Assessing Cognitive Linguistic Influences in the Assignment of Blame

Karen Zhou and **Ana Smith** and **Lillian Lee**
Cornell University, Ithaca, NY
{kz265, als476}@cornell.edu, llee@cs.cornell.edu

Abstract

Lab studies in cognition and the psychology of morality have proposed some thematic and linguistic factors that influence moral reasoning. This paper assesses how well the findings of these studies generalize to a large corpus of over 22,000 descriptions of fraught situations posted to a dedicated forum. At this social-media site, users judge whether or not an author is in the wrong with respect to the event that the author described. We find that, consistent with lab studies, there are statistically significant differences in usage of first-person passive voice, as well as first-person agents and patients, between descriptions of situations that receive different blame judgments. These features also aid performance in the task of predicting the eventual collective verdicts.

1 Introduction

Dyadic morality theory proposes that the harm one party causes another is an important component in how other people form judgments of the two parties as acting morally or not. Under this framework, perpetrators (agents) are perceived as blameworthy, whereas victims (patients) are not (Gray and Wegner, 2009; Schein et al., 2015). This effect appears to transfer to how active (agentive) a party is described to be, even if the activity was in the past — a phenomenon described by Gray and Wegner's (2011) paper titled, "To Escape Blame, Don't be a Hero — Be a Victim".

The online forum https://reddit.com/r/AmItheAsshole collects first-person descriptions of (purportedly) real-life situations, together with commentary from other users as to who is blameworthy in the situation described; two examples are shown in Figure 1. (Additional examples may be found in Appendix A.) This data allows us to evaluate findings from dyadic morality theory on a corpus involving over 22,000 events and 685,000 passed judgments.

The research questions we address with this data in this paper include:

(1) Do authors refer to themselves in passive voice more often in descriptions of situations where they are judged to be morally incorrect?

(2) How does an author's framing of themselves as an "agent" or "patient" in describing a moral situation affect the judgments they receive?

The first question is motivated by Bohner (2002), who found that using passive voice, by placing someone who was actually a victim in subject position (e.g., "X was threatened by Y"), causes the victim to seem more responsible for the event. (See also Niemi and Young (2016) on the effect of syntactic-subject position for perpetrator vs. victim descriptions.) Importantly, our two questions *together* separate passive voice from agentiveness.

We find that while the agentive aspect of dyadic morality theory is upheld in our data, passive voice theory does not align empirically. We also incorporate these theories as features in a verdict prediction task.

2 Data

The subreddit from which we draw our data is self-described as follows:

> A catharsis for the frustrated moral philosopher in all of us, and a place to finally find out if you were wrong in an argument that's been bothering you. Tell us about any non-violent conflict you have experienced; give us both sides of the story, and find out if you're right, or you're the [jerk].

It has served as the basis of prior computational analysis of moral judgment by Botzer et al. (2021) and Lourie et al. (2021). (The Moral Stories dataset

Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pages 61–69
Online Workshop, June 10, 2021.

Title: AITA for telling a kid to shut up on a plane

Situation: I was on a trip to Michigan a yearback and some kid was crying on the plane for 30 minutes. I yelled for the kid to shut up and the kid quieted down but the parents started freaking out.

Top verdict: YTA, call the flight attendant if this is going on. You yelling for a child to shut up is only going to escalate tensions and, like exactly what happened, get the parents pissed off at you and start freaking out.

(a)

Title: WIBTA if I had someone's car towed?

Situation: My building has pretty limited parking and we've been having an issue with people who don't live here taking up all the parking. I asked one guy if he lived here, and when he said he didn't I told him he couldn't park there. His car is back again, WIBTA if I had him towed without a warning?

Top verdict: NTA. Resident parking is included in what you pay towards rent, typically - it helps keep insurance lower than having to park on the street or away from the complex.
You gave him fair warning. If he wants to visit, he needs to find guest parking or park on the street himself.

(b)

Figure 1: Two example situations and the top-rated comment attached to each. Other comments are omitted for space. (a) In this case, the top-rated comment starts with **YTA** ("You're the [jerk]"), indicating a judgment that the post author is at fault, and no other participant is. (b) In this case, the top-rated comment starts with **NTA** ("You're not the [jerk]"), indicating a judgment that the post author is not at fault, but rather the other party is. (In the post title, "WIBTA" stands for "Would I be the [jerk]".)

