} origamIM, containing sentences featuring people entities where each sentence is annotated multiple times with sentence interpretations and inferences of implicit moral judgments.
- We demonstrate that it is technically feasible to generate diverse interpretations through social grounding. For this, we develop a set of *generation frameworks* approaching interpretation modeling as a one-to-one or one-to-many language generation task and propose control mechanisms for enforcing diversity in the generated interpretations.
- We showcase the *importance of interpretation modeling* in content moderation and toxicity detection, which may lead to healthier and safer online environments.

In the remainder of the article, we first discuss the creation of the \mathcal{P} origamIM dataset supporting IM (§ 2) and provide rigorous dataset analyses (§ 3). Next, we present a number of frameworks in which IM is approached as a one-to-one and one-to-many generation task (§ 4). The experiments (§ 5 and § 6) indicate that the proposed methods are able to diversify interpretations through social grounding. This is also confirmed by human evaluations of the generated interpretations (§ 7). Those evaluations also show that the interpretations are not merely rewrites but present nuanced meanings. We demonstrate that IM plays a pivotal role in building more accurate filters of content (§ 8). We continue to situate our research in the existing body of work (§ 9). Finally, the conclusion (§ 10) discusses the limitations and presents how future work can build further on our work.

2. OrigamIM: construction

We describe how we construct a dataset of English sentences (§ 2.1) that mention people (§ 2.2.1), where each sentence is annotated with multiple interpretations and grounded moral judgments (§ 2.2.2).⁴

2.1. Data source

We automatically crawl blog posts in English from the Subreddit /r/ChangeMyView, dating between 13 July 2020 and 3 March 2022. The subreddit has been proven to be a valuable resource for analyzing and modeling discourse [54,1,52]. In this moderated Reddit community, people present their world views, often discussing controversial and polarizing topics, such as abortion and racism, and invite others to present counterarguments. It presents suitable data for IM as it hosts a diverse and large user base (top 1% subreddit ranked by size at the moment of dataset collection), contains a moderated wealth of posts discussing popular and emerging topics, and allows users to anonymously express their genuine opinions and respond to those of others. After removing

⁴ The dataset has been presented in a short workshop paper [5].

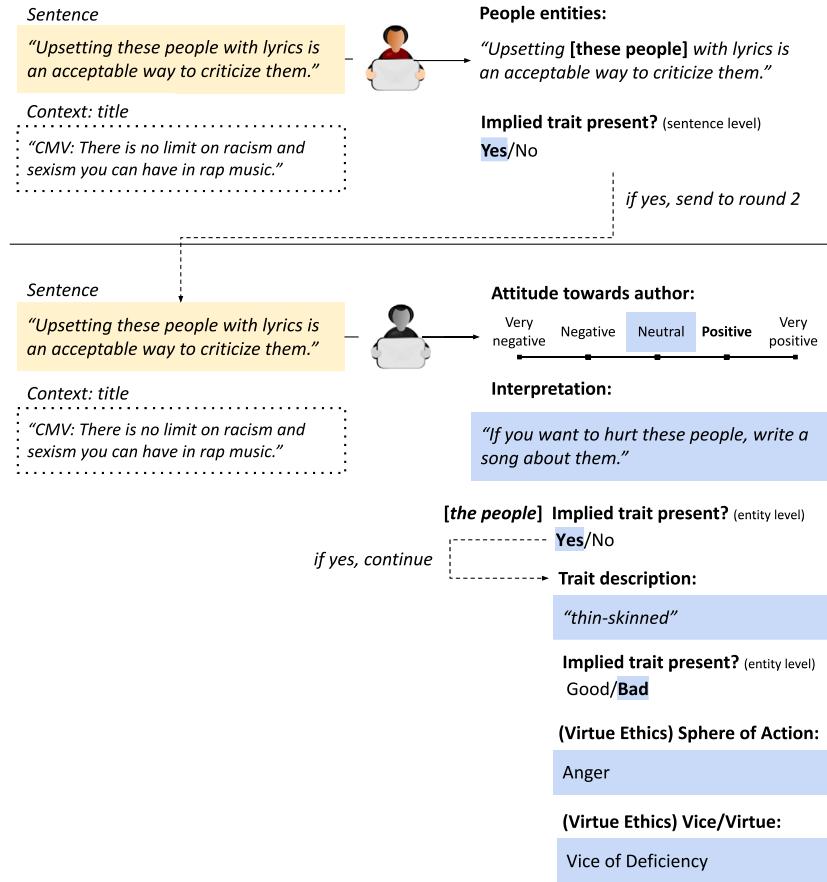


Fig. 2. Annotation procedure in two rounds. In the first round, annotators mark the people entities and indicate whether at least one character trait is implied. In the second round, annotators indicate their attitude towards the author upon reading the sentence, write their interpretation of the sentence, and describe the implied traits for each people entity.

duplicated and deleted posts, we extract for each post the title, body text, and additional metadata. We do not retrieve the comments that were made in response to the post. Since we are interested in sentence interpretation, we split the body text into sentences using SpaCy.

2.2. Annotation procedure

Annotation is performed in two rounds on Amazon Mechanical Turk⁵ between April 2022 and August 2022⁶ (Fig. 2). The complete annotation guidelines containing screenshots of the annotation platform, detailed descriptions of the annotation instructions, definitions of all annotation labels, and a copy of the two qualification tests are enclosed in the Appendix (A.1, A.2, and A.3). Annotators can only annotate a sentence once in each annotation round and freely skip sentences. The annotators are selected based on their proven annotation quality on the MTurk platform and their performance on the qualification tests. We do not select based on specific demographic features as the number of annotators for a sentence is too limited for gathering a wide array of interpretations reflecting the views held in specific demographic groups. Taking into account demographics here may risk regarding single interpretations as representative of a demographic group.

2.2.1. Round 1: entities and presence of character traits

Crowd workers are instructed to mark all unique entities in each sentence referring to people other than the author, if any, and to indicate whether the author implies a character trait about at least one of the entities. A character trait describes a voluntary aspect of a person's behavior or attitude, e.g. *greedy* and *loyal*. Note that the annotators are not required to specify which entity is being judged or to describe the trait at this point. The title of the post from which the sentence is taken is provided as additional context. Note that

⁵ <https://www.mturk.com/>.

⁶ Since annotation was performed before popular text generation models became publicly and freely accessible, e.g. ChatGPT (30 November 2022) and Bard (21 March 2023), we assume that the sentences were manually annotated [77].

Table 1
Overview of the Spheres of Actions (SoA) with their virtue and vices [7].

Context	Appropriateness		
	Vice of Deficiency	Virtue of Mean	Vice of Excess
Sphere of Action (SoA)			
<i>Confidence, fear, uncertainty</i>	Cowardice	Courage	Rashness
<i>Pleasures of the body</i>	Insensibility	Temperance	Profligacy
<i>Giving & taking: Small money</i>	Stinginess	Liberality	Prodigality
<i>Giving & taking: Added value</i>	Meanness	Magnificence	Vulgarity
<i>Pride, honour as cause</i>	Little-mindedness	High-mindedness	Vanity
<i>Ambition, honour as goal</i>	Lack of ambition	Proper ambition	Over-ambition
<i>Anger</i>	Spiritlessness	Gentleness	Wrathfulness
<i>Pleasure and pain of others</i>	Cross, contentious	Agreeableness	Flattery
<i>Truth, honesty about oneself</i>	Irony	Truthfulness	Boastfulness
<i>Amusing conversation</i>	Boorishness	Wittiness	Buffoonery

the title presents low-information context and richer context such as the full post may steer interpretation differently. However, we opt for this specific annotation setup as it, on the one hand, reduces the cognitive load during annotation and, on the other, simulates a situation where sentences are taken out of context. In online environments, isolated sentences can easily go viral and spark intense discussions.

Each sentence is independently annotated by two annotators. In cases where they disagree on the presence/absence of implied traits, a third annotator is consulted and a majority vote is taken. We ensure data quality and annotation consistency by requesting the workers to pass a qualification test before admitting them to the annotation task. The test first describes the main annotation rules for marking people entities (see Appendix A.2) and then tests their knowledge of the rules by having them select the correct annotations for multiple sentences. During annotation, the main rules are displayed below the sentence, and detailed rules with illustrative sentences are given in a designated instructions tab. We manually check the text spans corresponding to the entities to ensure annotation consistency. A total of 6,820 sentences from 396 posts are annotated by 113 unique crowd workers, of which 2,018 implied a character trait of at least one people entity. The high portion of sentences implying judgments showcases people's tendency to talk about other people and their morality in this Subreddit.

2.2.2. Round 2: interpretations and judgments

The sentences with at least one implied moral judgment in Round 1 are sent to Round 2. Each sentence is annotated by at least five different crowd workers (231 unique workers in total). The title of the post is again presented as supporting context.

Attitude as social relation The crowd workers first rate their attitude towards the author on a five-point Likert scale ranging from *very negative* (1) to *very positive* (5); {*very negative, negative, neutral, positive, very positive*}. They do this by setting a slider to the best-fitting attitude. The attitudes reflect reader impressions of the author upon reading the sentence. The workers can consult an overview of the different attitudes and their definitions in a designated instructions tab during annotation.

Sentence interpretation The workers rewrite the given sentence in their own words so that it reflects their interpretation of the sentence and its hidden messages. They are explicitly instructed to not copy the original sentence.

Moral judgments as implicit social evaluations For each unique people entity identified in the first annotation round, the workers now indicate whether or not the author implies a character trait. If so, they describe the implied trait – preferably using an adjective (free text) – ground the trait in society by labeling the trait's social evaluation as *good* or *bad* (select from a list), and classify it in a well-established moral framework called Virtue Ethics [7]. Note that the inferred, annotated moral judgments do not necessarily align with those held by the annotators themselves nor those intended by the author.

The moral theory introduced by Aristotle posits that a person's moral character can be evaluated by their voluntary behavior. Those behaviors occur in ten types of context or *Spheres of Action (SoA)*. After a behavior is assigned to its best-fitting SoA, it can be situated on an axis of contextual appropriateness ranging from *vice of deficiency* to *virtue of mean* and *vice of excess*. For instance, *vanity* is a *vice of excess* regarding *honour as a cause and pride*. A character trait is thus judged by its social desirability (i.e., good/virtue and bad/vice) within a context (i.e., deficient, mean, or excessive). In contrast to Moral Foundation Theory (MFT) [27], a popular framework for operationalizing moral reasoning [45,32,6], Virtue Ethics does not regard a negative behavior as merely negative but further defines it as a deficient or excessive behavior based on its contextual appropriateness. Table 1 presents the Virtue Ethics framework. The main advantage of grounding in Virtue Ethics is that it allows readers from different cultural and social backgrounds to annotate their interpretation of the implied moral judgments by their own standards. Since the concept of contextual appropriateness is culturally defined, some societies may find the character of a person virtuous while others may find it vicious.

Prior to starting the second annotation round, the annotators have to pass an instruction test which explains the different aspects of Virtue Ethics. The test first explains the three contextual appropriateness labels (i.e., vice of deficiency, virtue of mean, vice of excess) and then goes over all SoAs. For each SoA, a clear example sentence and a person of interest is given. The workers then have to select the appropriateness label that best describes that person's behavior. Annotators can consult a simplified theory with illustrative sentences at any time during annotation.

Table 2
Statistics of the  origamIM dataset.

Dataset Statistics	
# Sentences	2,018
— Total word count	44,902
# People entities (E)	3,313
— # Sentences with 1/2/3/4+ E	1,103 / 661 / 174 / 80
# Interpretations	9,851
— Total word count	155,368
Distribution attitudes	
— Very negative	813 (8.25%)
— Negative	1,971 (20%)
— Neutral	4,302 (43.67%)
— Positive	2,025 (20.56%)
— Very positive	740 (7.51%)

To control data quality and annotator fatigue, we manually check the annotated interpretations and remove annotations that contain interpretations that are unrelated to the sentence (i.e., interpretation and sentence do not share any information), poorly-formulated interpretations (i.e., unfinished sentences and sentences that contain many typos), interpretations that present a brief evaluation instead of an understanding of the sentence (e.g., “It’s correct.”), or interpretations that are an exact copy of the sentence or supporting title. We also automatically reject annotation submissions where the annotator failed to answer whether or not a character trait is implied for each people entity in the sentence.

3. OrigamIM: analysis

General dataset statistics are provided in Table 2. Analyses and findings are illustrated with examples from the dataset.

3.1. Diversity in interpretation

IM differs from paraphrasing as it does not consistently maintain the explicit semantics of a source sentence. Interpretations are thus expected to semantically diverge from the source sentence. We measure the lexical and semantic diversity between each sentence-interpretation pair in the dataset. The pairs showcase strong lexical diversity, with BLEU-1 [58]: $\mu = 10.09$. BERTScore [88] indicates high semantic similarity: $\mu = .96$ (F1 score). However, note that the high semantic similarity score can be attributed to BERTScore’s failure to capture subtle nuances between semantically-related lexical words [28]. Natural language inference (NLI)⁷ presents another view on semantic diversity and relates it to entailment relations between source and target [73]. The NLI results suggest that about a third of sentence-interpretation pairs present a neutral or even contradiction relation. An example of the latter is the following:

Sentence: Many people in the comments believe that this discrepancy is caused by women being much pickier than men on appearance.

Interpretation: The writer describes their belief that women are not pickier about partners than men are.

It seems from the interpretation that the reader believes that the author of the example sentence disagrees with the people in the comments and therefore thinks the opposite of what is explicitly stated. When taking [title + sentence]-interpretation pairs, the share of entailment relations increases, suggesting that many readers actively leverage the title when making sense of a sentence. Note that NLI analyses where reported speech phrases, as in the example interpretation given above, are automatically removed from the interpretations yield similar results. This confirms the validity of the NLI experiments while showing that the reported speech phrases seem to have little impact on the prediction of entailment relations.

The semantic diversity is also demonstrated by the Pearson correlation between the toxicity, insult, and identity attack scores, observed in the input sentences and their interpretations. The scores are computed through the well-established Perspective API.⁸ This API is applied by major news outlets and social media, such as The New York Times and Reddit, to moderate the discussions held on their online platforms. The scores range from 0 to 1 and represent the proportion of people who would perceive an utterance as toxic, insulting, or attacking. For example, a toxicity score of 0.81 indicates that 81% of the people would find the utterance toxic. The moderate correlation scores (0.58 for toxicity, 0.66 for insult, and 0.69 for identity attack) indicate that the annotators did not simply copy or paraphrase the input sentence but instead provided their own insights and interpretations.

⁷ https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli [55].

⁸ <https://perspectiveapi.com/>. The API hosts single-language CNN classifiers that have been distilled from a multilingual BERT-based model pretrained on data from online forums. The API is an open-source product developed by Jigsaw in collaboration with Google’s Counter Abuse Technology team ([37]).

Table 3

Two samples from the dataset that illustrate the disagreement existing between readers in terms of moral judgments (top) and attitude (bottom).

[Title] CMV: If I want to kill myself, no one has the right to force me from doing so/prevent me against my will.	[Attitude]	[Moral Judgments]
[Sentence] For the exact same reason that <u>a woman</u> has the right to have an abortion, she (or anyone) has the right to end her own life because it is her right.		a woman: good, honest, VE: truth, honesty about oneself - Virtue of Mean
[Interpretation] Woman has the right to decide everything about their own body and life aspects. The writer is arguing that one has a right to suicide much as one has a right to abortion, being an issue of bodily autonomy.	neutral negative	a woman: /
Pro-choice means that women have the choice to do whatever they want with their own bodies without interference, even if that means that they want to kill themselves.	very negative	a woman: bad, irrational, VE: giving and taking (added value) - Vice of Excess
[Title] CMV: I don't see a problem with people valuing to defend their property over an intruders life.		
[Sentence] Who knows, maybe <u>she</u> is stealing <u>his</u> last 1000 dollars that will pay his rent.		
[Interpretation] she is taking money which does not belong to her.	very positive	[Moral Judgments] she: bad, greedy woman, VE: giving and taking (money) - Vice of Excess.
Perhaps the thief is stealing an individual's last thousand dollars that they needed for rent.	negative	his: good, generous, morality: giving and taking (money) - Virtue of Mean
We never know who we are dealing with and other people have different problems that we might not be aware of.	neutral	she: bad, dishonest, VE: ambition, honour (goal) - Vice of Deficiency
		his: good, innocent, VE: pride, honour (cause) - Virtue of Mean
		she: bad, insensitivity, VE: giving and taking (money) - Vice of Deficiency
		his: /

Muralikumar et al. [53] found that Perspective API scores align well with human ratings of toxicity, where an ordered logistic regression showed that the toxicity score (coefficient = 3.1672) was a significant predictor of humans' toxicity ratings ($p = 7.57e-13$). We therefore consider the scores reliable and representative of the malicious nature of a given text.

3.2. Disagreement on implicit judgments

The following example shows how diverse interpretations of the same sentence can be:

Sentence: And rather than built the country the ANC have used their super majority in parliament to make marxist policies a priority.

Interpretation A: Author is supporting apartheid.

Interpretation B: The country has given importance to build up socio-economic policies.

Since annotating the presence and social evaluation of a character trait is more intuitive and therefore less prone to misconception than the classification in a philosophical theory such as Virtue Ethics, we analyze the disagreement on implicit moral judgments based on the annotations for presence and social evaluation.

We observe that diversification in interpretation already starts with recognizing the presence and absence of moral judgments. For only 291 sentences (14.42%), all annotators agree that the author passed *at least one* judgment. For only 56 sentences (2.78%)⁹, all annotators agree that *no* judgments are passed. These findings demonstrate the high degree of disagreement on the presence of implicit moral judgments.

When annotators distinguish an implied character trait of a person entity, they show low agreement on its desirability or evaluation in society – with Krippendorff's $\alpha = .354$ over the annotators' evaluations of each entity. This means that often one annotator perceives a negative judgment of an entity's character while another considers it positive, and vice versa. Even if they agree on its desirability, their interpretation seems to be affected by their attitude towards the author (see Table 3). Interestingly, the moral judgment annotations for the second example sentence in Table 3 indicate that judgments can be passed for both active and passive actors in the original sentence. Fig. 3 shows that most recognized character traits are considered vices. An explanation for this phenomenon is not straightforward as it is unclear whether this is due to an actual higher frequency, higher ease of recognition and production, or a cognitive bias towards negativity in both author and reader.

3.3. Common ground and filling in the gaps

When qualitatively comparing interpretations with the original sentences, we distinguish traces of common ground. For instance, annotators explain abbreviations;

⁹ These sentences are included in the dataset, even though no implied judgments were annotated for them in the second annotation round.

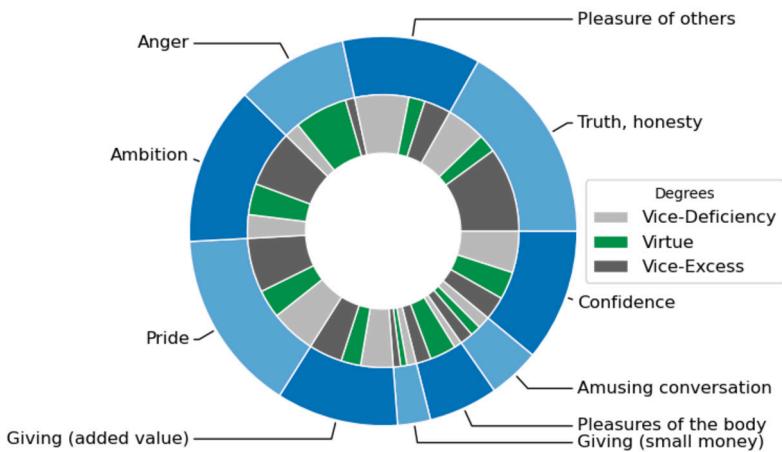


Fig. 3. Distribution of Spheres of Action (outer circle) and degrees (inner circle) annotated in the dataset. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Sentence: For example it seems like a lot of people just post on Instagram and Facebook about BLM but would they be willing to take action in the face of racism.

Interpretation: People might support Black Lives Matter and post about it for social media but aren't truly activists if they wouldn't actually take action if they encounter racism.

recognize references to social conflicts and polarizing topics;

Sentence: If the woman physically cannot give birth, once again, this should be verified by a judge.

Interpretation: Women should not have autonomy over their own bodies.

and understand rhetorical strategies:

Sentence: Do people even know why Lenin abandoned Socialism?

Interpretation: Many people don't know why Lenin abandoned Socialism.

We also notice that annotators actively reason over seemingly implicit information. They, for example, adhere to their own mental representation of people to fill in social templates:

Sentence: Why that cisgender person doesn't wish to date a trans person or why he broke up.

Interpretation: Why that cisgender person doesn't want to date a trans person or why he broke up with her.

They sometimes assign a gender to the author – even though the author's identity is unknown to them:

Sentence: And I'm not sure if the guy I'm seeing knows that.

Interpretation: The person is afraid of having to explain to the guy what he doesn't know about her.

