



The Royal Infirmary of Edinburgh
Anderson



TEXT DATA ANALYSIS

SUMMER SCHOOL

EDINBURGH, JUNE 05-09 2023

SPONSORED BY



Sgoil Cheumnaichean Saidheans



HOUSE KEEPING



- Toilets
- Food Consumption
- Water Fountains
- Fire Alarm
- Code of Conduct





TODAY'S SCHEDULE

Seminar: Extracting Latent Moral Information from Text with ChatGPT

Hands-on session 1: Text Analysis 1

Hands-on session 2: Text Analysis 2

BYOD Session: 2



EXTRACTING LATENT MORAL INFORMATION FROM TEXT WITH CHATGPT

<https://justinchuntingho.github.io/DCS-slides/>

Dr Justin Chun-Ting Ho,

Postdoctoral fellow at Academia Sinica, the national
academy of Taiwan



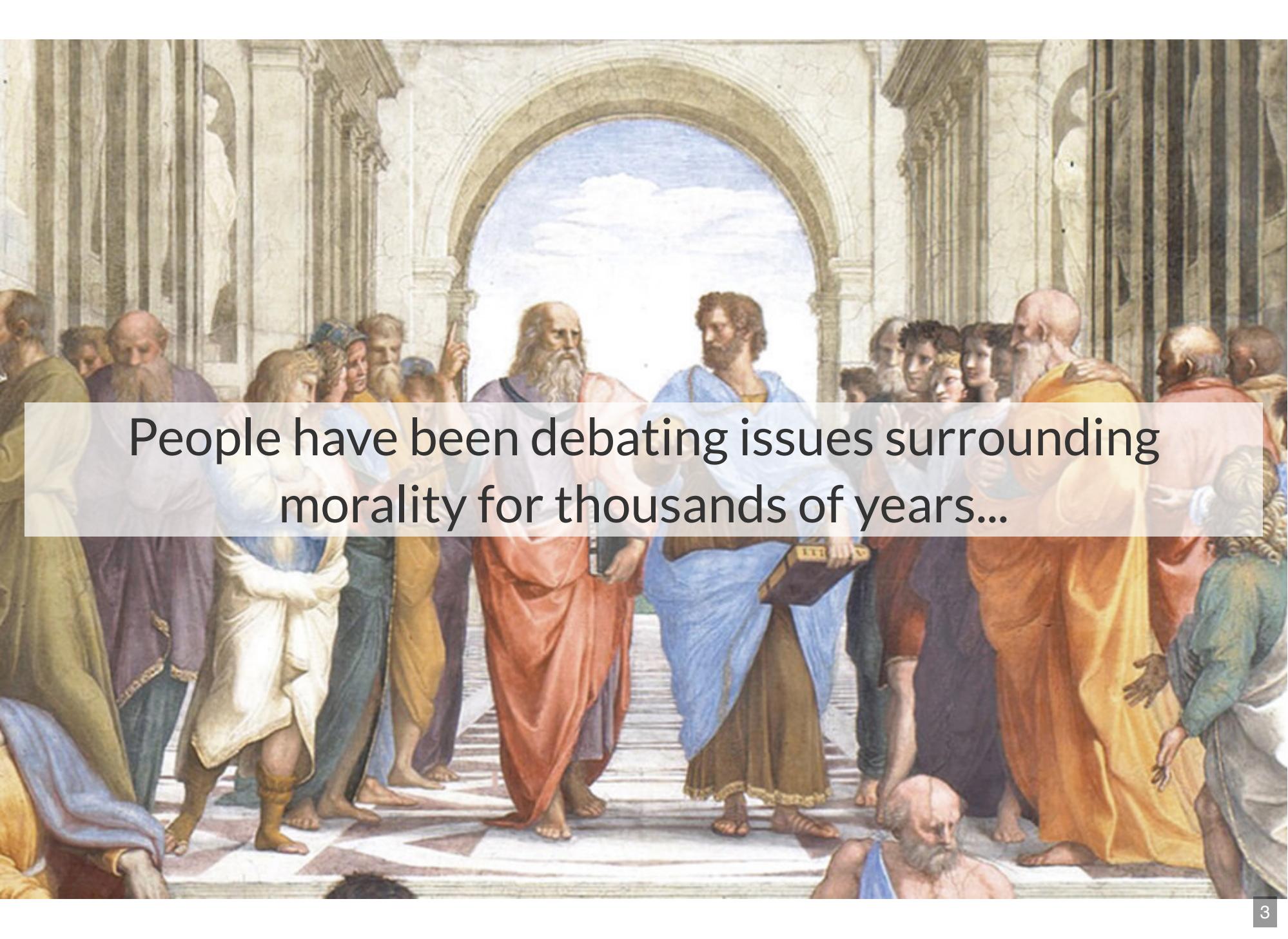
EXTRACTING LATENT MORAL INFORMATION FROM TEXT WITH CHATGPT

JUSTIN CHUN-TING HO
ACADEMIA SINICA

IN COLLOBRACTION WITH

**YOU-JIAN WU, YUAN HSIAO, CHINGCHING CHANG, HEN-HSEN HUANG,
SHIH-PENG WEN, HAO-HSUAN WANG, YU-MING HSIEH**

LET'S TALK ABOUT MORALITY

A classical painting depicting a group of philosophers in a portico. In the center, an elderly man with a long white beard, wearing a purple robe, stands with his hands clasped. To his right, another man in a blue robe points upwards. To his left, a man in a red robe looks towards the center. The background features a large archway leading to a bright sky with clouds. The foreground is filled with other figures, some looking towards the central group, others engaged in conversation. The style is characteristic of Renaissance or Neoclassical art.

People have been debating issues surrounding
morality for thousands of years...

孔子向道圖
相傳孔子曾問道於老子
作于西漢
畫于西漢



in every corner of the world...



HERE'S A PROBLEM

PEOPLE RARELY AGREES ON...

- What is the right thing to do?
- How to do?
- What is relevant to morality?
-

HOW ABOUT WE FOCUS ON THE FUNCTION?

Moral systems are interlocking sets of values, practices, institutions, and evolved psychological mechanisms that work together to suppress or regulate selfishness and make social life possible.

-- Jonathan Haidt, 2008

TWO KINDS OF MORAL SYSTEMS

- Individualizing: protecting individuals directly (eg using legal system)
- Binding: binding individuals into roles and duties to constrain their selfishness

HOW ABOUT WE TAKE ONE STEP FURTHER?

MORAL FOUNDATION THEORY

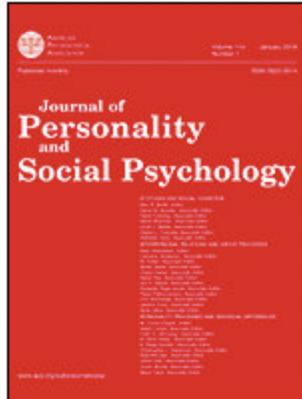
Moral foundations are...

- Innate
- Universal
- Intuitive



MORAL FOUNDATIONS

- Care/harm: care, sympathy, and compassion
- Fairness/cheating: rights and justice
- Loyalty/betrayal: patriotism and group identity
- Authority/subversion: social order and traditions
- Sanctity/degradation: disgust and spiritual concerns about the body



[Journal Information](#)
[Journal TOC](#)

[Search APA PsycNet](#)

APA PsycArticles: Journal Article

Liberals and conservatives rely on different sets of moral foundations.

[© Request Permissions](#)

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029–1046.

<https://doi.org/10.1037/a0015141>

How and why do moral judgments vary across the political spectrum? To test moral foundations theory (J. Haidt & J. Graham, 2007; J. Haidt & C. Joseph, 2004), the authors developed several ways to measure people's use of 5 sets of moral intuitions: Harm/care, Fairness/reciprocity, Ingroup/loyalty, Authority/respect, and Purity/sanctity. Across 4 studies using multiple methods, liberals consistently showed greater endorsement and use of the Harm/care and Fairness/reciprocity foundations compared to the other 3 foundations, whereas conservatives endorsed and used the 5 foundations more equally. This difference was observed in abstract assessments of the moral relevance of foundation-related concerns such as violence or loyalty (Study 1), moral judgments of statements and scenarios (Study 2), "sacredness" reactions to taboo trade-offs (Study 3), and use of foundation-related words in the moral texts of religious sermons (Study 4). These findings help to illuminate the nature and intractability of moral disagreements in the American "culture war." (PsycInfo Database Record (c) 2020 APA, all rights reserved)

POLITICAL PHILOSOPHY

- Liberals: optimistic view of human nature, people should be left free to pursue personal development
- Conservatives: pessimistic view of human nature, people are inherently selfish and imperfectible, constraints of authority, institutions, and traditions are necessary

PERSONALITIES

- Liberals: open to experience, more inclined to seek out change and novelty
- Conservatives: stronger preference for the familiar, stable, and predictable

BUT WHY?

