Loaded Language and Conspiracy Theorizing

Emily Klein (emily.klein001@gmail.com)
Department of Cognitive Science, Rensselaer Polytechnic Institute
Troy, NY 12180 USA

James Hendler (hendler@cs.rpi.edu)
Department of Computer Science, Rensselaer Polytechnic Institute
Troy, NY 12180 USA

Abstract

Loaded language is an umbrella term for words, phrases, and overall rhetorical strategies that have strong emotional implications and intent to sway others. Belief in conspiracy theories is tied to a range of strong emotions (van Prooijen and Douglas, 2018). Accordingly, language with strong emotional and persuasive content may be expressed by people experiencing the strong emotions associated with conspiracy theorizing. In this research, we examine multiple types of loaded language in two online parenting forums: one historically against vaccination, and another historically accepting of vaccination. It is well-established that conspiracy theories are the most influential contributor to anti-vaccination views (Hornsey et al., 2018) and anti-vaccination beliefs are strongly correlated with belief in unrelated conspiracy theories (Goldberg & Richey, 2020). Results indicate that users of an anti-vaccination forum use a greater frequency of loaded language to express themselves than users of a vaccination-neutral forum.
.

Keywords: conspiracy theories, language, pragmatics, social media analysis

Introduction

Conspiracy theorizing is not particularly new, unnatural, or uncommon, but it can be deeply harmful to individuals and communities (Douglas et al., 2017; van Prooijen & Douglas, 2017). Endorsement of conspiracy theories is associated with lowered intention to engage in politics (Jolley & Douglas, 2014a), resistance to follow medical advice (Jolley and Douglas, 2014b), a tendency to reject important scientific findings (Lewandowsky et al., 2013), increased intention to engage in everyday crime (Jolley et al., 2019), and increased violent extremist intentions (Rottweiler & Gill, 2020).

As a consequence of the rise of social media and online communication platforms, people share conspiracy theories faster and farther than ever before, predominantly via written text (Uscinski et al, 2018). Are there linguistic commonalities that appear consistently in conspiracy theorizing language? There is extensive literature on the psychology of conspiracy theorists, but there is little directed research on how the psychology of conspiracy theorists may relate to the language they use, especially at the level of pragmatic linguistic analysis of online forums.

On forums and social media networks, detection remains a formidable challenge, even with the help of automated AI content moderation. Large and small platforms struggle to reduce the spread of conspiracy theory content, despite increased investment in automated content moderation and fact-checking (Scott & Kern, 2022; Fetters Maloy & Oremus, 2021). Part of the challenge is that conspiracy theorizing language does not always contain detectable keywords. Consider how the QAnon movement attempted to circumvent increased content moderation efforts from major social media platforms by dropping Q-related labels (Collins, 2020), and how anti-vaccine activists on the pregnancy app What to Expect learned to understand and evade the app's keyword detection tools (Fetters Maloy & Oremus, 2021). Further, conspiracy theories are not exclusively shared by conventional users. While many believe conspiracy theorists are typically alt-right fringe forum users, research shows that the people who share conspiracy theory content online are not constrained to fringe forums, and they are not exclusively from the far-right (Morris, 2021). Rather, conspiracy theorists include a broad swath of people who resonate with anti-establishment rhetoric or the social issues alleged by conspiracy theories like child sex trafficking. Lifestyle bloggers on Instagram who would typically post exclusively about fashion, beauty, and parenting were lured into the QAnon conspiracy theory by concerns about child sex trafficking (Tiffany, 2020). This broader set of traits describing conspiracy theorists suggests that user-based profiling strategies for conspiracy theory detection may fall short, as conspiracy theories are not shared only by conventional propagators. Lastly, conspiracy theories are international in reach, making detection a cross-linguistic problem (Bruns et al., 2020).

Further complicating the detection problem is the need for algorithms to identify conspiracy theorizing language accurately and interpretably. The importance of accuracy and interpretability in this domain is paramount: people are unlikely to accept automated content moderation technology and its decisions without explanation of its choices. As a result of changing community standards and proprietary algorithmic content moderation systems, social media users develop folk theories about how and why content is flagged or accounts are suspended, attributing fault to an unidentified "they," other users, or bias on the part of the social media company. (Myers West, 2018; Vaccaro et al., 2020). Transparency and interpretability are the underpinnings of trust; a lack of trust (in governments, institutions, organizations, and others) is a distinguishing feature of

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

conspiracy theories (Douglas & Leite, 2016; Douglas et al., 2017, van Prooijen et al, 2021).

Unfortunately, the current generation of AI systems falls short on the ability to explain decisions and behavior in human terms (McShane & Nirenburg, 2021). The machine learning and neural network models of current content moderation systems are often considered black-box solutions. One approach to overcome this problem is to use interpretable variables *such as loaded language* as features. These variables are manually engineered features that are predictive of the outcome of interest. While manually engineering features is a labor-intensive task, combining certain hand-picked features with automatically learned features may make for more accurate, and more understandable AI systems for content moderation.

