The Blame Game: Understanding Blame Assignment in Social Media

Ruijie Xi and Munindar P. Singh *IEEE Fellow*
Department of Computer Science, North Carolina State University
rxi@ncsu.edu, mpsingh@ncsu.edu

***Abstract*—Cognitive and psychological studies on morality have proposed underlying linguistic and semantic factors. However, laboratory experiments in the philosophical literature often lack the nuances and complexity of real life. This paper examines how well the findings of these cognitive studies generalize to a corpus of over 30,000 narratives of tense social situations submitted to a popular social media forum. These narratives describe interpersonal moral situations or misgivings; other users judge from the post whether the author (*protagonist*) or the opposing side (*antagonist*) is morally culpable. Whereas previous work focuses on predicting the polarity of normative behaviors, we extend and apply natural language processing (NLP) techniques to understand the effects of descriptions of the people involved in these posts. We conduct extensive experiments to investigate the effect sizes of features to understand how they affect the assignment of blame on social media. Our findings show that aggregating psychology theories enables understanding real-life moral situations. Moreover, our results suggest that there exist biases in blame assignment on social media, such as males are more likely to receive blame no matter whether they are protagonists or antagonists.**

***Index Terms*—Moral language, User-generated content, moral understanding**

EXAMPLE : SAMPLE POST AND COMMENTS ON IT.

Title:
"AITA for snitching on my sister?"

Body:
"...I told my parents that my sister was staying up late with her tablet even though they had said she couldn't do it anymore. Now she's mad ..."

Top-level Comment:
"OTHER. While you shouldn't be parenting her, you didn't go to your parents until she repeatedly ignored them as well as your warnings ..."

Top-level Comment:
"AUTHOR. If your sister was doing something really bad that hurt someone ... You have undermined her trust in you ..."

Flair: "OTHER"

I. INTRODUCTION

How do people judge whether someone deserves blame for their actions? This question has been extensively studied in social science. Malle et al. [31] find that people assign blame to individuals they observe violating norms. Gray and Wegner [17] show that victims can escape from being blamed, whereas heroes may cause blame. Guglielmo and Malle [18] suggest that moral blame is more complex than moral praise. Such social science experiments were conducted mainly through questionnaires and surveys, which stylize the social situations and limit the number of participants. In contrast, online social systems enable people worldwide to share viewpoints about a spectrum of social situations on various topics, allowing researchers to explore real-life blame with the aid of computational tools.

This paper examines a popular subreddit (i.e., forum), /r/AmITheAsshole (AITA),[1] where users describe interpersonal conflicts and other users (i.e., audience) comment and judge who deserves blame. Example I shows a post and associated comments from AITA. The *title* and *body* are of the post, *top-level comment* comes from the audience (often including a verdict), and *flair* is the verdict of the top-voted comment. The most common verdicts are *author* and *other*. We use the term *blame assignment* to represent a post's verdict. Section II-A provides additional details.

Previous works take AITA as a resource for studying crowdsourced blame assignments on first-person moral situations. Much attention falls on accurately predicting verdicts [12, 24, 29]. These works apply neural networks such as transformers [10] to obtain high accuracy of prediction performance. However, these models focus on accuracy but do not shed light on what linguistic characteristics affect the audience's decisions on assigning blame. Moreover, these models may be flawed since they don't consider social psychology research [15].

Previous empirical work has not studied how social psychology theories generalize to real-life situations posted in AITA. According to the Theory of Dyadic Morality (TDM), blame is assigned to an *agent* when behaviors are causing damage to a vulnerable *patient* [43]. Under TDM, agents are perceived as blameworthy, where their agentiveness depends on how they are described [17]. The posts in AITA are first-person narratives that involve multiple individuals' and social identities (e.g., genders). Accordingly, the audience assigns blame to the individuals that they think are described as agents. However, what descriptive features of individuals affect the audience's recognition of agentiveness remains unstudied.

This work studies the features of AITA's posts that affect the audience's blame assignment. We especially focus on social psychology research relating to language and social features. Language features affect social media data in many ways. For instance, Beel et al. [3] find sentiment is powerful in predicting the contentiousness of conversations on Reddit. In addition, social factors, such as gender, affect social media interactions [3] and can lead to biases. For instance, De Candia et al. [9] find that males have a higher possibility to receive blame on social media. Ferrer et al. [13] find that Reddit posts' topics are

[1] https://www.reddit.com/r/AmItheAsshole/

TABLE I
SUMMARIZING RECENT WORK ON MORAL-DECISION MAKING MODELS AND AITA.

Type	Paper	Description	Dataset
Moral-decision making	Lourie et al. [29]	Building neural models to predict moral scenarios	Scraped from AITA
	Forbes et al. [14]	Building neural models to predict morality of social norms	Crowd-sourced dataset
	Emelin et al. [12]	Building neural models to predict intents, actions, and consequences of social norms	Crowd-sourced dataset
	Jiang et al. [24]	Building neural models to provide moral advisor	Multi-sourced datasets
Statistical Analysis	Nguyen et al. [37]	Taxonomizing the structure of moral discussions	Scraped from AITA
	De Candia et al. [9]	Analyzing demographic information of blame assignments	Scraped from AITA
	Zhou et al. [50]	Analyzing linguistic features in blame assignments	Scraped from AITA
	Botzer et al. [5]	Analyzing morality by building a moral judgment classifier	Scraped from AITA

gender biased, for instance, *power*-related posts are associated with males. Moreover, social scientists observe that gender and age affect morality in many ways [6, 48]. For instance, Reynolds et al. [41] find that moral typecasting stereotypes females into the role of suffering patients.

Malle et al. [31] proposes that assigning blame is a cognitive process that requires individuals to foresee the negative outcomes of agentive behaviors. Therefore, we define *cognitive-affective features* as language features that can shape the audience's blame assignment decisions. To this end, this paper aims to address two research questions:

$RQ_{Feature}$: What cognitive-affective language features are crucial in blame assignment?

RQ_{Social}: What biases, if any, arise in blame assignment on social media?

A. Methods

To answer $RQ_{Feature}$, we operationalize a set of novel factors using Natural Language Processing (NLP) that have explanatory power. We propose a novel **entity-centric** approach that partitions individuals involved in a situation as the protagonist (author) and antagonists (others). Then, we collect language features describing the entities based on existing social science research, such as emotions conveyed from attributive and predicative words [35]. The language features are categorized into contextual, psycholinguistic, and linguistic features, where psycholinguistic features are entity-based and others are situation-based. Although previous research uses these features to analyze social media data [25, 42], it doesn't apply them to morality with entity-based approaches. We use the proposed features to build machine learning classifiers to predict whether an entity will receive blame. Using these classifiers, we examine the features' effect sizes to understand how an entity causes blame given the description.

To answer RQ_{Social}, we consider gender and age as social factors that may lead to biases in blame assignment [5, 9]. We extend the previous works by conducting qualitative and quantitative analysis using a post's textual information, which helps understand a situation in linguistic terms. We extract the demographics of the entities from the posts (they are marked with expressions such as [25F]). We apply statistical methods to measure the association strengths between blame assignment and these social factors.