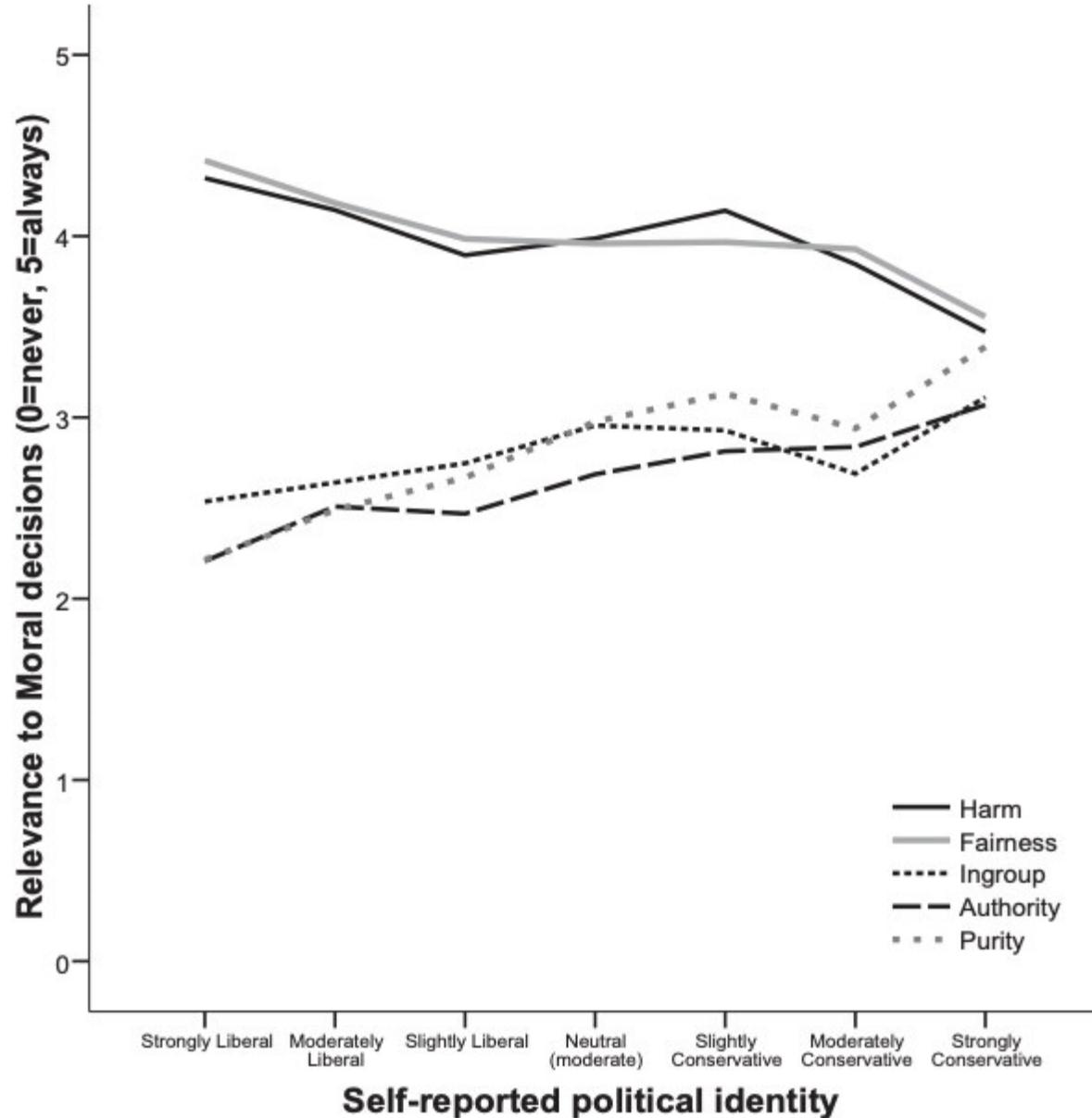


Figure 1. Relevance of moral foundations across political identity, Study 1.

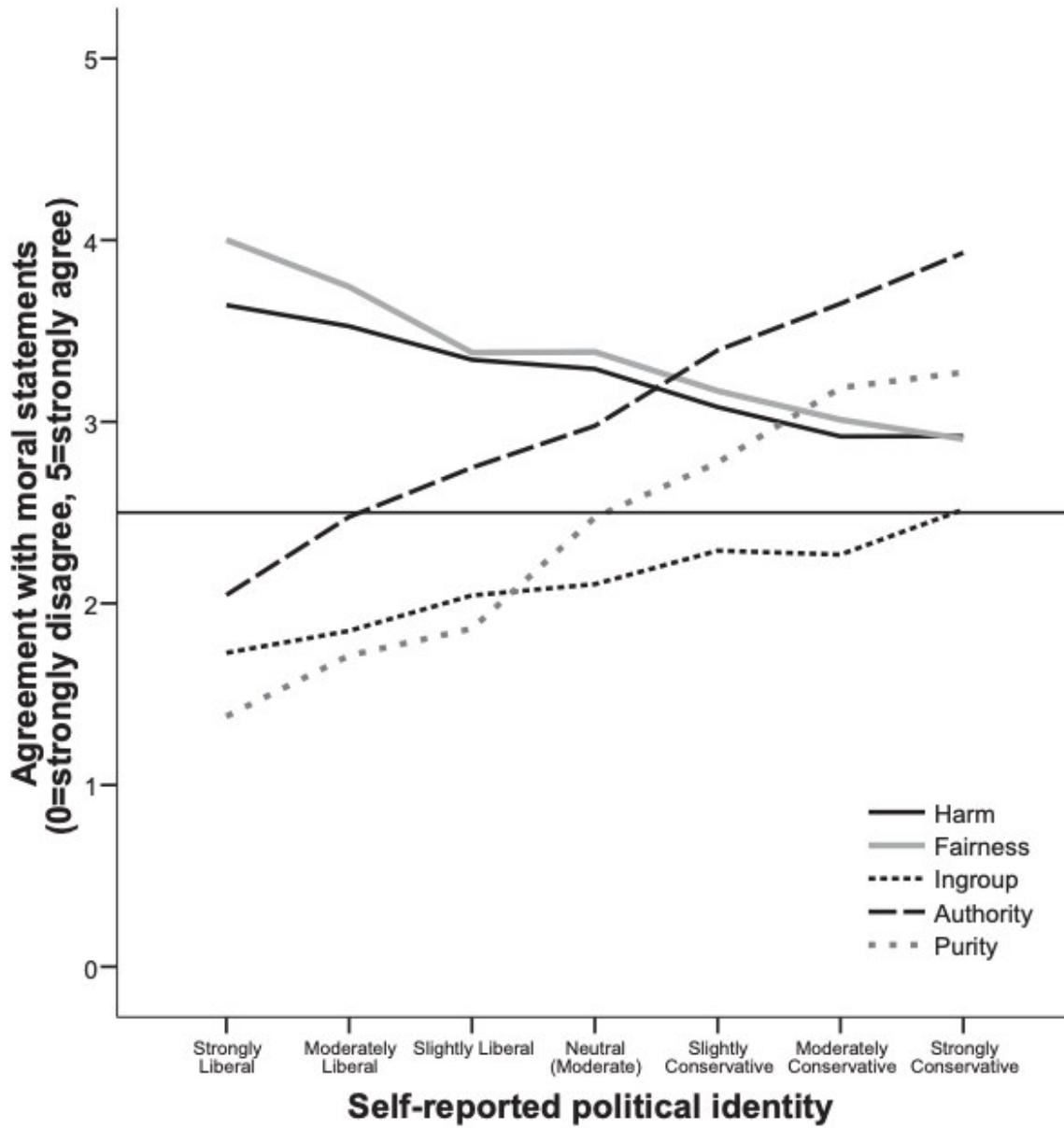


Figure 3. Agreement with moral statements across political identity, Study 2. The horizontal line at 2.5 indicates division of agreement and disagreement (2 indicates slight disagreement and 3 indicates slight agreement).