Although the pursuit of interpretable AI is a longer-term one than the scope of this paper, scholarship in linguistics can be a boon to advancing that program of research. Much of non-computational linguistics research is not directly applicable to research in artificial intelligence, other than the annotation of corpora in service of black-box machine learning. There exists a yet unrealized opportunity for non-computational linguistics researchers to collaborate with computational linguistics researchers to develop linguistic knowledge for application in AI in the near-term.

Setting aside AI, this research contributes to linguistic and cognitive science scholarship, helping us understand how conspiracy theorizing manifests in language and what language reveals about conspiracy theories and the people who believe them.

Loaded language in the present study

Each type of loaded language investigated in the present study is motivated by literature in the psychological and cognitive sciences. The two types explored in the present study are (1) Thought-terminating clichés and (2) Euphemistic and dysphemistic language.

Thought-terminating clichés Thought-terminating clichés, also known as semantic stop-signs, are a form of loaded language commonly used to quell cognitive dissonance (Lifton, 1989; Chiras, 1992). They de facto tell the interlocutor, "Let's not think about or discuss this further." Examples include *it is what it is*, *it's God's will*, *where we are*, *such is life*, *do the math*, etc**.** Thought-terminating clichés impede further critical thinking. Impeded critical thinking and faulty reasoning lead to *crippled epistemologies* – a factor in conspiracy theorizing (Sunstein & Vermeule, 2009). Crippled epistemologies, as Sunstein and Vermeule define them, arise not from irrationality or mental illness, but from critical thinking based on a limited number of relevant informational sources. Thus, people who endorse conspiracy theories are behaving rationally given the sparse or incorrect information available to them. Thought-terminating clichés, which implore others not to think about things further, may be a linguistic symptom experienced by someone who unknowingly reasons using flawed epistemologies.

Euphemistic and dysphemistic language Euphemistic language is language usage in which a neutral or inoffensive word/expression is substituted with one that is considered more pleasant or less derogatory, as in *pass away* for *die* or *between jobs* for *unemployed*. Euphemisms make the emotional impact of a word softer. In contrast, dysphemisms make the emotional impact of a word blunter, as in *worm food* for *dead*. Euphemism is a language tool and potentially "injurious weapon" (Bandura, 1999, p. 195) that can have serious ramifications when used for persuasion (Stein, 1998). Euphemistic language operates via ambiguity. Euphemisms and dysphemisms respectively diminish or exacerbate the emotional impact of a word by substituting words or phrases for alternatives with different connotations. The resulting ambiguity is a key aspect of the rhetorical strategies associated with conspiracy theorizing. Byford (2011) identifies several recurring rhetorical features of conspiracy theories that function by attempting to refute an official story. These features include obfuscation of the conspiracy theory's own flaws by means of excessive focus on the alleged problems of the official story, and diversion of attention by forcing conspiracy theory opponents to defend themselves instead of allowing them to attack conspiracy theory proponents' claims (p. 88-93). Creating confusion about both the official explanation and the conspiratorial explanation serves to muddy listeners' understanding of events. Euphemisms and dysphemisms may contribute to these strategies to create ambiguity**.**

Related Work

Conspiracy Theory Detection

While research on automatic fake news detection and hate speech detection dates to the mid-2010s, automatic conspiracy theory detection has only received recent attention by the natural language processing (NLP) community. Approaches thus far have largely focused on leveraging network features and user features for detection.

For example, Shahsavari et al. (2020) used machine-learning methods to extract narrative graphs about COVID-19 conspiracy theories in online posts and news articles, and network community detection algorithms to discover conspiracy communities. Tangherlini et al. (2020) pursued a similar network graph approach that compares the narrative frameworks of conspiracy theories to the narrative frameworks of actual conspiracies. Most recently, Giachanou et al. (2021) carried out a comparative analysis of pro- and anti-conspiracy theorizing posts that leverages user-based and psycholinguistic features. They used this data to create a Convolutional Neural Network (CNN) model that combines word embeddings and psycholinguistic characteristics to predict whether a user is a conspiracy propagator or not. Notably, this model leverages semantic information for detection. There is little research to date on how pragmatic

information may contribute to the automatic detection of conspiracy theories.

In linguistics, semantics is concerned with the literal, context-free meanings of words, while pragmatics is concerned with intended meaning in context. There is a paucity of pragmatics in AI/NLP research in general – not only in the domain of conspiracy theory detection. Semantics has been the focus of NLP research in recent years after the remarkable success of neural vector representations (word embeddings) including the Word2vec model (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), which can learn high quality vector representations of words. These models and their variations have achieved state-of-the-art progress on semantic similarity tasks. Although these tasks are a benchmark suited to the capabilities of the models, the field's progress on semantics cannot be dismissed. Pragmatics, on the other hand, has received little attention.

Loaded language functions as a persuasive technique, reflecting the speaker's/writer's purposeful choice of vocabulary intending to sway an audience or align with a stance. Loaded language has clear pragmatic function. This research puts pragmatics first in terms of exploring how language relates to the detection of conspiracy theories.