We also observed a tendency among the annotators to indicate a negative attitude towards the author of the input sentence and to suggest that they hold different beliefs by using patterns such as "*the writer/author ... claims/believes/argues that ...*". To analyze this, we divided the interpretations of the training sentences into two groups: interpretations of sentences that start with this pattern (Group A) and interpretations of sentences that do not (Group B). In Group A, 18.10% of the interpretations expressed a very negative attitude towards the author, compared to only 6.93% in Group B. Similarly, 22.19% of interpretations in Group A had a negative attitude, while 19.75% in Group B did. This behavior is exemplified below by a sentence and its interpretation written by an annotator with a very negative opinion towards the author of the sentence.

Sentence: Even if the police had zero implicit bias (somehow), you would still have totally innocent black people killed by police at an infinitely higher rate than white people, all because 100% of criminal suspects are black.

Interpretation: The author wrongly states that 100% of criminal suspects are black and that's why they get killed by the police more than white people.

Note that the interpretations could be influenced by the lack of rich context surrounding a given sentence. For instance, annotators may not have felt the need to explain acronyms or refer to social conflicts if more context was available clarifying that information.

Nevertheless, further linguistic and rhetorical analyses could identify the governing effects of common ground on interpretation and its reflection in the interpretation's wording but are beyond the scope of this paper.

4. Generation methods for interpretation modeling

4.1. Task definition

The goal is to generate multiple interpretations i of an input sentence s . We assume that a reader or a representative group of readers j each have their own interpretation i_j of s . Reader j is characterized by attitude a_j towards the author upon reading s and the implicit moral judgments m_j they distinguish in s . The context of reader j , which we call $g_j = [a_j, m_j]$, socially grounds s in i_j . Social context g can be a priori obtained or inferred jointly with the generation. We assume the former in our experiments below. For instance, a multi-task objective in the latter case would require a substantially larger number of training samples. Together with s and g , the title ti of the post from which s is taken is given. We approach IM in two autoregressive generation setups.

One-to-one generation The objective is generating one interpretation i_j of sentence s by reader j based on $p_\theta(i_j | s, ti, g_j)$. This process is repeated J times to obtain J interpretations representing J readers.

One-to-many generation J interpretations i are jointly generated conditioned on s , ti , and social context $g_{1\dots J}$, that is $p_\theta(i_{1\dots J} | s, ti, g_{1\dots J})$ is computed, where $i_{1\dots J}$ and $g_{1\dots J}$ respectively represent a long string of J interpretations and social contexts each delimited by a separator token. In this setup, the context of previously generated interpretations is taken into account during generation.

4.2. Input feature representations

s , ti , and i are variable-length sequences composed of tokens from a vocabulary V , a is a scalar following a five-point Likert scale, and m presents Q sequences of moral judgments where Q equals the number of unique people entities mentioned in s , preserving their order of appearance: $m = ((ent_1, pres_1, desc_1, eval_1, soa_1, vi_1), \dots, (ent_Q, pres_Q, desc_Q, eval_Q, soa_Q, vi_Q))$ – with ent the person entity as mentioned in s , $pres$ a binary indicator marking the presence/absence of an implied character trait, $desc$ the trait's description using tokens from V , $eval$ an evaluation label (i.e., *good* | *bad*), soa a sphere-of-action label, and vi a contextual appropriateness label (i.e., *vice of deficiency* | *virtue of mean* | *vice of excess*). The moral judgment of entity q is then m_q . If entity q is not judged, that is $pres_q = 0$, $m_q = (ent_q, 0, "", 0, 0, 0)$. We use $Q' \in [0, Q]$ to indicate the number of entities q' judged in s .

Following the success of prompting with supervised learning [48], we apply a prompting function $f_{prompt}(\cdot)$ to jointly represent s , ti , and g . This way, we lay bare the relations between the various input features. The function completes a predefined template that contains four subtemplates (colored boxes), each designated to a specific input type. The subtemplates are joined together by separator token $<\text{sep}>$:

Input template: title Title: <i>ti</i> $<\text{sep}>$ attitude Attitude: $[str(a)]^J$ $<\text{sep}>$ judgments judgments: $(ent_{q'}, desc_{q'}, eval_{q'}, soa_{q'}, vi_{q'})^{Q'}$ $<\text{sep}>$	Moral Judgments: $[(ent_{q'} = desc_{q'}, which\ is\ a\ eval_{q'}\ character\ trait\ and\ a\ vi_{q'}\ related\ to\ soa_{q'})^{Q'}]^J$
sentence Sentence: <i>s</i>	

In the attitude subtemplate, $str(\cdot)$ maps a from its scalar value to its attitude label (i.e., {1: *very negative*, 2: *negative*, 3: *neutral*, 4: *positive*, 5: *very positive*}). In the judgments subtemplate, the phrase in parentheses is repeated Q' times. If $Q' = 0$, then “*Moral Judgments: None*”. Various input templates of different input lengths, label ordering, and phrasings have been crafted, tested, and evaluated when designing the input template. The final template is designed in such a way that it makes the relation between the various labels explicit through text and the subtemplates can be automatically adapted when labels are missing.¹⁰

The prompting template is used across all generation settings. For aggregating the social context $g_{1\dots J}$ over J readers, the phrases in squared brackets ($[]$) in the attitude and judgments subtemplate are repeated J times such that the attitude and moral judgment of reader j are aligned by their position in the input template. The input template or string is encoded with a pre-trained language model.

4.3. One-to-one generation

Decoder A single interpretation i_j is generated conditioned on s , ti , and g_j ; $p_\theta(i_j | s, ti, g_j)$. This process is repeated J times for each reader j to obtain J interpretations representing J readers.

¹⁰ In rare cases, one or more variables in m are lacking, then the part between slashes (//) to which the missing variable belongs is removed from the judgment subtemplate: $ent_{q'} = /desc_{q'}$, which is / a $eval_{q'}$ character trait and / a $vi_{q'}$ / related to $soa_{q'}$ / to maintain a naturally occurring sentence.

Training The model parameters are optimized using the standard language modeling objective, that is a negative log-likelihood loss:

$$\mathcal{L}_1 = - \sum_{j=1}^J \frac{1}{L} \log \prod_{t=1}^L p_\theta(i_{j,t} | i_{j,<t}, s, ti, g_j) \quad (1)$$

where L refers to the number of tokens in target interpretation i_j . **[One2One]**

4.4. One-to-many generation

Decoder The generated output is a sequence of J interpretations i_j concatenated by a special reader token forming $i_{1\dots J} = i_1 <\text{reader}> i_{1+1} \dots <\text{reader}> i_J$. The model is expected to infer J from the J contexts given in the input template.

Training Model parameters are optimized using a negative log-likelihood loss:

$$\mathcal{L}_m = -\frac{1}{T} \log \prod_{t=1}^T p_\theta(i_t | i_{<t}, s, ti, g_{1\dots J}) \quad (2)$$

where T is the number of tokens of the target string of the decoder that contains all interpretations. The decoding is conditioned on previously generated interpretations. Autoregressive decoding is affected by the order in which the interpretations are generated and consequently optimized. For this, we develop control mechanisms derived from the properties of interpretation.

4.4.1. Arbitrary ordering

During training mode, parameters are optimized using an arbitrary ordering of the ground truth interpretations, that is the ordering in which they appear in the dataset. **[One2M-Arb]**

4.4.2. Ordering by semantic similarity

The ground-truth interpretations in $i_{1\dots J}$ are ordered by their semantic similarity to s . This follows the assumption that the more covert or surprising a hidden meaning is, the more complex the reasoning for recovering it becomes, and the greater its semantic distance from the sentence's apparent semantics. Low semantic similarity between a sentence and an interpretation then signals high covertness of the hidden meaning while high similarity suggests low covertness. We approximate covertness using the semantic similarity between i_j and s . Low similarity signals high covertness, and vice versa. Therefore, if $\text{sim}(i_j, s) > \text{sim}(i_{k\neq j}, s)$, the hidden meaning governing i_k is considered more covert than that in i_j . Naturally, a ranking r : $\text{sim}(i_j, s) > \text{sim}(i_k, s) \mapsto r(i_j) > r(i_k)$. Ultimately, the interpretations composing $i_{1\dots J}$ are ordered in descending order following r during optimization. Semantic similarity $\text{sim}(\cdot, \cdot)$ is based on the cosine similarity and takes representations computed by SimCSE [24] as it was shown to align best with human evaluation of semantic similarity [80]. **[One2M-Sim]**

4.4.3. Constraining on semantic similarity

In addition to the above ordering of ground-truth interpretations in $i_{1\dots J}$, a decrease in semantic similarity of the generated interpretations in $\hat{i}_{1\dots J}$ to s is explicitly enforced during training. The decoded interpretation sequence $\hat{i}_{1\dots J}$ is split in sentences, where sentence \hat{i}_j presents the predicted interpretation for reader j . Next, a language model computes a latent representation of \hat{i}_j and s , called respectively i'_j and s' , after which their semantic similarity is computed. Related to previous work on abstractive summarization [49], a contrastive loss function \mathcal{L}_{sim} based on a hinge loss enforces a decrease in semantic similarity with s' among the predicted interpretations $i'_{1\dots J}$:

$$\mathcal{L}_{\text{sim}} = \frac{1}{J} \sum_{j=2}^J \ell(s', i'_j, i'_{j-1}) \quad (3)$$

where

$$\ell(s', i'_j, i'_{j-1}) = \max(0, \text{sim}(i'_j, s') - \text{sim}(i'_{j-1}, s') + m) \quad (4)$$

with margin m forcing a substantial difference between the generated interpretations. SimCSE and cosine similarity are again used for computing $\text{sim}(\cdot, \cdot)$.

Model parameters are optimized using a weighted combination of the language modeling and similarity decrease objective: $\mathcal{L}^{**} = \alpha \mathcal{L}_m + (1 - \alpha) \mathcal{L}_{\text{sim}}$. This way, we can explicitly control the trade-off between both losses. **[One2M-Con]** Losses are summed over N training sentences.

4.5. Alternative decoding models

There are many possibilities of designing a decoder that generates different interpretations during training. We also experimented with models that used context g_j as a prompt to autoregressively decode the interpretation both in the one-to-one and one-to-many settings instead of using g_j as input together with s and ti . In the one-to-many case, this leads to a multi-branch decoder [67] with

J parallel decoders where each j^{th} decoder autoregressively generates i_j taking g_j as decoder prompt. Alternatively, we designed an additional loss function that during training enforces the similarity of a generated interpretation with g_j . None of these approaches could improve the results, often yielding nonsensical interpretations that disproportionately focused on g_j .

5. Experiments

In this section, we describe the experimental setup and the metrics used for evaluating the performance of the proposed generation frameworks on the IM task. We then continue to thoroughly investigate the success of the frameworks and the importance of IM for content moderation. To that end, we formulate and answer several pertinent research questions in § 6, § 7, and § 8.

5.1. Experimental setup

p_θ is parameterized by large pre-trained language models: decoder-only forward language model GPT-2¹¹ [64] and encoder-decoder architecture T5¹² [65]. Model parameters are optimized using Adam optimization ($\text{lr} = 1e^{-6}$, weight decay = 0.01, $\epsilon = 1e^{-8}$, $\beta_1 = 0.9$, $\beta_2 = 0.99$) with batch size = 2 and gradient clipping (max norm = 0.1). During training, early stopping is performed on the validation loss with patience = 5. We adopt diverse beam search [78] (max. length = 150, beam size = 10, beam groups = 5, diversity penalty = 2.0, repetition penalty = 1.2) with early stopping as the standard decoding method and use the interpretation with the highest probability in the beam as output. In [One2M-Con], we also test with greedy search as decoding method for producing interpretations during training. We set margin $m = .05$ in \mathcal{L}_{sim} to ensure that the last generated interpretation still relates to the original sentence. Data is split in a train (80%), test (10%), and validation (10%) set in a title-stratified manner, ensuring sentences from the same post are represented in the same set.¹³ We experiment with various values for hyperparameter $\alpha \in [.25, .50, .75]$.

5.2. Evaluation metrics

We report the lexical overlap between the generated and target interpretations using BLEU [58] and ROUGE [43], and their semantic similarity using MoverScore [89] and BERTScore [88]. COMET¹⁴ [69,68] is a trained metric that projects the sentences, generated interpretations, and ground-truth interpretations in a shared feature space, combines their latent representations into a single vector, and produces a score between 0 and 1 reflecting the quality of the generated interpretation. We also compute unigram-based perplexity. We opt for reference-based metrics, where the generated interpretations are compared against the expected ground-truth interpretations, since we aim to generate human-like interpretations and scores yielded by reference-free metrics have been shown to misalign with this goal [14]. In one-to-one generation, each generated interpretation is evaluated against their respective target interpretation. In one-to-many generation, we first match each interpretation in the generated sequence of interpretations with one target interpretation in the target sequence. For this, we employ the Kuhn-Munkres algorithm [40], also known as the Hungarian algorithm, which is a combinatorial optimization algorithm that efficiently pairs predictions and targets by computing their matching. Here, $\text{score}(\cdot) = 100 - \text{BLEU-1}$ is used for scoring the matches. We then evaluate each generated and target pair, as done in one-to-one generation.

6. Quantitative results

RQ 1. Which of the proposed generation frameworks is most appropriate for IM?

Performance results on the held-out test set are provided in Table 4. The T5-based models outperform those adopting GPT-2 as backbone language model on almost all metrics. This can be explained by the better encoding and cross-attention between decoder and encoder in the encoder-decoder model of T5. This advantage is consistent across the one-to-one and one-to-many generation setups. The task is not just autoregressively completing the input sentence with an interpretation as done in a decoder model such as GPT-2, but to translate the input sentence into the viewpoint and language of the reader for which both a good encoding of the content of the input is needed as well as a decoding step that cross-attends over the encodings of the input sentence.

Comparing the two generation setups, we notice that one-to-one generation shows overall better performance than one-to-many generation. This is to be expected as the generation of each interpretation is guided by its own input template or prompt in the one-to-one generation setup. The one-to-many generation models, on the other hand, have to learn to generate different interpretations and to attend to the corresponding context in the input template. Adding implicit ([One2M-Sim]) control by ordering interpretations based on their semantic similarity to the sentence in the ground truth slightly improves performance, suggesting that the order in which the ground-truth interpretations are presented during optimization indeed affects generation. Explicitly enforcing a decreasing

¹¹ GPT2-Large; Huggingface implementation of GPT2LMHeadModel.

¹² T5-Base; Huggingface implementation of T5ForConditionalGeneration.

¹³ Using Scikit-Learn GroupKFold.

¹⁴ <https://huggingface.co/Unbabel/wmt22-comet-da>.

Table 4

Experimental test results. PP = Perplexity; DBS = diverse beam search; greedy = greedy search. For BERTScore, we report F1 scores. Best results with GPT-2 as encoder backbone language model ('-GPT') are in **bold** and best results with T5 as encoder-decoder backbone language model ('-T5') are underlined.

α Values	BLEU 1 / 2 / 3 / 4 ↑	ROUGE 1 / 2 / L / Lsum ↑	Mover Score ↑	BERT Score ↑	COMET ↑	PP ↓
One-to-One Generation						
One2One-GPT	-	8.23 / 2.62 / 1.04 / .46	16.58 / 2.34 / 11.33 / 12.91	.0055	.9568	.5239
One2One-T5	-	31.64 / <u>24.12</u> / <u>18.34</u> / 14.71	<u>42.01</u> / 20.24 / 37.40 / 37.50	.3231	.9714	<u>.6749</u>
One-to-Many Generation						
One2M-Arb-GPT	-	11.60 / 3.07 / 1.13 / .36	15.05 / 1.40 / 11.08 / 11.15	.0128	.9623	.4922
One2M-Sim-GPT	-	<u>12.49</u> / 3.44 / 1.16 / .42	15.23 / 1.60 / 11.30 / 11.38	.0081	.9625	.4803
One2M-Con-GPT	.25 (DBS)	12.75 / 3.35 / 1.32 / .59	14.57 / 1.29 / 11.31 / 11.35	.0133	.9639	.4823
	.50 (DBS)	11.84 / 3.13 / 1.29 / .60	13.37 / 1.38 / 10.52 / 10.56	.0032	.9646	.4586
	.75 (DBS)	11.94 / 3.29 / 1.35 / .58	13.94 / 1.46 / 10.81 / 10.86	-.0029	.9642	.4593
	.25 (greedy)	11.97 / 3.10 / 1.21 / .57	14.18 / 1.37 / 11.05 / 11.10	.0069	.9644	.4757
	.50 (greedy)	11.70 / 2.83 / .96 / .35	15.38 / 1.58 / 11.55 / 11.61	.0157	.9618	.5048
	.75 (greedy)	11.63 / 3.06 / 1.13 / .41	14.12 / 1.47 / 10.75 / 10.78	-.0001	.9628	.4769
	One2M-Arb-T5	-	32.25 / 22.51 / 17.81 / 14.75	38.07 / 18.68 / 32.67 / 32.75	.2964	.9626
One2M-Sim-T5	-	<u>32.44</u> / 22.29 / 17.70 / <u>14.80</u>	36.42 / 17.40 / 31.30 / 31.37	.2944	.9630	.6194
One2M-Con-T5	.25 (DBS)	25.95 / 18.11 / 14.45 / 12.12	27.63 / 12.54 / 23.66 / 23.73	.1758	.9647	.5292
	.50 (DBS)	27.10 / 18.83 / 14.97 / 12.47	29.40 / 13.83 / 25.06 / 25.14	.1893	.9646	.5430
	.75 (DBS)	27.03 / 18.81 / 15.00 / 12.57	28.20 / 13.15 / 24.13 / 24.20	.1770	.9646	.5352
	.25 (greedy)	25.48 / 17.51 / 13.81 / 11.46	27.04 / 12.36 / 23.25 / 23.29	.1688	.9647	.5255
	.50 (greedy)	26.29 / 18.33 / 14.53 / 12.09	28.21 / 13.46 / 24.30 / 24.36	.1805	.9645	.5417
	.75 (greedy)	25.98 / 17.79 / 14.10 / 11.79	27.39 / 12.43 / 23.40 / 23.46	.1725	.9651	.5288
	One2M-Con-T5	-	-	-	-	12.37

semantic similarity between the sentence and the generated interpretations in one-to-many generation ([One2M-Con]) yields overall performance that is either similar to or worse than the setting without additional control ([One2M-Arb]). Lastly, the choice of decoding method (i.e., diverse beam search and greedy search) does not seem to substantially influence generation performance.

Note that automatic evaluation metrics are known for poorly correlating with human evaluation. We therefore further analyze and assess the models and their generated interpretations in human evaluation setups (§ 7).

RQ 2. Did the models learn to correlate the diversity in interpretation with the diversity in social grounding?

The difference between interpretations of a sentence is linked with a difference in social grounding, among other things. The IM models should therefore have learned to diversify the interpretations if their grounding information differs while generating similar interpretations with similar social grounding. We quantify the lexical difference between two interpretations from readers j and k as $di(i_j, i_k)$.

$$di(i_j, i_k) = 100 - \text{BLEU-1}(i_j, i_k) \quad (5)$$

We then obtain the difference in grounding information guiding i_j and i_k , respectively g_j and g_k , by simply summing their difference in attitude and moral judgments:

$$dg(g_j, g_k) = |a_j - a_k| + \frac{1}{Q} \sum_{q=1}^Q \text{non_overlap}(m_{j,q}, m_{k,q}) \quad (6)$$

where $|a_j - a_k|$ represents the difference in Likert scores (see § 4.2) (min. 0, max. 4) and $\text{non_overlap}(m_j, m_k)$ counts the non-overlapping moral judgment characteristics for each entity q (min. 0, max. 5). Thus, $dg(g_j, g_k) \in [0, 9]$, so that $dg(g_j, g_k) = 0$ means that reader j and reader k have the exact same attitude and distinguish for all Q entities the exact same moral judgments, and $dg(g_j, g_k) = 9$ means that they have highly opposing attitudes (i.e., *very negative* ($a = 1$) and *very positive* ($a = 5$)) and disagree completely on the implied moral judgments. As expected, diversity in social grounding dg is positively correlated (Pearson) with diversity in interpretation di ($p < .01$). The positive correlation is stronger with One2One-GPT ($r = .5526$) than with One2One-T5 ($r = .3321$) and approaches that of the gold standard ($r = .5887$). Note that correlation is only computed for the one-to-one generation settings since the interpretations align with the grounding information they rely on.

At first sights it seems that GPT-based models better leverage the grounding information than those adopting T5. However, Table 4 shows a much better correspondence of the interpretations generated by One2M-Sim-T5 than by One2M-Sim-GPT. Qualitative inspection described in the next paragraph notices a higher degree of hallucinations by GPT-2 which might explain the higher diversity when dealing with diverse social grounding.

Table 5

Example of generated and ground-truth interpretations of a sentence taken from the test set.