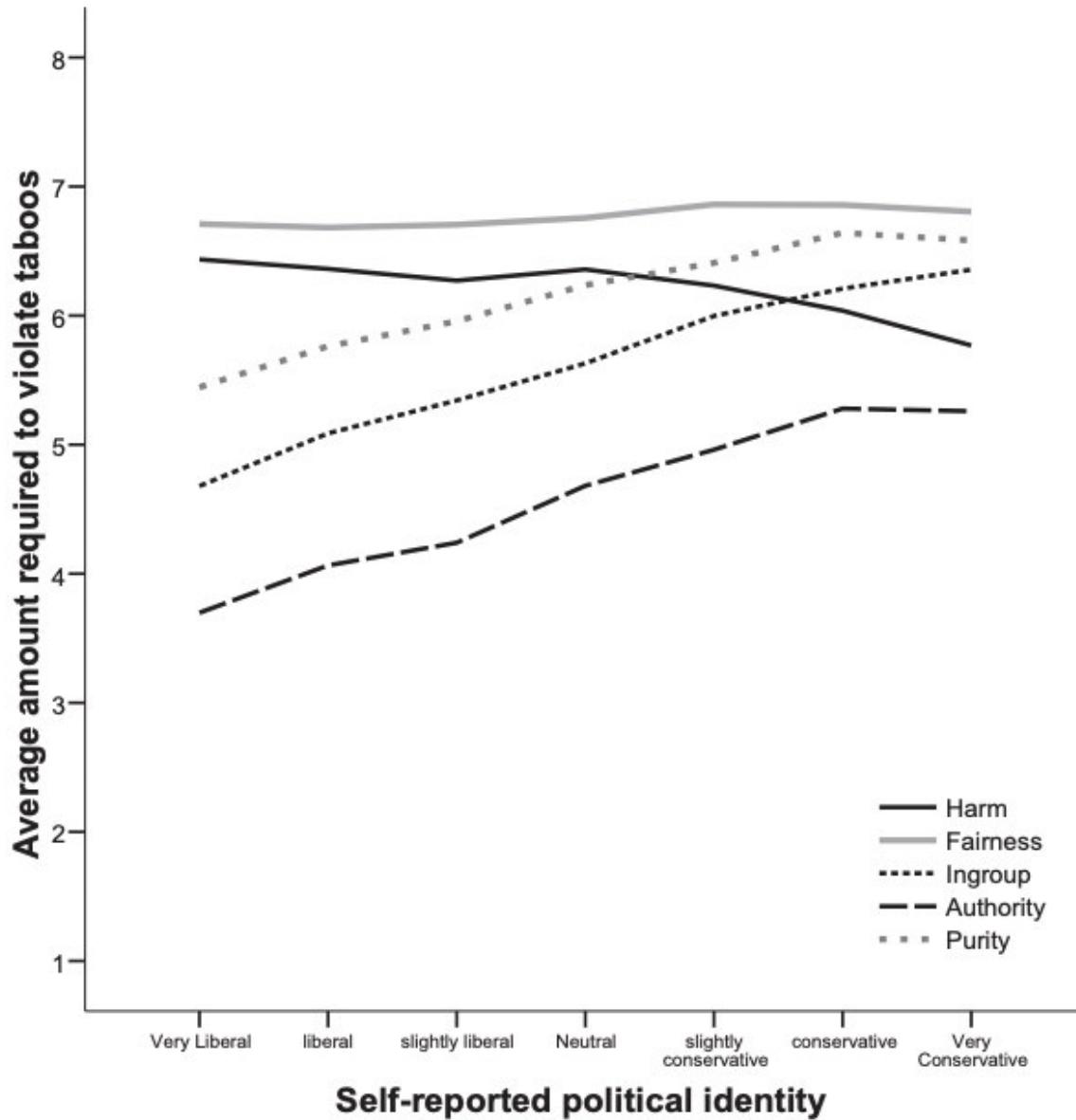


Figure 4. Average amount required to violate foundation-related taboo trade-offs across political identity, Study 3. Response scale was as follows: 1 = \$0 (*I'd do it for free*), 2 = \$10, 3 = \$100, 4 = \$1,000, 5 = \$10,000, 6 = \$100,000, 7 = *a million dollars*, 8 = *never for any amount of money*.

Table 1
*Raw Percentages of Foundation-Related Words in Liberal and
 Conservative Sermons, Study 4*

Foundation	Percentage		<i>t</i>	Effect size (<i>d</i>)
	Liberal sermons	Conservative sermons		
Harm	0.44	0.26	-2.71**	0.56
Fairness	0.22	0.10	-3.04**	0.65
Ingroup	0.52	0.22	-6.65***	-1.27
Authority	0.46	0.98	3.42***	0.81
Purity	0.25	0.64	4.40***	0.99

Note. Percentages represent the number of times any word in the dictionary for that foundation was used, divided by the total number of words in the entire liberal or conservative corpus, then multiplied by 100. A negative effect size indicates the effect was opposite to prediction.

** $p < .01$. *** $p < .001$.

MEASURING MORAL FOUNDATIONS (FROM TEXT)

COMMON APPROACHES

1. Moral Foundations Dictionary
2. Content Analysis
3. Crowdsourcing

MORAL FOUNDATIONS DICTIONARY

- Accessible **Here**
- Theoretical concern:
Intuitive Reactions vs Thorough Deliberation
- Domain-specificity Problem
- Language Dependence

CONTENT ANALYSIS

- Coder training (intensive vs minimal)
- Low intercoder reliability
- Single ground-truth of morality?
- Moral judgement as fast and unconscious process

CROWDSOURCING

- Crowd-truth paradigm
- Human codings are inherently subjective
- Crowdsourcing platform is not always available for everywhere

WHAT ABOUT CHATGPT?

Step 1

Collect demonstration data and train a supervised policy.

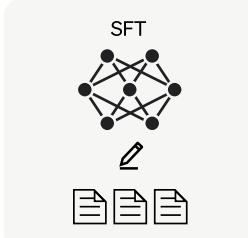
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

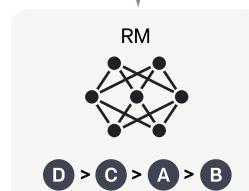
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



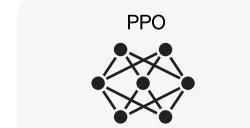
Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

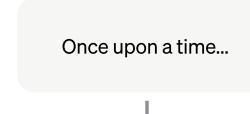
A new prompt is sampled from the dataset.



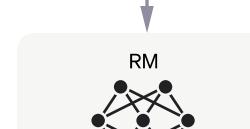
The PPO model is initialized from the supervised policy.



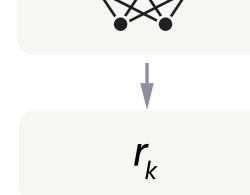
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



DOES CHATGPT KNOW MORAL FOUNDATION THEORY?

J Hi



Hello! How can I assist you today?



J do you understand moral foundations theory



Yes, as an AI language model, I have been trained on a vast range of topics, including moral foundations theory. Moral foundations theory is a social psychological theory that attempts to explain the origins of human moral reasoning. The theory posits that there are five primary moral foundations, which are:



1. Care/harm: the desire to protect and care for vulnerable individuals, as well as to reduce suffering and promote well-being.
2. Fairness/cheating: the concern for justice and reciprocity in social exchanges.
3. Loyalty/betrayal: the importance of group cohesion, loyalty, and patriotism.
4. Authority/subversion: the respect for legitimate authority and the desire for order and stability.
5. Sanctity/degradation: the aversion to physical and spiritual impurities, and the valuation of purity and cleanliness.

According to this theory, people prioritize these moral foundations differently, depending on factors such as culture, upbringing, and individual differences.

J

show me an example for each foundations



Sure, here are some examples of behaviors or situations that might be related to each of the five moral foundations:



1. Care/harm: A person who volunteers at a homeless shelter or donates to a charity that helps sick children is prioritizing the care/harm moral foundation.
2. Fairness/cheating: A person who refuses to take credit for someone else's work or who speaks out against discrimination is prioritizing the fairness/cheating moral foundation.
3. Loyalty/betrayal: A person who fiercely defends their country or political party, or who remains loyal to their family or friends even when it's difficult, is prioritizing the loyalty/betrayal moral foundation.
4. Authority/subversion: A person who follows the rules and laws, respects police officers or other authority figures, or values social order and tradition, is prioritizing the authority/subversion moral foundation.
5. Sanctity/degradation: A person who avoids eating certain foods for religious reasons, who insists on cleanliness and hygiene, or who is strongly opposed to pornography or other forms of "impurity," is prioritizing the sanctity/degradation moral foundation.

J

for each foundations, draft a tweet that show it.