Vaccination Stance Detection

To narrow the domain for this research, we focused on vaccine-related conspiracy theories as a test case of conspiracy theorizing. Vaccine hesitancy – the reluctance or refusal to be vaccinated or to have one's children vaccinated against contagious diseases – has existed since the advent of formal vaccines in the late 1700s and is closely tied to conspiracy theories. Romer and Jamieson (2020) found that belief in COVID-19 conspiracy theories is negatively associated with the perceived safety of vaccination and intention to be vaccinated against COVID-19. Hornsey et al. (2018) showed conspiracy theories to be the most influential contributor to anti-vaccination views.

Researchers studying anti-vaccination stances on social networks are working on detecting a range of phenomena, including filter bubbles, misinformation, and conspiracy theories, often by leveraging techniques from NLP. Approaches to studying the language used by those who express anti-vaccination attitudes online include network-based approaches (Memon et al. 2020), syntactic and lexical approaches (Faasse et al., 2016; Mitra et al., 2016), narrative approaches (Tangherlini et al. 2016), and content analysis approaches (Hoffman et al. 2019; Hughes et al. 2021). As in conspiracy theory detection, there is little work that compares anti-vaccination stances to vaccination-neutral stances by leveraging pragmatic language features for detection.

Loaded Language Detection

The NLP literature on loaded language generally focuses on propaganda detection in news articles. Da San Martino et al. (2019b, 2020b) carried out analyses of propaganda techniques (including loaded language) in the news. Nakov et al. (2021) analyzed propaganda techniques (including loaded language) in tweets related to the COVID-19 vaccine, but their analysis is based on the Prta system (Da San Martino, 2020b) which is designed for detecting and highlighting the use of propaganda techniques in online news. The NLP4IF-2019 Shared Task on Fine-Grained Propaganda Detection was a classification competition based on a corpus of news articles annotated with propagandist techniques (2019a).

News articles have historically been the domain for research on loaded language detection. But people increasingly get the news from friends on web forums and social media feeds, *not* from news outlets. This research focuses on loaded language detection in user-generated posts. Examining user-generated posts is particularly important when looking at vaccine-related conspiracy theories because people increasingly seek health-related information online from peers (Kata, 2010; Chu et al., 2017). Vrdelja et al. (2018) found that mothers most often seek information about vaccines from friends or online, not from their pediatricians. The traditional presentation of facts regarding vaccination is often not enough to sway the perspective of vaccine-hesitant parents (Kaufman et al. 2018), so it is imperative to understand the beliefs and attitudes of parents who believe in vaccine-related conspiracy theories if we aim to design better public health campaigns, prevent outbreaks of vaccine-preventable diseases, and ensure forward progress on vaccine hesitancy. Loaded language offers a window not only into propaganda detection, but into vaccine-related conspiracy theory detection as well.

Methodology

We pursue a well-established research paradigm by Bessi et al. (2015) that compares and contrasts conspiratorial and science-based narratives on social media. This juxtaposition has been pursued in many previous studies on social media dynamics because conspiracy and science groups form highly segregated online communities that promote narratives with little to no overlap (Fong et al., 2021). In this case, we contrast an anti-vaccination forum with a vaccination-neutral parenting forum. We use the term *vaccination-neutral* (as opposed to *pro-vaccination*) because most parents with vaccinated children qualify as vaccination-neutral, passively accepting rather than actively demanding vaccination (Milton & Mercier, 2015; Cell Press, 2015).

Data Collection

Two datasets of posts were scraped from two different online parenting forums: one historically anti-vaccination and one historically vaccination-neutral. The anti-vaccination forum is located at mothering.com, the companion website to the discontinued *Mothering* magazine, which described itself as "the magazine of natural family living". The vaccination-neutral forum is from the subreddit "r/parenting," a parenting forum described as "the place to discuss the ins and outs as well as the ups and downs of child-rearing." The anti-vaccination forum is a forum with the specified topic "vaccination." The vaccination-neutral forum does not

specify topic, so we extracted posts containing strings related to "vaccine," "inoculate," "jab," and "shot." The two datasets do not exclusively consist of respective anti-vaccination and vaccination-neutral sentiment, so some noise is to be expected. However, a manual random sampling and evaluation of posts shows the forums to be consistent with their historical positions on vaccination. The near non-existent level of noise was deemed acceptable. The data is as recent as June 28, 2021. At the time of collection, children ages 12-18 were eligible to receive COVID-19 vaccines in most states. See Table 1 for a side-by-side comparison of the forums.

The number of posts in the vaccination-neutral dataset is lower than the number of posts in the vaccination-neutral forum overall. This can be attributed to the fact that vaccination was an uncommon topic of conversation in this parenting forum. In contrast, the anti-vaccination forum is one of the most popular forums on its website: It was 14th out of 69 total forums for views, surpassing comprehensive, general forums like Preteens & Teens, and Baby Health.