B. Contributions and Findings

This paper contributes in two aspects. First, we characterize blame assignment with novel features inspired by social psychology literature. We show these features have sufficient accuracy in predicting blame assignment while being interpretable. Second, our proposed methods go beyond theoretical models and provide insights that can benefit psychological research, such as optimizing language used in surveys for laboratory experiments of studying morality.

Our analyses show that certain generic linguistic characteristics are highly correlated with blame assignment across the board: for instance, the protagonist can reduce blame by eliciting positive perspectives (e.g., supportive) towards themselves. Additionally, authors can reduce blame when they describe themselves using less powerful words, whereas using dominance-related words triggers blame. Furthermore, our analysis suggests that authors in the 15–45 age group are more likely to attract bias than others. In addition, males have a higher possibility to receive blame whether they are protagonists or antagonists, especially when they post specific situations, such as discussing *medicines and medical treatment* and *judgment of appearance*.

C. Literature Review

Recently, researchers have considered the possibility of improving moral decision-making through AI by understanding (im)moral social norms and behavioral rules using NLP. Table I summarizes recent relevant research, in two groups. The first group deals with predicting moral judgments. Lourie et al. [29] use the title of a post in AITA as a social norm and develop a large dataset including human described situations based on the social norms. Forbes et al. [14] break down blame assignments of one-liner scenarios into rules of thumb and ask annotators to write moral and immoral stories based on the rules. Similarly, Emelin et al. [12] build crowd-sourced dataset based on the social norms [29], which include actions, intentions, and consequences of a moral situation. Other works build moral judgment classifier to apply to other social media [5] and predict blame assignment on one-line natural language snippets from possibilities such as understandable, wrong, bad, and rude [24].

However, social scientists point out that the general enterprise of training morality models on crowd-sourced data with no underlying moral framework is deeply flawed, as is the case for the Delphi system [24] [15, 46]. Besides, psychologists note the necessity for such AI systems to have a coherent understanding of human moral psychology [28]. Hence, we do *not* aim to build the most accurate judgment predictor—and thus do not compete with state-of-art neural models. Instead, we expand the computational modeling of moral understanding based on how social psychology constructs are apparent in language.

The second group in Table I concerns analyzing AITA using statistical methods, such as creating a taxonomy of moral discussions [37], analyzing the correlation between users' demographics

and blame assignments [9], and identifying linguistic features in moral judgment [50]. However, no work has studied the effects of the descriptions on individuals' agentiveness reflected in social media.

II. Proposed Method

Our framework is divided into four phases as shown in Figure 1.

1) *Dataset collection* involves collecting data from a subreddit and preprocessing the data into a proper format.
2) *Entity-centric implementation* involves separating entities, generating subject-verb-object (SVO) tuples, collecting semantic roles, extracting gender and age, and generating adjectives-noun pairs (ANP).
3) *Feature Extraction* includes measuring psycholinguistic, contextual, and linguistic features.

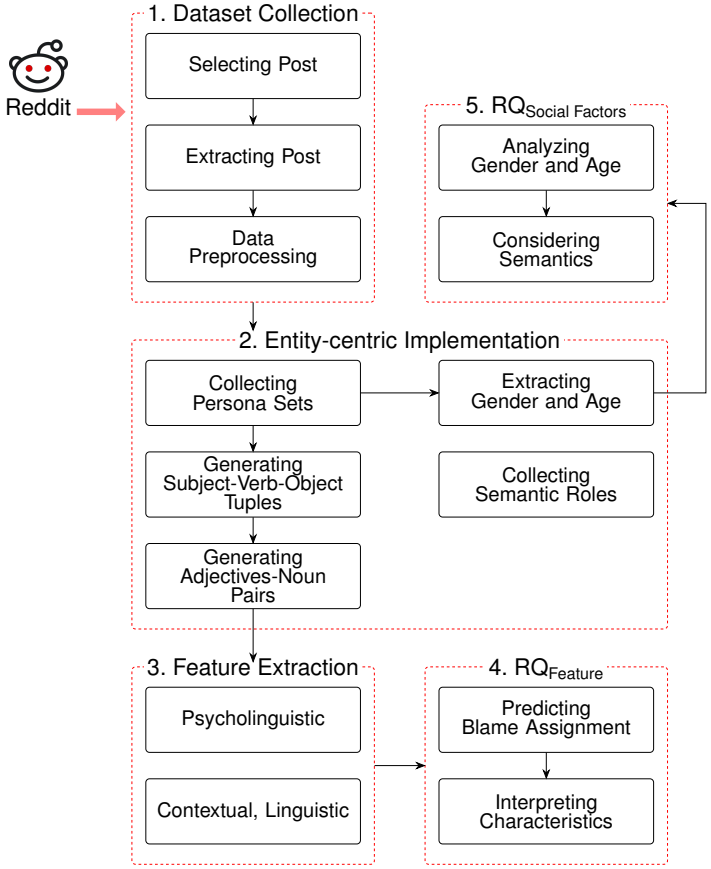


Fig. 1. Overall pipeline of the proposed method.

A. Dataset Collection

1) Selecting posts from Reddit: Although AITA has been used in previous studies, they are either not public [9, 50] or insufficient for answering our research questions [29, 37]. Our work needs the selected posts to contain predicate-argument structures that previous works haven't mentioned. To improve relevance and accuracy, we constrain our dataset (FAITA) (details are in Section II) to include posts that have:

- Been given flairs (the determined verdict).
- At least 50 top-level comments (judgments).
- Majority votes the same as the flair.
- At least ten extractable subject-verb-object tuples and ten extractable adjectives-noun pairs.

2) Extracting Post: We use the PushShift API[2] and Reddit API[3] to extract data over July 2020–July 2021. Some AITA stories may be faked to solicit outrage. The moderator deletes posts that are not truthful or not about interpersonal conflicts, which violate the subreddit rules.[4] We remove undesirable posts—those deleted, from a moderator, or too short—to ensure that the posts in our dataset decrease the conflicts between two parties and avoid discrepancies between data from Reddit and the archived data from PushShift.

Each judgment in the comments takes the form of a code: YTA, NTA, ESH, NAH, and INFO, which correspond to the classes AUTHOR, OTHER, EVERYONE, NO ONE, and MORE INFO. However, labeling the post with the majority votes may be inaccurate because morality is relative. Instead, we extract the *title*, *text*, and *flair* of each post. The *Flair* of each post is determined by the verdict of the top-voted comment 18 hours after submission (or the majority computed from its ten top-level comments if there is no flair field). We assign labels to YTA as 1 and NTA as 0 and discard other codes. This process yields 32,696 posts. We randomly split posts into 80% as the training, 10% as the development (dev), and 10% as the test sets. Table II-A2 shows the distributions of FAITA.