Sure, here are some sample tweets that illustrate each of the five moral foundations:



1. Care/harm: "It breaks my heart to see children suffering in hospitals. That's why I'm donating a portion of my paycheck every month to help fund research and care for sick kids. #care #charity #givingback"
2. Fairness/cheating: "It's not right that some people have more opportunities than others just because of their skin color or gender. We need to stand up against discrimination and fight for equality. #fairness #justice #equality"
3. Loyalty/betrayal: "I may not always agree with my family, but I'll always have their back. Blood is thicker than water, and loyalty means everything to me. #family #loyalty #brotherhood"
4. Authority/subversion: "I believe in respecting our leaders and institutions, even if we don't always agree with their decisions. We need order and stability to keep society functioning. #authority #lawandorder #respect"
5. Sanctity/degradation: "Our bodies are temples, and we need to treat them with respect. That's why I choose to eat clean, avoid drugs and alcohol, and strive for purity in all aspects of my life. #sanctity #purity #cleanliving"

METHODOLOGY

- Data: Headlines of major Taiwanese newspaper
- Dictionary: Translated version of MFD
- Content Analysis: Annotation by two highly trained experts
- Crowdsourcing: Annotation by 5 crowd-annotators, majority rule
- ChatGPT: Request through ChatGPT's API

INTERCODER RELIABILITY

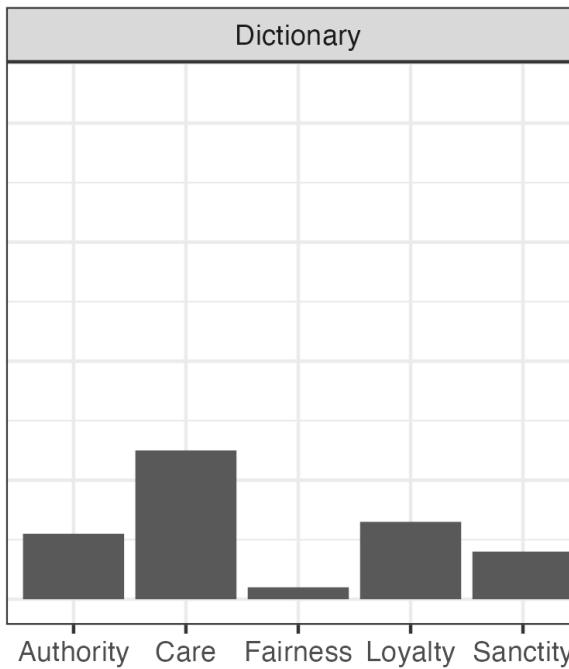
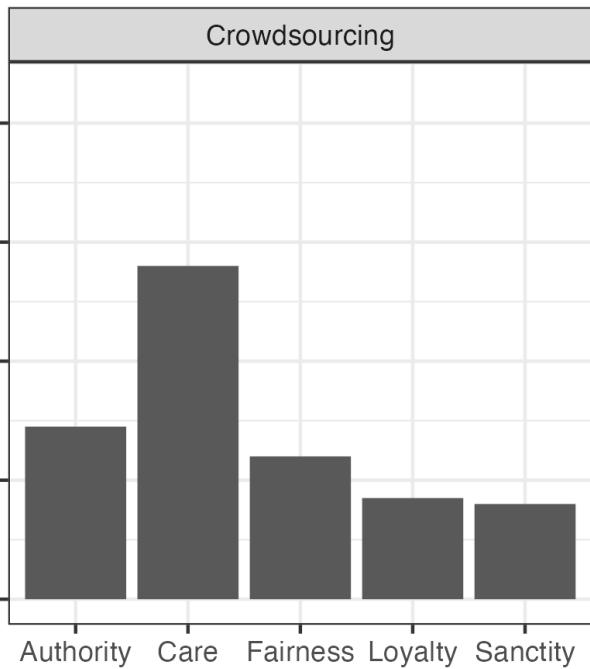
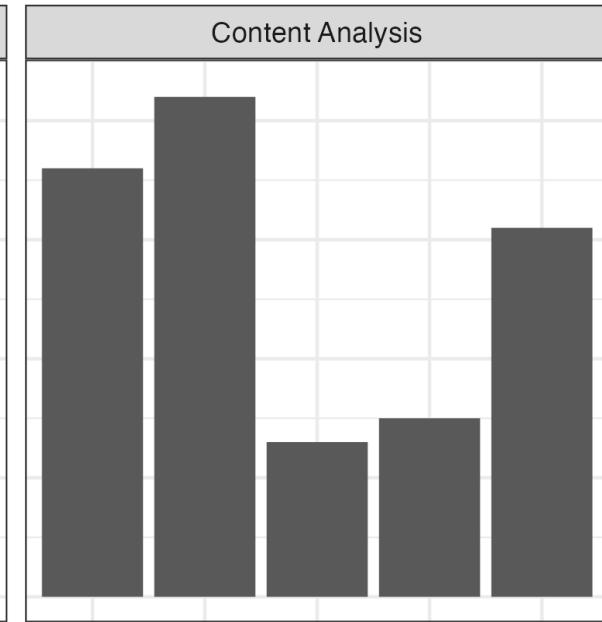
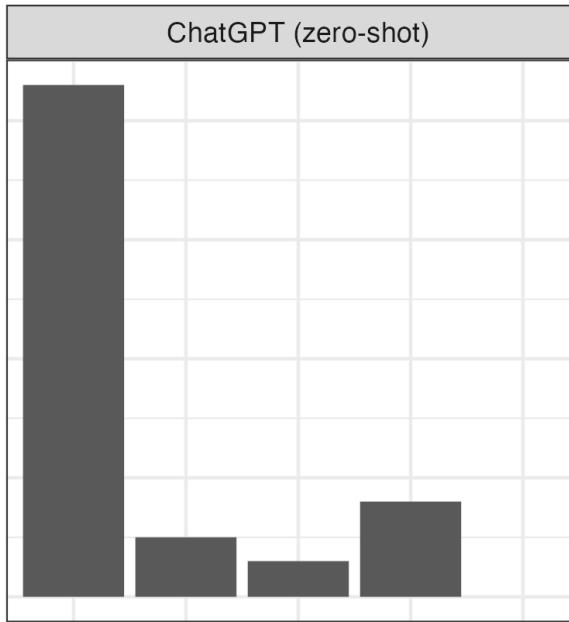
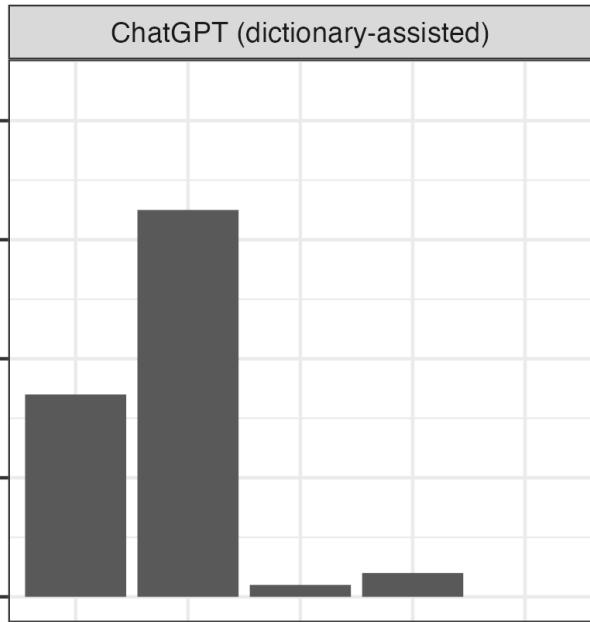
	Percentage Agreement	Krippendorff's alpha
Authority	0.66	0.295
Care	0.78	0.551
Fairness	0.87	0.478
Loyalty	0.66	0.0979
Sanctity	0.74	0.27

PROMPT (ZERO-SHOT):

According to the Moral Foundations Theory, does the following sentence express any of the moral foundations, provide explanation for each foundation. Answer in 0 and 1 as a csv containing these columns (Text, Care, Fairness, Loyalty, Authority, Sanctity, Explanation, Confidence), also report confidence score, put double quotes around string, return only the content of the csv, no extra text:

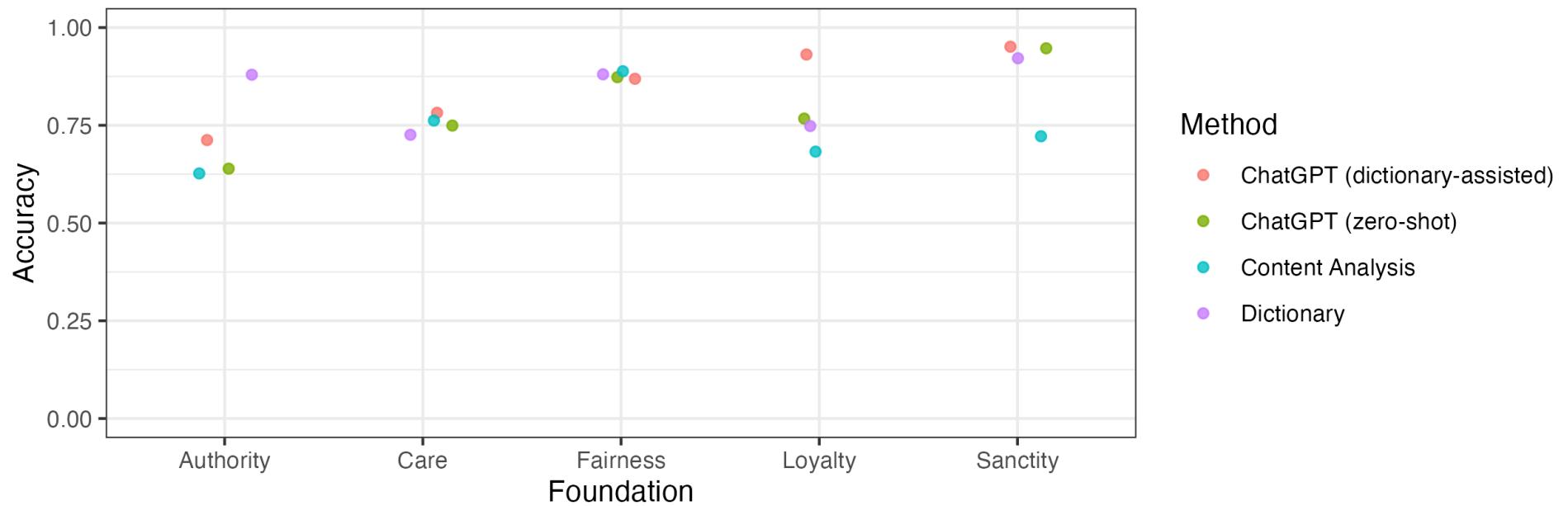
PROMPT (DICTIONARY-ASSISTED):