(Emelin et al., 2020) would have been an interesting alternative corpus to work with. It also draws some of its situations from the same subreddit.)

Since the SCRUPLES dataset (Lourie et al., 2021), also based on the aforementioned subreddit, does not include corresponding full comments, which we wanted to have as an additional source of analysis,[1] we scraped the subreddit ourselves. Our dataset (henceforth AITA) includes posts from the same timeframe as SCRUPLES: November 2018-April 2019.

The winning verdict of each post is determined, according to the subreddit's rules, by the verdict espoused by the top-voted comment 18 hours after submission. We aim to only include posts with meaningful content, so we discard posts with fewer than 20 comments and fewer than 6 words in the body, as manual appraisal revealed that these were often uninformative (e.g., body is "As described in the title").

For simplicity, we only consider situations with the YTA (author in the wrong, other party in the right) and NTA (author in the right, other party in the wrong) verdicts, although other verdicts (such as "everyone is in the wrong", and "no one is in the wrong") are possible. We still use over 75% of the data since these are the most prevalent outcomes on the forum, and the theories we assess align with having binary outcomes (comparing victim vs. perpetrator responsibility). This selection results in 22,795 posts, fewer than the over 32,000 in SCRUPLES (Lourie et al., 2021). The corpus contains more NTA posts, which are longer in word length on average (see Table 1).

Verdict	Average post length	Class proportion
YTA	333 tokens	40.3%
NTA	384 tokens	59.7%

Table 1: AITA corpus statistics.

3 Methodology

Passive subject identification To model the use of passive voice in moral situations, a dependency parser is used to match spans of passive subjects in sentences. We use spaCy's Matcher object to extract tokens tagged `nsubjpass`. Cases where the extracted passive subject is in first person (1P) are also tracked, as indication of the author being referred to passively. Some examples include:

[1] For each post, up to 100 of the most "upvoted" comments were also retrieved; these were not used here but could be useful for further cognitive theory reinforcement or nuanced controversy analysis.

- 1P passive subject: **I** was asked to be a bridesmaid and then she changed her mind last minute and **I** was removed from the bridal party in favor of one of her husbands cousins.
- Other passive subject: She obliged but **she** was pissed off the rest of the night.

For manual evaluation, we randomly selected 500 posts, containing a total of 675 uses of passive voice. The tagger achieved 0.984 precision on these posts. Among the 199 first-person passive subjects tagged, the precision achieved was 0.971.

Because Niemi and Young (2016) and Bohner (2002) find that passive voice is associated with greater perception of victims' causal responsibility, we hypothesize that situations with the YTA verdict may have higher rates of 1P passive subject usage.

Thematic role identification To approximate moral agents vs. patients, Semantic Role Labelling (SRL) is used to extract agents and patients. Semantic, or thematic, roles express the roles taken by arguments of a predicate in an event; an agent is the volitional causer of the event, while the theme or patient is most affected by the event (Jurafsky and Martin, 2019). The AllenNLP BERT-based Semantic Role Labeller (Gardner et al., 2017; Shi and Lin, 2019) is employed to extract spans that are tagged `ARG0` for agents and `ARG1` for patients. We also tag uses of 1P- agents and patients. Here are two examples:

- 1P agent: **I** don't want my fiance to take care these freeloaders anymore.
- 1P patient: He called **me** names, threatened divorce, and told me he's a saint for staying married to me.

As a sanity check, we manually evaluated a subset of 579 verb frames, corresponding to 15 posts, identified by the SRL tagger. The tagger achieved a precision of 0.934 on all verb frames. The precision on the 193 verb frames in this subset that contained a first-person `ARG0` or `ARG1` was 0.891. Examples where the tagger failed include sentence fragments (e.g. "Made my MMA debut today.") and use of first-person pronouns to describe other parties (e.g. "Everyone we know").

Gray and Wegner (2011) concluded that "it pays to be a [patient] when trying to escape blame. [Agents],... depending on the situation, may actually earn increased blame." Thus, we hypothesize that NTA may be associated with higher 1P-patient usage and YTA with higher 1P-agent usage.