Dataset	Train	Dev	Test
Posts	26,156	3,270	3,270
Sentences	376,846	125,766	125,332
Author Wrong (label 1)	9,874	1,182	1,238
Others Wrong (label 0)	16,282	2,088	2,031

3) Data Preprocessing: We combine the title and text of posts in FAITA. We preprocess the text using the NLTK toolbox.[5] We remove all emojis, punctuation (except periods for separating sentences), symbols, and special characters and replace contractions with patterns (e.g., replace *can't* with *can not*). We tokenize the sentences and lemmatize tokens using WordNet Lemmatiser [38]. We identify a "sentence" as *words separated by a period in the original post.*

B. Entity-Centric Implementation

We build a set of syntax-aware methods for extracting the protagonist (author) and antagonist (others) of each post using entity coreference and the syntactic dependency parse. These **entity-centric methods** require partitioning entity tokens into the protagonist and antagonist persona sets, understanding how the authors are portrayed the "cast of main characters" in the narratives, and how these characters behave. We use Semantic Role Labeling (SRL) [26] to identify the protagonist and antagonist in each post.

1) Collecting Persona Sets: The protagonist and *antagonists* persona sets, respectively, contain first-person pronouns (e.g., I, me, and we), and third-person pronouns (e.g., she, he, and they). We add the pronouns to the persona sets as key tokens. We use the Spacy[6] dependency parser to extract more candidate terms by identifying part of speech tags (e.g., PRON, PROPN, and NOUN). We filter the nouns and proper nouns by a total of 3,125 people-related words from prior research, such as characters in history

[2]https://github.com/pushshift/api
[3]https://www.reddit.com/dev/api
[4]https://www.reddit.com/r/AmItheAsshole/about/rules/
[5]https://www.nltk.org/
[6]https://spacy.io

textbooks [30]. Thus, we can collect all the people-related nouns and proper nouns. Then we use Huggingface[7] `neuralcoref` for coreference resolution, and append all tokens from spans that corefer to the pronouns in protagonist or antagonist persona sets, respectively.

2) Collecting SRL: Unlike syntax-aware methods, SRL analyzes sentences with respect to predicate-argument structures such as "*who* did *what* to *whom* and *when* and *how* and *why*." We employ the AllenNLP BERT-based Semantic Role Labeller [16] to extract spans tagged `ARG0` for *agent* and `ARG1` for *patient*. As the following example shows, each sentence may have multiple tagged spans; thus, we first identify the SRL-tagged sentences.

1) **They** (`ARG0`) claimed **me** (`ARG1`) a dependent even though **I** (`ARG0`) have been financially independent for about a year.

In each post, we match the entities in persona sets with SRL labels `ARG0` or `ARG1`. This enables us to find when the author describes themselves or others as *agent* or *patient* in each narrative.

3) Generating Subject-Verb-Object (SVO) Tuples: Beginning from the persona sets, we first identify verbs (VERB) that have dependencies with entities in the persona sets using a syntactic dependency parse tree. We consider entities typed `nsub` (nominal subject), `nsubjpass` (passive nominal subject), `csubj` (clausal subject), `csubjpass` (passive clausal subject), `xsubj` (controlling subject), to the verbs as subjects; we consider entities typed `dobj` (direct object), and `iobj` (indirect object), to the verbs, as objects. Then we add the SVO tuples for persona sets accordingly. Besides the directly generated SVO tuples, we also generate new SVO tuples by finding entities from spans that corefer to the subject or object. Using a dependency parser, it is possible to handle the negation of the verbs and add a "not" before the verb as shown in Figure 2.

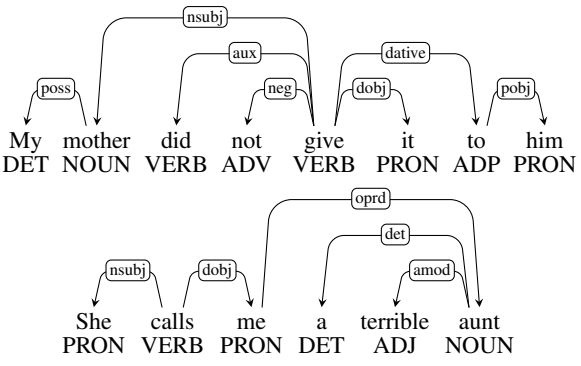


Fig. 2. Dependency parsers for the example sentences.

Figure 2 shows two sentences and their dependency trees in one post. Here, we consider four people, *my mother*, *him*, *she*, *me*, and two SVO tuples, which are *(my mother, not give, it)* and *(she call me)*. The coreference resolution finds *she* corefers to *my mother*, so we add *(my mother, call, me)* to the SVO tuples.

4) Generating Adjectives-Noun Pairs (ANP): Adjective-noun pair is a semantic construct for capturing the effect of an adjectival modifier to modify the meaning of the nouns such as "cute dog" or "beautiful landscape." Similarly, we use a dependency parse tree to identify adjectives for the entities in the persona sets. We use `amod` (adjectival modifier), `acomp` (adjectival complement), and `ccomp` (clausal complement) dependencies to select the adjectives modifying the entities. As shown in Figure 2, after we add *aunt* to the protagonist persona set, we find *terrible, aunt* pair because of the `amod` tag. We also add *terrible, me* because *aunt* corefers to *me*.

5) Extracting Gender and Age for Persona Sets: Note that posts on AITA are interpersonal stories; the complexity of the description makes it difficult and expensive to extract the social factors of the antagonist. Therefore, we extract and explore the social factors of the protagonist to improve the accuracy of the generated social factors. Gender and age identities are not typically available on Reddit, allowing for anonymous posting. Fortunately, the social media template for posting gender and age, e.g., [25f] (25-year-old female) or `(?i:i|i am a)([mf]|(?:fe)?male))`, enables us to use regular expressions to extract the information. We extract age by string matching on the gender modifier or the numeric age. To improve the accuracy of the extraction, we consider two string matching patterns: *I [25m] and my wife [25f]*. Besides, we consider gendered alternatives where available; for example, male can be estimated by `\b(boy|father|son)\b)` and female by (`\b(girl|mother|daughter)\b`). We do not match nonbinary genders because we do not have ground truth labels for nonbinary targets. To evaluate the regular expression, we took a random sample of 300 submissions and checked the result. We found no false positives and 2% false negative cases. In addition, when gender and age are extracted by regular expression, it matches the manually labeled one 94% of the time.

C. Feature Extraction

We categorize the features into **contextual**, **psycholinguistic**, and **linguistic** features. We measure the psycholinguistic features for subject-verb-object tuples and adjectives-noun pairs of the protagonist and antagonist persona sets separately. We calculate scores for other features of each post. Table II summarizes the categories.

1) Contextual Features: Content is essential in analyzing social media posts [19, 50]. We extract the content at two levels: term frequency-inverse document frequency (TF-IDF) weighted n-grams vectors (n = 1, 2) and post-level topics. TF-IDF weights combine term frequency $tf(t, d)$ (the occurrence of a term t in a document d) and inverse document frequency $idf(t, D)$ (the rating of t in a corpus D). It reflects how important a word is to a document in a corpus.