Table 1. A comparison of characteristics of the anti-vaccination and vaccination-neutral parenting forums.

	Anti-vax forum	**Vax-neutral forum**
Number of posts in forum	26,321	119,176
Number of posts in dataset	26,321	952
Number of users	275,400	3,400,000
Year started	1996	2008

Methods for Finding Thought-Terminating Clichés

Since there are no large, community-sanctioned lists of thought-terminating clichés (hereafter TTCs), we compiled lists from print and online resources – including books, articles, and blog posts by linguists, rhetoricians, journalists, and philosophers. TTCs that appeared more than once were given priority. TTCs that appeared only once or were criticized or downvoted in any user discussions were not included. We rewrote each TTC as a case-insensitive keyword search, excluding those containing references to He/Him in the religious sense, where capitalization is a feature. Straightforward keyword searches were carried out for each TTC in each of the two datasets. Since all the TTCs were either multiword expressions or contained punctuation, there was little chance that hits could be mistakenly counted as TTCs when they were not actually used as TTCs in context.

Methods for Finding Euphemistic and Dysphemistic Language

To identify words as euphemisms or dysphemisms of a target word, we first compiled lists of near-synonyms from human-generated sources. These euphemisms and dysphemisms were considered candidate euphemisms and dysphemisms to be rated as euphemistic or dysphemistic by three graduate student native-speaker annotators. We obtained near-synonyms for three concepts: DIE, LIE, and STEAL. We did not constrain the part of speech in our near-synonym search, so verbs, nouns, and adjectives were all included (e.g. *steal*, *theft*, *stolen*).

This methodology is inspired by Felt & Riloff's paper "Recognizing Euphemisms and Dysphemisms Using Sentiment Analysis" (2020). While they used the Basilisk bootstrapping algorithm for weakly supervised semantic lexicon induction, we use WordNet and Wiktionary as thesauri since they are community-sanctioned and sufficiently populated. First, we compile lists of near-synonyms using WordNet and Wiktionary as thesauri. To obtain *near*-synonyms (not just direct synonyms), we expand the inventory of synonyms to include those captured in the "See also" pages on Wiktionary. Thus, synonyms for dead/die/dying/death include results from the entry Thesaurus:kill, which is a linked entry listed under "See also" on the Thesaurus:dead entry. This method captures euphemistic and dysphemistic nuance that may be found in the parenting corpora with regard to vaccines, such as whether vaccines are *resulting in deaths* (a more neutral, if spurious statement), or whether they are *killing people* (a statement with greater negative polarity).

We obtained annotator ratings for each of the three concepts on a scale from 1 to 5, where 1 is most dysphemistic, 3 is neutral, and 5 is most euphemistic (κ for each concept $\geq$ 0.86). See Figure 1 for a visualization of the rating scale. In accordance with Felt & Riloff's methodology, for each phrase, we computed the average score across the three annotators and assigned each phrase to a "gold" x-phemism category: phrases with score < 2.5 were labeled dysphemistic, phrases with score > 3.5 were labeled euphemistic, and the rest were labeled neutral. Keyword searches were carried out for each set of euphemisms and dysphemisms for each concept. Unlike TTCs which are usually multiword expressions, many of the euphemisms and dysphemisms here are single words that could easily have literal and non-euphemistic/dysphemistic meanings in context. To ensure we were not counting words that should not be counted towards the total of euphemisms and dysphemisms, we checked the context of the actual post containing the hit for the meaning of the keyword. For example, when searching for instances of *crucify/crucifixion*, we omitted results that referred to the literal crucifixion of Jesus and counted only non-literal instances of *crucify/crucifixion* towards the total hit count for that word (e.g. "And please please please let's not turn this

into a crucifixion of me for vaccinating, or a place to try to talk me out of it....").

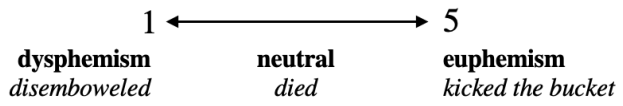

Figure 1. A visualization of the rating scale given to annotators. A score of 1 is most dysphemistic, 3 is neutral, and 5 is most euphemistic.

Results

Thought-terminating clichés

Anti-vaccination posts contained more thought-terminating clichés than vaccination-neutral posts, with 0.83% (8 out of every 1000 posts) of anti-vax posts containing at least 1 thought-terminating cliché, and 0.21% of vax-neutral posts containing at least 1 thought-terminating clichés (2 out of every 1000 posts). These results are significant at $p = .04$.

Table 2: Observed counts for thought-terminating clichés.

