Morality in Moderation: The Presence of Moral Language in Interventions Against Misinformation in Community Notes on Twitter.com

A. Wesel[1], M. Malik[2]

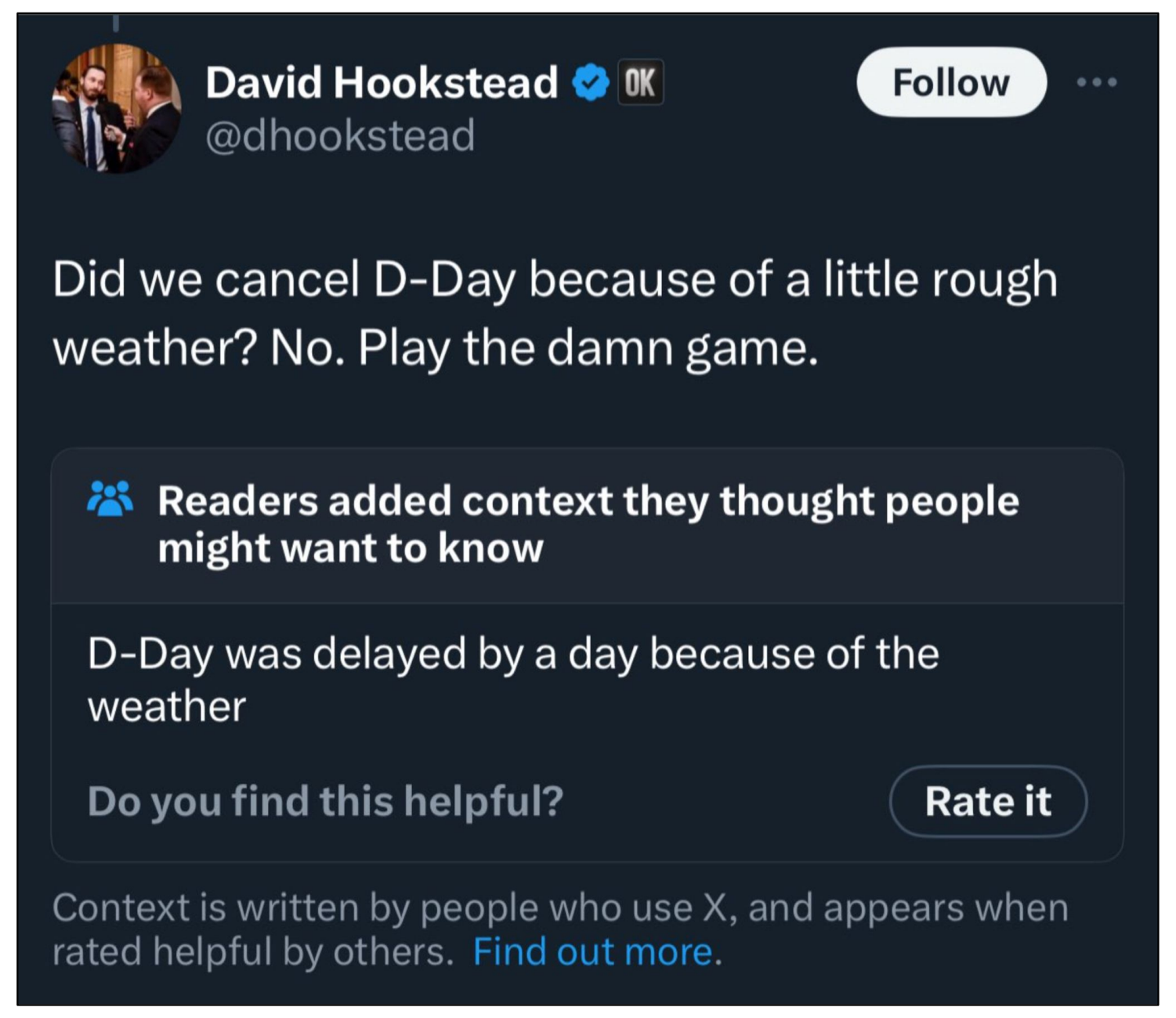
[1]Harvard-Westlake School, Los Angeles, CA 91604, [2]Media Neuroscience Lab, Department of Communication, University of California, Santa Barbara, Santa Barbara, CA 93106

Introduction

In recent decades, the internet has enabled information to be spread more rapidly than ever and enabled anyone to speak to the entirety of humanity. While this ability has led to vast increases in the wealth of human knowledge, any person can accidentally spread false beliefs or even intentionally deceive the public with a few clicks.