We extracted topics using Latent Dirichlet Allocation (LDA) [4]. LDA is a generative statistical model that assumes that each document (here, post) contains a distribution of topics; each topic is a distribution of words. We train a model on the *text* of posts in FAITA, exploring the number of topics ranging over 30–55 and finalized on 30, as it achieves the lowest perplexity. We then combine the topics that contain fewer than 200 posts. Table III shows a sample of eight hand-selected topics and ten example words belonging to each topic. In Table III, the "Topic Label" column is summarized manually by the authors according to all the words they are associated with; the percentage indicates how frequently each topic occurs in the dataset. We also show the ten most representative words for each topic. These topics show that posts in FAITA range from family to work issues. Additional topics with at least 100 posts include: driving safety (2.8%), gender communication differences (2.8%), games (2.6%), cooking

[7]https://huggingface.co

TABLE II
FEATURE CATEGORIES AND EXPLANATIONS.

Category	Feature	Explanation
Contextual	Topic	Lexicon-based topics measured by LDA [4]; each post has a list of topic it belongs to
Contextual	Content	TF-IDF weighted n-grams (n=1,2)
Psycholinguistic	Agent versus Patient	Ratio of author and others being an *agent* or a *patient* [43]
Psycholinguistic	Connotation Frames	Scores of connotation frames-related [42] words normalized by count of subject-verb-object tuples; the scores are calculated separately as writers' perspective, value, effect, mental state
Psycholinguistic	Agency and Power	Agency and power scores [39] normalized by count of subject-verb-object tuples
Psycholinguistic	Moral Content	Occurrences of the five virtue-vice paired Moral Foundation Theory lexicon [21] normalized by count of subject-verb-object tuples and count of adjectives-noun pairs
Psycholinguistic	Valence, Arousal, Dominance (VAD)	Occurrences of VAD lexicon [34] normalized by count of subject-verb-object tuples and count of adjectives-noun pairs
Psycholinguistic	Emotion	Occurrences of Emotion lexicon [35] normalized by count of subject-verb-object tuples or count of adjectives-noun pairs.
Linguistic	Subjectivity	Occurrences of subjectivity-related words [49] normalized by count of words
Linguistic	Hedge	Occurrences of hedge words [23] normalized by count of words
Linguistic	Modal	Occurrences of modal words normalized by count of words
Linguistic	Pronoun	Occurrences of first, second, and third pronouns
Linguistic	Sentiment	Averaged VADER [22] compound scores; nominal sentiment categories

TABLE III
SAMPLE TOPICS WITH REPRESENTATIVE WORDS.

Topic Label	Top Weighted Words
Relationship with family (20.8%)	life, relationship, mother, ex, child, father, life, wife, partner, son
Intimate relationship (17.3%)	girlfriend, boyfriend, relationship, dating, upset, feel, pretty, lot, love, guy
Living in shared accommodation (16.5%)	apartment, rent, live, room, living, house, lease, stay, bedroom
Money (7.3%)	pay, rent, saving, buy, job, account, car, loan, afford, cost
Pregnancy concerns in pets (5.5%)	dog, child, husband, child, pregnant, puppy, cat, law, animal, birth
Work (4.4%)	hour, work, boss, company, manager, job, employee, office, shift, week
Appearance judgment (4.2%)	hair, look, wear, white, black, comment, clothes, dress, looked, pretty
Neighborhood (3.3%)	neighbor, phone, email, post, account, people, use, street, yard, facebook

(2.7%), holiday gifts (2.7%), social media (2.3%), wedding plan (2.1%), medical treatment (1.6%), and school (1.1%).

2) Psycholinguistic Features: These refer to the lexico-semantic analysis of the cognitive association that a word carries and its literal meaning. The scores being introduced are separated into *agent*, *connotation frames*, *power and agency*, *moral content*, and *VAD*. From SVO tuples, we calculate scores for entities as subjects and objects. From ANP, we calculate scores for entities based on their adjective modifier. We normalize the scores calculated for entities in the persona sets to capture the values of the protagonist and antagonists.

a) Agent: The ratio of the time the protagonist and antagonists are assigned as *agents* or *patients* in a post using SRL labels.

b) Connotation Frames: A formalism for analyzing subjective roles and relationships implied by a given predicate [39]. To analyze nuanced dimensions of narratives in FAITA, we draw from a lexicon with annotations for 1,000 most frequently used English verbs across various dimensions, ranging from –1 to 1. A verb might elicit a positive sentiment for its subject but imply a negative sentiment for its object. For example, from "*Alice* betrayed *Bob*," the annotation contains the following dimensions:

- Writer's perspective. The writer (protagonist) elicits a negative perspective toward *Alice* as –0.67 (e.g., blaming) and a positive perspective toward *Bob* as 0.26 (e.g., supportive).
- Reader's perspective. (1) Values: the reader presupposes a positive value of *Bob* as 0.87 (strongly positive) and *Alice* as 0.47 (neutral to positive). (2) Effects: the reader presupposes the harms towards *Bob* as –0.93 (strongly negative) compared to *Alice* as 0.067 (neutral). (3) Mental states: the reader presupposes *Bob* is most likely to feel negative (–0.67) as a result of the event, but *Alice* it not likely to be affected (–0.03).

c) Power and Agency: A pragmatic formalism organized using frame semantic representations [42] to model how different levels of power and agency are implicitly projected on people through their actions. We use Sap et al.'s [2017] extension lexicon of Connotation Frames to measure the agency and power scores of *author* and *others*. This extension lexicon contains more than 2,000 transitive and intransitive verbs to model how different levels of power and agency are implicitly projected on the entities through their behaviors. Entities with high agency (subjects of *attack*) are active decision-makers, whereas entities with low agency (subjects of *doubts* and *needs*) are passive. This lexicon contains binary labels of each verb, which are positive (1), equal (0), and negative (–1).

d) Moral Content: The Moral Foundation Theory [20] has been widely adopted in the computational social community, which is critical in understanding how the psychological influence of social content unfolds, such as quantifying moral behaviors in Twitter [25] and taxonomizing the structure of moral discussions in Reddit [37]. We adopt the extended Moral Foundations Dictionary (eMFD) [21], which is a crowdsourced dictionary-based tool for extracting moral content from textual corpora. The eMFD contains 2,041 unique words, which are categorized into five broad domains based on MFT: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. Each word in the dictionary has a composite valence score ranging from –1 to 1.

e) VAD (Valence, Arousal, Dominance): The three affective dimensions are used to measure affective meanings from words that convey the author's attitudes toward the events and people referenced. We obtain the valence scores for 20,000 words from the NRC VAD lexicon [34], which contains real-valued scores ranging from 0 to 1 for each category.

f) Emotions: Emotions conveyed in words represent sentiment from the authors toward the described entities [35], which may place a considerable cognitive load on the audience [11]. The NRC Emotion lexicon [35] provides the emotion of around 20,000 words, indicating whether a word is associated with an emotion category. The categories are joy, sadness, anger, fear, trust, disgust, surprise, and anticipation.