4 Statistical Analysis

Due to the post length discrepancy between verdicts, we attempt to control for length in the analysis by assessing significance at the sentence level. While NTA posts average approximately 50 words more than YTA posts, sentences from NTA posts average only 0.5 words more than YTA sentences (17.0 vs. 16.6 words respectively).

We assess statistical significance as follows. We use a simple binomial test: let r be the rate of the given feature of interest (say, 1P-passive voice) over the entire collection of posts. We then compute the probability according to the r-induced binomial distribution — i.e., the null hypothesis that there is no difference between the YTA posts and the body of posts overall — of the observed number of occurrences of the feature in just the YTA posts. Similarly, we compute this probability for just the NTA posts.

4.1 Passive Subject Identification

We find that NTA situations have a higher rate of 1P passive subject usage than YTA situations, and that the deviation of both the rate in the YTA posts and in the NTA posts from the overall data is statistically significant. As shown in Table 2, 45.8% of NTA posts' passive voice uses are 1P, while 37.4% of YTA posts' passive voice uses are 1P.

The rate difference across verdicts is significant, with NTA posts having a higher 1P-passive rate (see Table 2). This could account for the 0.5-words-longer sentence average of NTA posts; since, for example, "I hit John" is shorter than its passive counterpart, "John was hit by me." This contradicts our hypothesis, as we expected higher 1P-passive rates for YTA posts.

We do not discount a possible explanation for this differing result being that the cognitive researchers had better control over narrative structure, content of their situations, and participants that provided judgment. On the other hand, it is also possible that the forum setting is, at least in certain respects, more natural (and definitely larger-scale) than the lab setting in which the original experiments took place.

Verdict	Rate	Binomial Significance Test
YTA	0.374	$p = 2.14\text{e-}22$
NTA	0.458	$p = 2.42\text{e-}15$
Overall	0.424	

Table 2: Rate of 1P-passive voice use, i.e. where the author is the passive subject.

	YTA	NTA	Overall
# Agents	32.1	37.8	35.5
1P-Agent Rate	**0.502**	0.482	0.492
# Patients	35.9	42.5	39.8
1P-Patient Rate	0.232	**0.238**	0.235
Verbs per Post	47.3	55.9	52.5

Table 3: SRL post average semantic role usage. Higher rates per row are bolded. Recall that on average, NTA posts are longer than YTA posts.

4.2 Thematic Role Identification

The NTA posts use more agents and patients by raw count and also have more verbs per post, since they are generally longer than YTA posts (see Table 3). When we examine proportions of uses, we find that the NTA posts have a higher rate of 1P-patient usage, while YTA posts have a higher rate of 1P-agent usage.

While the verdicts do not differ significantly in overall agent and patient usage, there are significant differences in rates of 1P (see Table 4). The rate of 1P-patient usage in NTA posts is significantly higher than that of YTA posts ($p < 0.005$), while the rate of 1P-agent usage in YTA posts is significantly higher than that of NTA posts ($p < 0.001$). These results seem to align with our hypothesis based on Gray and Wegner (2011)'s findings.

5 Verdict prediction task

In the previous section, we examined statistical correlations between features of interest in the previous literature to the verdicts presented in our data.

	YTA	NTA
Agent/Verbs	$p = 0.06$	$p = 0.15$
1P/Agent	$p = 1.70\text{e-}26$	$p = 7.54\text{e-}16$
Patient/Verbs	$p = 0.965$	$p = 0.972$
1P/Patient	$p = 1.06\text{e-}4$	$p = 3.49\text{e-}3$

Table 4: SRL sentence-level binomial significance tests. The differences for first-person/agent and first-person/patient rates of usage are noticeable.

Title: AITA for telling a kid to shut up on a plane

Situation: I was on a trip to Michigan a yearback and some kid was crying on the plane for 30 minutes. I yelled for the kid to shut up and the kid quieted down but the parents started freaking out.