The spread of misinformation is not solely about the presence of false information or lack of true information. In reality, there are a variety of cognitive processes that inform misinformation beliefs that are far more complex than simply comparing the validity of conflicting views (Ecker et al. 2022). The presence of moral and emotional language, repeated exposure to the same false information, or the perception of authority or credibility can all influence information belief beyond practical judgment.

Twitter.com's misinformation mitigation strategy, a crowd-sourced annotation platform called Community Notes, places an annotation containing correct information below false or misleading posts. Unfortunately, this strategy depends on believers of false information simply not having access to the truth, despite evidence that holders of false beliefs have a more complex relationship with the truth. As an instance of the complicated psychological processes involved with information perception, this study investigated the use of *moral* language in annotations.



David Hookstead ✓ OK
@dhookstead

Did we cancel D-Day because of a little rough weather? No. Play the damn game.

Readers added context they thought people might want to know

D-Day was delayed by a day because of the weather

Do you find this helpful?

Context is written by people who use X, and appears when rated helpful by others. Find out more.

An example of a Community Note

Dataset

Twitter.com makes a dataset of Community Notes and their associated Post IDs available online, but that dataset does not contain the text of the original post. We used Selenium ChromeDriver to autonomously navigate to the Post that the note was referencing and record the text of the post as well as other analytics such as the number of views, likes, and reposts. **Figure 1** shows the number of notes shown per Tweet ID, sorted by whether that Tweet ID has a note shown on the website. Notes are only shown if enough users agree that the note is necessary and useful. The vast majority of notes are never shown.