3) Linguistic Features: We estimate linguistic scores for *subjectivity*, *hedge*, *sentiment*, and *modal* in each post.

a) Subjectivity: arises when people express personal feelings or beliefs, e.g., in opinions or allegations [49], which comprises the authors' perspectives towards the descriptive situations, contributing to the audience's judgments. We compute the subjectivity of a post as the average score of words based on the Subjectivity lexicon [49] (nonneutral words of "weaksubj" = 0.5 and "strongsubj" = 1). Additionally, we count the numbers of first-person, second-person, and third-person pronouns because words such as "you" and "we" engage the audience with the discourse.

b) Hedge: is associated with indirection in politeness theory [7], which may affect the audience's judgments.

c) Sentiment: indicates emotions by conveying the polarity of an opinion. A negative tone may imply more immorality than a neutral tone. We calculate each post's compound sentiment scores and sentiment categories with the VADER package [22].

d) Modal: words affect the sentiment of the words they modify [27].

III. $RQ_{FEATURE}$: Blame Assignment Analysis

To answer $RQ_{Feature}$, we perform two statistical analyses: (1) prediction—can description frames predict blame assignment? (2) language characteristics analysis—can linguistic features affect blame assignment? Here, we focus on using the prediction as a tool for analyzing, not for the purpose of making an accurate prediction. Note that we do not take gender and age as features when conducting experiments as they are not available in some of the posts in FAITA.

A. Predicting Blame Assignment

Now we examine how well computational models can predict blame assignments in moral situations. For machine learning models, we explore two logistic regression models (LR) to compute the probability of a positive label for each sentence. All the models are built using scikit-learn[8] toolkit in Python. An LR classifier computes the probability of a discrete outcome given an input variable. BERT-LR is logistic regression where our features are replaced with the BERT [10] embeddings of input sentences. We evaluate the performance of different models in terms of recall, precision, and F1 scores. All computation models were run 10 times and we measure the standard deviation of the scores for each method. For LR, we set the class weights to "balanced" to account for the label imbalance. And we explore feature selection using the L1-norm and regularization using L2-norm. Other hyperparameters for LR include setting the weight ranging over $(1\mathrm{e}-4\,,\ 1\mathrm{e}-3\,,\ 1\mathrm{e}-2\,,\ 1\mathrm{e}-1\,)$. We propose two baseline models. Random predicts the verdict randomly. Length predicts using the lengths of the sentences in a post, which has been shown to be effective in predicting blame [50].

The quantitative results of our methods are shown in Table IV. BERT-LR and LR outperform the baselines significantly, while BERT-LR performs best. It is worth noting that our features, despite the lower performance than BERT-LR, are clearly informative of morality prediction because they directly capture the information contributing to the audience's decision on blame. Transformer models such as BERT encode linguistic characteristics in a more sophisticated manner and may include additional information. But it is less clear exactly what transformers capture and whether they capture irrelevant statistics. To examine the contribution of each feature category, we conducted ablation tests based on the LR model. Regarding F1 scores, psycholinguistic features have the highest contribution, followed by contextual and linguistic. This result reaffirms the importance of analyzing the lexical semantics of attributive and predicative words in first-person moral narratives.

TABLE IV
PREDICTION ACCURACY (MACRO-AVERAGE SCORES). ALL SCORES HAVE STANDARD DEVIATIONS BETWEEN 0.01 AND 0.03. THE BEST SCORES ARE IN BOLD. WE ONLY REPORT LR AND BERT-LR RESULTS BECAUSE THEY YIELD THE PERFORMANCES OF OTHER MODELS SUCH AS MULTILAYER PERCEPTRON, SVM, AND RANDOM FOREST.

Method	F1		Recall		Precision	
	DEV	TEST	DEV	TEST	DEV	TEST
Random	0.49	0.50	0.50	0.50	0.50	0.49
Length	0.39	0.38	0.50	0.50	0.49	0.52
LR	0.66	0.65	0.65	0.64	**0.66**	**0.65**
(×) Linguistic	0.63	0.62	0.60	0.61	0.63	0.62
(×) Contextual	0.53	0.53	0.54	0.54	0.60	0.60
(×) Psycholinguistic	0.48	0.49	0.52	0.53	0.59	0.59
BERT	**0.71**	**0.72**	**0.68**	**0.69**	**0.66**	**0.65**

B. Interpreting Characteristics

We measure the effect size and statistical significance of each feature. The effect of each feature is conditioned on the domain of each post using logistic regression. For interpretation purposes, we use the Odds Ratio (OR) (the exponent of the effect size). Odds represent the ratio of the probability of an author being blamed to the probability of not being blamed; OR is the ratio of odds when the effect size increases by one unit. The OR is calculated using the equation $OR = exp(\beta_i), i \in N$, where β_i is the coefficient of attribute i obtained by the trained LR model and N denotes the attribute set. Moreover, we estimate statistical significance by Spearman's Rank Correlation Coefficient to avoid assuming normality or other distributions for FAITA.

1) Contextual Features: We begin by looking at OR between topics and blame assignments. Table V shows the OR values and correlation coefficients for the authors being blamed corresponding to Table III. These results show that posts related to relationships, pregnancy concerns in pets, and games are positively correlated with blame assignment. Other topics mentioned in Section II-C1 that may decrease the probability of an author

[8]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

being blamed are school, holiday gifts, and cooking; the rest are positively correlated.

TABLE V
ODDS RATIO (OR) AND SPEARMAN'S CORRELATION COEFFICIENT OF TOPICS CALCULATED FROM THE TEST SET. AN EFFECT IS POSITIVE (BLUE) IF OR > 1 AND NEGATIVE (RED) IF OR < 1.

Topic	Moral Blame	P-value
Relationship with family	1.11	0.02
Intimate relationship	1.07	0.07
Living in shared accommodation	0.79	0.02
Money	0.82	0.20
Pregnancy concerns in pets	1.46	0.03
Work	0.98	0.03
Appearance judgment	1.16	0.07
Neighborhood	0.71	0.14

2) Psycholinguistic Features: Table VI shows the features that influence at least a 1% probability of the author being blamed. Table VI reveals that psycholinguistic features are informative. The results for agent and patient are consistent with social psychology that "being a victim can help escape blame" [17]. These features do not store lexical information but affect the audience's judgments in the cognitive aspect. In addition, our analysis indicates the protagonist can reduce blame by eliciting positive perspectives (e.g., supportive) towards themselves. Additionally, authors can reduce blame when they describe themselves as suffering more from harm than the antagonist. The Agency and Power features are consistent with the above findings because the high values imply the agent's high-level authority and powerful capability, which can trigger blame.