Features:

```
{
  'text.word_length': 41,                          LENGTH
  'text.passive_fp_i': 0,
  'text.passive_fp_we': 0,
  'text.passive_fp': 0,                            +PASSIVE
  'text.passive_total': 0,
  'text.passive_fp_rate': 0,
  'text.unqiue_passive_subjs': 0,
  'text.srl_num_sents': 2,
  'text.srl_avg_verbs_per_sents': 3.0,
  'text.srl_arg0_fp': 1,
  'text.srl_arg1_fp': 0,                           +SRL
  'text.srl_unique_arg0': 5,
  'text.srl_unique_arg1': 5,
  'text.srl_unique_verbs': 6
}
```

Figure 2: Features extracted for a sample post. The verbs identified by the tagger are highlighted in yellow, and the 1P ARG0 is highlighted in cyan. Note that the +Passive features are very sparse. First-person "me" and "us" were not added as +Passive features, as their inclusion yielded about 1% worse performance (likely since they added additional noise).

In this section, we turn to prediction as another way to examine the magnitude of potential linkages between these features and judgments of blame. In particular, we see how incorporating these quite small set of features compares against a baseline classifier that has access to many more (lexical-based) features, but where these features are not explicitly cognitively motivated.

Specifically, to analyze the significance of passive voice and thematic roles as features in making moral judgments, we model the task of predicting the verdict of a situation as binary classification (YTA or NTA). We compare the performance of a linear and non-linear model.

We stress that we are *not* striving to build the most accurate judgment predictor for moral-scenario descriptions, nor arguing the utility or importance of such a classification task. Rather, we are using prediction as a further mechanism for answering the research questions we delineated in the introduction to this paper. We do not use BERT since it is pre-trained, possibly containing encoded biases, and is not as interpretable as simpler models.

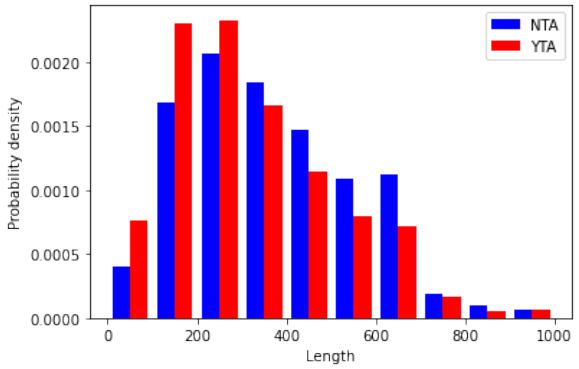


Figure 3: Normalized histogram of word length distributions across verdicts. The few posts with lengths $>$ 1000 tokens are omitted for clarity. The distributions were found to be significantly different by the t-test, Z-test, and Kolmogorov-Smirnov test ($p < 0.005$).

For an ablation study, we have four feature sets, with the corresponding number of features in brackets:

- Length [1]: word count of the post, since NTA posts average more words than YTA posts
- Length+Passive [7]: counts of passive and 1P passive subjects, and rate of 1P passive subjects.
- Length+SRL [8]: unique `ARG0` (i.e., agent), `VERB`, and `ARG1` (i.e., patient) token counts (where # `ARG1` tokens per frame ≤ 2), count of sentences including `ARG0` and/or `ARG1` tokens, average number of `VERB`s per sentence as identified by the tagger, and 1P agent and patient token counts.
- Length+Passive+SRL [14]: features of both Length+Passive and Length+SRL.

Figure 2 provides an example of features extracted for a real post.

We assess these feature sets against 43,110 lexical-based features from a TF-IDF transform of lowercased unigrams and bigrams with 0.1% minimum document frequency.

The configurations for the linear and non-linear models are described below. The AITA data is split 60/20/20 for the train/val/test sets, after random shuffling. Both models are trained on this same split of data.

Linear Model We opt for a simple model to begin with to avoid overfitting on the dataset and for purposes of interpretability. For a linear model, we use the scikit-learn logistic regression model (LR) (Pedregosa et al., 2011). Hyperparameters for the logistic regression model include setting random state to 0, choosing "liblinear" as the solver, and setting the class weights to "balanced" to account for the label imbalance.

Non-Linear Model We incorporate a non-linear model, as we observed that our feature count distributions were weakly bi-modal even after grouping instances under the NTA/YTA labels (see Figure 3). We use the scikit-learn random forest model (RF) (Pedregosa et al., 2011). Hyperparameters for the random forest model include setting class weights to "balanced", "sqrt" for the maximum features, and 100 for number of estimators. Through tuning over the range [5,15], we found that setting the maximum depth to 7 prevented overfitting on the training data.