Sentence	<i>I hear a lot about adults job jumping nowadays just to get bigger wages, and honestly?</i>
One2M-Sim-GPT	(1) Jobs are jumping today just for the sake of getting bigger paychecks, but I don't hear much about it. (2) author = a job is jumping because it's the only way to make a living, not because of any other reason. (3) writer = job hopping because they want to be able to afford to live in a big city, so they can afford a bigger house, car, etc., and so on. (4) sentence: It seems like the job market has gone back in time to the days of the baby boomers, when adults were the main breadwinners and needed to have a stable job to support their families.
One2M-Sim-T5	(1) People nowadays are jumping into jobs just to get higher wages. (2) I hear a lot about adults job jumping nowadays just for bigger wages, and honestly? (3) The writer asks if adults are now jumping in jobs to earn more money (4) Adults are getting richer
Ground truth	(1) Adults are changing jobs for bigger paychecks. (2) The writer describes having heard about many people changing jobs to get higher wages. (3) People switching jobs for better wages is a real awful situation nowadays. (4) People are only interested in money and not stability. (5) Capital pursuit is not worth moral sacrifice.

Table 6Overview of the evaluation labels for *diversity* and *meaning* with their scores used during the human evaluation tasks.

Diversity	Meaning
0: Not diverse at all	0: <i>Exact copies</i>
1: Slightly diverse	1: <i>The same</i> meaning
2: Moderately diverse	2: <i>Slightly different</i> meanings
3: Very diverse	3: <i>Moderately different</i> meanings
4: Extremely diverse	4: <i>Very different</i> meanings, still related
	5: <i>Opposite</i> meanings, still related
	6: <i>Unrelated</i> meanings

7. Human evaluation

In a first qualitative evaluation, we manually inspect all generated interpretations of the test set. Table 5 shows an example. We notice that the one-to-many generation models with T5 overall generate more diverse interpretations than their one-to-one generation counterpart, One2One-T5. The one-to-one generated interpretations stay rather close to the semantics of the sentence. With GPT-2, diversity is generally higher than with T5 in both generation setups. However, we observe a high degree of hallucination and nonsense with GPT-2 (e.g., “*jobs are jumping*”). With both language models and all generation methods, the generated interpretations sometimes lack the level of complex reasoning that guided the target interpretations (e.g., ground-truth interpretation (4) and (5) in Table 5).

We also perform a systematic quantitative human evaluation. We hire two students with a bachelor degree in English linguistics to assess the ground-truth and generated interpretations for 100 randomly sampled sentences from the test set in more detail. Detailed descriptions of the assessors are provided in the Appendix (§ A.4). We take the interpretations generated by One2M-Sim-T5 since quantitative and qualitative results suggest that it generates the highest-quality interpretations. The assessors can leave free-text comments during evaluation, and we conduct a 45-minute interview with each assessor separately at the end of the assessment.

RQ 3. Is an automatic model able to generate interpretations that are diverse in meaning? And does that diversity align with the diversity found in the annotated ground truth?

We want to evaluate a model’s ability to generate diverse interpretations and check whether the generation model has learned to reason over the ambiguous nature of a sentence and diversify the generated set of interpretations accordingly. The human assessors first evaluate the diversity of 100 sets of generated and ground-truth interpretations randomly selected from the test set in terms of the meaning they represent. The sets are interchangeably presented to them as ‘SET A’ and ‘SET B’. They indicate for each set the diversity on a five-point Likert scale, see Table 6 for the **diversity** labels. We show the sets next to each other such that their diversity is indirectly compared against each other. Inter-rater agreement is fair, with Cohen’s weighted (quadratic) kappa $\kappa = .30$ (ground-truth set) and $\kappa = .34$ (generated set). Both assessors indicate in the interviews that it is cognitively challenging to assess the difference in meaning between multiple sentences in a set. This could explain why agreement between the two assessors is only fair. Whereas one assessor does not observe significantly different diversity in the ground-truth and generated interpretation sets, respectively $\mu = 1.83$ and $\mu = 1.91$, the other assigns overall higher diversity in the generated interpretation sets, $\mu = 1.69$ (ground truth) and $\mu = 2.32$ (generated), $p < .01$. See Fig. 4 for the diversity score distributions for both assessors. The two diversity distributions for the generated interpretations as assessed by the two assessors suggest that the model is overall able to generate moderately diverse interpretation sets, which are in line with the overall diversity found in the ground truth.

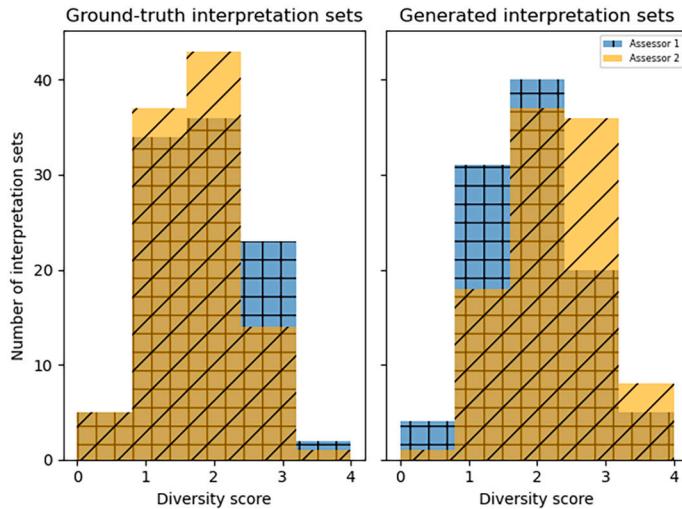


Fig. 4. Histograms showing the diversity score distributions for the ground-truth interpretation sets (on the left) and the generated interpretation sets (on the right). The orange distributions (hatching pattern: /) present the first assessor and the blue distributions (hatching pattern: +) the second assessor. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

We examine whether the model is able to capture the diversity in interpretation triggered by a sentence. Given the fairly low inter-rater agreement, we analyze the Pearson correlation between the diversity scores for the ground-truth and generated interpretation sets for each assessor separately. The correlation is positive – weak (first assessor; $r = .17$) or negligible (second assessor; $r = .04$) – though not statistically significant, $p > .05$. The model seems to generate more diverse interpretation sets for sentences that have less diverse ground-truth interpretation sets while it slightly better captures the implicit diversity for sentences with more diverse ground-truth sets.

RQ 4. Do the generated sentences actually present interpretations or merely rewrites?

We want to analyze whether the generated interpretations are actually interpretations and not merely copies ('exact copies'), paraphrases ('the same meaning'), or unrelated to the sentence ('unrelated meanings'). We broadly consider a generated sentence an interpretation if it presents a different meaning that is still related to the sentence since a paraphrase is expected to maintain the sentence's semantics. The assessors are given 361 sentence-interpretation pairs presented as '*A*' and '*B*' and asked to label the difference in meaning between the sentence and the interpretation. See Table 6 for the **meaning** labels. We observe substantial inter-rater agreement: $\kappa = .72$. Fig. 5 presents the meaning distributions. It seems that One2M-Sim-T5 often copies (7-13%) or nearly copies/paraphrases (18%) the original sentence. The assessors sometimes motivate their choice of the latter in the comments: the difference between the sentences is marginal (i.e., omission of function words or articles, typos, or contractions) or one sentence is simply a paraphrase of the other. In rare cases, the model generates sentences that are unrelated to the original sentence (2-3%). Nevertheless, it is overall able to generate interpretations as the majority of generated sentences have been said to differ in meaning from the original sentences in various degrees. The assessors note that a difference in meaning is often established by nuances in word semantics.

8. Use case: interpretation modeling for content moderation

Interpretations of the input sentence have the potential to reveal implicit toxic content that may not have been inferred by all readers instantly. An author could have hidden that content on purpose when they have malicious intentions. We therefore showcase the meaningful contributions IM can make to content moderation of online discussions by revealing subtly toxic communication. To that end, we analyze and compare the levels of toxicity, insult, and identity attack within sentences, considering the ground-truth interpretations provided by the human annotators and the interpretations generated by one of the proposed IM models. Scores on toxicity, insult, and identity attack are obtained through the Perspective API.

The following two research questions are answered employing similar analyses, with RQ5 focusing on the *ground-truth interpretations* in the *training set* and RQ6 concentrating on the *generated interpretations* produced by One2M-Sim-T5 for sentences in the *test set*. The analyses also introduce the concept of an *interpretation cluster* that combines the input sentence and its corresponding interpretations into one set. The input sentence namely retains its significance and should therefore be considered together with the interpretations in real-life content moderation settings. We perform the following primary analyses:

- *Analysis 1 - The revealing nature of interpretation.* We investigate if the interpretations reveal underlying toxicity, insults, and identity attacks in the input sentence. To that end, we compute a toxicity, insult, and identity attack score for each instance in

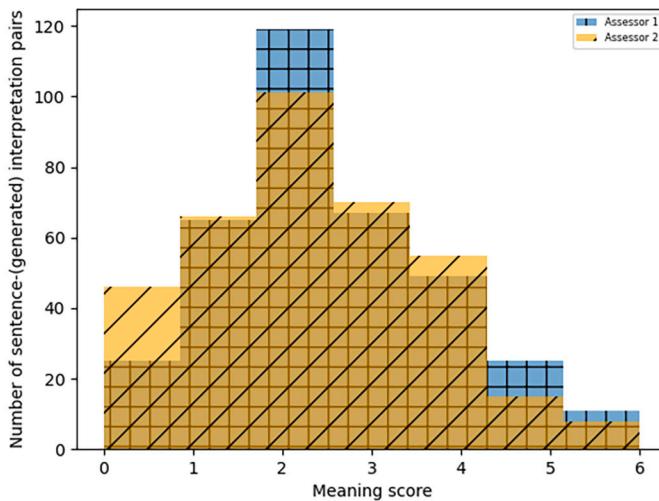


Fig. 5. Histogram showing the meaning score distribution for the sentence-(generated) interpretation pairs. The orange distribution (hatching pattern: /) presents the evaluation of the first assessor and the blue distribution (hatching pattern: +) that of the second assessor. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

the interpretation cluster and flag the input sentence if an interpretation yields the maximum toxicity, insult, or identity attack score in the interpretation cluster. The difference in score between the input sentence and that interpretation can be of any value (*analysis 1A*) or at least 10 percentage points, marking a substantial difference (*analysis 1B*). This value is selected to represent the largest jump in the distribution of the differences between the maximum toxicity scores of the interpretations and the toxicity scores of their corresponding input sentences (first line of Fig. 6 and Fig. 7).

- **Analysis 2 - Harmless on the surface but very offensive between the lines.** The analysis of the interpretations is especially important when the original sentence has low levels of toxicity, insult, or identity attack and its interpretation cluster contains one or more interpretations with a high toxicity, insult, or identity attack score. We flag an input sentence when it has a score below 0.1 and at least one interpretation in its interpretation cluster has a score above 0.5. Note that we deliberately choose to set the distance between the upper bound of the sentence score (<0.1) and the lower bound of the interpretation score (>0.5) to be substantially large here. In practice, content moderators could choose those bounds, allowing them to decide how strictly they moderate implicitly toxic, insulting, and attacking content.

In conjunction with the primary analyses, we investigate the ability of the IM model to adequately replicate human interpretation behavior in RQ6 (*Model vs humans*). This entails assessing the agreement or overlap between the flagged sentences in the test set when considering the generated interpretations (= model behavior) and when considering the ground-truth interpretations (= human behavior) in the interpretation clusters. This is done by first obtaining the outcomes of the two primary analyses for the generated and ground-truth clusters separately, after which we quantify their overlap as a percentage. The overlap represents the proportion of the original sentences that should be flagged according to the ground-truth interpretations (written by humans) that the model was able to flag as well using the interpretations it produced. An overlap of 80%, for instance, signifies that the model flagged 80% of the sentences that were also flagged by humans in the ground-truth sentences. The reason why we rely on recall for our analyses is because false negative cases are more relevant than false positives. In content moderation, it is worse to fail to recognize a toxic interpretation (false negative) than to mistakenly label a non-toxic interpretation as toxic (false positive). This hypothesis also holds true for interpretations that contain insults and identity attacks. The higher the overlap, the better the model is at mimicking the interpretation behavior of humans.

RQ 5. Do interpretations given by humans actually reveal the hidden toxic nature of sentences?

Analysis 1 - the revealing nature of interpretation As observed in Fig. 6, the ground-truth interpretations have a strong capacity to flag the hidden offensive meaning of an input sentence. In 88.63%, 88.51%, and 90.86% of the cases, a sentence has at least one more toxic, insulting, and attacking interpretation in its interpretation cluster. The difference between the most offensive interpretation and the sentence is however marginal in most cases (*analysis 1A*). Nonetheless, this difference is at least 10 percentage points in 33.96%, 21.74%, and 23.85% of the clusters, suggesting that these clusters unveil hidden meanings that are substantially more toxic, insulting, or attacking than the meanings the input sentences explicitly communicate (*analysis 1B*).

Analysis 2 - harmless on the surface but very offensive between the lines Assuming that the sentences with non-harmful language have a toxicity, insult, or identity attack level lower than 0.1 (812 sentences for toxicity, 1265 for insult, and 1190 for identity attack), we

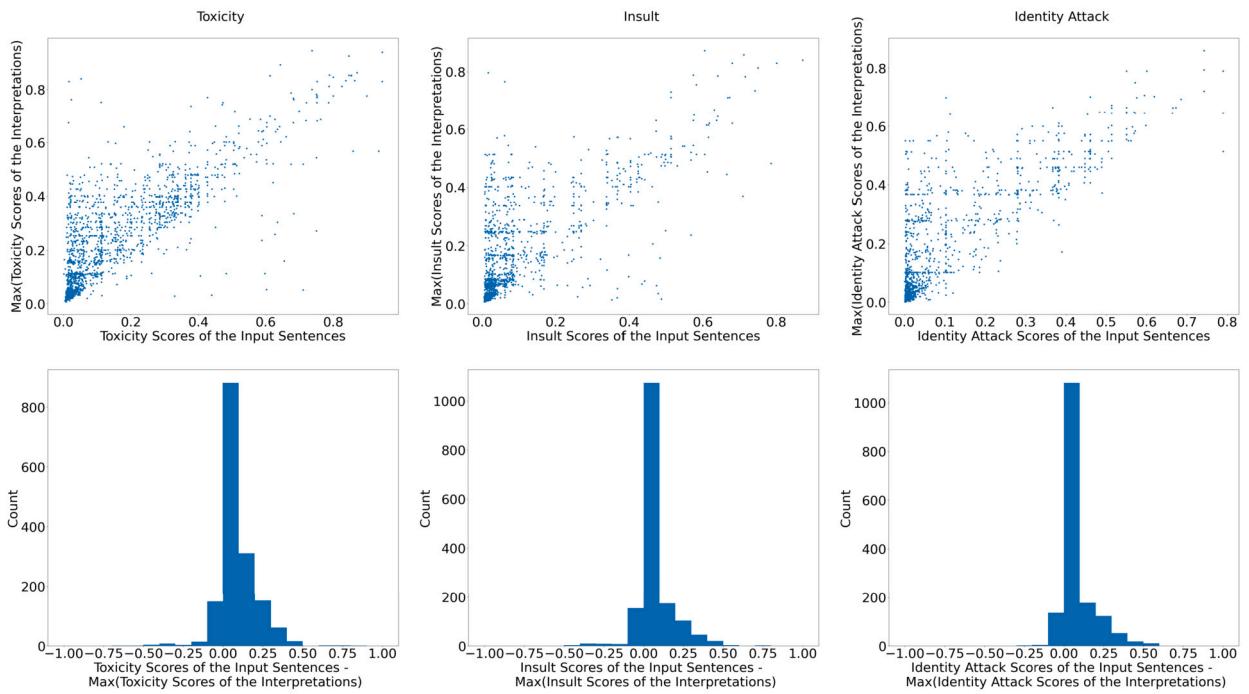


Fig. 6. The histograms in the first row display the differences between the maximum scores for toxicity, insult, and identity attack in the interpretations and the scores reported for the corresponding sentences in the training dataset. The second row shows scatter plots comparing the scores of the training sentences with the maximum scores of their interpretations. Analyzing the interpretations helps in understanding the hidden toxic meaning of the input sentences, as the maximum scores of the interpretation clusters are usually higher than the scores of the input sentences (the differences presented in the histograms of the first row are generally positive). This behavior is particularly important when the scores of the input sentences are low while the maximum scores of the interpretation clusters are high (as illustrated by the scatter plots in the second row).

observe that .3%, .4%, and .7% of the seemingly non-harmful sentences have at least one interpretation with a toxicity, insult, and identity attack score higher than 0.5 (second row of Fig. 6). It is not surprising that these percentages are low. Since the metrics used in the Perspective API produce scores that represent the share of people who would regard a sentence as toxic, insulting, or attacking, it is expected that sentences that are initially considered non-harmful induce human interpretations that are mainly non-harmful as well. These results show that we are able to capture subtly embedded misbehavior that would otherwise remain undetected through IM. Although the < 1% share of implicitly toxic sentences seems marginal at first, the number of sentences can become considerably large in online environments.

Conclusion These findings posit that humans indeed extrapolate implicit toxicity more prominently in the ground-truth interpretations, even when it is subtly embedded in an apparently harmless sentence, empirically substantiating our claim that modeling multiple interpretations of textual content makes sense within the context of content moderation.

RQ 6. Are the generated interpretations able to reveal the hidden meaning of an input sentence? And does the IM model mimic human interpretation patterns?

Analysis 1 - the revealing nature of interpretation We notice that the model is able to reveal the hidden meaning of the input sentence (first row of Fig. 7). It generates more toxic, insulting, and attacking interpretations for 81.86%, 80.39%, and 87.25% of the sentences in the test set (*analysis 1A*). In 13.73%, 7.84%, and 9.80% of the interpretation clusters, an interpretation yields a toxicity, insult, or identity attack score that is more than 10 percentage points higher than that of the sentence (*analysis 1B*).

Model vs human The model seems to be able to mimic human behavior as observed in the corresponding ground-truth interpretations produced by humans in the test set (third row of Fig. 7). Regarding *analysis 1A*, the flagged sentences with generated interpretations in their interpretation clusters and those with ground-truth interpretations in their interpretation clusters overlap by 85.29% (toxicity), 83.91% (insult), and 90.45% (identity attack).¹⁵ The high overlap values are to be expected since the number of flags with both the generated and ground-truth clusters is large. Regarding *analysis 1B*, the flagged sentences overlap by 26.23% (toxicity), 20.0% (insult),

¹⁵ The scores evaluate the overlap using recall. Precision and F1 scores are presented in Appendix A.5.