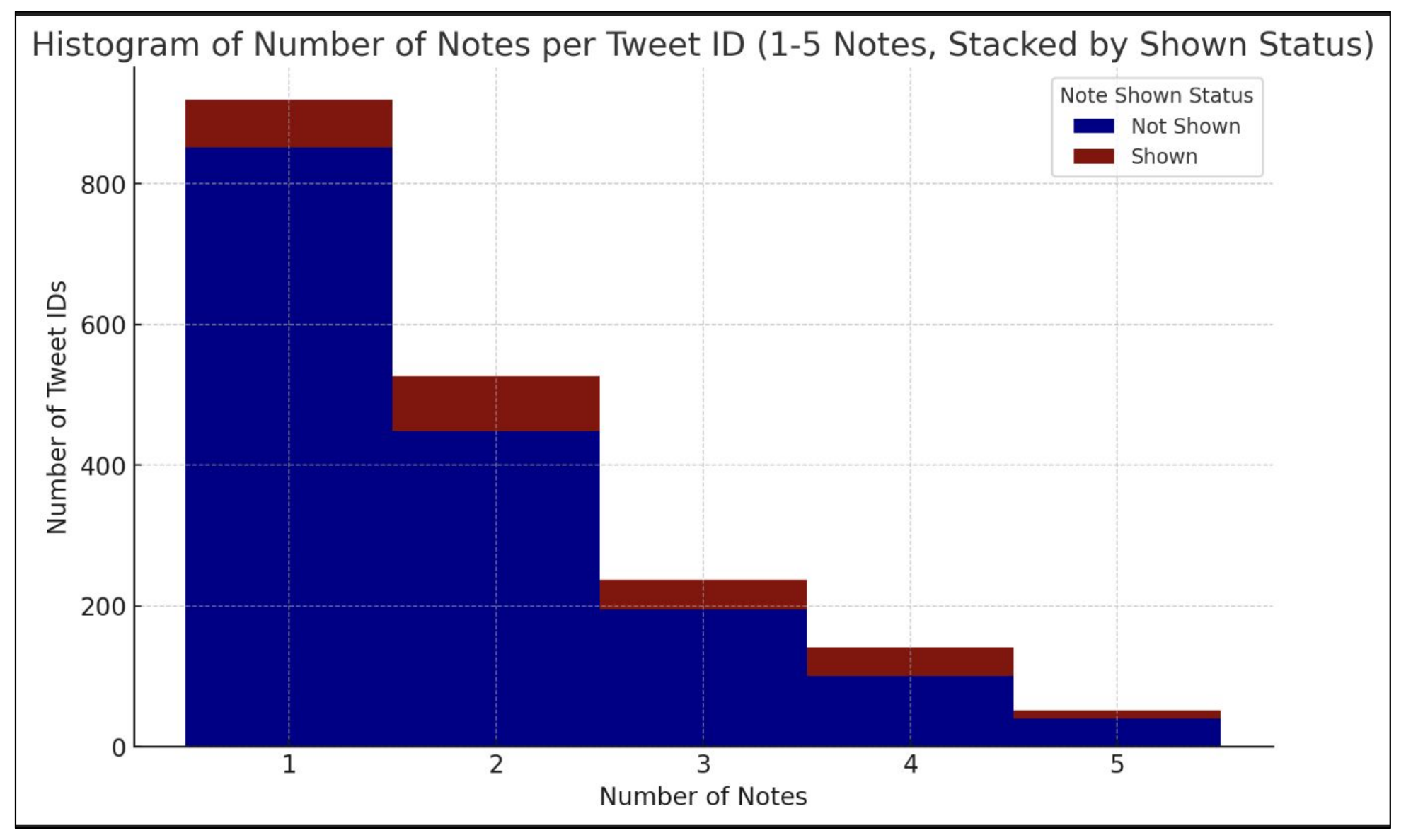


Figure 1. Many posts only have a single suggested note.