6 Task Results

To give the imbalanced labels equal importance, we evaluated macro-average scores. Weighted average scores were usually around 1% higher than the macro-average scores. Overall, the non-linear model achieves higher F1 scores for each of our feature sets, though the linear model does better with TF-IDF features (see Tables 5 and 6).

6.1 Linear Model Results

Compared to the random forest, the linear model achieves better performance with TF-IDF features (0.58 vs. 0.62 F1 score). Length+SRL has the best performance of our feature sets, with 0.56 precision and recall and 0.54 F1 score (see Table 5). The distinction in performance across feature sets is less clear than with the non-linear model, suggesting that the logistic regression model is not able to learn as well from these particular features.

From the ROC curves, we see that Length+SRL shows a little improvement over Length alone at higher thresholds, but does around equally at lower thresholds (see Figure 4a). The performance gap between TF-IDF features and our features is greater with the linear model.

We also see from the confusion matrix in Figure 4a that the best model version tends to predict YTA. Depending on desired use case — and recalling that we are not necessarily promoting judgment prediction as a deployed application — it may be better to err on the side of predicting one side or the other. If the priority is to catch all possible occurrences of the author being judged to be in the

	Prec	Rec	F1	AUC
Majority	0.30	0.50	0.37	—
TF-IDF	0.62	0.62	0.62	0.667
Length	0.55	0.55	0.53	0.576
+Passive	0.55	0.55	0.53	0.574
+SRL	**0.56**	**0.56**	**0.54**	**0.587**
+Passive+SRL	0.55	0.55	0.54	0.585

Table 5: Macro-average results of verdict prediction task with the LR model. Best scores among the feature sets are bolded.

	Prec	Rec	F1	AUC
Majority	0.30	0.50	0.37	—
TF-IDF	0.59	0.59	0.58	0.620
Length	0.55	0.55	0.55	0.568
+Passive	0.55	0.55	0.54	0.565
+SRL	**0.56**	**0.56**	**0.56**	**0.586**
+Passive+SRL	**0.56**	**0.56**	**0.56**	0.585

Table 6: Macro-average results of verdict prediction task with the RF model. Best scores among the feature sets are bolded.

wrong, this model would be better suited than the non-linear model. However, this model would also yield more false accusations, which could be more undesirable.

6.2 Non-Linear Model Results

Like the linear model, Length+SRL does best overall, with 0.56 for precision, recall, and F1 score (see Table 6). Length+Passive+SRL performs similarly.

From the ROC curves, we see that Length+SRL shows some improvement over Length alone (see Figure 4b). With these feature sets, we achieve performance close to that of a model trained with TF-IDF, with much fewer features: 43,110 vs. a mere 8. In addition, TF-IDF may overfit to topics (e.g., weddings), whereas our features are easier to transfer across domains.

From the confusion matrix in Figure 4b, we notice that even with balanced class labels, the best model still slightly favors predicting NTA.

7 Discussion

Despite noting the significant difference in first-person passive voice usage between verdicts, the feature set of Length+Passive yields slightly lower performance than the Length baseline for both models. This could be due to not having enough instances of passive voice, as each post has on average 1.39 counts of passive voice, of which 30.5% are first-person. Regex searches confirmed that the dependency-parser did not simply have poor recall, though the methods for passive voice extraction are not exact. Thus, the passive features may be acting as noise.

Length+SRL builds off of more SRL instances per post, so these features provide less noisy information. This feature set's performance beats that of the Length baseline for both models, suggesting that SRL features do play a role in making moral judgments. The SRL features do not store lexical information, which helps remove the influence of the content of the posts. Length+Passive+SRL performance likely suffers from the additional passive features' noise.