Does the following sentence express the moral foundation of Care/harm, defined as safe*, peace*, compassion*, empath*, sympath*, care, protect*, shield, shelter, amity, secur*, benefit*, defen*, guard*, preserve, harm*, suffer*, war, fight*, violen*, hurt*, kill, endanger*, cruel*, brutal*, abuse*, damag*, ruin*, ravage, detriment*, crush*, attack*, annihilate*, destroy, stomp, abandon*, spurn, impair, exploit, and wound. Answer in 0 and 1 as a csv containing these columns (Text, Harm), put double quotes around string, return only the content of the csv, no extra text:

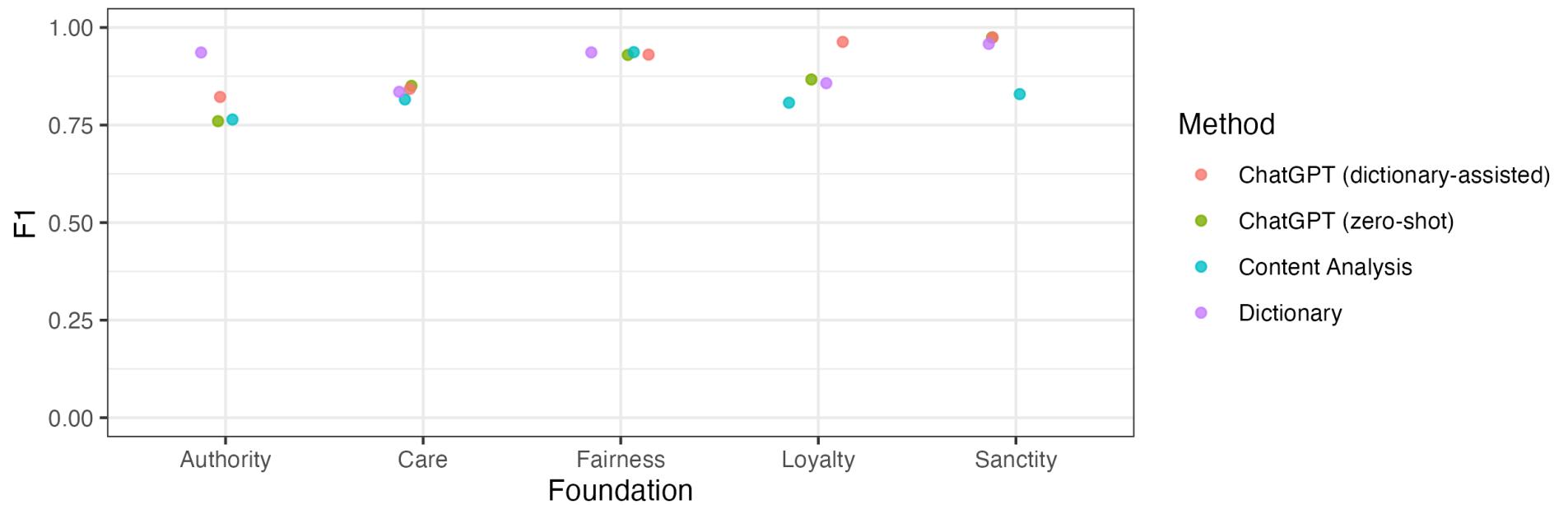


Authority Care Fairness Loyalty Sanctity

ACCURACY



F1 SCORE



REPRODUCIBILITY

	Percentage Agreement	Krippendorff's alpha
Authority	0.96	0.87
Care	0.91	0.791
Fairness	1	1
Loyalty	0.99	0.662
Sanctity	1	NA

THANK YOU!

TWITTER: @JUSTIN_CT_HO

GITHUB: JUSTINCHUNTINGHO

EMAIL: JUSTINHO@GATE.SINICA.EDU.TW



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



COFFEE BEAK

**WE ARE GOING TO RESTART AT
11:00**



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



TEXT ANALYSIS

JESSICA WITTE

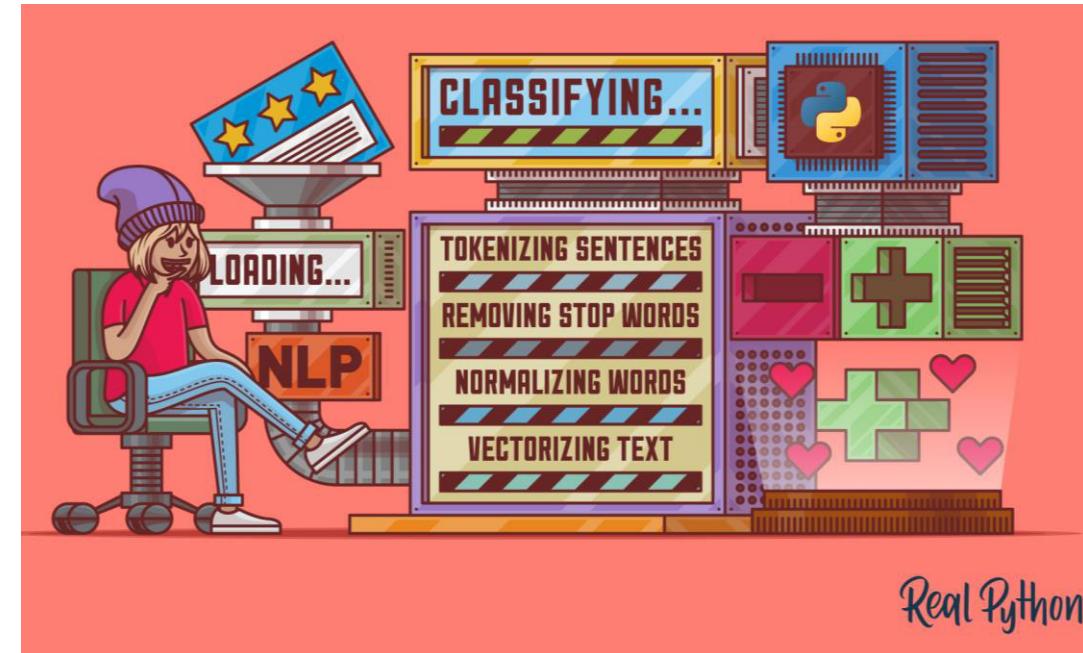


TEXT ANALYSIS

- Computationally evaluating, investigating, and exploring textual (natural language) data
- Supervised and unsupervised methods
- Topic modelling, named entity recognition (NER), sentiment analysis, text classification