Dataset	TTCs	Non-TTCs	Total
anti-vax	218	26144	26362
vax-neutral	2	952	954
total	220	27096	27316

Examples:

agree to disagree

- One thing we learned was that there are times when we simply have to **agree to disagree** and move on.
- can't we just **agree to disagree**? i realize she is concerned for the health of my child, but we have done SO much research into this and we are well-educated in health issues in general.

do (your/your own/the) research

- For those of you that may harbor a bit of fear for vaccine preventable diseases...Fear not...**do your research** and sail through it...In my experience, my DD's teething was more work than the measles.
- It's a lesson learned: always trust your instinct and **do your own research**!

so it goes

- It's odd because while they were premature (seven weeks), they've been around many many people and numerous illnesses and never caught them at this rate. But **so it goes**…

it is what it is

- I am passionate about not vaxxing...I am, its an instinct and I just dont know why. **It is what it is**.
- Not my first choice, or really my choice at all, but **it is what it is**.

everything happens for a reason

- I am convinced that **everything happens for a reason** and am sure that if we had vaccinated her she would not have been okay.

anyway.

- She was just born and you want to make her sick? **Anyway**.
- If a grey and unresponsive babe is an unusual event, just was the heck is an adverse reaction? **Anyway**.

Euphemistic and dysphemistic language

Across all concepts, dysphemisms are more commonly used than euphemisms in the anti-vaccination dataset, with 1.42% of anti-vaccination posts containing euphemistic or dysphemistic language, and 0.84% of vaccination-neutral posts containing euphemistic or dysphemistic language. When comparing the frequency of euphemism vs. dysphemism usage across concepts in the vaccination-neutral dataset, dysphemisms are more commonly used for the concepts LIE and STEAL, but not for the concept DIE. Euphemisms were more commonly used than dysphemisms for DIE in the vaccination-neutral dataset, specifically the euphemism *passed away*.

Table 3. Euphemism (euph) and dysphemism (dys) counts for DIE, LIE, and STEAL in each dataset.

	DIE euph	DIE dys	LIE euph	LIE dys	STEAL euph	STEAL dys
anti-vax	30	207	21	88	1	28
vax-neutral	3	2	0	1	0	2

The top 3 dysphemisms for each concept were *kill*, *murder*, and *drop dead* for DIE; *fraudulent*, *deceptive*, and *deceive* for LIE; and *crime*, *rob*, and *kidnap* for STEAL. Dysphemisms for DIE occurred significantly more frequently in the anti-vaccination dataset than in the vaccination-neutral dataset at $p = .006$. Consider the following examples containing the top dysphemisms for each concept:

- What doesn't make sense to me is that people actually believe in companies such as Merck, GlaxoSmithKline, and other pharmaceutical companies when they have engaged in heinous business practices, been sued, knowingly **killed** people...
- "The scientific publication system is far from golden standard, and a process that can be easily **fraudulent** and abused."
- some of these people in the medical profession have no conscience. i look forward to the day when they're tried for their **crimes**.

Discussion and Conclusion

We found that users of an anti-vaccination parenting forum are more likely to use loaded language to express themselves

than users of a vaccination-neutral parenting forum. Since many anti-vaccination beliefs are based on various conspiracy theories (that vaccines do not work, vaccines are harmful, vaccines contain microchips, vaccines are part of a government-sponsored depopulation program, etc.) and anti-vaccination beliefs are strongly correlated with belief in unrelated conspiracy theories (Goldberg & Richey, 2020), these findings support the hypothesis that loaded language may be indicative of conspiracy theorizing.

For thought-terminating clichés, anti-vaccination posts contained significantly more of this type of loaded language than vaccination-neutral posts. Thought-terminating clichés impede further critical thinking, and impeded critical thinking and faulty reasoning lead to crippled epistemologies (Sunstein et al., 2009), a factor in conspiracy theorizing. A discussion forum comparably replete with thought-terminating clichés may be more disposed to to conspiracy theorizing.

For euphemistic and dysphemistic language, anti-vaccination posts contained significantly more dysphemisms about death than vaccination-neutral posts. Euphemistic and dysphemistic language produce a certain amount of ambiguity in meaning since they substitute words or phrases for alternatives with different connotations. Ambiguity regarding the nature and cause of death (e.g. implying that children are killed rather than dying) is in service of a conspiratorial narrative which posits a powerful group of people as actors. Historically, dysphemisms have been studied as tools to conceptualize political enemies (Bakhtiar, 2016) and to divide the world into "good" and "bad" for persuasive ends (Veisbergs, 2006). Dysphemisms swap out neutral/basic words and phrases for more negative ones to express hatred, neglect, or irritation (Krysin, 1994). In the case of death, which is already a taboo topic with negative polarity, dysphemisms may serve to emphasize the perceived cruelty or barbarity of dying as a result of vaccination or at the hands of the pharmaceutical industry.