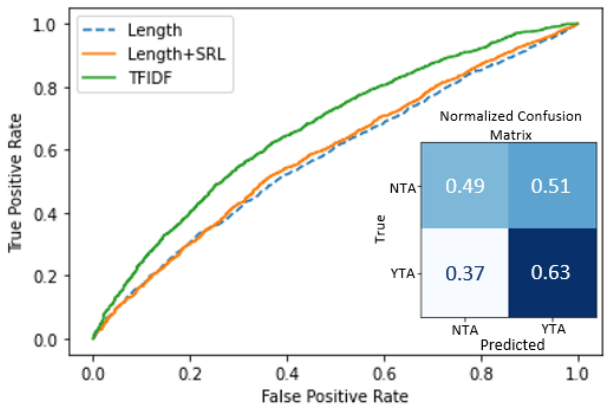
A notable difference is the non-linear model's tendency to favor NTA and the linear model's preference for YTA. A possible explanation for this is that the features corresponding to YTA situations are more linearly separable than those corresponding to NTA situations.

Comparing scores for the Length baseline, we see that the random forest has a 3.8% improvement in F1 score over logistic regression. This may suggest that post length is not a linear feature, which would account for nuances such as long YTA and short NTA posts (see Figure 1b for an example of a short NTA post).

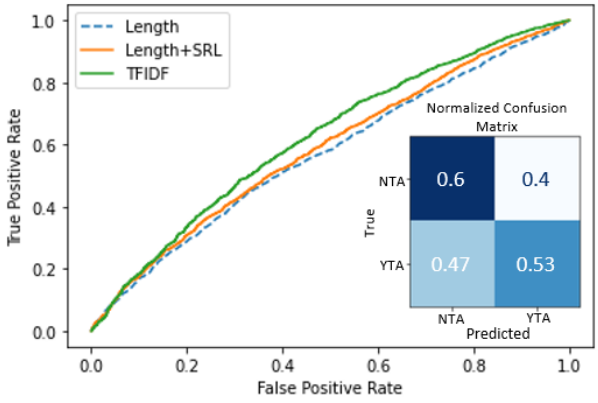
Caveats We are certainly not saying that blameworthiness can be reduced to use of first-person descriptors. There are a multitude of features and factors at play, and there may be alternative parameters to consider for the task.

Even if we restrict attention to linguistic signals, there are quite a few confounds to point out. As just one example: it is possible that authors purposefully manipulate their use of first-person pronouns to appear less guilty. Another possibility to consider: there may be correlations between whether an author *believes* they are guilty and how they describe a situation, so that commenters are not picking up on the actual culpability in the described scenario so much as the author's self-blame.

Also, we can look beyond linguistic factors. For example, when deciding whether to "upvote" a particular judgment comment, voters may be affected by the (apparent) identity of the commenter (or, for that matter, the original post author) and the content of other comments. We have not accounted for such factors in our study.



(a) LR model.



(b) RF model.

Figure 4: The ROC curves for verdict prediction task, for the feature sets described in the key, and the confusion matrices for the verdict prediction task results of the Length+SRL feature set.

We must also keep in mind that users of the forum constitute a particular sample of people that is likely not representative of many populations of interest.

8 Conclusion and Future Work

We introduce findings from moral cognitive science and psychology and assess their application to a forum of user-generated ethical situations. Statistical tests confirm that there are significant differences in usage of first-person passive voice along with first-person agents and patients among situations of different verdicts. Incorporating these differences as features in a verdict prediction task confirms the linkage between first-person agents and patients with assigned blame, though passive voice features appear too sparse to yield meaningful results.

From this study, we conclude that the manner in which a situation is described does appear to influence how blame is assigned. In the forum we work with, people seem to be judged by the way they present themselves, not just by their content, which aligns with previous cognitive science studies. Future endeavors in ethical AI could incorporate such theories to promote interpretability of models that produce moral decisions.

There are several areas of this project that could be refined and pursued further. We can repeat these experiments with the other verdicts, incorporating situations where all parties or no parties are blamed. We can use stricter length control than the sentence-level comparison, since the average sentence length still differs between posts of different verdicts. We should also incorporate validation that the SRL methodology effectively extracts the moral agents and patients we are trying to analyze. Another direction we would like to pursue, and one also mentioned by a reviewer, is to group situations by topic to try to control for other confounds in the moral situations. Finally, we hope to be able to incorporate the range of votes from the comments accompanying each post to allow for more nuanced verdict prediction, as done with SCRUPLES in Lourie et al. (2021).

Acknowledgements

We thank the anonymous reviewers, Yoav Artzi, and Rishi Advani for their generous feedback and suggestions. This work was supported in part by ARO MURI grant ARO W911NF-19-0217.