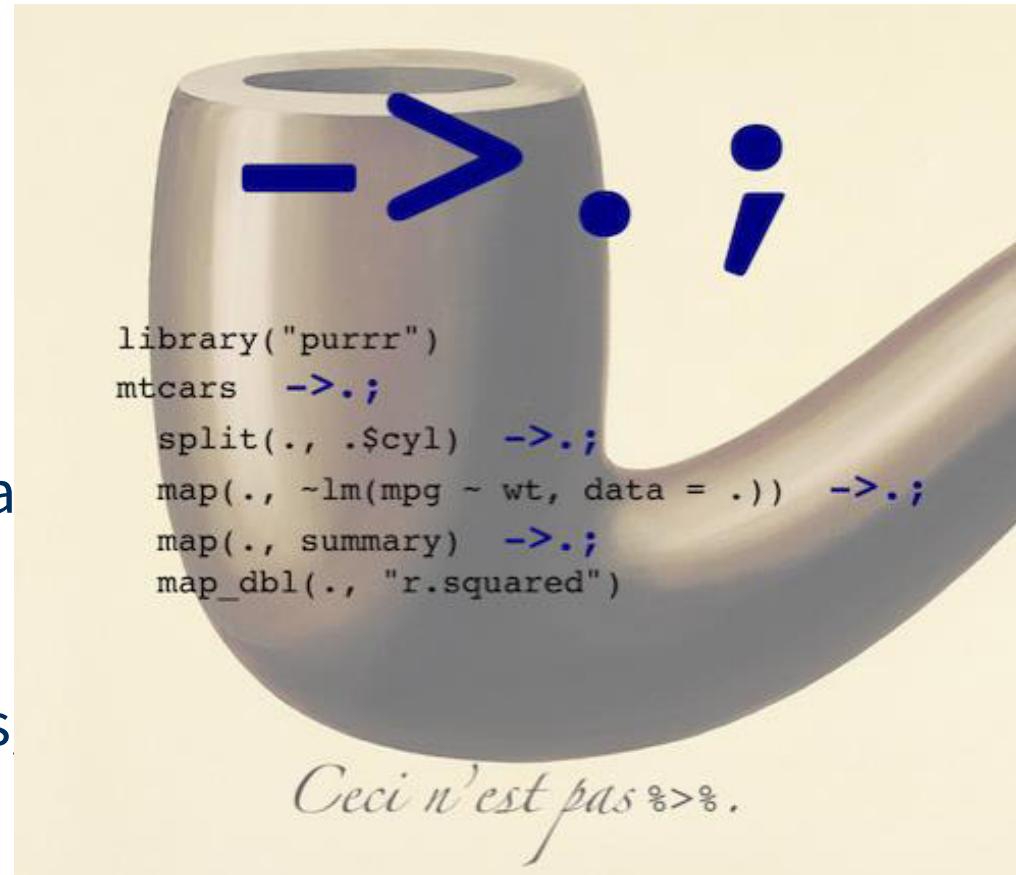
Some caveats:

- Training data matters
- Computers “read” differently than we do
- Potential for quantitative/qualitative mismatch in tools, results, statistical tests, etc.



WHAT IS TEXT MINING?

- Analysing textual data computationally
- A type of natural language processing (NLP)
- Natural languages=human languages (e.g. English, Arabic)
- Aims to identify patterns and trends in natural language data
- Methods from machine learning, linguistics, statistics, information studies, and other disciplines



SOME METHODS FOR TEXT MINING

- No-code web applications (e.g. [Voyant tools](#))
- Keyword search/term frequency
- Term frequency-inverse document frequency (TF-IDF)
- Named entity recognition (NER)
- Topic modelling
- Sentiment analysis

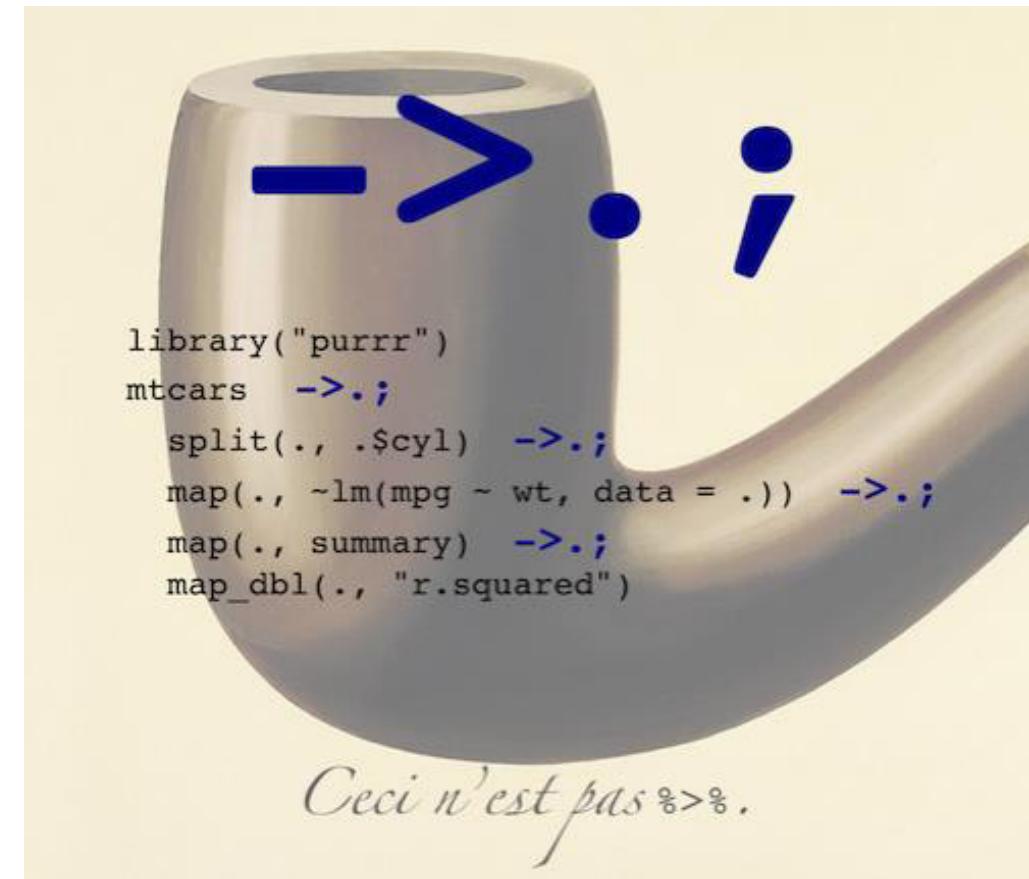
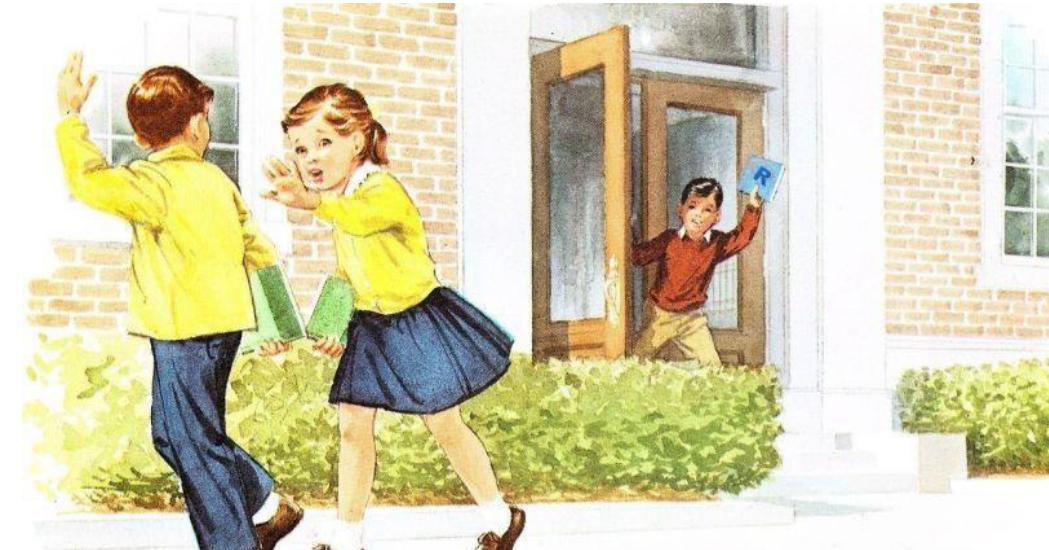


TABLE DISCUSSION

In the context of your current or future research, can you think of a potential use case where text mining methods might help you answer a research question?



Run, or he's going to tell us about again!

R



LESSON OVERVIEW

- **Datasets:** text about the cost of living scraped from the Scottish and UK governments' websites
- **Research questions:**
 - 1.What are some common themes that appear in the UK government's publications about the CoL?
 - 2.Comparing the UK and Scotland data sets, do we see any patterns/similarities/differences in themes?
- **Topics covered:** data cleaning and pre-processing, constructing word clouds, corpus construction



PRE-PROCESSING

Tokenization, Text Cleaning and Normalization



TOKENISATION

- Tokenization involves breaking down a piece of text into smaller units called tokens.
- Tokens can be individual words, sentences, or even characters, depending on the level of granularity desired.
- Tokenization helps in standardizing and organizing text data, making it easier to analyse and process.
- Word-based tokenization breaks down text into individual words, treating each word as a separate token