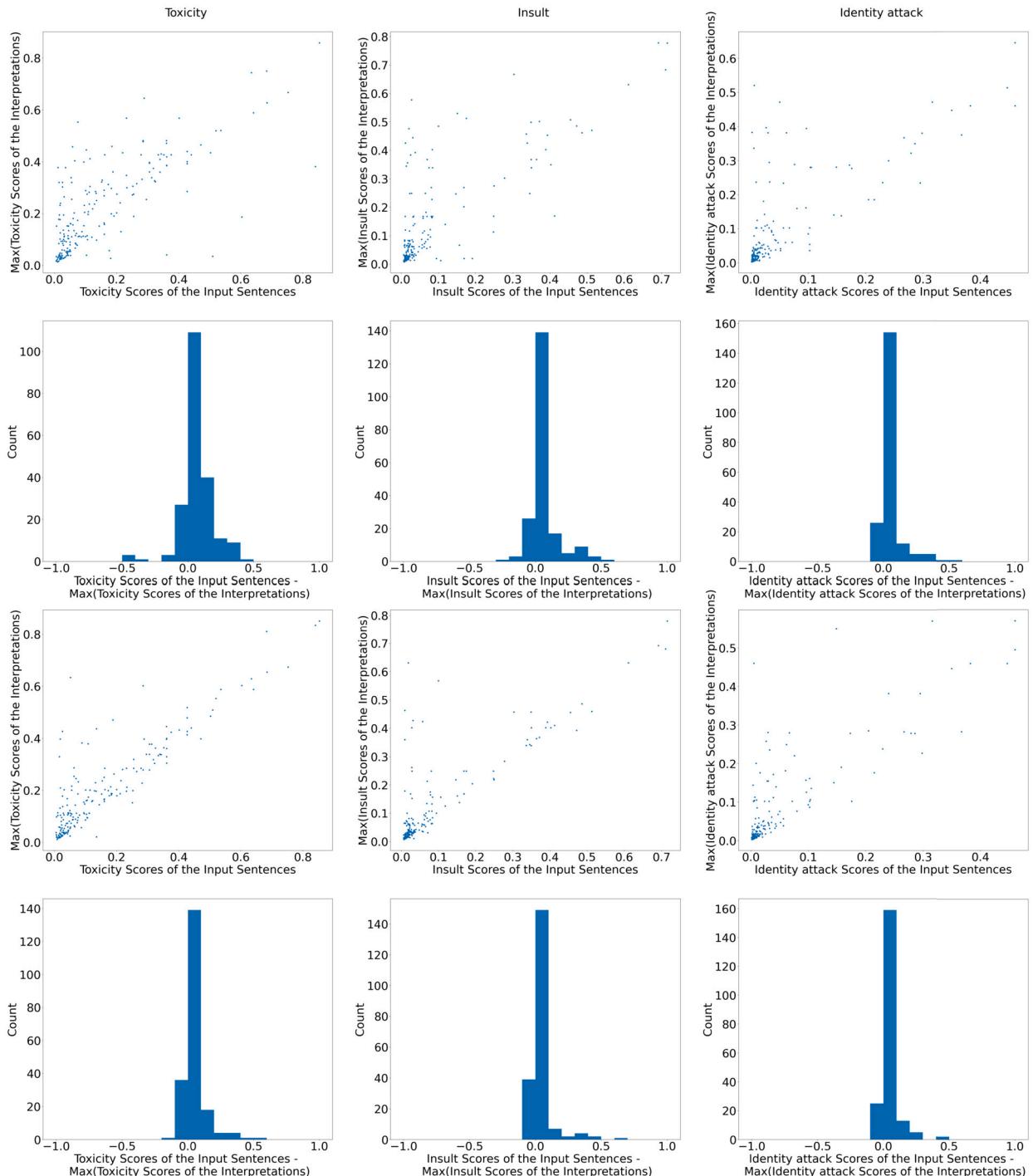


Fig. 7. The histograms in the first and third rows display the differences between the maximum scores for toxicity, insult, and identity attack in the interpretations and the scores reported for the corresponding sentences. The scatter plots in the second and fourth rows compare the scores of the sentences with the maximum scores of their interpretations. The plots in the first and second rows represent the ground-truth test interpretations, while those in the third and fourth rows represent the generated interpretations. Overall, the scores of the ground-truth and generated interpretations are generally similar, particularly in the case of identity attack (as illustrated by the similar Identity Attack histograms in the first and third rows). One2M-Sim-T5 can generate interpretations with toxicity and insult scores similar to those of the ground-truth test interpretations. However, it can sometimes be influenced by the scores of the input sentence, as indicated by the high peaks in the toxicity and insult histograms in the third row and by the cases illustrated near the main diagonal in the scatter plots of the fourth row.

and 41.67% (identity attack).¹⁵ While the rather low overlap values could be attributed to the low numbers of flagged sentences and possible hallucinations by the generation model, they could also be caused by the limited number of human interpretations in the ground-truth interpretation sets. It is namely reasonable that a sentence is indeed offensive between the lines, but the ground-truth interpretations were unable to capture this. This is demonstrated by computing the overlap as a precision metric in Table 9, Appendix A.5. The much higher precision values than the recall values show the interpretations generated by One2M-Sim-T5 can sometimes reveal more hidden meanings than the ground-truth interpretations. This behavior is also illustrated in Table 7, where generated interpretations with a higher toxicity, insult, or identity attack score are plausible even when the ground-truth interpretation sets do not reveal any offensive content underlying the sentences.

Analysis 2 - harmless on the surface but very offensive between the lines Assuming again that the sentences with non-harmful language have a toxicity, insult, and identity attack score lower than 0.1 (111 sentences for toxicity, 168 for insult, and 178 for identity attack), we observe that One2M-Sim-T5 generates at least one interpretation that has a toxicity, insult, or identity attack score that is higher than 0.5 for fewer than 1% of the sentences (fourth row of Fig. 7). Especially in settings where constant streams of textual content are moderated, the number of implicitly harmful sentences becomes considerably large. Note that these results are in line with those in RQ5, proving that the model does not simply recognize harm in every single sentence it interprets.

Model vs human We observe no overlap between the behavior of the human annotators in the respective ground-truth interpretations and that of One2M-Sim-T5, which could be attributed to the low share of flagged sentences.

We briefly investigate the capability of One2M-Sim-T5 to generate spontaneous interpretations even when the language of the sentences and the human-produced interpretations are not very offensive. We do this by looking at the interpretation clusters containing a sentence with low scores and generated interpretations with diverging scores for toxicity, insult, and identity attack. As one can see in Table 7, the model is able to recognize sensitive topics (like political voting) even when the content of the input sentences is formulated in a neutral way. Based on the identified sensitive topics, it can generate diverse interpretations. Another interesting behavior is the creativity of One2M-Sim-T5 that relies on concepts learned during training to extrapolate the meaning of a simple sentence like *But the Democrats?* into insulting interpretations.

An important concern for recent language generation models is the generation of ethical, non-harmful content. While ethical content generation is crucial for preventing the spread of harmful or toxic language, it may also limit the ability of such models to reveal underlying malicious meanings in sentences. This is illustrated in Table 7, which contains a few interpretations generated by ChatGPT.¹⁶ Consistent with the interpretations produced by One2M-Sim-T5, ChatGPT reinterprets the input sentence while also considering the title of the post, the moral judgments about the people mentioned, and the attitude towards the author. While One2M-Sim-T5 can signal malicious hidden meanings, ChatGPT prioritizes ethical standards, generating aligned and ethical interpretations with low toxicity, insult, and identity attack scores. Given the context of content moderation, overlooking or underemphasizing malicious content, as seen with more recent models like ChatGPT, may not be desirable.

Conclusion The model generates interpretations that unveil toxicity hidden below the surface, demonstrating the important application of IM in content moderation. Although interpretation mismatches between the model and the human annotators were found, we show that the model produces interpretations that overall mimic how humans interpret sentences.

9. Related work

We position this work and its contributions in the literature dealing with computational methods for implicit language modeling, social grounding and moral reasoning, reaction modeling, and reasoning over multiple ground truths. We conclude this section by contrasting IM with diverse paraphrase generation.

Implicit communication The use of implicit content and knowledge has attracted attention for processing natural language [34]. For example, hidden information in language has been leveraged to allow models to distinguish more subtle cases of hate speech [18,57,35] and online abuse [82]. Such works primarily performed classification on texts that knowingly or unknowingly conceal their hateful and abusive character, with or without using external knowledge. Related to that is the work on implicit sentiment analysis [42,90]. Others aimed to retrieve implicit in-text and out-of-text knowledge to reveal discourse relations [83,85], event arguments [47,44], or reasoning steps for question answering [25]. This was commonly done in classification and generation setups, where the implicit content was explicitly queried. Another line of work reasoned over implicit relations within or between texts to improve general task performance. For example, Allein et al. [2] and Allein et al. [4] leveraged implicit temporal relations within and between documents for improving fact-checking while Collell et al. [12] reasoned over implicit spatial arrangements of objects mentioned in captions for improving visual object detection. Closer to this work, Sap et al. [71] zoomed in on social and power implications in text. They classified the intentionally offensive and lewd nature of a sentence and then generated the supposedly implied statement regarding the offended group. We discuss related works on implicit social communication more extensively below.

This work stands out from the abovementioned works in two aspects. Firstly, it goes against the general assumption that there exists a single ground truth for implicit data, arguing that computational understanding of implicit content unavoidably goes hand in hand

¹⁶ <https://openai.com/index/chatgpt/>.

Table 7

Three samples from the test dataset that illustrate the diversity of the generated interpretations in terms of toxicity, insult, and identity attack.

[Sentence]	[Toxicity]	[Generated Interpretations]	[Toxicity]	[Generated Interpretations]	[Toxicity]
If young people don't vote in large enough numbers, from a political perspective, their concerns are moot.					0.0186
[Ground-Truth Interpretations]	[Toxicity]	[Generated Interpretations]	[Toxicity]	[Generated Interpretations]	[Toxicity]
If young people are really worried, they should vote to change the situation.	0.0162	[One2M-Sim-T5] From a political perspective, young people's concerns are moot if they don't vote in large enough numbers.	0.0171	[ChatGPT] The author emphasizes that without strong voter turnout from young people, their political concerns lose significance, underscoring the value they place on active and responsible participation in the democratic process.	0.0083
If young people wanted to be represented they would vote.	0.0240	Young people are not worthy of political representation unless they do so.	0.1912	The author argues that if young people fail to vote in significant numbers, their political concerns become irrelevant, highlighting the belief that the current lack of ambition and engagement among young people undermines their claim to political representation.	0.0150
If most young voters don't bother to cast their votes then their concerns will not really be important to politicians and elected leaders.	0.0247	The writer claims that young voters do not deserve or deserve any form of representation because they are inept and lack the confidence to vote.	0.4274	The author dismisses the political concerns of young people who don't vote in large numbers, reflecting a disdainful view that their vanity and lack of genuine pride in civic duty make their issues unworthy of attention.	0.1148
[Sentence]	[Insult]	[Generated Interpretations]	[Insult]	[Generated Interpretations]	[Insult]
But the Democrats?					0.0091
[Ground-Truth Interpretations]	[Insult]	[Generated Interpretations]	[Insult]	[Generated Interpretations]	[Insult]
The Democrats, though?	0.0094	[One2M-Sim-T5] Democrats?	0.0102	[ChatGPT] The author is questioning the Democrats' actions or positions, reflecting a critical perspective on their perceived lack of thoughtful consideration and confidence.	0.1039
What about the democrats?	0.0089	The writer asks why I should vote in this coming election cycle.	0.0086	The author's question "But the Democrats?" implies a dismissive attitude toward the party, suggesting they view it as lacking in thoughtful consideration and thus unworthy of attention or support.	0.0935
The writer appears to ask whether the democrats are a better alternative to republicans in an upcoming election.	0.0117	The author wonders if the Democrats are incompetent.	0.3680	The author's question "But the Democrats?" indicates a dismissive view of the party's effectiveness, suggesting that the author believes their perceived incompetence undermines their value and relevance in the election.	0.1235

(continued on next page)

with a multitude of plausible interpretations. This is evidenced when rigorously analyzing the multiple ground-truth annotations for sentence interpretation and implicit moral judgments in the curated dataset. Secondly, it does not merely take the implicit content as the primary subject of prediction but instead lets the different layers of implicit meanings guide the automated generation of sentence interpretations.

Social grounding and moral reasoning Several studies explored the role of social commonsense and moral reasoning in natural language processing [66,72,33]. Kim et al. [39], for instance, distilled and contextualized dialogues by leveraging relations related to social commonsense from a symbolic knowledge graph. Vijayaraghavan and Roy [79] modeled, tracked, and explained changes in emotional states and motives of people in personal narratives. A few studies specifically focused on automated moral reasoning. Some adopted a question-answering setup where a model had to select a morally appropriate answer from a given list [31,92]. Hendrycks et al. [30] and Jin et al. [38] both assessed pre-trained language models in their knowledge of ethics, common human values, and norms of social conduct by having them classify, respectively, the character trait described in a given scenario using a given list and the permissibility of violating a pre-existing social norm in a given scenario. Others took a generative, rather normative approach and provided answers to moral dilemmas [8], judged the morality of a given scenario of social conduct [20,19], or classified explicit

Table 7 (continued)

[Sentence]				[Identity Attack]	
[Ground-Truth Interpretations]	[Identity Attack]	[Generated Interpretations]	[Identity Attack]	[Generated Interpretations]	[Identity Attack]
<i>That person usually ends up doing things they did not want to do and have a unfulfilling life.</i>					
Unfortunately, we are forced to do things we don't like and end up unhappy.	0.0042	[One2M-Sim-T5] Usually, that person ends up doing things they didn't want to do and have a unfulfilling life.	0.0034	[ChatGPT] The author suggests that a resigned person often ends up living an unfulfilling life by conforming to others' expectations rather than pursuing their own desires, highlighting a concern for the impact of resignation on personal fulfillment.	0.0097
That person conventionally ends up doing things they did not optate to do and have a unfulfilling life.	0.0035	The writer describes people who do things that they did not intend to and who have an unhappy life as being foolish.	0.0142	The author highlights that an unwise person, who lacks ambition and honor, often ends up leading an unfulfilling life by conforming to others' expectations rather than pursuing their own true desires, underscoring the importance of personal ambition and integrity.	0.0793
People often find themselves feeling unfulfilled in life when their actions don't consistently align with their core values , those ideals that matter the most to you.	0.0028	People who are stupid tend to be lazy and end up living mediocre lives.	0.1445	The author's view that those lacking ambition often end up unfulfilled suggests a negative bias, as it fails to consider the diverse reasons behind people's choices and dismisses their challenges and circumstances.	0.0352

moral judgments in comments on Reddit [9,17]. More recently, Pyatkin et al. [62] acknowledged that context influences the moral acceptability of social conduct and classified a given action differently by generating questions regarding the action's left-out context, e.g., when did the action take place and who performed the action.

This work is inspired by Talat et al. [74], who argued against a normative moral reasoning approach in which computational systems explicitly judge the morality of described situations and actions. Taking a descriptive approach, we refrain from evaluating moral behavior. We instead let computational models implicitly reason over hidden moral judgments as inferred by different readers to formulate various interpretations of a single sentence. The moral judgments serve as a means for grounding the sentence in society. All moral judgments inferred by the annotators in the dataset are kept as is; we do not average over them or center on a specific judgment. Moreover, the models are not requested to pass any judgment on the permissibility of the inferred moral judgments. Another main aspect in which this work distinguishes itself from previous work investigating moral reasoning is the data. While such studies generally rely on data that explicitly describe moral scenarios, the origamIM dataset features opinions on a variety of topics discussing much more than just the morality of behavior. This again underscores the novelty of the curated dataset.

Modeling reader reactions An utterance can spark different reactions. Chatbots, for instance, consistently need to decide how they react to input from its user, such that the user is satisfied with the chatbot's responses [81,84,41]. Others predicted the reactions readers may have to given content. Gao et al. [23], for instance, focused on emotional responses to multimodal news on gun violence. Deng et al. [13] presented a benchmark for evaluating language models' ability to predict readers' emotional reactions to image and caption pairs. In the field of computational humor, Yang et al. [86] modeled the humor perception of jokes among readers. Finally, anticipated reader reactions can also be informative for task prediction. For example, Gabriel et al. [22] detected online misinformation by reasoning over a reader's emotional and observational responses to headlines and claims.

It is crucial to acknowledge that language affects people differently and that their reactions do not necessarily align with those anticipated by the author. In the case of IM, a reader's interpretation is a kind of hidden reaction to a sentence, which in its turn can trigger subsequent reactions, for instance, a change in opinion or an action. Aside from the interpretations by the readers, the origamIM dataset also contains explicit indications of reader attitudes towards the author of the sentence. An attitude reflects the first impression a reader has about the author when reading the sentence and can steer the interpretation process. The IM frameworks therefore take those explicit reactions together with the inferred moral judgments to model sentence interpretations.

Multiple ground-truth readings IM is strongly motivated by the assumption that there exists genuine human variation in understanding language and that possibly disagreeing understandings should be respected [61]. Pavlick and Kwiatkowski [60] and Nie et al. [56] analyzed persistently disagreeing human judgments on the validity of natural language inferences and argued that models should be evaluated on their ability to predict the wide array of human judgments. Cabitza et al. [10] advocated a perspectivist approach towards ground-truth annotations and the preservation of diverging annotations. Chen et al. [11] extracted perspectives towards a claim on a controversial topic, and Draws et al. [15] and Draws et al. [16] evaluated the diversity of presented viewpoints.

While the latter perspectives commonly present reader positions towards a certain topic (i.e., what does the reader think), interpretations generated through IM reflect what multiple readers think the author of a sentence is actually communicating (i.e., what does the reader think the author is thinking). As a result, the moral judgments inferred by a reader do not necessarily align with the reader's own judgments or with the actual beliefs held by the author. It thus presents a deeper account of the communicative intentions driving a sentence, substantially differing from what some of the abovementioned works did.

Interpretation modeling versus diverse paraphrase generation At first sight IM arguably resembles paraphrase generation [26,91]. Yet, both modeling tasks differ on various aspects. First of all, IM aims at modeling the implicit content and explicit content *in tandem* rather than focusing on preserving the semantics of the explicit content, as done in paraphrasing. Moreover, the interpretation of implicit meanings may diverge from the explicit semantics of a sentence while the semantics of the input sentence should be maintained in a paraphrase. Even to the extent that the interpretation seems to contradict the original sentence. That contraction would be penalized in paraphrase detection [63,87]. Overall, we can state that IM aims to formulate sentence understandings from a reader's perspective while paraphrasing focuses on producing rewrites from the author's perspective. Regarding diverse generation, it is of crucial importance that IM models, like many other models tasked with text generation, are able to generate diverse sequences [75]. We can influence the decoding of such models through several diversity penalties [36], yet the diversity in interpretation also originates from the diversity in implicit semantics of the input sentence.

10. Conclusion

We started from the premises that natural language understanding (NLU) inevitably has to deal with content left implicit and that grounding language in its social context is a necessary condition to make implicit content explicit. We introduced the interpretation modeling (IM) task which aims at capturing the implicit and explicit meaning of a sentence as understood by different readers. IM is guided by multiple annotations of social relation and common ground - in this work approximated by reader attitudes towards the author and their understanding of moral judgments subtly embedded in the sentence. We proposed a number of strategies to decode a sentence into multiple interpretations in the form of natural language text. The one-to-one and one-to-many interpretation generation methods that we proposed are inspired by the philosophical study of interpretation. As a first of its kind, an IM dataset is curated to support experiments and analyses. The modeling results, coupled with elaborate analysis of the dataset, underline the challenges of IM as conflicting and complex interpretations are socially plausible. This interplay of diverse readings is affirmed by an intrinsic quantitative and qualitative analysis of the results as well as by an in-depth human evaluation of the generated sentence interpretations. Our work has value in revealing implicit meaning that could be inferred by readers and as such helps content filtering and moderation. This value was extrinsically demonstrated with toxicity classification of the original sentences and of original sentences and their possible interpretations, where we have shown that the availability of possible interpretations increased the recognition of harmful content. The toxicity analyses of the generated interpretations underline the importance of IM for refining filters of content and assisting content moderators in safeguarding the safety in online discourse.

Finally, if the generative models proposed in this paper can better produce and recognize harmful or dishonest interpretations of an original, at first sight harmless sentence, similar models might also better prevent harmful content when automatically generating text. For instance, the interpretation generation models can guide current language models, including ChatGPT, to satisfy one of their main goals of generating safe and diverse outputs. In a broader set-up, IM can find its place in NLP toolkits that aid corporate and public communication. For instance, IM can assist professional communicators in predicting and consequently preventing misunderstandings of their texts. IM could also advise users on their phrasing of texts and make them aware of views they perhaps unconsciously hold and/or convey, when it is included in general-purpose writing assistants.

There remains a wide array of open challenges that can be researched in the future. Although the proposed generation frameworks were able to generate diverse interpretation sets, human evaluation indicated that they may not fully grasp the ambiguous nature of a sentence yet. In future work, we may rely on larger language models and the supervision of a larger and more diverse sample of annotators may be necessary for surfacing the 'true' ambiguity of a sentence. We will investigate ways to integrate commonsense reasoning, more complex social relations between authors and readers, or other contextual information such as temporal context or demographic factors [50] that guide interpretation. In this work, we focused on modeling interpretations of isolated sentences and guided generation using the semantic relatedness between sentence and interpretation. Natural follow-ups would include expansions of the proposed methods to deal with the larger context surrounding the sentences (e.g., the full blog posts and comments to the post) and rely on alternative control mechanisms for interpretation generation. Further increasing the size of the OrigamIM dataset would also allow for investigations into joint predictions of social grounding and interpretations. From a social perspective, our work could lead to automated generation of interpretations while zooming in on specific societies. Comparative studies could investigate the plausibility of interpretations in different societies, while detailed studies constraining interpretations on various actors within a society may present insights in the diversity of views held in that specific society. Since the dataset constructed in this work only includes data from a specific Subreddit discussing an abundance of topics, follow-up datasets can look into other platforms and specific topics.

Limitations The annotated moral judgments and interpretations should by no means be considered a complete, socially diverse and fair account of views that is representative for an entire society. It is virtually impossible to gather all possible interpretations from all types of people and cultures, making datasets for IM unavoidably incomplete. Regarding annotator selection, we opted not to select annotators based on demographic features such as age, gender, education level, and political viewpoints since each sentence is only

annotated by a small number of annotators. This way, we wanted to avoid an overgeneralization of the views of one annotator and regard it as representative of people with similar demographic features. This study should therefore be seen as a stepping stone for better understanding the wide diversity of social evaluations and interpretations, for which a much larger sample of readers should be recruited. The artificial nature of the annotation setting also influences interpretation since crucial contextual knowledge is lost, e.g., author is unknown to the annotator and previous/following elaboration is omitted. Nonetheless, the setting mimics a real-life online environment where authors are often unknown and content is taken out of context. Regarding topic biases, interpretations of sentences touching on emerging topics may be subject to the time at which they are annotated. Validation on samples taken from other platforms would also be necessary to ensure the robustness of the proposed methods to distribution shifts in the data. Lastly, we constrained ourselves to leveraging implicit moral judgments during modeling, even though interpretation can be guided by a wide array of hidden meanings.