Care and harm are opposite concepts in Moral Foundation Theory, whereas increasing the use of words from the lexicon reduces the probability of the author being blamed. Moreover, we find that VAD features do not have significant p-values. However, they increase the probability of the author being blamed by 23% when increasing the use of the dominance lexicon when describing the protagonist. Different emotion categories have different effects on blame assignment. Specifically, using sadness-related words when describing the protagonist lowers the probability of the authors being blamed to one-third. Additionally, increasing the use of disgust-related words when describing the antagonist more than doubles the probability of the author being blamed. We highlight a possible explanation: the description frames of the protagonist and antagonist need to be captured as a whole, not as individual components.

3) Linguistic Features: As shown in Table VII, subjectivity is positively correlated to blame assignment in contrast to hedging, which indicates that subjective descriptions increases the possibility of the author being blamed with greater certainty. The frequent use of third-person pronouns triggers blame because the audience may think that the author is trying to escape from blame by avoiding describing themselves. Although second-person pronouns have a small p-value, they increase the probability only by 1% of the author being blamed. However, the negative sentiment category strongly affects blame assignment with an OR of 3.18, which may explain that extreme sentiment triggers blame assignment.

IV. RQ$_{\text{SOCIAL}}$: SOCIAL FACTORS ANALYSIS

In this section, we examine gender and age features in FAITA to investigate whether audiences exhibit differences in their assessments of moral situations.

TABLE VI
ODDS RATIO (OR) AND SPEARMAN'S CORRELATION COEFFICIENT OF PSYCHOLINGUISTIC FEATURES CALCULATED FROM THE TEST SET. WP REPRESENTS *writer's perspective*. AN EFFECT IS POSITIVE (BLUE) IF OR > 1 AND NEGATIVE (RED) IF OR < 1.

Feature	Protagonist		Antagonist	
	Moral Blame OR	p-value	Moral Blame OR	p-value
Agent	1.93	0.05	0.93	0.002
Patient	0.53	0.03	1.01	0.001
WP	1.01	0.006	0.81	0.031
Value	0.99	0.13	1.02	0.13
Power	1.04	0.006	0.97	0.003
Agency	2.00	0.003	0.96	0.002
Care	0.99	0.03	1.03	0.03
Harm	0.97	0.08	1.00	0.07
Betrayal	0.95	0.06	1.11	0.16
Loyalty	0.97	0.08	1.03	0.17
Valence	0.99	0.14	1.22	0.13
Arousal	1.04	0.11	1.21	0.15
Dominance	1.23	0.14	1.09	0.15
Joy	0.98	0.09	0.98	0.13
Sadness	0.31	0.05	1.28	0.005
Anger	1.05	0.01	0.11	0.03
Fear	2.33	0.01	1.06	0.03
Trust	1.10	0.08	0.20	0.09
Disgust	1.06	0.07	2.16	0.02
Anticipation	1.34	0.05	1.74	0.04

TABLE VII
ODDS RATIO (OR) AND SPEARMAN'S CORRELATION COEFFICIENT OF LINGUISTIC FEATURES CALCULATED FROM THE TEST SET. AN EFFECT IS POSITIVE (BLUE) IF OR > 1 AND NEGATIVE (RED) IF OR < 1.

Feature	Moral Blame	p-value
Subjectivity	1.09	0.006
Hedge	0.66	0.04
First pronoun	1.45	0.10
Second pronoun	1.01	0.0009
Third pronoun	1.96	0.003
Sentiment score	1.78	0.001
Sentiment: positive	1.18	0.005
Sentiment: neutral	0.99	0.08
Sentiment: negative	3.18	0.01

A. Analyzing Gender and Age Association

This section investigates whether the authors' self-reported gender and age lead to an imbalance in blame assignment. Using the method of Section II-C, 13,935 posts describe the genders of all entities involved, and 6,079 posts state the authors' age is between 15 and 65. To determine the association between blame and social factors, we perform the χ^2 significance test and compute Cramer's ϕ as the effect size. Here, 0.07–0.21, 0.21–0.35, and >0.35 respectively indicate small, moderate, and strong association [8].

We aggregate occurrences of entities being blamed when they are protagonists and antagonists. The overall χ^2 test result between genders and blame assignment is ($\chi^2(13,935) = 515.02, p < 0.001$) with $\phi = 0.17$. Whereas the effect size indicates a small association between gender and blame assignment in FAITA, the evidence indicates there is an association between the two ($p < 0.001$). Besides, we observe that males are 53% (the

log-odds-ratio of occurrences when authors of different genders receive blame) more likely to receive blame. The results allow us to discern the direction of the biases due to gender: male authors are more likely to be considered agentive no matter their position. The observation coheres with previous psychological research that some sets of biases stereotype females into the role of suffering *patient* on social media [41].

To further investigate the correlation between blame assignment and age, we divide authors' ages (antagonists' ages are scarce) into four groups ranging from 15 to 55 as the range accounts for almost 80% of active Reddit users. Table VIII illustrates blame assignment is associated with protagonists' ages when they are in the 15–45 age group ($p < 0.05$), especially when authors are in the 36–45 age group ($\phi = 0.18$).

TABLE VIII
THE COLUMNS ARE AGE RANGES. N REPRESENTS THE NUMBER OF CORRESPONDING POSTS. $p < 0.05$ INDICATES THE AGE GROUP AND BLAME ASSIGNMENT ARE ASSOCIATED.(** : $p < 0.05$, ***: $p < 0.001$.)

	Age Ranges			
Metrics	15-25	26-35	36-45	46-55
N	3,554	1,951	410	136
χ^2	76.56 (***)	50.89 (***)	13.46 (**)	2.96 ()
ϕ	0.15	0.16	0.18	0.15

B. Considering Semantics with Social Factors

We now examine how blame assignment differs between female and male protagonists in similar situations. We employ pretrained sentence-BERT models [40] to cluster the 13,935 posts based on semantic similarity. We learn embeddings of the posts' *titles* as they serve as summaries of posts. To remove the effect of gender-related tokens, we replace gender-identified words with "someone" using the resources mentioned in Section II-B5.

We adopt Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [33] because no external references identify topic numbers in FAITA. Then we perform dimension reduction with Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [32] to alleviate the problem of sparse embeddings. We fine-tune parameters for HDBSCAN and UMAP models [1] by increasing the model's Density-Based Clustering Validation (DBCV) score [36]. To enhance the semantic similarity in each cluster, we exclude the clusters containing fewer than 50 posts. This process yields 7,248 posts that clustered into 47 groups, where the counts of posts in a cluster range from 51 to 712. We measure χ^2 and ϕ to select the clusters where gender is strongly associated with blame assignment ($p < 0.001$ and $\phi > 0.35$) and find six clusters include 1,162 posts.