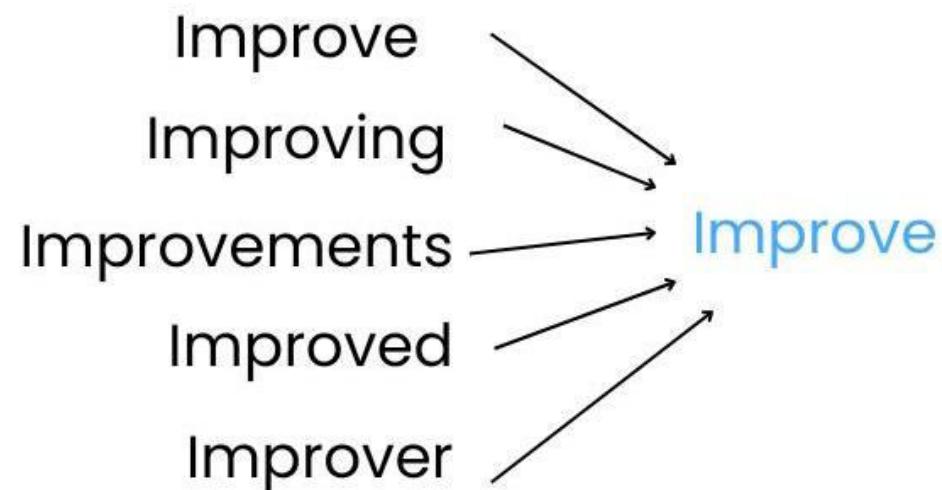
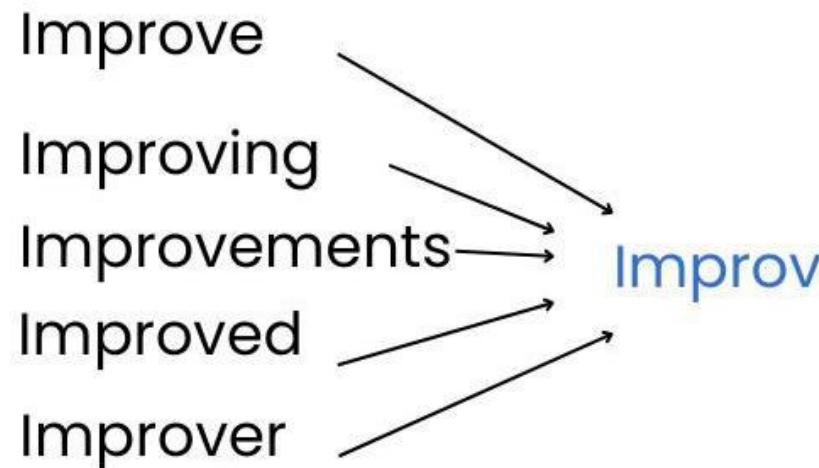


TEXT CLEANING & PREPROCESSING

- Formatting text for analysis and removing extraneous information
- R has several libraries for preprocessing (we will use tidytext, tm, and quanteda)
- Workflows vary depending on research objective, field, and dataset
- Common steps include standardising capitalisation, removing URLs and symbols, stopword removal, tokenisation, stemming, and lemmatization
- Stopwords include words like “a,” “the,” “of,” “an” that don’t add meaning to the dataset



STEMMING & LEMMATIZATION





THE UNIVERSITY *of* EDINBURGH
Centre for Data, Culture & Society



A blurred photograph of a person sitting at a desk, viewed from the side and back. They appear to be working on a laptop. The background is a warm, reddish-orange color.
TIME FOR R



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



LUNCH BEAK

**WE ARE GOING TO RESTART AT
13:30**



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



TEXT ANALYSIS

JESSICA WITTE



TOPIC MODELLING

- Unsupervised method
 - “Bag of words” (BoW) organised into a defined number of topics
 - Latent Dirichlet analysis (LDA) and latent semantic analysis (LSA)
 - Ideal for large, pre-processed datasets free from irony, sarcasm, humour, etc.
 - Example use cases:
 - Identifying common themes or concepts
Grouping documents based on similarity
 - Detecting changes in a dataset over time

Just bought a book from IKEA

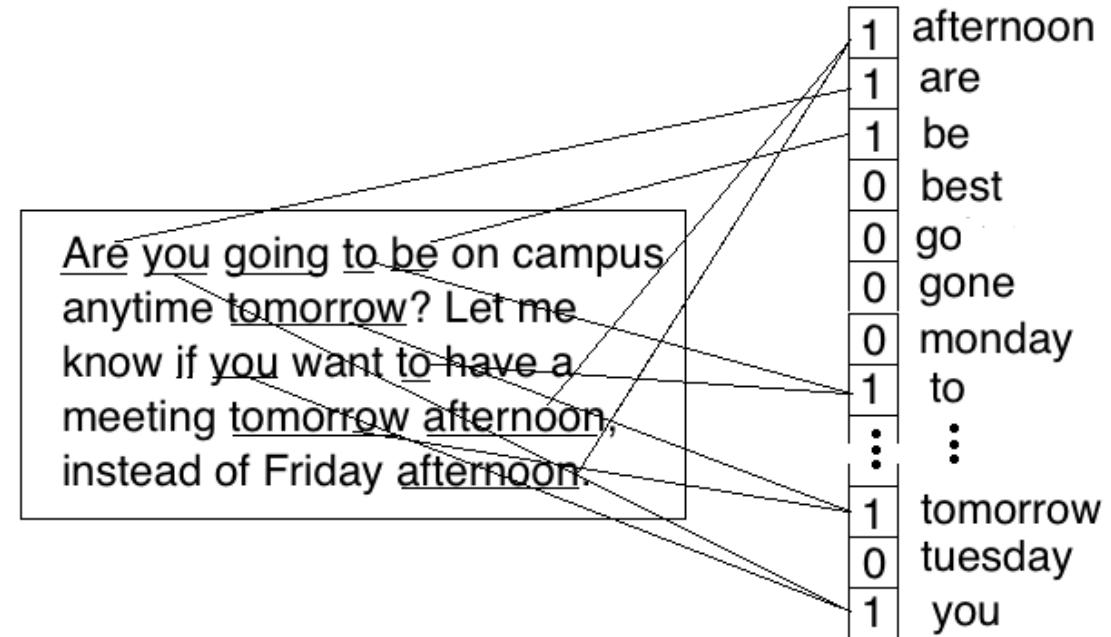


<https://ir.co.il/humor/humor001.htm>



CAN COMPUTERS READ?

- “Bag-of-words” ignores context
- Cleaning & pre-processing steps are essential, but workflows differ between datasets and research questions
- Linguistic nuance often ignored or misinterpreted
- Manually annotating a sample can diagnose errors and improve results



SUPERVISED & UNSUPERVISED LEARNING

- Two main methods for machine learning
- **Supervised learning:** trains algorithms using data annotated by humans
 - Classification (assigning text to specified categories)
 - Regression (predicting associations or relationships to specified categories)
- **Unsupervised learning:** trains algorithms using unlabelled raw data
 - Finding patterns in the text (grouping similar data, identifying deviations)
 - Exploratory overview of a large dataset

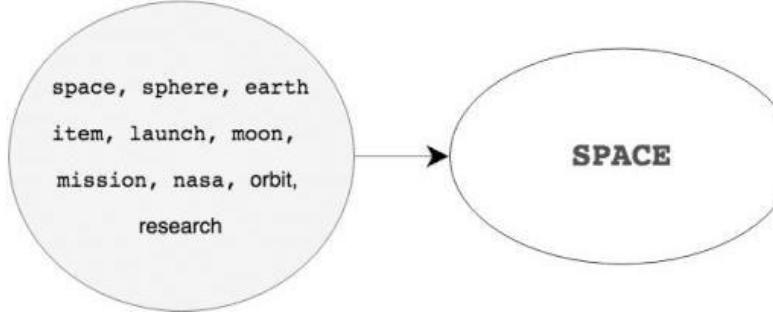
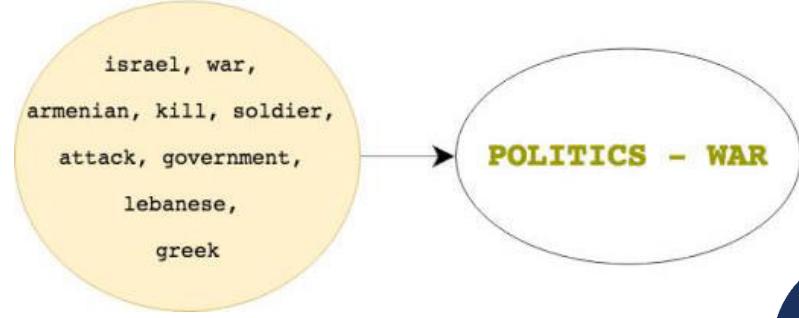
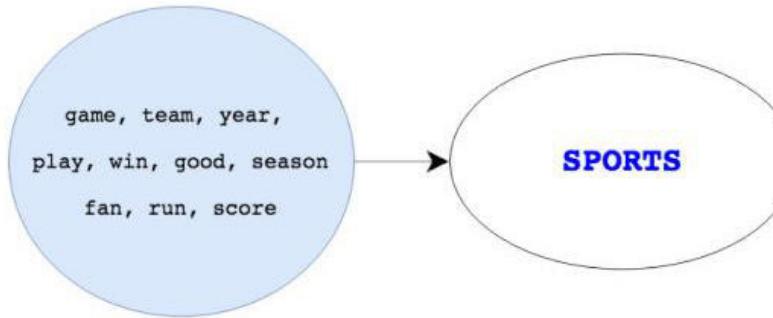
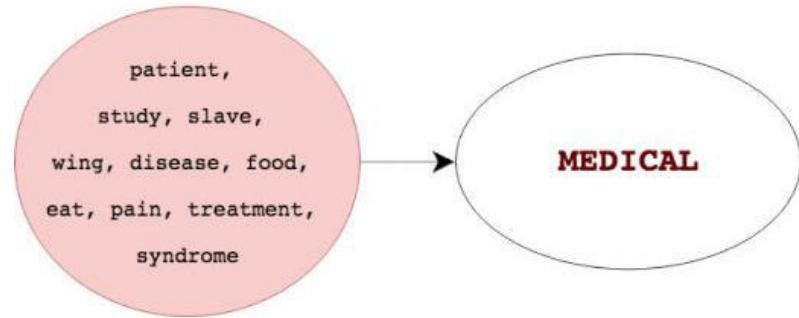
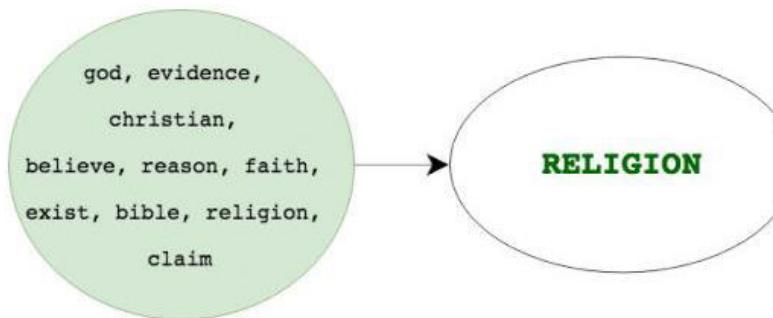
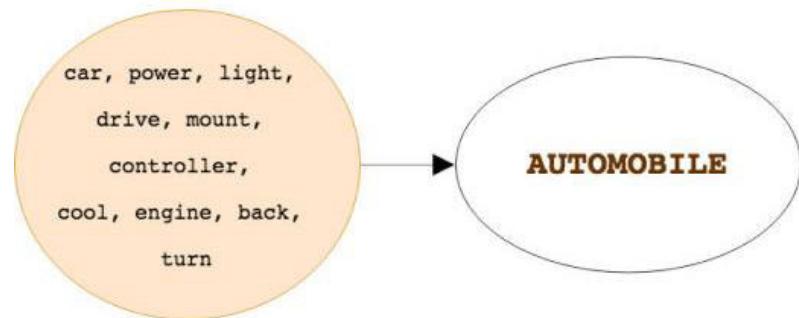