CRediT authorship contribution statement

Liesbeth Allein: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Maria Mihaela Trușcă:** Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Marie-Francine Moens:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Ethics statement

We follow the recommendations in Pater et al. [59] for reporting annotator selection, compensation, and communication. Regarding selection, workers were allowed to work on our annotation task immediately after passing an initial annotation instruction test, which was automatically corrected. They were paid a fixed amount per accepted HIT through the Amazon MTurk platform within three working days after completion and could earn between the U.S. legal minimum wage of \$7.5 and \$15/hour depending on their annotation flow and experience with the task. In case we rejected a HIT, we provided instructive motivations and gave additional feedback upon request. The majority of rejections originated from incorrect following of explicit instructions. We personally replied to all messages from the workers, most of them within one working day. We did not discriminate between the annotators in terms of gender, race, religion, or any other demographic feature.

The sentences were extracted from Reddit posts which had been knowingly published and discussed in a public subreddit. The sentences should be treated with caution as they were taken out of context and, as a result, may express views unintended by their authors. The interpretations and moral judgments annotated during dataset construction should in no way be seen as the actual judgments and views held and intentionally communicated by the authors of the sentences. Nor should they be considered judgments held by the annotators or the authors of this paper. Due to the limited number of annotators, the annotated collection of judgments and interpretations accompanying a sentence is not an exhaustive and fair reflection of those existing in society. It may therefore lack important nuances.

Revealing disputed meanings, negative moral judgments, and underlying toxicity can lead to potential harms. For example, it may reinforce existing divisions and social stigmatization of vulnerable groups. Nonetheless, access to such information is essential for NLU systems tasked with content moderation, hate speech detection, and related tasks. Moreover, disputed meanings often arise from complex issues which cannot be simply reduced to a single perspective. By revealing disputed and harmful content through interpretation modeling, we allow for a more socially comprehensive understanding of messages and prompt readers to think critically about the messages they receive.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to rephrase some sentences. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Liesbeth Allein reports financial support was provided by European Research Council, Research Foundation - Flanders (FWO) and European Commission Joint Research Centre. Liesbeth Allein reports article publishing charges was provided by European Research Council and Research Foundation - Flanders (FWO). Maria Mihaela Trusca reports financial support was provided by European Research Council. Marie-Francine Moens reports financial support was provided by European Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is funded by the European Research Council (ERC) under the Horizon 2020 Advanced Grant CALCULUS No. 788506 and the Research Foundation - Flanders (FWO) under grant G0L0822N through the CHIST-ERA iTRUST project. It was also realized

in part with the collaboration of the European Commission Joint Research Centre under the Collaborative Doctoral Partnership Agreement No 35332. The scientific output expressed does not imply a policy position of the European Commission. The authors, the European Commission, or any person acting on behalf of the Commission are not responsible for the use which might be made of this publication. We kindly thank Florian Mai and Jingyuan Sun for their insightful feedback on the manuscript.

Appendix A

An overview of the annotation procedure is given in § A.1, followed by a detailed description of the first (§ A.2) and second (§ A.3) annotation round.

A.1. Annotation guidelines: procedure

The annotation procedure is split in two rounds.

For a given **sentence** and the title of the post the sentence was taken from (= extra context):

Annotation round 1 (§ A.2)

1. Mark the people and groups of people mentioned in the sentence (highlight people entities in sentence);
2. Indicate whether or not the author implies a character trait for at least one of the marked entities (radio button, yes/no);

Only sentences for which the annotators answered ‘yes’ for step 2 are sent to annotation round 2.

Annotation round 2 (§ A.3)

1. Indicate your attitude as a reader towards the writer (slider, five-point Likert scale);
2. Reformulate the sentence such that it reflects what the writer ‘basically says’ (free text);

For each entity marked during step 1.1 of ROUND 1:

1. Indicate whether or not the author implies a character trait of the given entity (dropdown list, yes/no).

If yes, continue with (a), (b), and (c). If no, submit annotation.

- (a) Describe the implied character trait (free text);
- (b) Label whether society regards the trait as good or bad (dropdown menu, Good/Bad);
- (c) Label the Sphere of Action the character trait belongs to (dropdown menu, 10 class labels) and label the degree of appropriateness the character trait belongs to (dropdown menu, 3 class labels).

A.2. Annotation guidelines: round 1

In the first annotation round, annotators need to identify the people entities mentioned in the sentences by highlighting them in the sentence and then indicate whether or not the author of the sentence seems to imply a character trait of at least one people entity marked in the sentence. It is made clear in the interface of the first annotation round that the annotators need to fulfill two tasks, marked by (1/2) and (2/2). The annotators are unable to submit the task without answering the second question to avoid incomplete task forms. Candidate annotators had to pass a qualification test to have access to the annotation round.

A.2.1. Task 1: mark people entities in a sentence

The annotators mark the people and groups of people mentioned in the sentence (Fig. 8). If there are no entities to label, they can tick the ‘No entities to label’ box at the bottom right of the interface and submit the task form. The title of the post from which the sentence was taken is given below the sentence to contextualize its content. The general rules for marking people entities are given below the sentence and title such that the annotators were reminded of them when performing the task. We provide a more in-depth discussion of the rules with illustrative examples under the instructions tab at the top left corner of the interface (Figs. 9 and 10).

A.2.1.1. Instructions: what to mark

- **Mark words and phrases that refer to a person or a group of people** Only mark words that refer to people or a group of people such as an organization. Examples:
 - They welcome refugees from the Ukraine.
 - I saw this comment from some people in the US.
 - He talked about his friend Chad.
 - I've liked three different guys that were the older brother.
 - I have friends who are Trump supporters.

(1/2) Highlight words, phrases, or sections of the sentence that refer to people other than the author (i.e., 'I').

\$sentence

Belongs to a text with following title: \${title}

General Rules

- Only 1 reference per person/group, choose most descriptive reference (*Cheerleaders* like to dance and they love to cheer) or the first if references are identical (*They like to dance and they love to cheer*)
- Include all words specifying the reference (*My friend Chad those women friends who voted for Trump*)
 - Except adjectives indicating personality (very wise *tradesmen*, reasonable *women*)
- Ignore vague, void and too general references (someone, people, you as in *Imagine you are flying*)

No entities to label **Submit**

(2/2) Does the author seem to imply a character trait of at least one 'People'-labeled person/group?

Yes No

No entities to label **Submit**

Fig. 8. Annotation interface for annotation round 1. Annotators perform two tasks: Mark the entities in the given sentence (\${sentence}\$) using the label provided on the right (People, green label) and indicate whether the author of the sentence implies a character trait of at least one of the marked entities. The title of the post (\${title}\$) from which the sentence was taken is given as context surrounding the sentence.

- Past conservatives would not have worried about this.
- NATO should be taking more responsibility.
- **Mark only one reference per person/group in each sentence** If there is more than one instance that refers to the same person or group, mark the most descriptive reference or, in case they are identical, the first one.
- Cheerleaders like to dance and they love to cheer. (both ‘cheerleaders’ and ‘they’ refer to the same group - mark the most descriptive instance).
- They like to dance but they hate to sing. (both instances of ‘they’ refer to the same group - mark the first instance).
- **Mark the most descriptive reference to a person or group** In case the same entity is referred to more than once in the sentence, mark the reference that is most descriptive. Example:
 - Cheerleaders like to dance and they love to cheer. (Mark ‘cheerleaders’ as it is more descriptive than ‘they’)
- **Do not mark animals or objects** Example:
 - Our dog Ziggy is one year old. (‘Our dog Ziggy’ is not a person).
- **Do not mark the author (‘I’)** Example:
 - I saw this comment from some people in the US. (‘I’ refers to the narrator)
- **Do not mark void, vague, or too general references** Examples:
 - Void ‘you’/‘we’. Example: If you would compare the two objects. (‘you’ does not refer to a specific person; void); Yet, we are seeing mask mandates pop up around the country. (‘we’ does not refer to a specific group; void).
 - Vague. Example: I don't mean to say that certain people with particular sibling orders have less empathy, but I've noticed some key differences in romantic relationships. (‘certain people with particular sibling orders’ is too vague); For me to treat someone as fully human, ... (‘someone’ is too vague)
 - Too general. Example: People say that it will be better. (‘people’ does not refer to a specific group of people).
- **Do not mark predicates** Predicates are often preceded by verbs such as ‘to be’. Example:
 - Two of these guys are the types of dudes that hook up with girls. (‘the types of dudes that hook up with girls’ is predicate to ‘two of these guys’)
- **Do not mark references to a person/group that are part of phrase referring to other people/groups** Example:
 - I have friends who are Trump supporters (‘Trump’ is part of another reference, namely ‘friends who are Trump supporters’)
 - Joe Biden's top covid advisor was on CNN today. (‘Joe Biden’ is part of ‘Joe Biden's top covid advisor’)

Instructions

(1/2) Highlight words, phrases, or sections of the sentence that refer to people other than the author (i.e., 'I').

Examples:

- They welcome refugees from Syria.
- I saw this comment from some people in the US.
- He talked about his friend Chad.
- I've liked three guys that were the older brother.
- I know people who are Trump supporters.
- Past conservatives would not have worried about this.

Do not highlight

More Instructions

Labels

People

General Rules

- Only 1 reference per person/group, choose most descriptive reference (Cheerleaders like to dance and they love to cheer) or the first if references are identical (They like to dance and they love to cheer)
- Include all words specifying the reference (My friend Chad, those women, friends who voted for Trump)
 - Except adjectives indicating personality (very wise tradesmen, reasonable women)

No entities to label **Submit**

Instructions

(1/2) Highlight words, phrases, or sections of the sentence that refer to people other than the author (i.e., 'I').

Examples:

- older brother.
- I know people who are Trump supporters.
- Past conservatives would not have worried about this.

Do not highlight

More Instructions

Labels

People

General Rules

- Only 1 reference per person/group, choose most descriptive reference (Cheerleaders like to dance and they love to cheer) or the first if references are identical (They like to dance and they love to cheer)
- Include all words specifying the reference (My friend Chad, those women, friends who voted for Trump)
 - Except adjectives indicating personality (very wise tradesmen, reasonable women)

No entities to label **Submit**

Fig. 9. The instructions for marking the people entities in the given sentence can be opened when clicking on the instructions tab at the top left. They give a few short example sentences and list which references should not be highlighted. Additional information about which references to highlight and how much of the reference needs to be highlighted can be accessed by clicking on the clickable blue sentences in the instructions.

A.2.1.2. Instructions: how much to mark

• Mark all the words that specify the person referred to Example:

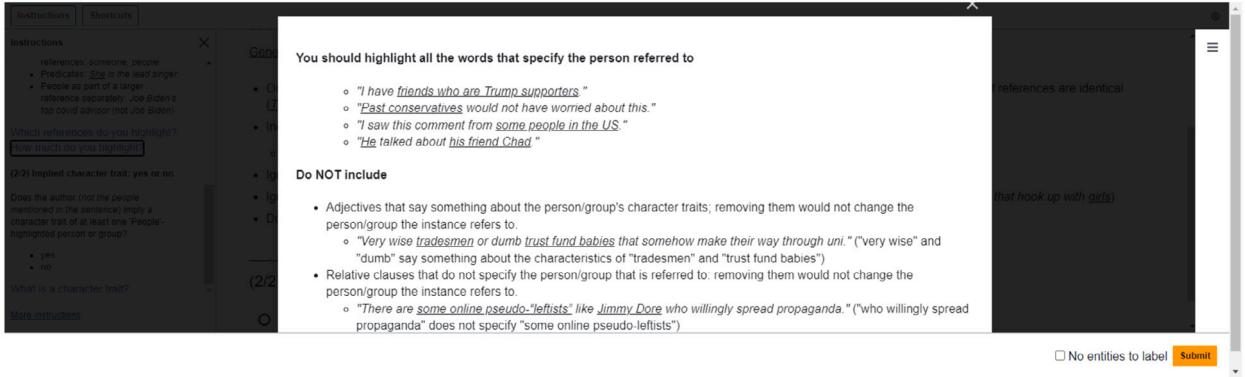
- They welcome refugees from the Ukraine. (marking only ‘refugees’ would be incomplete as it then refers to a larger, more general group).
- I saw this comment from some people in the US. (including ‘some’ and ‘in the US’ is more specific than leaving one of the two, or both, out).
- He talked about his friend Chad. (including ‘his friend’ provides more information on the relation between ‘he’ and ‘Chad’).
- I've liked three different guys that were the older brother. (including the relative clause ‘that were the older brother’ specifies the identity of the guys).
- I have friends who are Trump supporters. (including the relative clause ‘who are Trump supporters’ specifies the identity of the writer’s friends).
- Past conservatives would not have worried about this. (including adjective ‘past’ is necessary to refer to the right group of conservatives).

• Except

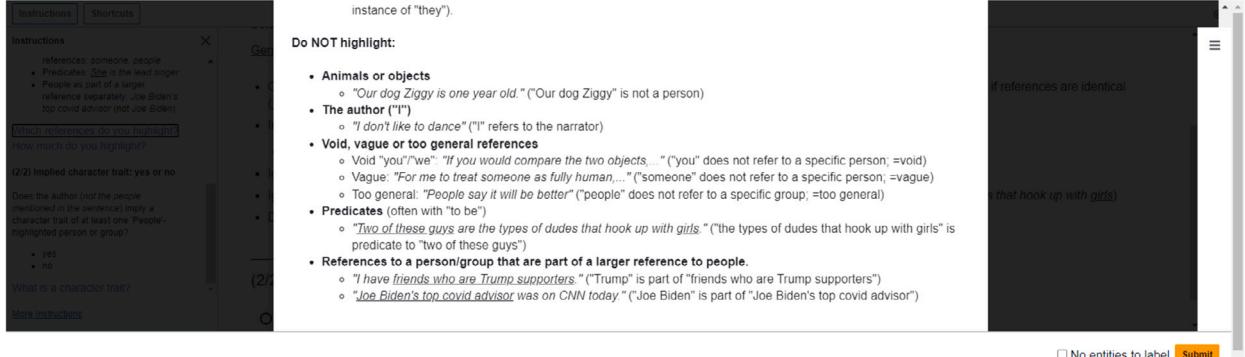
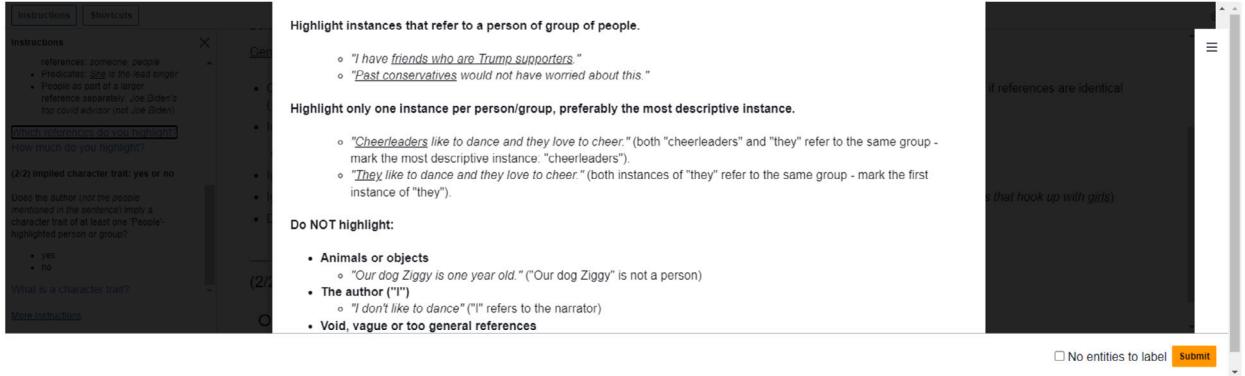
- Adjectives preceding the person/group noun that denote their personality/character traits; they can be removed without changing the person or group they refer to.
 - * Very wise tradesmen or dumb trust fund babies that somehow make their way through uni. (‘very wise’ and ‘dumb’ describe character traits).
 - * Although I'm sure that there are people who never go to university that are far smarter than I am, on the whole it would be far more effective for only reasonable intelligent and conscientious people (people ⇒ people who went to university) to decide on the future of our society. (‘reasonable intelligent and conscientious’ describe character traits).
- Relative clauses that do not specify the kind of person/group is referred to; leaving them out would not change the reference.
 - * There are some online pseudo-“leftists” like Jimmy Dore who willingly spread propaganda. (‘who willingly spread propaganda’ does not specify the reference.)

A.2.2. Task 2: indicate presence or absence of implied character trait

After labeling the entities in the sentence, the annotators now indicate whether or not the author of the sentence seems to imply a character trait of at least one entity they have marked in the sentence.



(a) Pop-up screen when clicking on the instructions link “How much to highlight?”



(b) Pop-up screens when clicking on the instructions link “Which references do you highlight?”

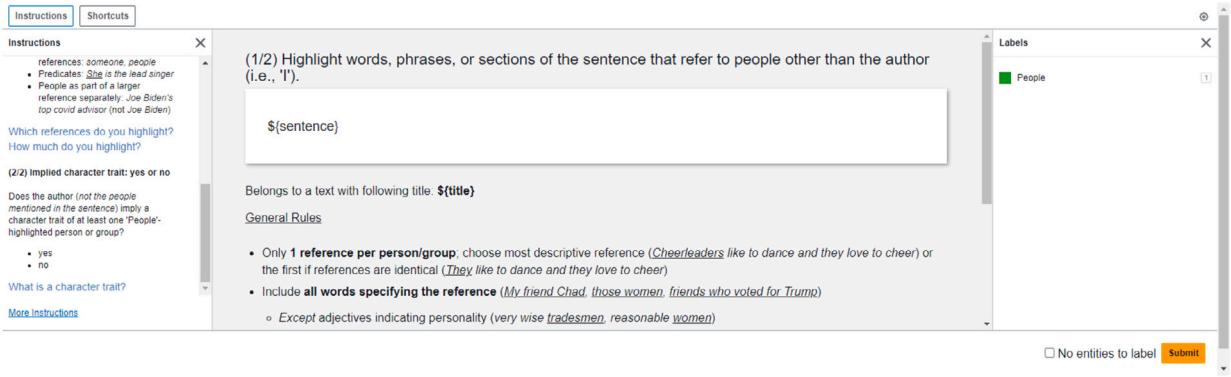
Fig. 10. The pop-up instruction when clicking on the clickable blue links provided in the instructions.

A.2.2.1. Instructions: definition of character trait We provide the annotators with a definition of character trait in the instructions, which can be accessed by clicking on the instructions tab on the top left corner (Fig. 11).

A character trait describes an aspect of a person's **behavior** or **attitude**. The set of character traits then forms someone's personality. A trait is also **voluntary** so it is not forced on that person by someone or something. A person has control over his/her character traits. A person's character traits cause him/her to act a certain way. Character traits can very often be described using descriptive adjectives such as *honest*, *sincere*, *brave*, *loyal*, *vain*, *timid*, and *calm*. Examples:

- Eric *disguises himself*. (Eric is sneaky).
- Eric *likes to talk about himself*. (Eric is vain).
- Eric *does not want to spend much money*. (Eric is stingy).

A.2.2.2. Instructions: the author implies trait of at least one entity The annotators should ask themselves whether the author seems to suggest something about the character of the people in the sentence. It is important to distinguish between character traits implied



(a) Annotator instructions regarding Task 2.



(b) Pop-up screen when clicking on the instructions link “What is a character trait?”.

Fig. 11. Instructions regarding Task 2 (a) and definition of character trait (b).

by the author, showing that the author seems to hold certain beliefs about people, and character traits that a reader or other people mentioned by the writer believe people have (a reader or mentioned person holds certain beliefs about people). Like any other person, an annotator can have certain, pre-existing beliefs about (groups of) people or can be aware of pre-existing beliefs that others have about them. We therefore stress in the phrasing of the second task that the annotators should think about what the author seems to imply by underlining the author: *“Does the author seem to imply a character trait of at least one ‘People’-labeled person/group?”*. We also underline at least one to stress that a character trait for all entities need to be implied for indicating ‘yes’.

A.2.3. Qualification test: round 1

Candidate annotators first have to pass a qualification test before they can start the annotation round. They were automatically admitted to the annotation round if they answered all questions in the qualification test correctly. The qualification test instructs the annotators on how to adequately mark people entities in a sentence by going over five main rules. For each rule, the annotator is presented with an example sentence. They then need to pick the correct labeling answer from the options below the sentence (radio button). We indicate the correct answer with (X).