References

Gerd Bohner. 2002. Writing about rape: Use of the passive voice and other distancing text features as an expression of perceived responsibility of the victim. *British Journal of Social Psychology*, 40:515–529.

Nicholas Botzer, Shawn Gu, and Tim Weninger. 2021. Analysis of moral judgement on Reddit. https://arxiv.org/abs/2101.07664.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *CoRR*, abs/2012.15738.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A deep semantic natural language processing platform.

Kurt Gray and Daniel M. Wegner. 2009. Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of Personality and social Psychology*, 96 3:505–20.

Kurt Gray and Daniel M. Wegner. 2011. To escape blame, don't be a hero—Be a victim. *Journal of Experimental Social Psychology*, 47(2):516–519.

Daniel Jurafsky and James H. Martin. 2019. *Speech and Language Processing*, 3 edition. Prentice Hall.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *AAAI*.

Laura Niemi and Liane Young. 2016. When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin*, 42.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Chelsea Schein, Amelia Goranson, and Kurt Gray. 2015. The uncensored truth about morality. *The Psychologist*, 28(12):982–985.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. https://arxiv.org/abs/1904.05255.

A Additional Examples

Warning: Some content in these examples may be offensive or upsetting.

Figure 5 shows an example where there was relative disagreement about the guilty party. Figure 6 shows an example where there was general consensus about the verdict.

Title: **AITA For not caring about gay pride?**

Situation: I had thought I recently met two lovely people who happen to be homosexual. However, as I got to know them I soon began to dislike them. It seemed like they had to make it painfully obvious that they were gay. I felt like every other sentence they tried to squeeze in "Being a gay woman." and various other things such as that. it made me very uncomfortable, I don't have an issue with gay people but I wouldn't go as far as to say as I support them. We're all human, if you're gay then you're gay. Just keep me out of it, is all I ask. After this continued for awhile I decided to try and subtly hint at them that they were mentioning they were gay far too frequently.

When that didn't work, I resorted to telling them my beliefs on the subject and I was quickly resented, being called homophobic and unaccepting of them. (Even though I /clearly/ said I do accept them for their sexuality. I would just appreciate it if they didn't annoyingly mention it ever few minutes! Am I the asshole for this?

Top verdict: **YTA**. Start paying attention to how often hetero people talk about their dating, mating, and flirtations. Spoiler: ALL the time, but no one ever thinks they're "flaunting their heterosexuality".

Additional comment: **[NTA],** it's annoying when people make sexuality... Or any singular trait as the sole defining characteristic they have.

Figure 5: A situation where there was noticeable disagreement among comments. Depicted is the top-rated comment and one additional, contrary opinion.

Title: **AITA for disagreeing with my husband in front of my son?**

Situation: We (husband 30s, me 20s, and my son 8) were sitting on the couch talking about how technology has changed the world.

My husband says, “Technology is really bad.”

I say, “It’s not that bad, as long as you don’t overuse it.” That’s literally all I said.

After the kids went to bed, my husband blew up at me. He told me not to demean him in front on the kids. He called me names, threatened divorce, and told me he’s a saint for staying married to me.

His blow up was, in my eyes, completely unwarranted. He did apologize, but said if I hadn’t voiced my disagreement in front of my son, he wouldn’t have blown up.

I’ve been profusely apologizing for disagreeing with him in front of the kids. He’s convinced me that I was in the wrong.

I should have watched my mouth. If I had, he wouldn’t have gotten mad at me.

Top verdict: **NTA** - That’s called a normal conversation. If you husband thinks you can’t have your own opinions there is a problem. Your son should see that there are different perspectives and that his dad isn’t the be all end all.

Additional comment: **NTA**. Your husband sounds nuts. Does he keep his mouth shut when he disagrees with you in front of the kids? It's normal to have healthy disagreements in front of kids and it's good to model that for them. If your husband has a pattern of this kind of behavior, he sounds emotionally abusive. If this is an isolated incident, watch closely to see if it repeats itself. This is not normal. And he isn't really going to divorce you over that. He wants you to be scared to disobey his rule again.

Figure 6: A situation wherein other comments in general agreement with the final verdict (i.e., that of the top-rated comment). We show only one additional comment due to space constraints.