This study has two important limitations. First, the sample size of the vaccination-neutral dataset is small. While this does not mean it is not representative, there is an increased chance for a sampling bias to occur. Future work can address this issue by evaluating the presence of loaded language in a forum dedicated to discussing a broader range of conspiracy theories and comparing it to a forum dedicated to discussing science topics, such as the subreddits r/conspiracy and r/science. This methodology avoids encountering the sample size issue in the non-conspiratorial narrative by using the entirety of a forum (or a recent subsection), rather than extracting posts pertinent to a topic. In the present study we focused on vaccine-related conspiracy theories as a test case to study loaded language and conspiracy theorizing while planning to expand the domain in future research. Second, the methods described for finding euphemisms and dysphemisms in text are fairly labor-intensive and only semi-automatic. As described earlier, we checked the context of each post containing the hit for the meaning of the euphemism or dysphemism to ensure we were not counting words whose contextual meaning is not euphemistic or dysphemistic, and therefore should not be counted towards the totals. To make this process more efficient, we will develop filters to remove non-euph/dysphemistic patterns of usage. For example, developing a filter to omit senses of *slaughter* that are followed by (beef/poultry/pork/…) removes common conflating senses of *slaughter*. That way, the remaining hits for *slaughter* are more likely to be euph/dysphemistic like "Johnson & Johnson is slaughtering children."

Despite these limitations, this study demonstrates that users of anti-vaccination parenting forums are more likely to use loaded language to express themselves than users of a vaccination-neutral parenting forum. Since many anti-vaccination beliefs are based on various conspiracy theories (that vaccines do not work, vaccines are harmful, vaccines contain microchips, vaccines are part of a government-sponsored depopulation program, etc.), these findings support the hypothesis that loaded language may be indicative of conspiracy theorizing.

This research lays the groundwork for an expanded examination of loaded language in forums dedicated to discussing a broader range of conspiracy theories. In accordance with the paradigm described by Bessi et al. (2015), anti-vaccination attitudes correspond to conspiracy theorizing narratives and vaccination-neutral attitudes correspond to science-based narratives. We intend to examine additional types of loaded language and evaluate their presence in forums according to this established paradigm. In particular, we are beginning studies relating to (1) biblical references and phrases, and (2) ingroup/outgroup language – two other forms of loaded language identified in the literature. Biblical literalists are significantly more likely to believe in a variety of conspiracy theories (Baylor University, 2021), and may occasionally speak in *Christianese*, a religiolect with distinct terms and jargon used within many branches and denominations of Christianity. Ingroup identification – the desire to belong to and maintain a positive image of the ingroup – is important in conspiracy ideation (Douglas et al., 2017). Distancing oneself/one's ingroup from the outgroup (of powerful people who carry out the conspiracy) is also important. Ingroup and outgroup language has been studied in the context of conspiracy theories and anti-vaccination views, but only in the online space of Twitter which has a distinctive character limit (Mitra et al., 2016; Fong et al., 2021).

The present study examines the relationship between loaded language and conspiracy theorizing in a test case scenario, and provides evidence for whether and to what extent loaded language appears in conspiracy theorizing language. Findings contribute to scholarship in linguistics and the cognitive sciences, as well as to the type of complex, expert knowledge needed to build more interpretable AI systems for tasks such as content moderation.

Acknowledgments

We thank our anonymous reviewers for their helpful feedback. This work was supported by the DARPA KAIROS

program, contract no. FA8750-19-C-0206. Professor Hendler's work is supported in part through the Rensselaer-IBM Artificial Intelligence Research Collaboration, a member of the IBM AI Horizons network.

References

Bakhtiar, M. (2016). "Pour water where it burns": Dysphemistic conceptualizations of the enemy in Persian political discourse. *Metaphor and the Social World*, *6*(1), 103-133.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. Personality and Social Psychology Review, 3, 193-209.

Baylor University. (2021). *Baylor Religion Survey: Conspiratorial Beliefs*. [Data set]. Retrieved from https://www.baylor.edu/baylorreligionsurvey/index.php?id=978882

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE*, *10*(2). https://doi.org/10.1371/journal.pone.0118093

Bruns, A., Harrington, S., & Hurcombe, E. (2020). 'Corona? 5G? or both?': The dynamics of covid-19/5G conspiracy theories on Facebook. *Media International Australia*, *177*(1), 12–29. https://doi.org/10.1177/1329878x20946113

Byford J. (2011) The Anatomy of the Conspiracy Theory. In: Conspiracy Theories. Palgrave Macmillan, London. https://doi.org/10.1057/9780230349216_4

Cell Press. (2015, October 29). What blocks pro-vaccine beliefs?. ScienceDaily. Retrieved January 18, 2022 from www.sciencedaily.com/releases/2015/10/151029134122.htm

Chiras, D. (1992). Teaching Critical Thinking Skills in the Biology & Environmental Science Classrooms. *The American Biology Teacher*, 54 (8), 464–468, https://doi.org/10.2307/4449551

Chu, J. T., Wang, M. P., Shen, C., Viswanath, K., Lam, T. H., & Chan, S. S. C. (2017). How, when and why people seek health information online: qualitative study in Hong Kong. *Interactive journal of medical research*, *6*(2), e7000.