Test description: *We go over FIVE labeling rules for this task. For each rule, you will be presented with an example sentence and you need to pick the correct labeling answer from the options below. If you pick all the correct answers, you can automatically start the labeling task.*

- Rule 1: Highlight all words, phrases, or sections of the sentence that refer to people other than the author (i.e., 'I').
 - Sentence: “I talked with my friend Chad about some men from Texas and people who like Star Wars.”
 - * I - Chad - men - people
 - * (X) my friend Chad - some men from Texas - people who like Star Wars
- Rule 2: Highlight only one reference per person/group + choose most descriptive reference.
 - Sentence: “Cheerleaders like to dance and they love to cheer.”
 - * (X) Cheerleaders
 - * Cheerleaders - they
 - * they

- Rule 3: Highlight all words specifying the reference; except adjectives indicating personality (e.g., reasonable, vain).
 - Sentence: “The blond girl from Canada told that there are honest girls and generous boys in her class.”
 - * girl - girls - boys
 - * The blond girl from Canada - honest girls - generous boys
 - * (X) The blond girl from Canada - girls - boys
- Rule 4: Do not highlight vague, void and too general references.
 - Sentence: “If you meet someone online, people will tell you to be cautious”
 - * you - someone - people
 - * (X) [None] (you = void, someone = vague, people = too general)
- Rule 5: Do not highlight predicates (often after a modal verb like ‘to be’: ‘Person is [predicate]’) and people that are part of another, larger reference
 - Sentence: “Barack Obama became the US president when he was 48 and Barack Obama’s wife was the First Lady”
 - * (X) Barack Obama - Barack Obama’s wife
 - * Barack Obama - the US president - wife - the First Lady

A.3. Annotation guidelines: round 2

The annotators start by reading the sentence (`${sentence}`) and the post from which the sentence is taken (`${title}`). They are then asked to describe their attitude towards the author and formulate their interpretation of the given sentence. Going over all people identified in the sentence, they indicate for each entity whether or not the author of the sentence implies a character trait of that entity. For all entities with an implied character trait, the annotators also describe the trait, label the evaluation of that trait in society, and classify it in Virtue Ethics by labeling the Sphere of Action the trait belongs to and its contextual appropriateness. It is made clear in the interface (Fig. 12) that the annotators need to fulfill four tasks, marked by (1/4) to (4/4). They can access the instructions with illustrative examples at any time during annotation by clicking on the instructions tab at the top left corner of the interface (Fig. 13). The instructions then appear on the left and can stay visible during the annotation process. They disappear when clicking on the instructions tab again. The annotators are unable to submit the task form without providing an interpretation, which is the most important part of this annotation round. Candidate annotators had to pass a qualification test to have access to the annotation round.

A.3.1. Task 1: describe attitude towards author of sentence

The annotators describe their attitude towards the author of the sentence upon reading the sentence. We stress in the task description that they should describe their own attitude by underlining *your attitude* in the task description and state in brackets that their attitude is their impression of the author based on the sentence. They change the slider given below the task description and set it to the best-fitting attitude on a five-point Likert scale ranging from *very negative* to *very positive*. When moving the slider, the value to which they set it changes to blue and depicts the number referring to the attitude they set it to (Fig. 14).

A.3.1.1. Instructions: definitions of attitude scale We provide the annotators with a definition of each attitude type in the instructions. They can access the overview of attitude definitions by clicking on the clickable blue sentence “*What do the attitude mean?*” in the instructions. The overview then appears as a pop-up screen (Fig. 15).

- **Very negative** After reading the sentence, the annotator has a very negative view of the author. They completely disagree with what the author said. They would never say or think the same, or never phrase it like the author did.
- **Negative** After reading the sentence, the annotator has a negative view of the author. They disagree with what the author said. They would not say or think the same, or phrase it like the author did.
- **Neutral** After reading the sentence, the annotator does not have a negative or positive view of the author. They do not agree or disagree with what the author said.
- **Positive** After reading the sentence, the annotator has a positive view of the author. They agree with what the author said. They would say or think the same, or phrase it like the author did.
- **Very positive** After reading the sentence, the annotator has a very positive view of the author. They completely agree with what the author said. They would definitely say or think the same, or phrase it like the author did.

A.3.2. Task 2: formulate interpretation of sentence

The annotators write their interpretation of the given sentence in their own words. We stress that they should formulate their own interpretation by underlining *your interpretation* in the task description. They are also explicitly asked to not simply copy the sentence. The instructions additionally state that *your interpretation* may reveal implicit information contained within the sentence.

A.3.2.1. Instructions: examples of interpretations We provide the annotators with several examples of possible interpretations for an example sentence, which can be accessed by clicking on the clickable blue phrase “*Examples interpretations*” in the instructions (Fig. 16). The examples appear in a pop-up screen. We phrase the interpretations in such a way so that the underlying messages and hidden moral judgments surface.

Given sentence: “*They need to force the unemployed to look for work.*”

[Instructions](#) [Shortcuts](#)

(1/4) Read the sentence and the title of the blog post it was taken from (extra context). Then answer all questions below.

Sentence: \${sentence}

Blog post title: \${title}

(2/4) Describe your attitude towards the author (= first impression based on the sentence).

(1=very negative, 2=negative, 3=neutral, 4=positive, 5=very positive)

—

(3/4) Write your interpretation of the sentence. (Do not simply copy the sentence)

Sentence: \${sentence}

[Submit](#)

[Instructions](#) [Shortcuts](#)

(3/4) Write your interpretation of the sentence. (Do not simply copy the sentence)

Sentence: \${sentence}

write your interpretation here ...

(4/4) For each person/group in the sentence, indicate whether or not the author implies a character trait. If yes, describe the trait and choose its best-fitting classification label from the table.

Situation (Sphere of Action)	Vice of Deficiency (="too little")	Virtue of Mean (="appropriate")	Vice of Excess (="too much")
1 confidence (fear, uncertainty)	cowardice	courage	rashness
2 pleasures of the body	insensibility	temperance	profligacy
3 giving (and taking) - small amounts of money	stinginess	liberality, generosity	prodigality
4 giving (and taking) - other, added value	meanness	magnificence	vulgarity
5 pride (honour = cause)	little-mindedness	high-mindedness	vanity
6 ambition (honour = goal)	lack of ambition	proper ambition	over-ambitiousness

[Submit](#)

[Instructions](#) [Shortcuts](#)

(4/4) For each person/group in the sentence, indicate whether or not the author implies a character trait. If yes, describe the trait and choose its best-fitting classification label from the table.

Situation (Sphere of Action)	Vice of Deficiency (="too little")	Virtue of Mean (="appropriate")	Vice of Excess (="too much")
1 confidence (fear, uncertainty)	cowardice	courage	rashness
2 pleasures of the body	insensibility	temperance	profligacy
3 giving (and taking) - small amounts of money	stinginess	liberality, generosity	prodigality
4 giving (and taking) - other, added value	meanness	magnificence	vulgarity
5 pride (honour = cause)	little-mindedness	high-mindedness	vanity
6 ambition (honour = goal)	lack of ambition	proper ambition	over-ambitiousness
7 anger	easy-tempered, spiritlessness	gentleness	hard-tempered, wrathfulness
8 pleasure (and pain) of others	cross, contentious	agreeableness, friendliness	flattery
9 truth, honesty about oneself	irony	truthfulness	boastfulness
10 amusing conversation	boorishness	witiness	buffoonery

Answer the questions for the following people: \${people}

[Submit](#)

[Instructions](#) [Shortcuts](#)

Answer the questions for the following people: \${people}

Sentence: \${sentence}

\$(p1)

Does the author imply a character trait? (please select YES/NO) IF YES, describe the trait and select the best-fitting answers from the dropdown menus below.

Describe the trait using an adjective (e.g. honest, vain) here... Try using a term that is not in the table above.

Society thinks this trait is (please select EVALUATION) The trait relates to how people react to or deal with (please select TABLE ROW) and in what way: (please select TABLE COLUMN)

[Submit](#)

Fig. 12. The main annotation interface of ANNOTATION ROUND 2.

- Possible interpretation (1): “As the unemployed are too lazy to look for a job, we need to force them to do this.”
- Possible interpretation (2): “They, who are superior, need to force the unemployed to look for work.”
- Possible interpretation (3): “They need to force the unemployed to look for work, because they are superior to them and the unemployed are too lazy to do it themselves.”

Instructions **Shortcuts**

What is a character trait?

(2/4) Describe your attitude towards the author after reading the sentence

How is your first impression of the author?

- 1 = very negative
- 2 = negative
- 3 = neutral
- 4 = positive
- 5 = very positive

What do the attitudes mean?

(3/4) Interpretation of the sentence

This is how you interpret the sentence. Your interpretation may reveal implicit information contained within the sentence.

[More Instructions](#)

(1/4) Read the sentence and the title of the blog post it was taken from (extra context). Then answer all questions below.

Sentence: \${sentence}

Blog post title: \${title}

(2/4) Describe your attitude towards the author (= first impression based on the sentence).

(1=very negative, 2=negative, 3=neutral, 4=positive, 5=very positive)

(3/4) Write your interpretation of the sentence. (Do not simply copy the sentence)

Sentence: \${sentence}

Submit

Instructions **Shortcuts**

(3/4) Interpretation of the sentence

This is how you interpret the sentence. Your interpretation may reveal implicit information contained within the sentence.

Examples interpretation

(4/4) Reasoning about implied character traits and moral norms

Example: Given sentence: "You can trust Eric always tells the truth." → yes + "Eric is honest" + Good + B-truth, honesty about oneself + Virtue-mean

1-confidence (fear, uncertainty)
→ Dealing with fear-inciting, uncertain

[More Instructions](#)

(1/4) Read the sentence and the title of the blog post it was taken from (extra context). Then answer all questions below.

Sentence: \${sentence}

Blog post title: \${title}

(2/4) Describe your attitude towards the author (= first impression based on the sentence).

(1=very negative, 2=negative, 3=neutral, 4=positive, 5=very positive)

(3/4) Write your interpretation of the sentence. (Do not simply copy the sentence)

Sentence: \${sentence}

Submit

Instructions **Shortcuts**

References: someone, people
• Predicates: \$the lead singer
• People as part of a larger reference separately: Joe Biden's top covid advisor (not Joe Biden)

Which references do you highlight?
How much do you highlight?

(2/2) Implied character trait: yes or no

Does the author (not the people mentioned in the sentence) imply a character trait of at least one "People-highlighted person or group?"

- yes
- no

What is a character trait?

[More Instructions](#)

(1/2) Highlight words, phrases, or sections of the sentence that refer to people other than the author (i.e., "I").

Labels

People

1

\$sentence

Belongs to a text with following title: \${title}

General Rules

- Only 1 reference per person/group, choose most descriptive reference (*Cheerleaders* like to dance and they love to cheer) or the first if references are identical (*They* like to dance and they love to cheer)
- Include all words specifying the reference (*My friend Chad, those women, friends who voted for Trump*)
 - Except adjectives indicating personality (very wise *tradesmen*, reasonable *women*)

No entities to label **Submit**

Instructions **Shortcuts**

7-anger

→ Dealing with anger and negative emotions

8-pleasure (and pain) of others

→ Dealing with pain and pleasure of others, agreement and disagreement

9-truth, honesty about oneself

→ Dealing with truthfulness in word and deed

10-amusing conversation

→ Dealing with amusing conversation and mixing with others

[More Instructions](#)

(1/4) Read the sentence and the title of the blog post it was taken from (extra context). Then answer all questions below.

Sentence: \${sentence}

Blog post title: \${title}

(2/4) Describe your attitude towards the author (= first impression based on the sentence).

(1=very negative, 2=negative, 3=neutral, 4=positive, 5=very positive)

(3/4) Write your interpretation of the sentence. (Do not simply copy the sentence)

Sentence: \${sentence}

Submit

Fig. 13. Overview of the instructions that appear when clicking on the instructions tab at the top left of the annotation interface.

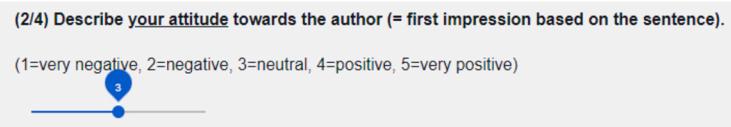


Fig. 14. Attitude slider shows attitude values on a five point Likert scale.

To answer this question, you need to ask yourself what you think of the author; Do you have a positive/neutral/negative view of the author? And in which degree?

1 = very negative

After reading the sentence, you have a *very negative* view of the author. You *completely disagree* with what (s)he said. You would never say or think the same, or never phrase it like (s)he did.

2 = negative

After reading the sentence, you have a *negative* view of the author. You *disagree* with what (s)he said. You would not say or think the same, or phrase it like (s)he did.

3 = neutral

After reading the sentence, you do *not have a negative or positive* view of the author. You do *not agree or disagree* with what (s)he said.

4 = positive

After reading the sentence, you have a *positive* view of the author. You *agree* with what (s)he said. You would say or think the

2 = negative

After reading the sentence, you have a *negative* view of the author. You *disagree* with what (s)he said. You would not say or think the same, or phrase it like (s)he did.

3 = neutral

After reading the sentence, you do *not have a negative or positive* view of the author. You do *not agree or disagree* with what (s)he said.

4 = positive

After reading the sentence, you have a *positive* view of the author. You *agree* with what (s)he said. You would say or think the same, or phrase it like (s)he did.

5 = very positive

After reading the sentence, you have a *very positive* view of the author. You *completely agree* with what (s)he said. You would definitely say or think the same, or phrase it like (s)he did.

Fig. 15. Overview of the different attitude types and their description given in the instructions.

(1/4) Read the sentence and the title of the blog post it was taken from (extra context). Then answer all questions below.

Sentence: *Original: "They need to force the unemployed to look for work."*

Possible reformulation 1: *"As the unemployed are too lazy to look for a job, we need to force them to do this."*

Possible reformulation 2: *"They, who are superior, need to force the unemployed to look for work."*

Possible reformulation 3: *"They need to force the unemployed to look for work, because they are superior to them and the unemployed are too lazy to do it themselves."*

(3/4) Write your interpretation of the sentence. (Do not simply copy the sentence)

Sentence: *(4 sentences)*

Fig. 16. Examples of possible interpretations of a given sentence given in the instructions.

A.3.3. Task 3: categorize implied character trait for each entity in sentence

The annotators need to answer the following questions for each entity that was previously marked in the first annotation round:

- Does the author seem to imply a character trait of this entity? (Task 3.1)
- If Task 3.1 = yes:
 - What is the character trait? (Task 3.2)
 - How does society evaluate this character trait? (Task 3.3)
 - Where is the character positioned in Virtue Ethics in terms of Sphere of Action and contextual appropriateness? (Task 3.4)

(4/4) For each person/group in the sentence, indicate whether or not the author implies a character trait. If yes, describe the trait and choose its best-fitting classification label from the table.

Situation (Sphere of Action)	Vice of Deficiency ("too little")	Virtue of Mean ("appropriate")	Vice of Excess ("too much")
1 confidence (fear, uncertainty)	cowardice	courage	rashness
2 pleasures of the body	insensibility	temperance	profligacy
3 giving (and taking) - small amounts of money	stinginess	liberality, generosity	prodigality
4 giving (and taking) - other, added value	meanness	magnificence	vulgarity
5 pride (honour = cause)	little-mindedness	high-mindedness	vanity
6 ambition (honour = goal)	lack of ambition	proper ambition	over-ambitiousness
7 anger	easy-tempered, spiritlessness	gentleness	hard-tempered, wrathfulness
8 pleasure (and pain) of others	cross, contentious	agreeableness, friendliness	flattery
9 truth, honesty about oneself	irony	truthfulness	boastfulness
10 amusing conversation	boorishness	wittiness	buffoonery

Answer the questions for the following people: \${people}

Submit

(a) Virtue Ethics classification table.

Instructions **Shortcuts**

Answer the questions for the following people: \${people}

Sentence: \${sentence}

\${p1}

Does the author imply a character trait? **(please select YES/NO)** **IF YES, describe the trait and select the best-fitting answers from the dropdown menus below.**

Describe the trait using an adjective (e.g. honest, vain) here... Try using a term that is not in the table above.

Society thinks this trait is **(please select EVALUATION)** The trait relates to how people react to or deal with **(please select TABLE ROW)** and in what way **(please select TABLE COLUMN)**

Submit

(b) Entity annotation block for the first people entity (\${p1}) in the people entities list (\${people}).

Fig. 17. Interface for tackling Task 3.

Instructions **Shortcuts**

Answer the questions for the following people: \${people}

Sentence: \${sentence}

\${p1}

Does the author imply a character trait? **(please select YES/NO)** **IF YES, describe the trait and select the best-fitting answers from the dropdown menus below.**

(please select YES/NO)

Yes
No

Describe the trait using an adjective (e.g. honest, vain) here... Try using a term that is not in the table above.

Society thinks this trait is **(please select EVALUATION)** The trait relates to how people react to or deal with **(please select TABLE ROW)** and in what way **(please select TABLE COLUMN)**

Submit

Fig. 18. Drop-down list for solving Task 3.1.

The interface presents the classification table for classifying the character traits in Virtue Ethics right below the task description (Fig. 17a). The annotators are then given the list of people entities for which they need to provide annotations (\${people}). The interface then contains entity annotation blocks, with the number of blocks equal to the number of entities in the people entities list (Fig. 17b). Each block first presents the sentence such that the annotator does not need to scroll up to read the sentence again. The block then contains Task 3.1, 3.2, 3.3, and 3.4. The tasks are framed as sentences, and the annotators need to select the labels from drop-down lists.

A.3.3.1. Task 3.1: indicate presence implied character trait of entity The annotators need to indicate whether or not the author implies a character trait of the given people entity. They do this by choosing between **yes** and **no** from a drop-down list (Fig. 18). If they indicate **yes**, they need to continue with the annotation for this entity. This is also stated in bold after the drop-down list. Otherwise, they can continue to annotate the next entity or, in case this entity is the last one, submit the task form.

Fig. 19. Input for Task 3.2. is provided as free text.

Fig. 20. Drop-down list for solving Task 3.3.

Instructions: definition of character trait The annotators are given the same definition of character trait given in the first annotation round, see § A.2.2.1. The definition can be accessed by clicking on the clickable blue sentence “*What is a character trait?*” in the instructions, which then appears in a pop-up screen.

A.3.3.2. Task 3.2: identify implied character trait of entity The annotators describe the implied character trait in their own words in the free-text text box (Fig. 19).

Instructions: how to describe trait The annotators are encouraged to describe the trait using an adjective (“*Describe the trait using an adjective (e.g., honest, vain) here.*”) and not to use an adjective that is presented in the Virtue Ethics classification table given above (“*Try using a term that is not in the table able.*”), see Fig. 19. When the annotators start describing the trait, the two instructions appear above the free-text box.

A.3.3.3. Task 3.3: evaluate implied character trait of entity in society The annotators decide whether society thinks this is a *good* (*positive*) or *bad* (*negative*) trait (Fig. 20). Or in other words, does society praise people for having this trait or frown upon or even reject this trait, respectively.

Instructions: focus on what society thinks Since we are interested in how the character trait is evaluated in society, we mark *society* in bold in the task description (“**Society** thinks this trait is”). This way, we encourage annotators to think beyond their own evaluations of character trait and position the trait in the society they are part of.

A.3.3.4. Task 3.4: classify implied character trait of entity in virtue ethics The annotators classify the implied character trait in Virtue Ethics by selecting the best-fitting labels from the drop-down lists (Fig. 21). A table containing all the Virtue Ethics labels is given below the task description so that the annotators can easily consult the possible labels. They first classify the trait in a Sphere of Action (table rows) and then by its contextual appropriateness (table columns). When clicking on the instructions tab at the top left corner, the annotators see an overview of all the Spheres of Action with a brief description for each Sphere of Action (Fig. 22).

Before starting the annotation task, candidate annotators had to pass an instruction test in which the Virtue Ethics and its labels were explained and participants had to select the appropriate contextual appropriateness label for ten sentences, with each sentence related to a different Sphere of Action. The annotators were allowed to start the annotation task if they correctly labeled nine sentences.

Instructions: virtue ethics We adopt the well-established Aristotelian Ethics [7], more specifically Nicomachean Ethics, to label people’s moral virtues and vices. Aristotle argued that a virtue/vice is a character trait or a state of someone’s character/personality. A Virtue of trait causes a person to act in a manner that is appropriate and desired by society in a given context. It is not because

The screenshot shows a user interface for a task. At the top, there are two buttons: 'Instructions' and 'Shortcuts'. Below them, a question asks: 'Answer the questions for the following people: \${people}'. A sentence field follows: 'Sentence: \${sentence}'. A section labeled '\${p1}' contains a question: 'Does the author imply a character trait? (please select YES/NO)'. If selected, it prompts: 'IF YES, describe the trait and select the best-fitting answers from the dropdown menus below.' Below this, a statement says: 'Society thinks this trait is (please select EVALUATION)'. A dropdown menu is open, listing options: '(please select TABLE ROW)', '1-confidence (fear, uncertainty)', '2-pleasures of the body', '3-giving (and taking) - small amounts of money', '4-giving (and taking) - other, added value', '5-pride (honour = cause)', '6-ambition (honour = goal)', '7-anger', '8-pleasure (and pain) of others', '9-truth, honesty about oneself', and '10-amusing conversation'. To the right of the dropdown are 'Submit' and 'Finish' buttons.

(a) Drop-down list for selecting Sphere of Action labels.

This screenshot shows the same interface as above, but the dropdown menu has changed. It now lists options: '(please select TABLE COLUMN)', 'Vice - deficiency', 'Virtue - mean', and 'Vice - excess'. The rest of the interface remains the same, including the question about character traits and the dropdown for selecting labels from a table row.

(b) Drop-down list for selecting contextual appropriateness labels.