Results

1: Difference between Notes and Posts

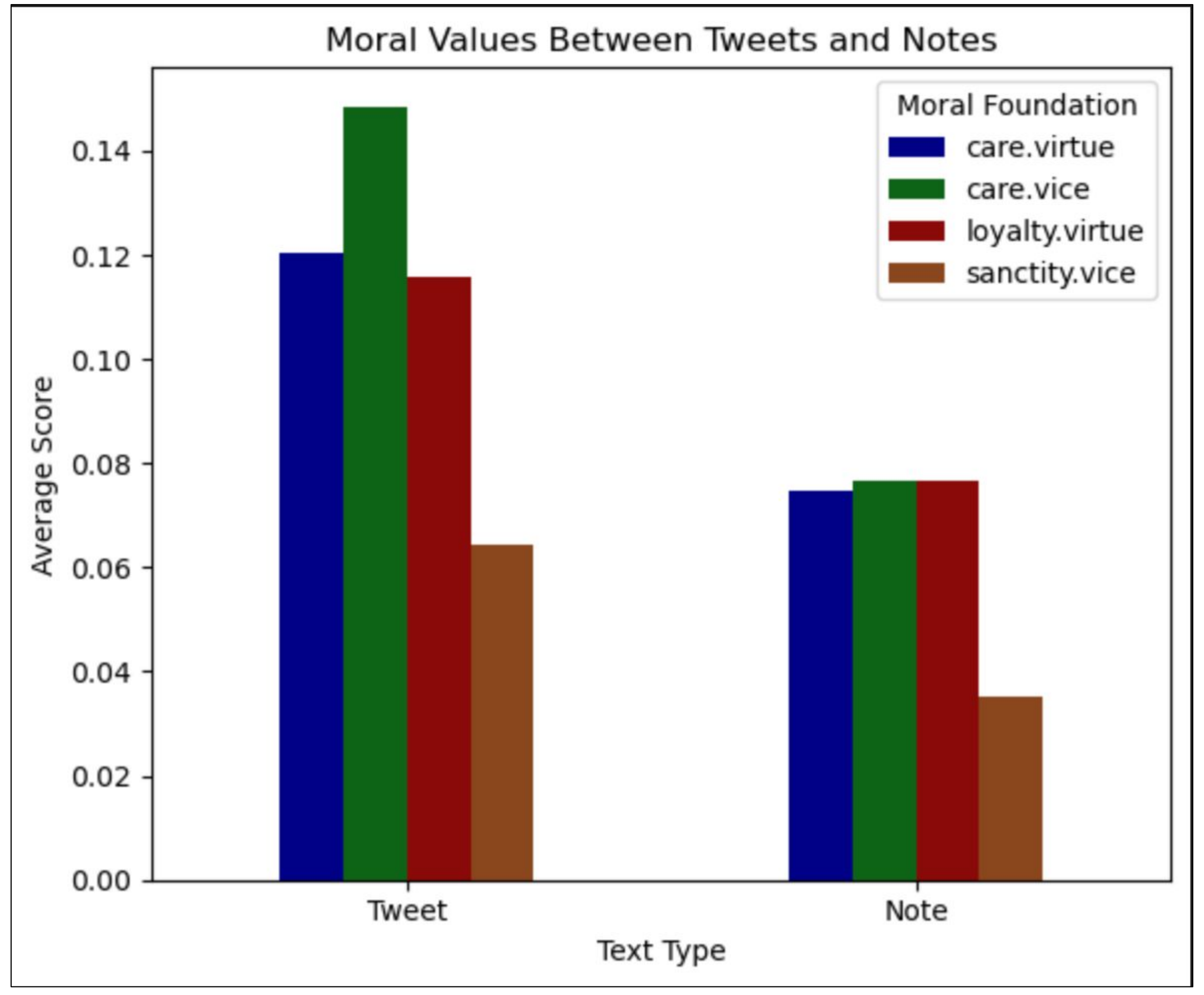