To categorize the semantics of the clusters, we use the UCREL Semantic Analysis System (USAS) [47], a framework for automatic semantic analysis and tagging of text, which is based on McArthur's Longman Lexicon of Contemporary English [45]. USAS has a multitier structure with 21 major discourse fields subdivided in fine-grained categories such as *People*, *Relationships*, and *Power*. Using USAS, we label each cluster with the most frequent tag (or tags) among the highest TF-IDF-scored nouns from the posts. We notice that the topics of the most gender-associated situations in FAITA corroborate previous work on categorizing language biases in Reddit [13]. For example, the most frequent gender-polarized situations on FAITA (ordered by frequency) are *kin*, *relationship: intimate/sexual*, *groups and affiliation*, *anatomy and physiology*, *work and employment*, *sports*, *games*, *money*, *medicines and medical treatment*, and *judgment of appearance*. These tags account for the most discussed topics, as Table III shows. It is important to note that our analysis of social factors in blame assignment is more suggestive than conclusive. Our analysis suggests that social biases exist in social media posts, which influences blame assignment, at least in some topics.

V. DISCUSSIONS

This paper contributes to studying the nature of morality by assessing social psychology insights on descriptive real-life situations. We incorporate a novel set of language features with machine learning models for predicting who is considered blameworthy. Statistical methods help visualize the effects of the features. The effective prediction performance confirms the linkage between blame assignments and psychological observations on social media. For example, entities described using less *care*-related words (a category from Moral Foundation Theory [20]) are more likely to receive blame. Furthermore, our findings suggest that gender and age are associated with blame assignment. For example, males are more likely to receive blame than females; biases in blame assignments are more likely to be presented when protagonists are in the 15–45 age group.

Our results can be explained by the fact that people perceiving themselves as deserving blame are subject to feelings of guilt [44]. In our case, these feelings conveyed from social media posts may affect the audience's decision making on who is blameworthy. In addition, psychological literature observes that social media might have typecasting towards male and female individuals [13, 41]. However, people of different genders might be subject to different social pressures and thus be different in choosing the language to describe conflict [2]. Our results are coherent with these observations, which reaffirms the significance of considering social psychology instruments in using computational methods to understand the nature of morality [15].

A. Implications

Our work contributes a new framework to demonstrate the significance of psychological theories in real-life situations, with theoretical and empirical guidelines to assist in studying morality. First, our research provides novel language features based on social psychology that enable textual and psychological insights into the nature of morality. Second, our proposed features can be applied to build interpretable models for blame assignment. Practically, this work could motivate the design of future AI systems to incorporate psychological findings to promote the interpretability of real-world practical morality.

In addition, our study contributes to theoretical research such as The Theory of Dyadic Morality (TDM) [43] by providing language features. For example, our analysis answers how individuals' agentiveness is affected by the associated descriptions. Our work can help design laboratory experiments. For example, since we demonstrate that language used to describe social situations may affect a participant's cognition, special attention could be paid when stylizing such situations. Particularly, these designs can be tailored to subject-matter experts for studying advanced components in theoretical social research that demand human validations.

B. Limitations and Future Work

As in any study dealing with social media data, there are some limitations. First, this study design has the advantage of a higher ecologic validity but presents critical causal inference challenges. There might be hidden confounders that we cannot measure given the lack of data, such as the demographics of the audience. In addition, the research may only be generalizable to some populations, which may not account for other factors influencing blame assignment on social media, such as age, culture, and education. Therefore, the coefficients we find cannot be interpreted as a direct causal effect, which means our research is more suggestive than conclusive.

This work could be extended in interesting ways. One direction is to incorporate language features from the comments accompanying each post to extract cognitive-affective features directly from the audience. The accompanying comments may help explain *what*, *why*, and *how* language features affect the audience's cognitive processes. In addition, to provide a causal explanation of how social factors appear in blame assignments and how they function, future work can leverage explicit and implicit social factors in the narratives.

Acknowledgments

Thanks to the reviewers for their helpful comments. Thanks to the NSF (IIS-2116751) for partial support.

References

[1] D. Angelov, "Top2vec: Distributed representations of topics," *arXiv preprint arXiv:2008.09470*, 2020.

[2] M. Asher, A. Asnaani, and I. M. Aderka, "Gender differences in social anxiety disorder: A review," *Clinical psychology review*, vol. 56, pp. 1–12, 2017.

[3] J. Beel, T. Xiang, S. Soni, and D. Yang, "Linguistic characterization of divisive topics online: Case studies on contentiousness in abortion, climate change, and gun control," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 16, 2022, pp. 32–42.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning research*, vol. 3, pp. 993–1022, 2003.

[5] N. Botzer, S. Gu, and T. Weninger, "Analysis of moral judgment on Reddit," *IEEE Transactions on Computational Social Systems*, 2022.

[6] J. Bracht and A. Zylbersztejn, "Moral judgments, gender, and antisocial preferences: An experimental study," *Theory and Decision*, vol. 85, no. 3, pp. 389–406, 2018.

[7] S. E. Brennan and J. O. Ohaeri, "Why do electronic conversations seem less polite? The costs and benefits of hedging," in *Proceedings of the International Joint Conference on Work Activities Coordination and Collaboration*. New York: Association for Computing Machinery, 1999, p. 227–235.

[8] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum Associates, 1988.

[9] S. De Candia, G. De Francisci Morales, C. Monti, and F. Bonchi, "Social norms on Reddit: A demographic analysis," in *14th ACM Web Science Conference 2022*, ser. WebSci. NY: Association for Computing Machinery, 2022, pp. 139–147.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[11] K. Dijkstra, R. A. Zwaan, A. C. Graesser, and J. P. Magliano, "Character and reader emotions in literary texts," *Poetics*, vol. 23, no. 1-2, pp. 139–157, 1995.

[12] D. Emelin, R. Le Bras, J. D. Hwang, M. Forbes, and Y. Choi, "Moral stories: Situated reasoning about norms, intents, actions, and their consequences," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online and Punta Cana, Nov. 2021, pp. 698–718.

[13] X. Ferrer, T. van Nuenen, J. M. Such, and N. Criado, "Discovering and categorising language biases in Reddit," *arXiv preprint arXiv:2008.02754*, 2020.

[14] M. Forbes, J. D. Hwang, V. Shwartz, M. Sap, and Y. Choi, "Social chemistry 101: Learning to reason about social and moral norms," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 653–670.

[15] K. C. Fraser, S. Kiritchenko, and E. Balkir, "Does moral code have a moral code? Probing Delphi's moral philosophy," *arXiv preprint arXiv:2205.12771*, 2022.

[16] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "AllenNLP: A deep semantic Natural Language Processing platform," in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne: Association for Computational Linguistics, Jul. 2018, pp. 1–6.

[17] K. Gray and D. M. Wegner, "To escape blame, don't be a hero—be a victim," *Journal of Experimental Social Psychology*, vol. 47, no. 2, pp. 516–519, 2011.

[18] S. Guglielmo and B. F. Malle, "Asymmetric morality: Blame is more differentiated and more extreme than praise," *PloS one*, vol. 14, no. 3, p. e0213544, 2019.