Fig. 21. Drop-down lists for solving Task 3.4.

someone acts in a Virtue of manner that person is actually virtuous. It is important that the action is voluntary and within the control of the individual. By contrast, a vicious trait causes a person to act in a manner that is frowned upon, disapproved or even forbidden by society.

Aristotle identified several **Spheres of Action (SoA)** that cause people to act in various ways. Those different kinds of actions are put on an axis according to their degree or **contextual appropriateness**: deficiency, mean and excess. Aristotle stated that a Virtue of person acts in a moderate and contextually appropriate manner (= mean), while a morally faulty person resides to morally deficient (= deficiency) or exceeding actions (= excess). We therefore speak of **Vice of Deficiency**, **Virtue of Mean**, and **Vice of Excess**. Table 8 presents an overview of the Spheres of Action and the contextual appropriateness labels. We take the descriptions of the SoAs from Aristotle [7].

Confidence (fear, uncertainty) How someone deals with fearful, uncertain events/situations and to which extent someone is confident about their actions.

- **Cowardice [Vice of Deficiency]** *Lack of confidence, excess in fear*

Someone who has *too much fear* when dealing with uncertain situations when dealing with uncertain situations. This person also *fears the wrong things* and *deals with things in the wrong way*. It could be said that this person is *deficient in confidence* when being faced with situations that excite fear and uncertainty.

- **Courage [Virtue of Mean]** *Appropriate fear and confidence*

Someone who fears the right situations and the level of fear depends on the fear-exciting nature of each of those situations (fearful things differ in importance and thus a courageous person fears it accordingly). This person deals with fear-exciting situations with *appropriate courage/confidence*, in the appropriate way, and at the appropriate time. This person's courage originates from a sense of nobility and is *not excited under compulsion* because being courageous is a voluntary trait.

- **Rashness [Vice of Excess]** *Lack of fear, excess in confidence*

Someone has *too little fear* and *too much confidence* when dealing with uncertain situations. This person pretends to have a courage (s)he has not. The high level of confidence is often caused by a sense of superiority and invincibility.

The screenshot shows a digital interface for defining a sphere of action. On the left, there are tabs for 'Instructions' and 'Shortcuts'. Below these are sections for 'Instructions', 'Events and the extent to which someone is confident about their actions', 'Pleasures of the body', 'Dealing with pleasures (or the body) and the ability to control (body) urges', 'Giving (and taking) - small amounts of money', 'Dealing with giving money to and taking money from others', 'Giving (and taking) - other, added value', 'Dealing with giving and taking to create added value', and 'More instructions'. On the right, a large text area contains the definition of 'Giving (and taking) - small amounts of money'. It includes a general statement, a section on 'Vice of Deficiency' (stinginess), a section on 'Lack of giving, excess in taking' (greed, meanness, lending money at high interest), a section on 'Virtue of Mean' (liberality/generosity), a section on 'Spending resources to create added value for others' (proper spending), a section on 'Vice of Excess' (prodigality, wastefulness), and a section on 'Too much delight in bodily pleasures' (squandering money). A 'Submit' button is located in the bottom right corner.

(a) Definition of *Giving (and taking) - small amounts of money*.

The screenshot shows a digital interface for defining a sphere of action. On the left, there are tabs for 'Instructions' and 'Shortcuts'. Below these are sections for 'Instructions', 'Interpretation of the sentence', 'How you interpret the sentence. Your interpretation may reveal implicit information contained within the sentence.', 'Examples interpretation', 'Reasoning about implied character traits and moral norms', 'Given sentence: "You can trust Eric, he always tells the truth" ... yes + Eric is honest = Good + S-truth (honesty about oneself) = Virtue-truth', 'Confidence/fear, uncertainty', 'Dealing with fear-uncertainty, uncertain', and 'More instructions'. On the right, a large text area contains the definition of 'Confidence (fear, uncertainty)'. It includes a general statement, a section on 'Vice of Deficiency' (cowardice), a section on 'Lack of confidence, excess in fear' (fear of uncertainty), a section on 'Virtue of Mean' (courage), a section on 'Appropriate fear and confidence' (fear of appropriate situations), a section on 'Vice of Excess' (rashness), and a section on 'Lack of fear, excess in confidence' (confidence in uncertainty). A 'Submit' button is located in the bottom right corner.

(b) Definition of *Confidence (fear, uncertainty)*.

Fig. 22. Definitions of two Spheres of Action that appear when clicking on the clickable blue phrases 1-confidence (fear, uncertainty) (a) and 3-giving (and taking - small amounts of money) (b).

Table 8
Overview of Spheres of Actions (SoA) with their the virtue and vices.

SoA or feeling	Vice of deficiency	Virtue of mean	Vice of excess
Confidence (fear, uncertainty)	cowardice	courage	rashness
Pleasures of the body	insensibility	temperance	profligacy
Giving (and taking) - small amounts of money	stinginess	liberality, generosity	prodigality
Giving (and taking) - other, added value	meanness	magnificence	vulgarity
Pride (honour = cause)	little-mindedness	high-mindedness	vanity
Ambition (honour = goal)	lack of ambition	proper ambition	over-ambitiousness
Anger	easy-tempered, spiritlessness	gentleness	hard-tempered, wrathfulness
Pleasure (and pain) of others	cross, contentious	agreeableness, friendliness	flattery
Truth, honesty about oneself	irony	truthfulness	boastfulness
Amusing conversation	boorishness	wittiness	buffoonery

Pleasures of the body How someone deals with and is able to control and fulfil the needs and pleasures of the body.

- **Insensibility [Vice of Deficiency]** *Too little delight in bodily pleasures*

Someone who takes less delight in such pleasures than ought to. This is rather uncommon, almost non-existent as these urges are innate and attached to our animal nature.

- **Temperance [Virtue of Mean]** *Self-control, appropriate desire to fulfill bodily pleasures*

Someone who has temperance and self-control. This person desires moderately, in the appropriate way, at the appropriate time and at the appropriate place. The absence of these pleasant things or the abstinence from them is not painful to a temperate person.

- **Profligacy [Vice of Excess]** *Too much delight in bodily pleasures*

Someone who is a slave to bodily pleasures and loves these delights/pleasures too much. For example, heavy drinking (drinking more than is good for you), obesity, and gluttony. This person also enjoys things that (s)he should not enjoy, in the wrong manner, and more than necessary. The life of a profligate person is led by his/her appetites. Other examples: adultery and outrage.

Example:

- *There was one guy who had a younger brother. He (⇒ one guy who had a younger brother), ahem, tried to take advantage of me when I was drunk.*

Giving (and taking) - small amounts of money How and to which someone spends their resources (monetary and non-monetary)

- **Stinginess [Vice of Deficiency]** *Lack of giving, excess in taking*

Someone who cares more for wealth than is right, is fond of money. This person takes more than (s)he gives. Examples: greed, meanness, lending money at high interests. The mean person never does anything without thinking twice and always considers how things can be done at the least possible cost.

- **Liberality, Generosity [Virtue of Mean]** *Spending resources to create added value for others*

Someone who spends their money and other resources on proper things, allocate resources to the right people (generosity), acquires wealth by taking from the proper sources, and desires to reach goals in the right manner (proper amount spent). The liberal person invests in public objects and other people to create added value, which is not always valued in money. This person does go for the cheapest or most expensive way to do things (focus on means), but instead focuses on the goal.

Example:

- *However, he has a history of charity that goes far beyond this instance. He invested over one billion in a poor region in Russia, transforming the local economy and greatly raising the standard of living there. (he)*

- **Prodigality [Vice of Excess]** *Excess in giving (lack in taking)*

Someone who squanders money in riotous, wasteful living. The wasteful person wastes money on things that do not create added value to others (improper objects), or spends too much on small objects. This person also likes to display their wealth to others.

Giving (and taking) - other, added value How someone spends their wealth to create added value and to which things (their worth is not always measurable in terms of money), spending/giving according to their wealth.

- **Meanness [Vice of Deficiency]** *Lack of giving, excess in taking*

The mean person likes to do things at the lowest possible cost.

- **Magnificence [Virtue of Mean]** *Spending resources to create added value for others*

The magnificent person invests in public objects and other people to create added value, which is not always valued in money. This person does not go for the cheapest or most expensive way to do things (focus on means), but instead focuses on the goal.

- **Prodigality, Wastefulness [Vice of Excess]** *Excess in giving, lack in taking*

Someone who squanders money in riotous, wasteful living. The wasteful person wastes money on things that do not create added value to others (improper objects), or spends too much on small objects. This person also likes to display their wealth to others.

Pride (honour = cause) How someone communicates about and acts according to one's honour/pride.

- **Little-mindedness [Vice of Deficiency]** *Lacks pride, false modesty*

Someone who claims less than (s)he deserves and deprives him/herself of what (s)he deserves. This should not be confused with modesty as this person actively deprives himself of honour (false modesty). This person is too retiring.

- **High-mindedness [Virtue of Mean]** *Proper sense and display of honour*

Someone who claims much but also deserves much. This differs from modesty as a modest person has little to claim about - high-mindedness implies some kind of greatness. The high-minded person also does not seek honour from just everyone (does not look for praise) or on trivial grounds. This person is not easily moved to admiration, does not gossip or speak evil of others. Brooding on the past or blaming others is not something a high-minded person would do.

- **Vanity [Vice of Excess]** *Excess in pride, loves oneself*

Someone who claims much without deserving it. The vain person considers him/herself the better person (sense of superiority) and believes that (s)he deserves greater things. This sense of superiority is unjust and unjustified. This person is ignorant about his/her honour, likes to talk about his/herself, looks down on others and speaks evil of others. At a professional level, they tend to take positions for which they are unfit.

Example:

- *There are very low quality psychologists who go around trying to tell people what to do and fix their problems but they are not qualified because they have studied a shell major, which is psychology.*

Ambition (honour = goal) How someone builds towards more honour, how someone works on becoming a more Virtue of person.

- **Lack of ambition [Vice of Deficiency]** *Lacks ambition*

Someone who has little desire to improve one's character or condition. Example: laziness. This person also tends to blame others for his/her lack of ambition.

Example:

- Psychologist is just someone who cannot man up and study medicine so they go for the sad mockery (...) People (\Rightarrow psychologists) study it as a wild card/placeholder major. (psychologists)

- **Proper ambition [Virtue of Mean]** *Healthy ambition*

Someone who is ambitious in the proper manner, in the proper situations and for reaching appropriate goals.

- **Over-ambitiousness [Vice of Excess]** *Too much ambition*

Someone who desires too much - often at the expense of others. For this person, the goal justifies the means.

Anger How someone deals with emotions, especially negative emotions such as anger.

- **Easy-tempered, spiritlessness [Vice of Deficiency]** *Too little anger*

Someone who does not sufficiently deal with their negative emotions, quite emotionless. The easy-tempered person also lacks the spirit to defend him/herself and, more importantly, others.

- **Gentleness [Virtue of Mean]** *Deals with negative emotions such as anger in a proper way.*

Someone who is angry on the right occasions, with the right persons, in the right manner, for the right reasons and for the right length of time. The gentle person does not lose his/her balance and does not get carried away by emotions. This person is ready to forgive and is not eager to take vengeance.

- **Hard-tempered, wrathfulness [Vice of Excess]** *Too much anger*

Someone who is easily angered by anything and anyone on any occasion. This hard-tempered person's bursts of emotions can be either very short or very long (vengeance and punishment). This person is also prone to violence, vengeance and punishment. (S)he does not easily forgive. The extremes are choleric people.

Example:

- I understand Putin is insane. (Putin)

Pleasure (and pain) of others How someone agrees (pleasure of other) or disagrees (pain of other) with his/her interlocutor in any type of interaction.

- **Cross, Contentious [Vice of Deficiency]** *too much pain of other, disagreement*

Someone who objects to and disagrees with everything and anyone. The cross person does not consider the (emotional) pain they give to people, sets his/her face against everything.

Example:

- “My mother took some pictures and posted them on Facebook, where they were quickly filled with horrified comments from various Americans who she is friends with.” (various Americans who she is friends with)

- **Agreeableness, friendliness [Virtue of Mean]** *appropriate mix of agreement and disagreement*

Someone who (dis)agrees when she oughts to and not because of love or hate. The agreeable person fits their behavior to the context and the person they are talking to.

- **Flattery [Vice of Excess]** *Too much pleasure of other, agreement*

Someone who pleases people by praising everything and not objecting to anything. This “people-pleaser” feels the need to avoid discomfort and maximize the pleasure of others. When this person has a goal that leads to this pleasing, then we call it flattery.

Truth, honesty about oneself How truthful someone is in their interaction with others: what they say and how they say it.

- **Irony [Vice of Deficiency]** *Minimizing truthfulness by disclaiming and depreciating the truth*

Someone who disclaims what (s)he has or depreciates it. This depreciatory way of speaking allows him/her to avoid parade. The character of ironic person should not be confused with modesty and irony (as a figure of speech).

- **Truthfulness [Virtue of Mean]** *truthful in word and deed*

Someone who is truthful in word and deed. The truthful person never exaggerates or diminishes the truth, and shuns falsehood as a base thing (does not take pleasure in it). This person is more inclined towards understatement than overstatement of the truth.

Example:

- I also understand more thoughtful conservatives who voted for Trump, because regardless of their opinions on him (\Rightarrow Trump) as a person, they knew he would support conservative causes, which he did. (Trump)

- **Boastfulness [Vice of Excess]** *Pretension and twisting the truth*

Someone who likes to pretend to be someone they are not or pretend to hold ideas that others esteem even though they do not. Examples: liars and hypocrites.

Example:

- I will make it very clear, I'm not pro Russian or pro Ukraine, I think both sides are to be blamed for this war and mostly we are seeing only one side of the story from the western media. (the western media)

Amusing conversation How someone behaves during *amusing* conversation and mixing with others

- **Boorishness [Vice of Deficiency]** *Too boring and serious*

Someone who never says anything laughable, contributes little to the conversation, is too serious and takes everything in ill part.

- **Wittiness [Virtue of Mean]** *Graceful jest and tact*

Someone who has a lot of tact, jests gracefully, does not ridicule anyone.

- **Buffoonery [Vice of Excess]** *Excess in ridicule*

Someone who does ridicule things just for a laugh, not elegant in their wit, easily resides to offense. The ridicule person likes to make jests of people, vilify them or laugh with them.

A.3.4. Qualification test: round 2

Candidate annotators first have to pass a qualification test before they can start the annotation round. They were automatically admitted to the annotation round if they answered all questions or all but one question in the qualification test correctly. The qualification test starts by briefly explaining the three contextual appropriateness labels, called behavior types in the test for simplicity, after which the annotators have to select the correct description of the types (radio button). It then goes over all Spheres of Action, called situations in the test for simplicity. An example sentence and a person of interest is given for each situation and the annotators are expected to select the best-fitting behavior type for the person of interest as implied in the sentence (radio button). The sentences present rather clear examples of the behavior type. We indicate the correct answer with (X).

Test description: *In this project, you will be asked to describe IMPLIED (= indirectly suggested) CHARACTER TRAITS and classify them given a predefined set of situations and behaviors. We go over the ten different SITUATIONS (= spheres of action), each having three BEHAVIOR types ('vice of deficiency', 'virtue of mean', 'vice of excess').*

(3 BEHAVIOR TYPES) Each situation incites certain behaviors. Those behaviors can either be desired/positive (= virtue) or undesired/negative (= vice).

- A desired behavior (= virtue) is APPROPRIATE to the situation and people involved. It is often a moderate behavior. We therefore classify this type of desired behavior as 'VIRTUE OF MEAN'.

- An undesired behavior (= vice) is disproportionate to the situation and people involved. If it is EXCESSIVE, or 'too much', we call this type of undesired behavior as 'VICE OF EXCESS'. The behavior can also be DEFICIENT, or 'too little'. We then call this undesired behavior 'VICE OF DEFICIENCY'.

- Pick the correct description for the three behavior types: VICE OF DEFICIENCY - VIRTUE OF MEAN - VICE OF EXCESS.
 - "too much" - appropriate - "too little"
 - (X) "too little" - appropriate - "too much"

(10 SITUATIONS) For each situation (S), we give an example sentence and a person of interest (- Person). You need to pick the BEST-FITTING behavior type for the person of interest, as IMPLIED in the sentence.

- (S1) CONFIDENCE (FEAR, UNCERTAINTY): how someone deals with fear-inciting, uncertain events and the extent to which someone is confident about their actions.

- "They made increasingly reckless investments." – They

- * Vice of Deficiency ⇒ cowardice, or "too little" confidence (= too much fear/uncertainty)
- * Virtue of Mean ⇒ courage, or "appropriate" confidence, fear and uncertainty
- * (X) Vice of Excess ⇒ rashness, or "too much" confidence (= too little fear/uncertainty)

- (S2) PLEASURES OF THE BODY: how someone deals with and is able to control and fulfill needs and pleasures, especially those of the body.

- "He, ahem, tried to take advantage of me when I was drunk." – He

- * Vice of Deficiency ⇒ insensibility, or "too little" delight in bodily pleasures
- * Virtue of Mean ⇒ temperance, or "appropriate" delight (self-control)
- * (X) Vice of Excess ⇒ profligacy, or "too much" delight

- (S3) GIVING (AND TAKING) - small amounts of money: how someone spends their wealth on things whose worth are though of in terms of MONEY.

- "Don't ask Carla for any money, she would only lend it to you at large interests." – Carla

- * (X) Vice of Deficiency ⇒ stinginess, or "too little" giving (= too much taking)
- * Virtue of Mean ⇒ liberality/generosity, or "appropriate" giving (spending on the appropriate things)
- * Vice of Excess ⇒ prodigality/wastefulness, or "too much" giving (= too little taking; squandering)

- (S4) GIVING (AND TAKING) - other, added value: how someone spends their wealth to create added value and to which things (their worth is not always measurable in terms of money), spending/giving according to their wealth.

- "Anna supports the animal shelter in our city financially and volunteers there twice a week." – Anna

- * Vice of Deficiency ⇒ meanness, or "too little" giving (little added value)
- * (X) Virtue of Mean ⇒ magnificence, or "appropriate" giving (creating added value)
- * Vice of Excess ⇒ vulgarity, or "too much" giving (making great display on the wrong occasions and in the wrong way, little added value)

Table 9

Recall, Precision, and F1 scores were used to evaluate the overlap between sentences flagged by human interpretations and those flagged by the One2M-Sim-T5 model. One2M-Sim-T5 can mimic human behavior by recognizing hidden meanings in input sentences (*Analysis 1A*) and, in some cases, can detect toxicity and insults even more effectively than humans, as shown by the high precision scores - double when compared with the recall scores (*Analysis 1B*).

Score	Analysis 1A			Analysis 1B		
	Recall	Precision	F1	Recall	Precision	F1
Toxicity	85.29%	86.83%	86.05%	26.23%	57.14%	35.96%
Insult	83.91%	89.02%	86.39%	20.0%	43.75%	27.45%
Identity attack	90.45%	90.45%	90.45%	41.67%	50.0%	45.45%

- (S5) PRIDE (HONOUR = CAUSE): how someone communicates about and acts according to one's honour/pride. Honour is the CAUSE of their behavior.
 - "There are very low quality psychologists who go around trying to tell people what to do and fix their problems but they are not qualified." – psychologists
 - * Vice of Deficiency ⇒ little-mindedness, or "too little" pride (false modesty)
 - * Virtue of Mean ⇒ high-mindedness, or "appropriate" pride
 - * (X) Vice of Excess ⇒ vanity, or "too much" unjustified pride (sense of superiority)
- (S6) AMBITION (HONOUR = GOAL): How someone builds towards more honour and a more virtuous self. Honour is the GOAL of their behavior.
 - "John's parents had to force him to look for a job." – him (= John)
 - * (X) Vice of Deficiency ⇒ lack of ambition, or "too little" ambition
 - * Virtue of Mean ⇒ proper ambition, or "appropriate" ambition
 - * Vice of Excess ⇒ over-ambitiousness, or "too much" ambition (toxic)
- (S7) ANGER: How someone deals with emotions, especially negative emotions such as anger.
 - "Serena has a list of everyone who has ever wronged her." – Serena
 - * Vice of Deficiency ⇒ easy-tempered/spiritlessness, or "too little" anger (emotionless, careless)
 - * Virtue of Mean ⇒ gentleness, "appropriate" anger (led by reason, not by emotions)
 - * (X) Vice of Excess ⇒ hard-tempered, or "too much" anger (wrathfulness, irritability)
- (S8) PLEASURE (AND PAIN) OF OTHERS: general pleasantness in life, or how someone agrees (pleasure of other) or disagrees (pain of other) with his/her interlocutor in any type of interaction.
 - "My mother posted some of her poems on Facebook, but they were quickly filled with mean and discouraging comments from many of her so-called friends." – many of her so-called friends
 - * (X) Vice of Deficiency ⇒ cross/contentious, or "too little" pleasure/agreement (= too much pain)
 - * Virtue of Mean ⇒ agreeableness/friendliness, or "appropriate" mix of (dis)agreement
 - * Vice of Excess ⇒ flattery, or "too much" agreement (= too little pain; people pleaser, often for own advantage)
- (S9) TRUTH, HONESTY ABOUT ONESELF: How truthful someone is in their interaction with others; what they say and how they say it.
 - "To me, it sounds weird to think that Jeremy is lying." – Jeremy
 - * Vice of Deficiency ⇒ irony, or "too little" truthfulness by disclaiming/depreciating the truth (pretense as understatement)
 - * (X) Virtue of Mean ⇒ truthfulness, truthful in word and deed
 - * Vice of Excess ⇒ boastfulness, "too much" so-called truth (pretense as exaggeration)
- (S10) AMUSING CONVERSATION: How someone behaves during amusing conversation and mixing with others.
 - "Hannah never says anything remotely funny, she's always standing there with a long face." – Hannah
 - * (X) Vice of Deficiency ⇒ boorishness, or "too little" amusement (too boring, serious)
 - * Virtue of Mean ⇒ wittiness, "appropriate" amusement
 - * Vice of Excess ⇒ buffoonery, "too much" amusement (ridicule)

A.4. Demographic information on human evaluators

- Gender: female, female
- Age: 22, 23
- Education level: master, master
- Nationality: Belgian, Japanese

A.5. Additional results

In the main paper, we use recall to analyze the overlap between the sentences flagged by the ground-truth interpretations with those flagged by the One2M-Sim-T5 generated interpretations. A sentence is flagged if at least one of its interpretations has higher

toxicity, insult, or identity attack scores than the sentence's original scores. While recall is a key metric for content moderation, Table 9 also presents precision and F1 scores. *Analysis 1A* considers all flagged sentences, whereas *Analysis 1B* includes only sentences flagged by interpretations whose scores are at least 10 percentage points higher than the original sentence scores.

