Data-driven Understanding of Real-life Moral Dilemmas

Josh Nguyen, Ziyu Chen, Georgiana Lyall, Alasdair Tran, Nicholas Carroll,
Minjeong Shin, Colin Klein and Lexing Xie

Australian National University

HMI

Background and Objectives

Moral dilemmas in daily life contain complexities and nuances typically overlooked in philosophy (e.g., trolley problem vs. marriage issues).

This project aims to explore:

- the **sources** of everyday moral conflicts;
- **how** moral issues are discussed and evaluated online; and
- the **differences** in moral framing and judgment from online discourse.

Study Domain

Reddit's r/AmItheAsshole

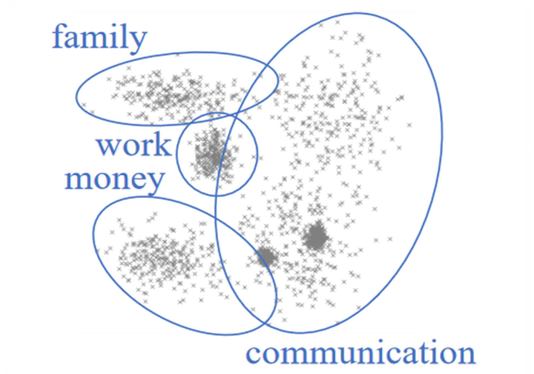
- 3M followers
- One of Reddit's most discussed forums.
- Author **posts** about an interpersonal conflict.
- Other Redditors **judge** the author in the comments: who is in the wrong and why?
- Community **upvotes** on comment.
- The most voted comment is the **verdict**.

Methodology

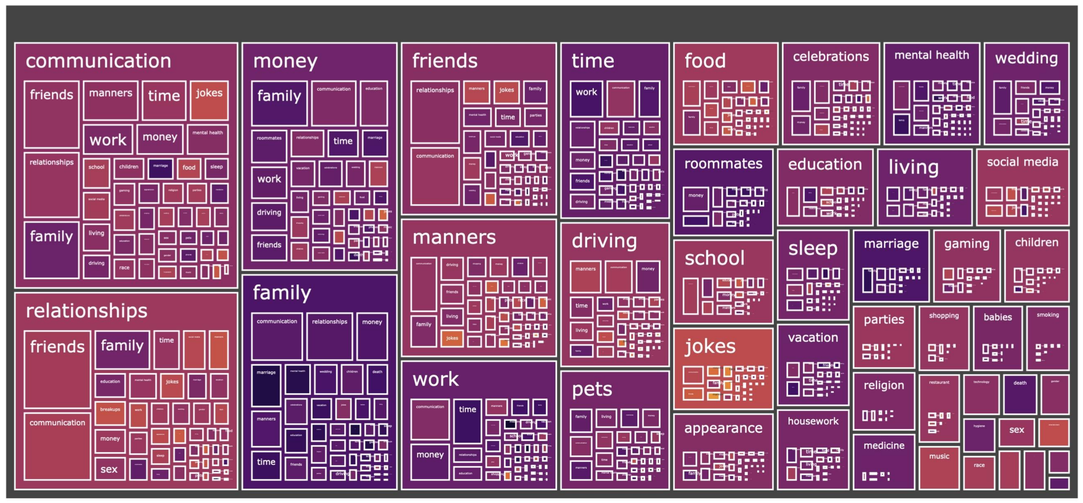
- Collected a dataset of **100K moral discussions** and 8M comments from Reddit's r/AmItheAsshole.
- Used **topic modeling** and designed two stages of **human validation** to discover and curate a list of prevalent topic/moral "domains."
- Measured and compared the **moral valence** in the topics using extensively validated lexicons.

Findings

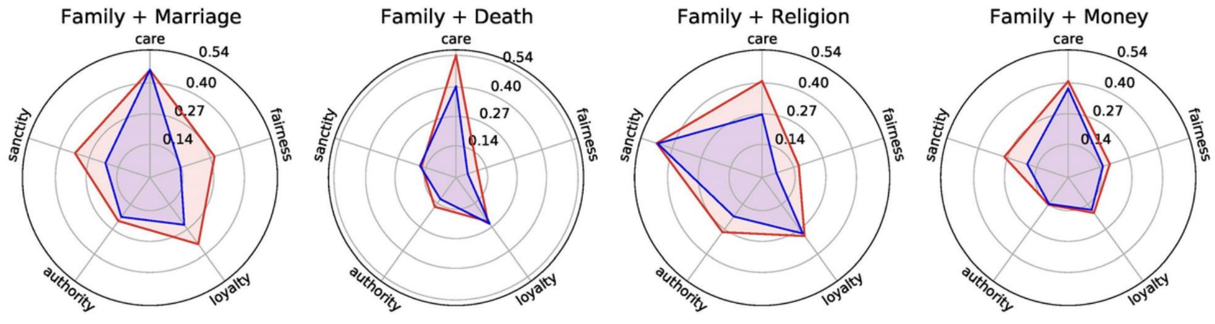
- Discovered 47 fine-grained **topics** of discussion from 100K posts
- Most topics typically fall into **non-traditionally moral domains**, e.g., *work*, *celebrations* and *communication*.



- Humans tend to perceive dilemmas in **pairs of topics**, e.g., *family* and *religion*. (See the tree map on the right.)
- The **language** used in moral framing in posts and judgments varies among topics and topic pairs. For example, there is much more emphasis on the notions of *loyalty* and *sanctity* in religion-themed judgments.



Topics and topic pairs (nested blocks) discovered on r/AmItheAsshole. Block size represents the number of posts and block shade represents moral valence (lighter = more often negatively judged).



Prevalence of moral foundations (*care*, *fairness*, *loyalty*, *authority*, and *sanctity)* in some moral judgments on r/AmItheAsshole.

Red = negative judgment
Blue = positive judgment

Conclusion

- Daily life presents moral conflicts that are much **broader** and **more low-stakes** than idealized moral dilemmas in philosophy.
- Most moral conflicts involve at least **two topics**, such as *manners* and *money*.
- Topics in a pair interact in **non-trivial** ways, especially w.r.t. the framing and judgment of a moral dilemma.
- Current extensively validated **lexicons** (moral foundations, LIWC, empath) are useful for analyzing moral dimensions in these discussions.

Ongoing Work

It remains to be seen how **effective** the moral foundations dictionary and related lexicons are on other large-scale datasets.

Hypothesis: existing **taxonomies** of moral intuitions (like the moral foundations theory or morality-as-cooperation) help explain people's diverse and often conflicting judgments on a range of moral issues.

Questions:

- How can we accurately and consistently **identify** moral foundations within real-world text data?
- Do certain moral foundations play any **role** in political leaning or reveal insights about demographic groups online?
- Do **other moral taxonomies** emerge from these datasets?

Reference

Nguyen, TD, Lyall, G, Tran, A, Shin, M, Carroll, NG, Klein, C and Xie, LX. "**Mapping Topics in 100,000 Real-Life Moral Dilemmas**". In: *Proceedings of the International AAAI Conference on Web and Social Media* 16.1 (2022), pp. 699–710.