[19] Z. Guo, Z. Zhang, and M. P. Singh, "In opinion holders' shoes: Modeling cumulative influence for view change in online argumentation," in *Proceedings of the 29th Web Conference (WWW)*. Taipei: ACM, Apr. 2020, pp. 2388–2399.

[20] J. Haidt and J. Graham, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," in *Social Justice Research*. Springer, Mar. 2007, vol. 20, pp. 98–116.

[21] F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber, "The Extended moral foundations dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text," *Behavior Research Methods*, vol. 53, pp. 232–246, 2020.

[22] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Ann Arbor, 2014, pp. 216–225.

[23] K. Hyland, *Metadiscourse: Exploring Interaction in Writing*. London, UK: Bloomsbury Publishing, 2018.

[24] L. Jiang, J. D. Hwang, C. Bhagavatula, R. L. Bras, M. Forbes, J. Borchardt, J. Liang, O. Etzioni, M. Sap, and Y. Choi, "Delphi: Towards machine ethics and norms," *arXiv preprint arXiv:2110.07574*, 2021.

[25] H. Joe, P. W. Gwenyth, Y. Leigh, H. Shreya, M. D. Aida, L. Ying, K. Brendan, A. Mohammad, K. Zahra, M. Madelyn, M. Gabriela, P. Christina, E. C. Tingyee, C. Jenna, L. Christian, L. Jun Yen, M. Arineh, and D. Morteza, "Moral foundations Twitter corpus: A collection of 35k tweets annotated for moral sentiment," *Social Psychological and Personality Science*, vol. 11, no. 8, pp. 1057–1071, 2020.

[26] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*. Upper Saddle River: Prentice-Hall, 2009.

[27] S. Kiritchenko and S. Mohammad, "The effect of negators, modals, and degree adverbs on sentiment composition," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. San Diego: Association for Computational Linguistics, Jun. 2016, pp. 43–52.

[28] Y. Liu, A. Moore, J. Webb, and S. Vallor, "Artificial moral advisors: A new perspective from moral psychology," in *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '22. New York: Association for Computing Machinery, 2022, pp. 436–445.

[29] N. Lourie, R. L. Bras, and Y. Choi, "Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes," *arXiv preprint arXiv: 2008.09094*, 2020.

[30] L. Lucy, D. Demszky, P. Bromley, and D. Jurafsky, "Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas U.S. history textbooks," *American Educational Research Association*, vol. 6, no. 3, p. 2332858420940312, 2020.

[31] B. Malle, S. Guglielmo, and A. Monroe, "A theory of blame," *Psychological Inquiry*, vol. 25, pp. 147–186, 04 2014.

[32] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv preprint arXiv:1802.03426*, Feb. 2018.

[33] L. McInnes and J. Healy, "Accelerated hierarchical density based clustering," in *IEEE International Conference on Data Mining Workshops (ICDMW)*, 2017, pp. 33–42.

[34] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Jul. 2018, pp. 174–184.

[35] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[36] D. Moulavi, P. A. Jaskowiak, R. J.G.B.Campello, A. Zimek, and J. Sander, "Density-based clustering validation," in *Proceedings of SIAM International Conference on Data Mining (SDM)*, 2014, pp. 839–847.

[37] T. D. Nguyen, G. Lyall, A. Tran, M. Shin, N. G. Carroll, C. Klein, and L. Xie, "Mapping topics in 100,000 real-life moral dilemmas," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 16, 2022, pp. 699–710.

[38] S. Poria, A. Gelbukh, E. Cambria, P. Yang, A. Hussain, and T. Durrani, "Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis," in *IEEE 11th International Conference on Signal Processing*, vol. 2, 2012, pp. 1251–1255.

[39] H. Rashkin, S. Singh, and Y. Choi, "Connotation frames: A data-driven investigation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1. Berlin: Association for Computational Linguistics, Aug. 2016, pp. 311–321.

[40] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Hong Kong: Association for Computational Linguistics, 11 2019, pp. 3982–3992.

[41] T. Reynolds, C. Howard, H. Sjåstad, L. Zhu, T. G. Okimoto, R. F. Baumeister, K. Aquino, and J. Kim, "Man up and take it: Gender bias in moral typecasting," *Organizational Behavior and Human Decision Processes*, vol. 161, pp. 120–141, 2020.

[42] M. Sap, M. C. Prasettio, A. Holtzman, H. Rashkin, and Y. Choi, "Connotation frames of power and agency in modern films," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen: Association for Computational Linguistics, Sep. 2017, pp. 2329–2334.

[43] C. Schein and K. Gray, "The theory of dyadic morality: Reinventing moral judgment by redefining harm," *Personality and Social Psychology Review*, vol. 22, no. 1, pp. 32–70, 2018.

[44] J. F. Scott, *Internalization of norms: A sociological theory of moral commitment*. Prentice-Hall, 1971.

[45] D. Summers and A. Gadsby, *Longman Dictionary of Contemporary English: The Complete Guide to Written and Spoken English*. Longman Group Limited, 1995.

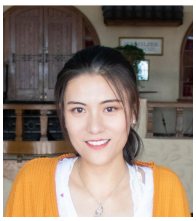
[46] Z. Talat, H. Blix, J. Valvoda, M. I. Ganesh, R. Cotterell, and A. Williams, "A word on machine ethics: A response to Jiang et al. (2021)," *arXiv preprint arXiv:2111.04158*, 2021.

[47] http://ucrel.lancs.ac.uk/usas/, 2002, online; accessed 30 November 2021.

[48] G. R. Wark and D. L. Krebs, "Gender and dilemma differences in real-life moral judgment." *Developmental Psychology*, vol. 32, no. 2, p. 220, 1996.

[49] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Vancouver: Association for Computational Linguistics, Oct. 2005, pp. 347–354.

[50] K. Zhou, A. Smith, and L. Lee, "Assessing cognitive linguistic influences in the assignment of blame," in *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media (SocialNLP)*. Online: Association for Computational Linguistics, Jun. 2021, pp. 61–69.



Ruijie Xi received a B.S. in software engineering from Harbin Institute of Technology University, Harbin, China, in 2015, and an M.S. in computer science from the University of Delaware, Newark, DE, USA, in 2017. She is currently pursuing a Ph.D. at the Department of Computer Science, North Carolina State University. Her research interests focus on applying Artificial Intelligence approaches for social good, with current interests on understanding morality in social media and improving prosocial behaviors in the real world. Contact her at rxi@ncsu.edu.

Munindar P. Singh is an Alumni Distinguished Graduate Professor in Computer Science at NC State University. His interests include sociotechnical systems and the uses of AI in society. Singh received a Ph.D. in computer sciences from the University of Texas at Austin. He is a former editor-in-chief of *IEEE Internet Computing* and the *ACM Transactions on Internet Technology*. He is a Fellow of AAAI, AAAS, ACM, and IEEE, and a member of Academia Europaea. His other recognitions include the ACM SIGAI Autonomous Agents Research Award and the IEEE TCSVC Research Innovation Award. Contact him at singh@ncsu.edu.