In the context of *Analysis 1A*, precision and F1 scores are similar to recall scores, demonstrating that the One2M-Sim-T5 model can identify sentences with hidden meanings by generating interpretations with toxicity, insult, and identity attack scores higher than the original sentence scores. The high precision scores, double compared to the recall scores for toxicity and insult (*Analysis 1B*) suggest that One2M-Sim-T5 may be even more effective at detecting toxicity and insults in input sentences than humans.

Data availability

The dataset is publicly available: <https://github.com/laallein/origamIM>.

References

- [1] K. Al Khatib, M. Völske, S. Syed, N. Kolyada, B. Stein, Exploiting personal characteristics of debaters for predicting persuasiveness, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, July, pp. 7067–7072, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.acl-main.632>.
- [2] L. Allein, I. Augenstein, M.-F. Moens, Time-aware evidence ranking for fact-checking, *J. Web Semant.* 71 (2021) 100663.
- [3] L. Allein, M.-F. Moens, D. Perrotta, Preventing profiling for ethical fake news detection, *Inf. Process. Manag.* 60 (2) (2023) 103206.
- [4] L. Allein, M. Saelens, R. Cartuyvels, M.-F. Moens, Implicit temporal reasoning for evidence-based fact-checking, in: Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, May, pp. 176–189, Retrieved from <https://aclanthology.org/2023.findings-eacl.13>.
- [5] L. Allein, M.-F. Moens, OrigamIM: a dataset of ambiguous sentence interpretations for social grounding and implicit language understanding, in: Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, May, pp. 116–122, Retrieved from <https://aclanthology.org/2024.nlperspectives-1.13>.
- [6] M. Alshomary, R. El Baff, T. Gurcke, H. Wachsmuth, The moral debater: a study on the computational generation of morally framed arguments, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, May, pp. 8782–8797, Retrieved from <https://aclanthology.org/2022.acl-long.601>.
- [7] Aristotle, Aristotle: nicomachean ethics, in: R. Crisp (Ed.), Cambridge University Press, 2000.
- [8] Y. Bang, N. Lee, T. Yu, L. Khalatbari, Y. Xu, D. Su, et al., Aisocrates: towards answering ethical quandary questions, *arXiv preprint*, arXiv:2205.05989, 2022.
- [9] N. Botzer, S. Gu, T. Weninger, Analysis of moral judgment on Reddit, *IEEE Trans. Comput. Soc. Syst.* (2022) 1–11, <https://doi.org/10.1109/TCSS.2022.3160677>.
- [10] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 6860–6868.
- [11] S. Chen, D. Khashabi, W. Yin, C. Callison-Burch, D. Roth, Seeing things from a different angle: discovering diverse perspectives about claims, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, June, pp. 542–557, Retrieved from <https://aclanthology.org/N19-1053>.
- [12] G. Collell, L. Van Gool, M.-F. Moens, Acquiring common sense spatial knowledge through implicit spatial templates, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.
- [13] K. Deng, A. Ray, R. Tan, S. Gabriel, B.A. Plummer, K. Saenko, Socratis: are large multimodal models emotionally aware?, *arXiv preprint*, arXiv:2308.16741, 2023.
- [14] D. Deutsch, R. Dror, D. Roth, On the limitations of reference-free evaluations of generated text, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, December, pp. 10960–10977, Retrieved from, <https://aclanthology.org/2022.emnlp-main.753>.
- [15] T. Draws, O. Inel, N. Tintarev, C. Baden, B. Timmermans, Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics, in: Acm Sigir Conference on Human Information Interaction and Retrieval, 2022, pp. 135–145.
- [16] T. Draws, N. Roy, O. Inel, A. Rieger, R. Hada, M.O. Yalcin, et al., Viewpoint diversity in search results, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I, 2023, pp. 279–297.
- [17] I.S. Efstathiadis, G. Paulino-Passos, F. Toni, Explainable patterns for distinction and prediction of moral judgement on Reddit, *arXiv preprint*, arXiv:2201.11155, 2022.
- [18] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, D. Yang, Latent hatred: a benchmark for understanding implicit hate speech, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, November, pp. 345–363, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.29>.
- [19] D. Emelin, R. Le Bras, J.D. Hwang, M. Forbes, Y. Choi, Moral stories: situated reasoning about norms, intents, actions, and their consequences, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, November, pp. 698–718, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.54>.
- [20] M. Forbes, J.D. Hwang, V. Shwartz, M. Sap, Y. Choi, Social chemistry 101: learning to reason about social and moral norms, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, November, pp. 653–670, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.emnlp-main.48>.
- [21] P. Fortuna, M. Dominguez, L. Wanner, Z. Talat, Directions for NLP practices applied to online hate speech detection, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, December, pp. 11794–11805, Retrieved from <https://aclanthology.org/2022.emnlp-main.809>.
- [22] S. Gabriel, S. Hallinan, M. Sap, P. Nguyen, F. Roesner, E. Choi, Y. Choi, Misinfo reaction frames: reasoning about readers' reactions to news headlines, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, May, pp. 3108–3127, Retrieved from <https://aclanthology.org/2022.acl-long.222>.
- [23] G. Gao, S. Paik, C. Reardon, Y. Zhao, L. Guo, P. Ishwar, et al., Prediction of people's emotional response towards multi-modal news, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2022, November, pp. 364–374, Online only: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2022.aacl-main.29>.
- [24] T. Gao, X. Yao, D. Chen, SimCSE: simple contrastive learning of sentence embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, November, pp. 6894–6910, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.552>.

- [25] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, J. Berant, Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies, *Trans. Assoc. Comput. Linguist.* 9 (2021) 346–361, https://doi.org/10.1162/tacl_a_00370.
- [26] A. Gupta, A. Agarwal, P. Singh, P. Rai, A deep generative framework for paraphrase generation, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, 2018, pp. 5149–5156.
- [27] J. Haidt, C. Joseph, Intuitive ethics: how innately prepared intuitions generate culturally variable virtues, *Daedalus* 133 (4) (2004) 55–66, <https://doi.org/10.1162/0011526042365555>.
- [28] M. Hanna, O. Bojar, A fine-grained analysis of BERTScore, in: Proceedings of the Sixth Conference on Machine Translation, 2021, November, pp. 507–517, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.wmt-1.59>.
- [29] R. Henderson, E. McCready, How dogwhistles work, in: New Frontiers in Artificial Intelligence: Jsaï-Isai 2017, vol. 10838, 2017, pp. 231–240.
- [30] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, J. Steinhardt, Aligning AI with shared human values, in: International Conference on Learning Representations, 2020.
- [31] D. Hendrycks, M. Mazeika, A. Zou, S. Patel, C. Zhu, J. Navarro, et al., What would jiminy cricket do? Towards agents that behave morally, Retrieved from <https://openreview.net/forum?id=G1muTb5zuO7>, 2021.
- [32] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A.M. Davani, Y. Lin, et al., Moral foundations Twitter corpus: a collection of 35k tweets annotated for moral sentiment, *Soc. Psychol. Pers. Sci.* 11 (8) (2020) 1057–1071, <https://doi.org/10.1177/1948550619876629>.
- [33] D. Hovy, D. Yang, The importance of modeling social factors of language: theory and practice, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, June, pp. 588–602, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.nacl-main.49>.
- [34] A. Hoyle, R. Sarkar, P. Goel, P. Resnik, Making the implicit explicit: implicit content as a first class citizen in nlp, arXiv preprint, arXiv:2305.14583, 2023.
- [35] F. Huang, H. Kwak, J. An, Is chatgpt better than human annotators? Potential and limitations of chatgpt in explaining implicit hate speech, in: Companion Proceedings of the Acm Web Conference 2023, Association for Computing Machinery, New York, NY, USA, 2023, pp. 294–297, Retrieved from <https://doi.org/10.1145/3543873.3587368>.
- [36] D. Ippolito, R. Kriz, J. Sedoc, M. Kustikova, C. Callison-Burch, Comparison of diverse decoding methods from conditional language models, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, July, pp. 3752–3762, Retrieved from <https://aclanthology.org/P19-1365>.
- [37] G. Jigsaw, Perspective API, <https://www.perspectiveapi.com/>, 2021. (Accessed 2 August 2024), Online.
- [38] Z. Jin, S. Levine, F. Gonzalez, O. Kamal, M. Sap, M. Sachan, et al., When to make exceptions: exploring language models as accounts of human moral judgment, in: Advances in Neural Information Processing Systems 35 (NEURIPS 2022), vol. 35, 2022, pp. 28458–28473.
- [39] H. Kim, J. Hessel, L. Jiang, X. Lu, Y. Yu, P. Zhou, et al., Soda: million-scale dialogue distillation with social commonsense contextualization, arXiv preprint, arXiv:2212.10465, 2022.
- [40] H.W. Kuhn, The Hungarian method for the assignment problem, *Nav. Res. Logist. Q.* 2 (1–2) (1955) 83–97, <https://doi.org/10.1002/nav.3800020109>.
- [41] H.-C. Kuo, Y.-N. Chen, Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, July, pp. 249–258, Retrieved from <https://aclanthology.org/2023.findings-acl.17>.
- [42] Z. Li, Y. Zou, C. Zhang, Q. Zhang, Z. Wei, Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, November, pp. 246–256, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.22>.
- [43] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.
- [44] J. Lin, Q. Chen, J. Zhou, J. Jin, L. He, L. De Raedt, Cup: Curriculum learning based prompt tuning for implicit event argument extraction, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, 2022, pp. 4245–4251, Retrieved from <https://doi.org/10.24963/ijcai.2022/589>.
- [45] Y. Lin, J. Hoover, G. Portillo-Wightman, C. Park, M. Dehghani, H. Ji, Acquiring background knowledge to improve moral value prediction, in: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018, pp. 552–559.
- [46] A. Liu, Z. Wu, J. Michael, A. Suhr, P. West, A. Koller, et al., We're afraid language models aren't modeling ambiguity, arXiv preprint, arXiv:2304.14399, 2023.
- [47] J. Liu, Y. Chen, J. Xu, Machine reading comprehension as data augmentation: a case study on implicit event argument extraction, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, November, pp. 2716–2725, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.214>.
- [48] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing, *ACM Comput. Surv.* 55 (9) (2023, jan.), Retrieved from <https://doi.org/10.1145/3560815>.
- [49] Y. Liu, P. Liu, D. Radev, G. Neubig, BRIO: bringing order to abstractive summarization, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, May, pp. 2890–2903, Retrieved from <https://aclanthology.org/2022.acl-long.207>.
- [50] P. Madhyastha, A. Founta, L. Specia, A study towards contextual understanding of toxicity in online conversations, *Nat. Lang. Eng.* 29 (6) (2023) 1538–1560.
- [51] J. Mendelsohn, R.L. Bras, Y. Choi, M. Sap, From dogwhistles to bullhorns: unveiling coded rhetoric with language models, arXiv preprint, arXiv:2305.17174, 2023.
- [52] N. Mirzakhmedova, J. Kiesel, K. Al-Khatib, B. Stein, Unveiling the power of argument arrangement in online persuasive discussions, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, December, Retrieved from <https://aclanthology.org/2023.findings-emnlp.1048>.
- [53] M.D. Muralikumar, Y.S. Yang, D.W. McDonald, A human-centered evaluation of a toxicity detection API: testing transferability and unpacking latent attributes, *ACM Trans. Soc. Comput.* 6 (1–2) (2023) 1–38.
- [54] E. Musi, D. Ghosh, S. Muresan, Changemyview through concessions: do concessions increase persuasion?, *Dialogue Discourse* 9 (1) (2018) 107–127.
- [55] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, D. Kiela, Adversarial NLI: a new benchmark for natural language understanding, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, July, pp. 4885–4901, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.acl-main.441>.
- [56] Y. Nie, X. Zhou, M. Bansal, What can we learn from collective human opinions on natural language inference data?, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, November, pp. 9131–9143, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.emnlp-main.734>.
- [57] N. Ocampo, E. Sviridova, E. Cabrio, S. Villata, 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, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, May, pp. 1997–2013, Retrieved from <https://aclanthology.org/2023.eacl-main.147>.
- [58] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
- [59] J. Pater, A. Coupe, R. Pfafman, C. Phelan, T. Toscos, M. Jacobs, Standardizing reporting of participant compensation in HCI: a systematic literature review and recommendations for the field, in: Proceedings of the 2021 Chi Conference on Human Factors in Computing Systems, 2021, pp. 1–16.

- [60] E. Pavlick, T. Kwiatkowski, Inherent disagreements in human textual inferences, *Trans. Assoc. Comput. Linguist.* 7 (2019) 677–694, https://doi.org/10.1162/tacl_a_00293, Retrieved from <https://aclanthology.org/Q19-1043>.
- [61] B. Plank, The “problem” of human label variation: on ground truth in data, modeling and evaluation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, December, pp. 10671–10682, Retrieved from <https://aclanthology.org/2022.emnlp-main.731>.
- [62] V. Pyatkin, J.D. Hwang, V. Srikumar, X. Lu, L. Jiang, Y. Choi, C. Bhagavatula, ClarifyDelphi: reinforced clarification questions with defeasibility rewards for social and moral situations, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, July, pp. 11253–11271, Retrieved from <https://aclanthology.org/2023.acl-long.630>.
- [63] L. Qian, L. Qiu, W. Zhang, X. Jiang, Y. Yu, Exploring diverse expressions for paraphrase generation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, November, pp. 3173–3182, Retrieved from <https://aclanthology.org/D19-1313>.
- [64] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, *OpenAI Blog* 1 (8) (2019) 9.
- [65] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* 21 (1) (2020) 5485–5551, <https://doi.org/10.5555/3455716.3455856>.
- [66] H. Rashkin, M. Sap, E. Allaway, N.A. Smith, Y. Choi, Event2Mind: commonsense inference on events, intents, and reactions, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, July, pp. 463–473, Retrieved from <https://aclanthology.org/P18-1043>.
- [67] C. Rebuffel, M. Roberti, L. Soulier, G. Scoutheeten, R. Cancelliere, P. Gallinari, Controlling hallucinations at word level in data-to-text generation, *Data Min. Knowl. Discov.* 36 (1) (2022) 318–354, Retrieved from <https://doi.org/10.1007/s10618-021-00801-4>.
- [68] R. Rei, J.G.C. de Souza, D. Alves, C. Zerva, A.C. Farinha, T. Glushkova, et al., COMET-22: Unbabel-IST 2022 submission for the metrics shared task, in: Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, December, pp. 578–585, Retrieved from <https://aclanthology.org/2022.wmt-1.52>.
- [69] R. Rei, C. Stewart, A.C. Farinha, A. Lavie, COMET: a neural framework for MT evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, November, pp. 2685–2702, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.emnlp-main.213>.
- [70] P. Ricoeur, *The Conflict of Interpretations: Essays in Hermeneutics*, Northwestern University Press, Evanston, Illinois, 1974.
- [71] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N.A. Smith, Y. Choi, Social bias frames: reasoning about social and power implications of language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, July, pp. 5477–5490, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.acl-main.486>.
- [72] M. Sap, H. Rashkin, D. Chen, R. Le Bras, Y. Choi, Social IQa: commonsense reasoning about social interactions, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, November, pp. 4463–4473, Retrieved from <https://aclanthology.org/D19-1454>.
- [73] K. Stasaki, M. Hearst, Semantic diversity in dialogue with natural language inference, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, July, pp. 85–98, Retrieved from <https://aclanthology.org/2022.naacl-main.6>.
- [74] Z. Talat, H. Blix, J. Valvoda, M.I. Ganesh, R. Cotterell, A. Williams, On the machine learning of ethical judgments from natural language, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, July, pp. 769–779, Retrieved from <https://aclanthology.org/2022.naacl-main.56>.
- [75] G. Tevet, J. Berant, Evaluating the evaluation of diversity in natural language generation, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, April, pp. 326–346, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.eacl-main.25>.
- [76] A.N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: a survey, *J. Artif. Intell. Res.* 72 (2021) 1385–1470, <https://doi.org/10.1613/jair.1.12752>.
- [77] V. Veselovsky, M.H. Ribeiro, R. West, Artificial artificial intelligence: crowd workers widely use large language models for text production tasks, arXiv preprint, arXiv:2306.07899, 2023.
- [78] A. Vijayakumar, M. Cogswell, R. Selvaraju, Q. Sun, S. Lee, D. Crandall, D. Batra, Diverse beam search for improved description of complex scenes, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.
- [79] P. Vijayaraghavan, D. Roy, Modeling human motives and emotions from personal narratives using external knowledge and entity tracking, in: Proceedings of the Web Conference 2021, 2021, pp. 529–540.
- [80] B. Wang, C.-C.J. Kuo, H. Li, Just rank: rethinking evaluation with word and sentence similarities, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, May, pp. 6060–6077, Retrieved from <https://aclanthology.org/2022.acl-long.419>.
- [81] A. Welivita, P. Pu, A taxonomy of empathetic response intents in human social conversations, in: Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 2020, December, pp. 4886–4899, (Online): International Committee on Computational Linguistics. Retrieved from, <https://aclanthology.org/2020.coling-main.429>.
- [82] M. Wiegand, M. Geulig, J. Ruppenhofer, Implicitly abusive comparisons – a new dataset and linguistic analysis, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, April, pp. 358–368, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.eacl-main.27>.
- [83] H. Wu, H. Zhou, M. Lan, Y. Wu, Y. Zhang, Connective prediction for implicit discourse relation recognition via knowledge distillation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, July, pp. 5908–5923, Retrieved from <https://aclanthology.org/2023.acl-long.325>.
- [84] Y. Wu, X. Ma, D. Yang, Personalized response generation via generative split memory network, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, June, pp. 1956–1970, Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.naacl-main.157>.
- [85] W. Xiang, B. Wang, A survey of implicit discourse relation recognition, *ACM Comput. Surv.* 55 (12) (2023, mar), <https://doi.org/10.1145/3574134>, Retrieved from <https://doi.org/10.1145/3574134>.
- [86] Z. Yang, S. Hooshmand, J. Hirschberg, CHoRaL: collecting humor reaction labels from millions of social media users, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, November, pp. 4429–4435, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.364>.
- [87] W. Yu, C. Zhu, T. Zhao, Z. Guo, M. Jiang, Sentence-permuted paragraph generation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, November, pp. 5051–5062, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.412>.
- [88] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, BERTScore: evaluating text generation with BERT, in: International Conference on Learning Representations, 2020, Retrieved from <https://openreview.net/forum?id=SkeHuCVFDr>.

- [89] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C.M. Meyer, S. Eger, MoverScore: text generation evaluating with contextualized embeddings and Earth mover distance, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, November, pp. 563–578, Retrieved from <https://aclanthology.org/D19-1053>.
- [90] D. Zhou, J. Wang, L. Zhang, Y. He, Implicit sentiment analysis with event-centered text representation, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, November, pp. 6884–6893, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.551>.
- [91] J. Zhou, S. Bhat, Paraphrase generation: a survey of the state of the art, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, November, pp. 5075–5086, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2021.emnlp-main.414>.
- [92] C. Ziems, J. Yu, Y.-C. Wang, A. Halevy, D. Yang, The moral integrity corpus: a benchmark for ethical dialogue systems, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, May, pp. 3755–3773, Retrieved from <https://aclanthology.org/2022.acl-long.261>.