Collins, B. (2020, September 26). *Qanon leaders look to rebrand after Tech Crack Downs*. NBCNews.com. Retrieved from https://www.nbcnews.com/tech/tech-news/qanon-leaders-look-rebrand-after-tech-crack-downs-n1241125

Da San Martino, G., Barron-Cedeno, A., & Nakov, P. (2019a). Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. In *Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda* (pp. 162-170).

Da San Martino, G., Yu, S., Barrón-Cedeno, A., Petrov, R., & Nakov, P. (2019b). Fine-grained analysis of propaganda in news articles. 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)* (pp. 5636-5646).

Da San Martino, G., Cresci, S., Barrón-Cedeno, A., Yu, S., Di Pietro, R., & Nakov, P. (2020a). A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*.

Da San Martino, G., Shaar, S., Zhang, Y., Yu, S., Barrón-Cedeño, A., & Nakov, P. (2020b). "Prta: A system to support the analysis of propaganda techniques in the news." In Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics, ACL 2020, Seattle, Washington, USA.

Douglas, K. M., & Leite, A. C. (2016). Suspicion in the workplace: Organizational conspiracy theories and work-related outcomes. *British Journal of Psychology*, *108*(3), 486–506. https://doi.org/10.1111/bjop.12212

Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The Psychology of Conspiracy Theories. Current Directions in Psychological Science, 26(6), 538–542. https://doi.org/10.1177/0963721417718261

Faasse, K., Chatman, C. J., & Martin, L. R. (2016). A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post. *Vaccine*, *34*(47), 5808–5814. https://doi.org/10.1016/j.vaccine.2016.09.029

Felt, C., & Riloff, E. (2020, July). Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing* (pp. 136-145).

Fetters Maloy, A., & Oremus, W. (2021, December 24). *Pregnancy apps have become a battleground of vaccine misinformation*. The Washington Post. Retrieved from https://www.washingtonpost.com/technology/2021/12/23/pregnancy-apps-covid-vaccine-misinformation/

Fong, A., Roozenbeek, J., Goldwert, D., Rathje, S., & van der Linden, S. (2021). The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter. Group Processes & Intergroup Relations, 24(4), 606–623. https://doi.org/10.1177/1368430220987596

Giachanou, A., Ghanem, B., & Rosso, P. (2021). Detection of conspiracy propagators using psycho-linguistic characteristics. *Journal of Information Science*. https://doi.org/10.1177/0165551520985486

Goldberg, Z. J., & Richey, S. (2020). Anti-Vaccination Beliefs and Unrelated Conspiracy Theories. World Affairs, 183(2), 105–124. https://doi.org/10.1177/0043820020920554

Hoffman, B. L., Felter, E. M., Chu, K. H., Shensa, A., Hermann, C., Wolynn, T., & Primack, B. A. (2019). It's not all about autism: The emerging landscape of anti-vaccination sentiment on Facebook. *Vaccine*, *37*(16), 2216-2223.

Hornsey, M. J., Harris, E. A., Fielding, K. S. (2018). Relationships among conspiratorial beliefs, conservatism

and climate scepticism across nations. Nature Climate Change, 8, 614–620. https://doi.org/10.1038/s41558-018-0157-2

Hughes, B., Miller-Idriss, C., Piltch-Loeb, R., White, K., Criezis, M., Cain, C., & Savoia, E. (2021). Development of a Codebook of Online Anti-Vaccination Rhetoric to Manage COVID-19 Vaccine Misinformation. *medRxiv*.

Jolley D., Douglas K. (2014a). The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint. British Journal of Psychology 105(1): 35–56. pmid:24387095

Jolley D., Douglas K. (2014b). The effects of anti-vaccine conspiracy theories on vaccination intentions. PLoS ONE 9(2): e89177. pmid:24586574

Jolley, D., Douglas, K., Leite, A., & Schrader, T. (2019). Belief in conspiracy theories and intentions to engage in everyday crime. Br. J. Soc. Psychol., 58: 534-549. https://doi.org/10.1111/bjso.12311

Kata, A. (2010). A postmodern Pandora's box: anti-vaccination misinformation on the Internet. *Vaccine*, *28*(7), 1709-1716.

Kaufman, J., Ryan, R., Lewin, S., Bosch-Capblanch, X., Glenton, C., Cliff, J., ... & Hill, S. (2018). Identification of preliminary core outcome domains for communication about childhood vaccination: An online Delphi survey. *Vaccine*, *36*(44), 6520-6528.

Krysin, L. (1994). Euphemisms in modern Russian speech. Russian Philology 1(2), (pp.28-49). Berlin.

Lewandowsky S., Gignac G., Oberauer, K. (2013). The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. PLoS ONE 8(10): e75637. pmid:24098391

Lifton, R. J. (1989). Chapter 16, The Older Generation: Robert Chao. In *Thought reform and the psycology of totalism: A study of "Brain washing" in China* (p. 429). The University of North Carolina.

McShane, M., & Nirenburg, S. (2021). *Linguistics for the Age of Ai*. The MIT Press. https://doi.org/10.7551/mitpress/13618.001.0001

Memon, S. A., Tyagi, A., Mortensen, D. R., & Carley, K. M. (2020, October). Characterizing sociolinguistic variation in the competing vaccination communities. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (pp. 118-129). Springer, Cham.