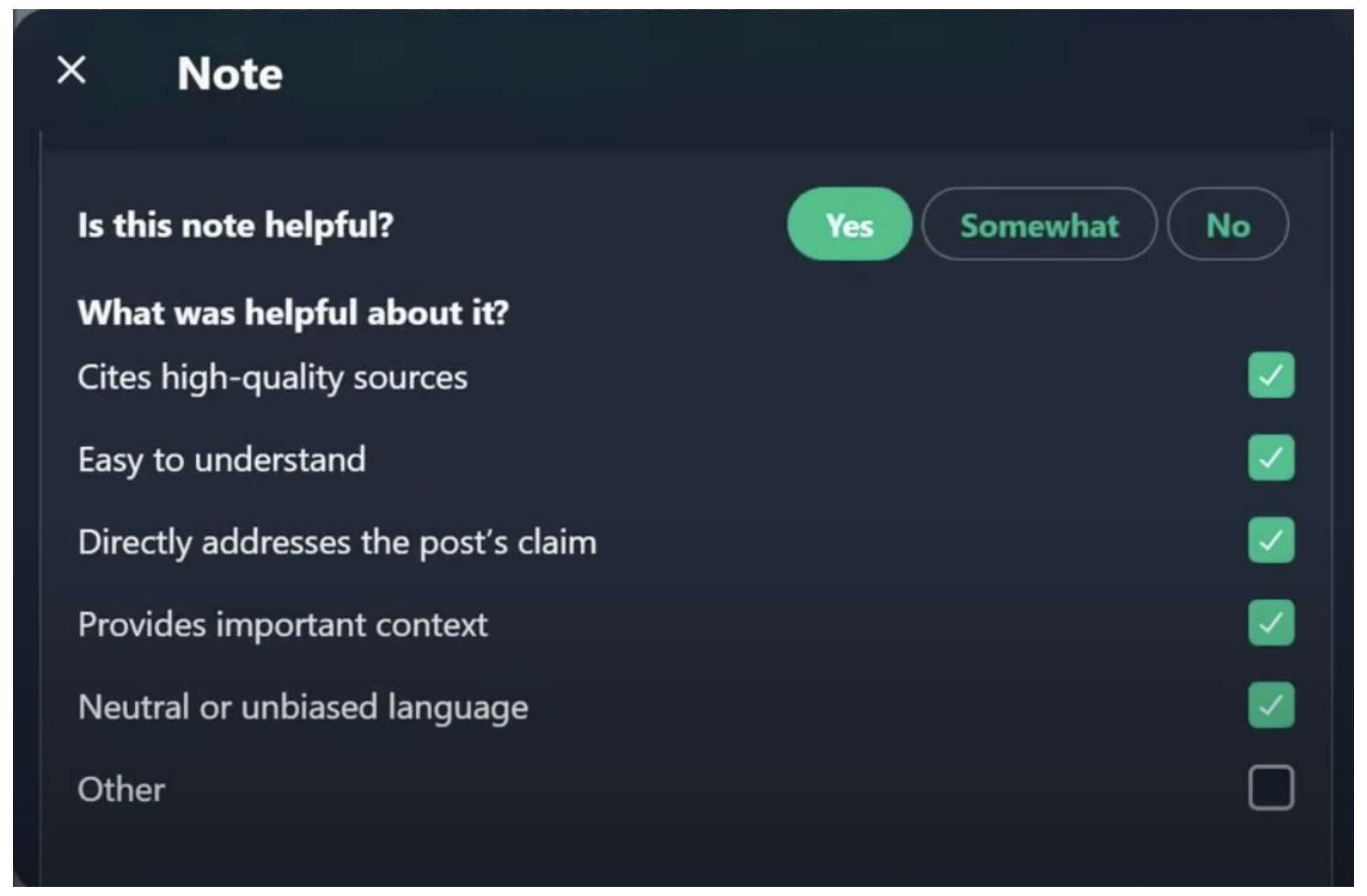


Figure 2. Tweets use more moral language than Notes.



Is this note helpful? ● Yes ○ Somewhat ○ No

What was helpful about it?

- Cites high-quality sources ☑
- Easy to understand ☑
- Directly addresses the post's claim ☑
- Provides important context ☑
- Neutral or unbiased language ☑
- Other □

Neutral or unbiased language is a stated goal of Notes

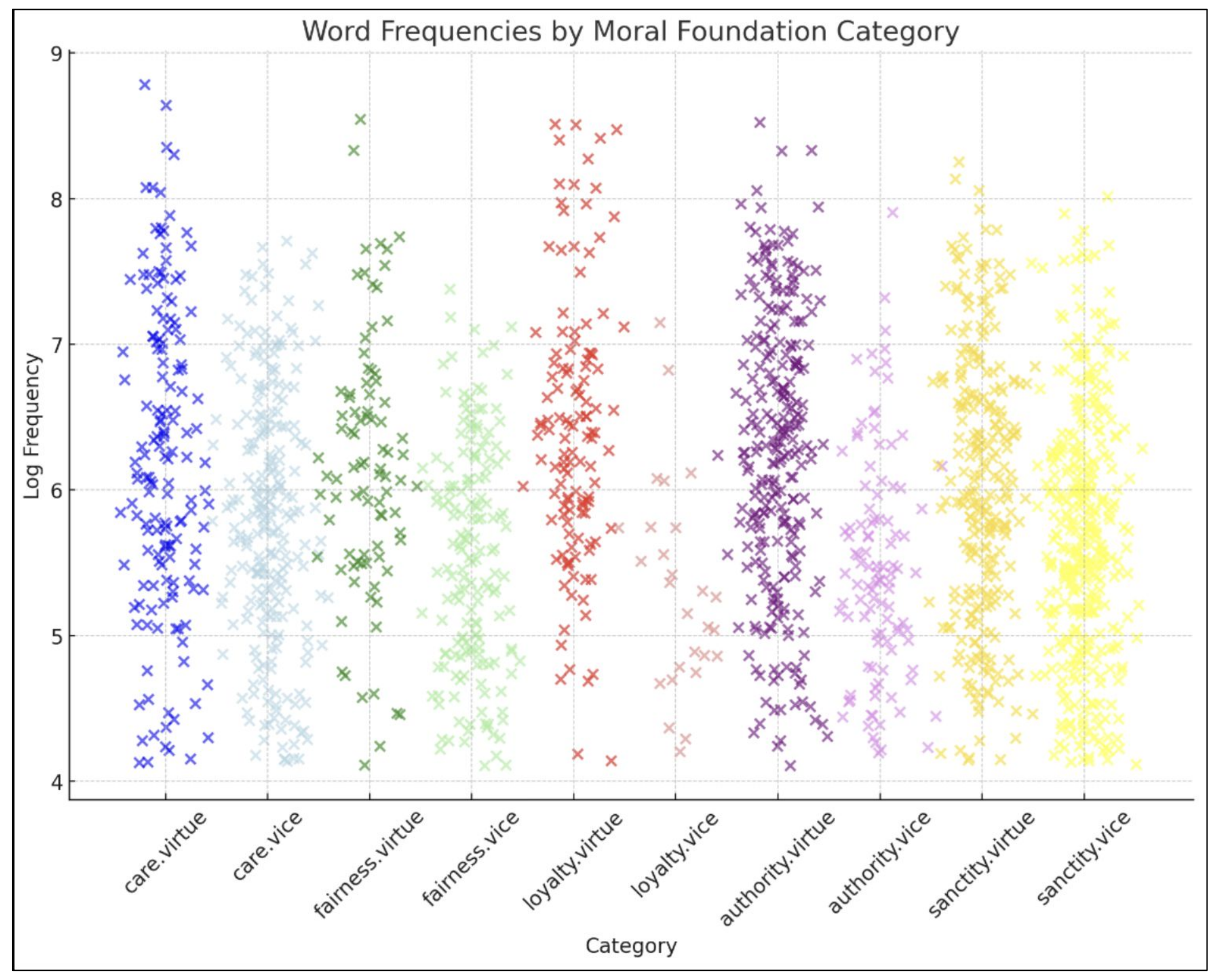


Figure 3. Error analysis: some moral foundations use less popular language.