Mikolov, T., Sutskever, I., Chen, K., Corrado., G. & Dean, J. (2013). "Distributed representations of words and phrases and their compositionality". *Advances in Neural Information Processing Systems*. arXiv:1310.4546.

Milton, H., & Mercier, H. (2015). Cognitive obstacles to pro-vaccination beliefs. Trends in Cognitive Sciences, 19(11), 633–636. https://doi.org/10.1016/j.tics.2015.08.007

Mitra, T., Counts, S., & Pennebaker, J. W. (2016). Understanding anti-vaccination attitudes in social media. In *Tenth International AAAI Conference on Web and Social Media*.

Morris, A. (2021, October 25). *It's not Q. it's you*. Rolling Stone. Retrieved from https://www.rollingstone.com/culture/culture-features/qanon-expert-joesph-uscinski-1242636/

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Nakov, P., Alam, F., Shaar, S., Da San Martino, G., & Zhang, Y. (2021). "A Second Pandemic? Analysis of Fake News About COVID-19 Vaccines in Qatar." arXiv preprint. arXiv:2109.11372

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.3115/v1/d14-1162

Romer, D., Jamieson, K. H. (2020). Conspiracy theories as barriers to controlling the spread of COVID-19 in the US. Social Science and Medicine, 263, Article 113356. https://doi.org/10.1016/j.socscimed.2020.113356

Rottweiler, B. & Gill, P. (2020) Conspiracy Beliefs and Violent Extremist Intentions: The Contingent Effects of Self-efficacy, Self-control and Law-related Morality, Terrorism and Political Violence, https://doi.org/10.1080/09546553.2020.1803288

Scott, M., & Kern, R. (2022, January 6). *The Online World Still can't quit the 'big lie'*. POLITICO. Retrieved January 16, 2022, from https://www.politico.com/news/2022/01/06/social-media-donald-trump-jan-6-526562

Shahsavari, S., Holur, P., Wang, T., Tangherlini, T. R., & Roychowdhury, V. (2020). Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science*, 1–39. Advance online publication. https://doi.org/10.1007/s42001-020-00086-5

Stein, H. F. (1998). Euphemism, spin, and the crisis in organizational life. Westport, CT: Quorum Books.

Sunstein, C, Vermeule, A (2009). Conspiracy Theories: Causes and Cures. *Journal of Political Philosophy* 17: 202–227.

Tangherlini, T. R., Roychowdhury, V., Glenn, B., Crespi, C. M., Bandari, R., Wadia, A., & Bastani, R. (2016). "Mommy blogs" and the vaccination exemption narrative: results from a machine-learning approach for story aggregation on parenting social media sites. *JMIR public health and surveillance*, *2*(2), e6586.

Tangherlini, T., Shahsavari, S., Shahbazi, B., Ebrahimzadeh, E., & Roychowdhury, V. (2020). An automated pipeline for the discovery of conspiracy and conspiracy theory narrative frameworks: Bridgegate, Pizzagate and storytelling on the web. PLOS ONE 15(6): e0233879. https://doi.org/10.1371/journal.pone.0233879

Tiffany, K. (2020, August 18). *The women making conspiracy theories beautiful*. The Atlantic. Retrieved from

https://www.theatlantic.com/technology/archive/2020/08/how-instagram-aesthetics-repackage-qanon/615364/

Uscinski, J., DeWitt, D., & Atkinson, M. (2018). A Web of Conspiracy? Internet and Conspiracy Theory. In A. Dyrendal, D. Robertson, & E. Asprem (Eds.), *Handbook of Conspiracy Theory and Contemporary Religion* (Vol. 17, pp. 106–130).

Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). "At the End of the Day Facebook Does What ItWants": How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–22. https://doi.org/10.1145/3415238

van Prooijen, J., & Douglas, K. (2017). Conspiracy theories as part of history: The role of societal crisis situations. Memory Studies, 10(3), 323–333. https://doi.org/10.1177/1750698017701615

van Prooijen, J., & Douglas, K. (2018). Belief in conspiracy theories: Basic principles of an emerging research domain. European journal of social psychology, 48(7), 897–908. https://doi.org/10.1002/ejsp.2530

van Prooijen, J., Spadaro, G., & Wang, H. (2021). Suspicion of institutions: How distrust and conspiracy theories deteriorate social relationships. *Current Opinion in Psychology*, *43*, 65–69. https://doi.org/10.1016/j.copsyc.2021.06.013

Veisbergs, A. (2006). Nazi and Soviet Dysphemism and Euphemism in Latvian. *Baltic Postcolonialism*, *6*, 139.

Vrdelja, M., Kraigher, A., Verčič, D., & Kropivnik, S. (2018). The growing vaccine hesitancy: exploring the influence of the internet. *European journal of public health*, *28*(5), 934-939.