2: What makes a Note chosen?

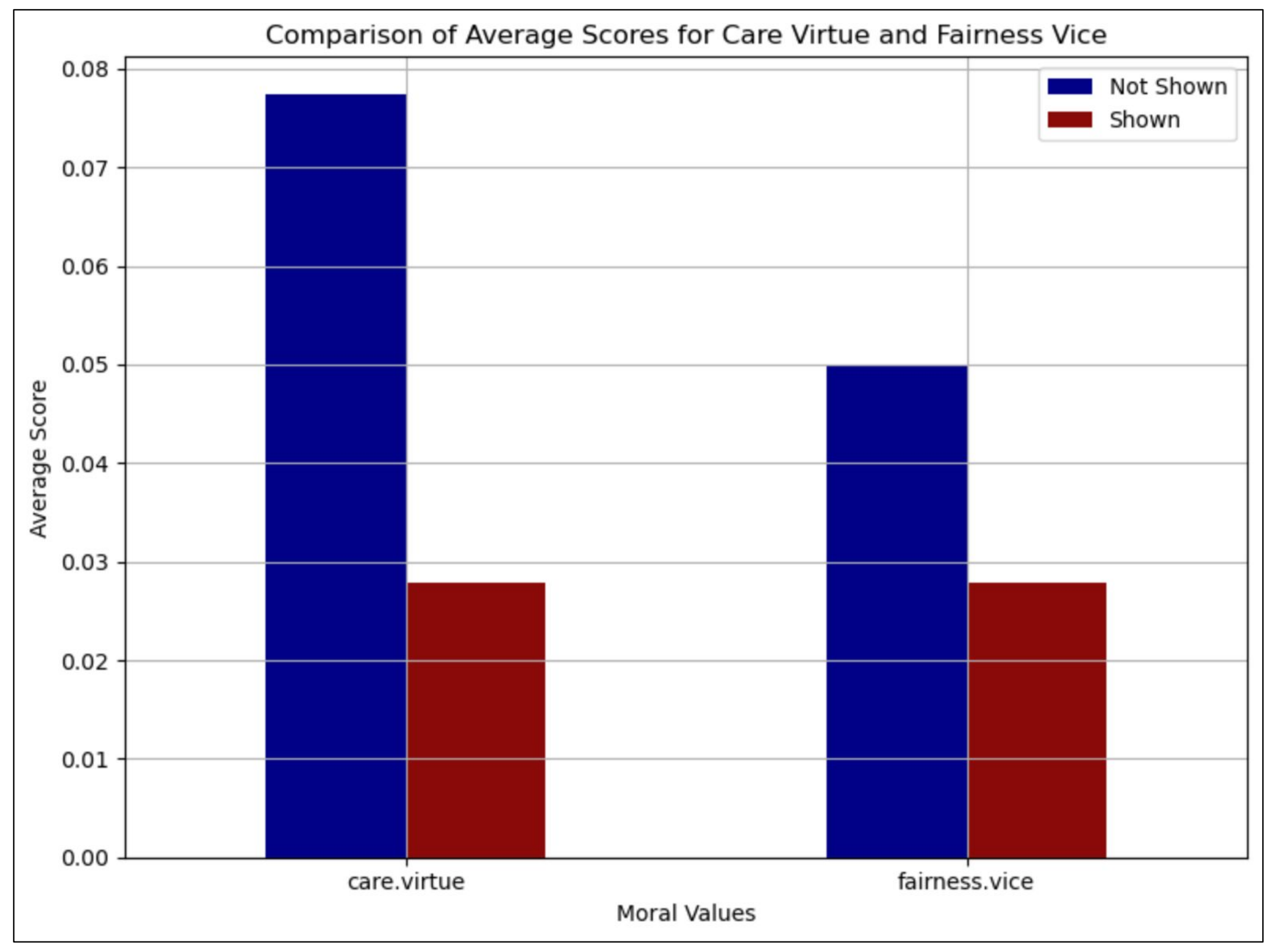
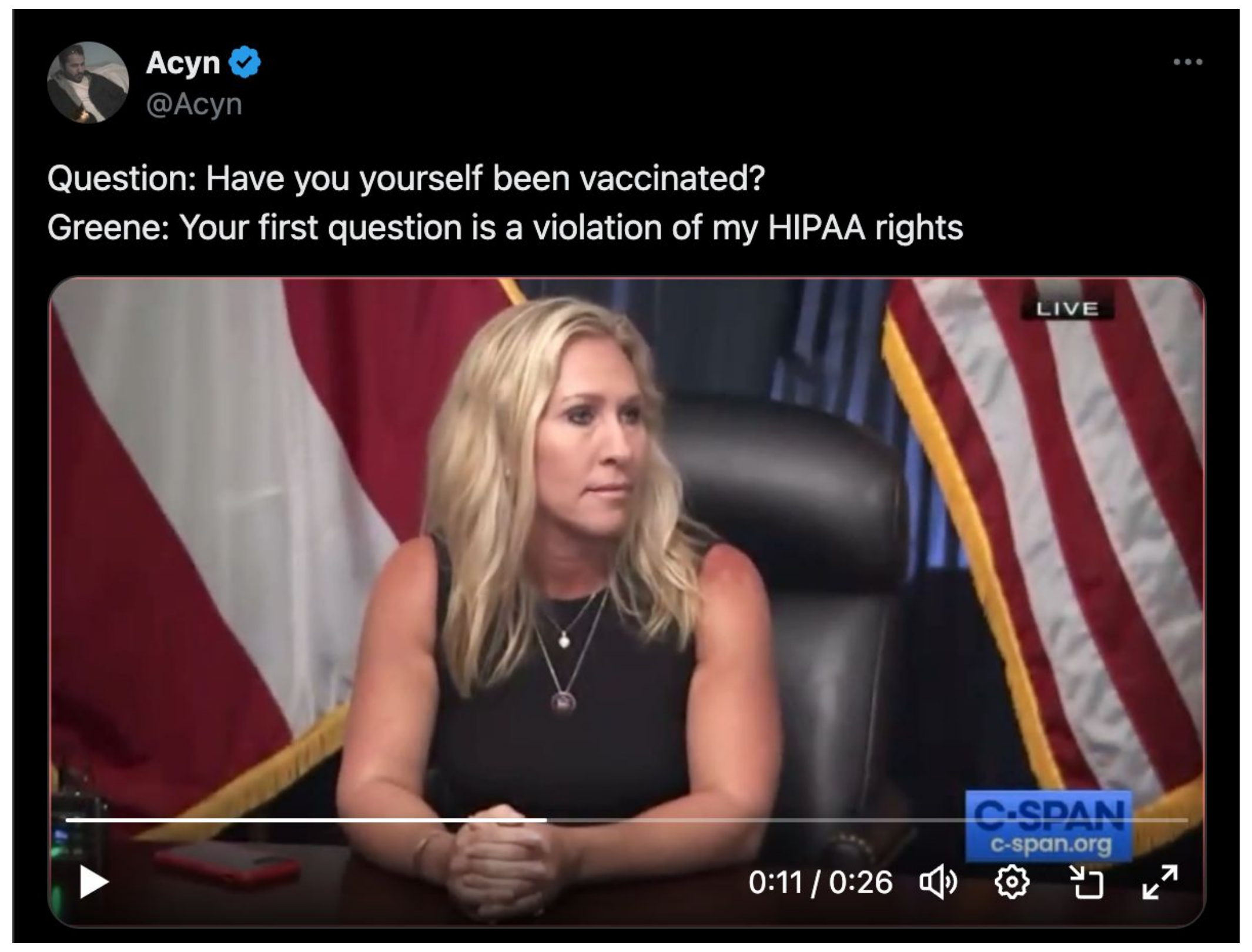


Figure 4. Tweets use more moral language than Notes.



Acyn ✓
@Acyn

Question: Have you yourself been vaccinated?
Greene: Your first question is a violation of my HIPAA rights

Which note is better? [Moral language highlighted]

"HIPAA does not protect an individuals rights to disclose their own medial history or to stop a 3rd party like a business from asking. HIPAA only stops healthcare related entity's and businesses from disclosing your medial history." [grammar preserved]

Or

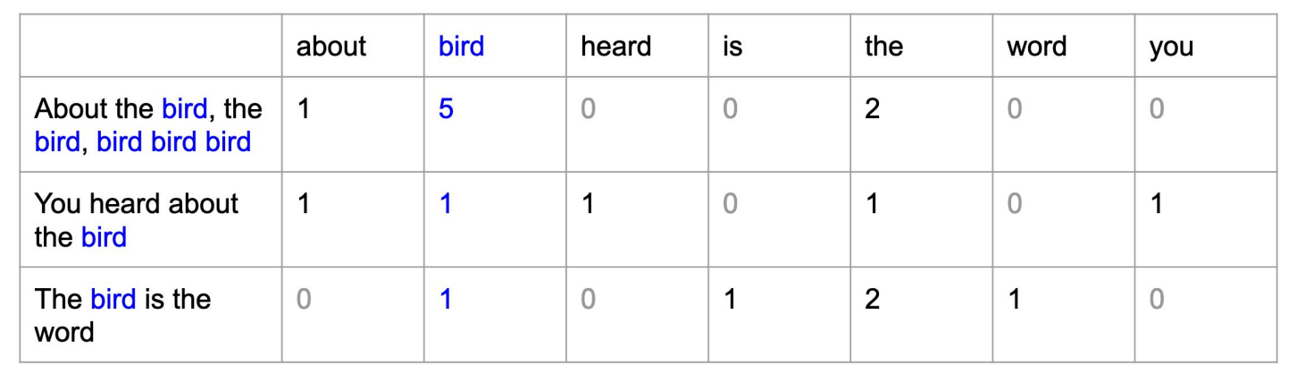
"HIPAA does not prevent employers, businesses, or individuals from asking someone about Covid-19 vaccination status."

The second option was chosen to be shown on Twitter.com. There are many reasons why someone would prefer the second option; clarity, concision, and grammar are all clear differentiating factors. But our results indicate the lack of moral language might be important too!

Methodology

To quantitatively study the moral value of certain statements, this study employs Moral Foundations Theory (MFT) (Graham et al. 2013). MFT seeks to divide moral judgments into five sets of foundational intuitions that explain more specific claims of morality and immorality: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation (Haidt and Joseph 2004).

To measure the moral content of a post or an annotation, we used a Bag-of-Words approach using Moral Foundations Dictionary 2.0 (MFD 2.0), a dictionary of words that correspond to each moral foundation developed in 2020 as a successor for the original Moral Foundations Dictionary developed in 2009 (Frimer 2020). MFD2.0 is widely used in recent research of moral appeals on social media.



	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

Demonstration of Bag-of-Words

Discussion

Twitter.com's official policies recognize that "Neutral or unbiased language" is a desirable attribute in an intervention against false information. Our results call into question what it means for language to be "neutral and unbiased," whether that should be a goal, and whether Community Notes are meeting that goal in the status quo.

Given that the same people using Twitter.com are the ones writing the notes, **Figure 2** indicates that the same people use different words when writing notes than when normally posting. They recognize the position as "fact-checker" as unique and requiring a different vocabulary. Contentious moral appeals may make a post on the normal platform more popular, but there is some recognition among users that annotations should have less moral content than the average post. Future research should investigate whether the use of a different moral vocabulary has a positive or negative effect on whether the annotations are effective blocks against misinformation.

Figure 4 indicates that people are more likely to choose notes if they use less moral language. This indicates that users are accurately selecting for community notes which use "Neutral" language. However, such neutrality might be counterproductive. Since people of different political backgrounds respond better to certain messaging, it might benefit the program to invoke binding foundations if the annotation intends to convince conservatives or avoid invoking binding foundations if the annotation intends to convince liberals.

We also found several statistically significant relationships relating to the Loyalty Vice moral invocation. However, in the entire dataset of thousands of posts, under ten invoked Loyalty Vice. **Figure 3** offers an explanation for this skew; words that invoke that moral are much rarer than those that invoke other foundations.

References

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., Amazeen, M. A. 2022. The psychological drivers of misinformation belief and its resistance to correction. Nature Reviews Psychology. 1 (1): 13–29

Frimer, J. A. 2020. Do liberals and conservatives use different moral languages? Two replications and six extensions of Graham, Haidt, and Nosek's (2009) moral text analysis. Journal of Research in Personality, 84.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. 2013. Moral Foundations Theory. Advances in Experimental Social Psychology (pp. 55–130).

Haidt, J., and Joseph, C. 2004. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. Daedalus, 133(4), 55–66.