TOPIC MODELLING

- Unsupervised machine learning method (doesn't need training)
- Clusters words and phrases by “topic,” or loose themes/patterns
- Human researchers interpret the results
- Latent Dirichlet Allocation (LDA)
 - “latent” meaning “yet to be uncovered” or “hidden”
 - “Dirichlet” is a form of Bayesian statistical distribution
 - “allocation” or “putting things into categories”



INFERRING THE TOPIC FROM KEYWORDS



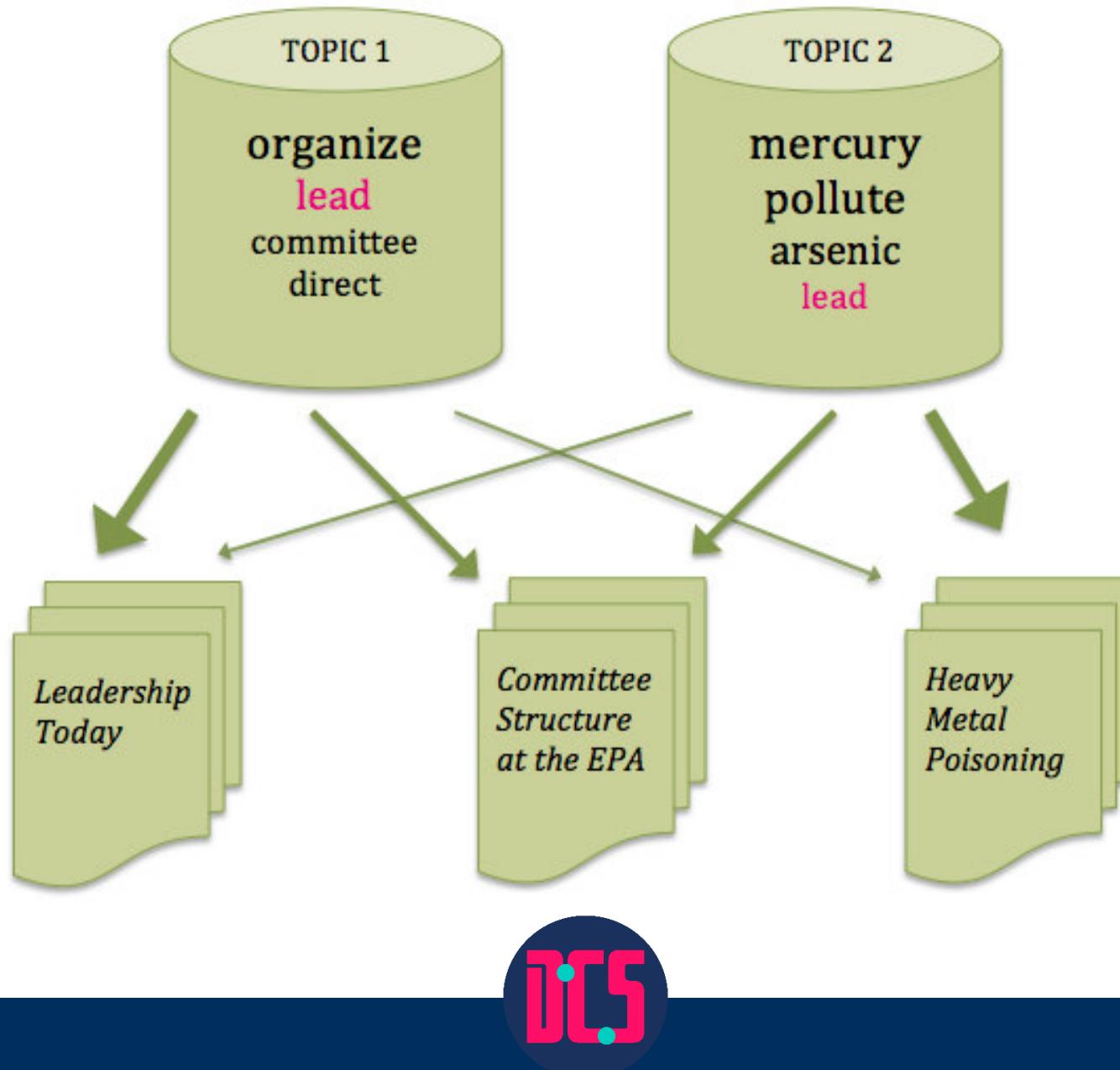
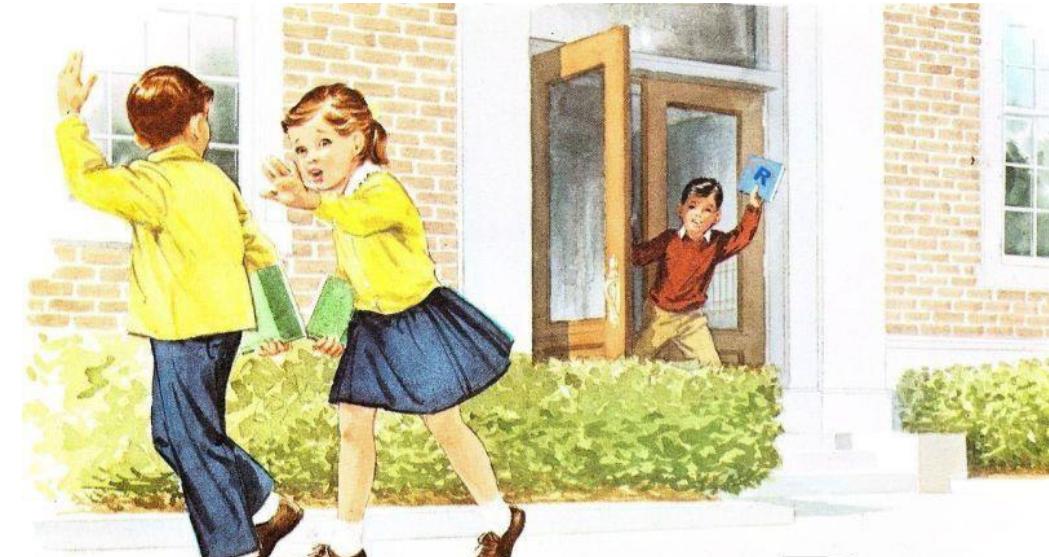


TABLE DISCUSSION

In the context of your current or future research, can you think of a dataset you might explore with topic modelling?



Run, or he's going to tell us about again!

R



LESSON OVERVIEW

- **Datasets:** text about the cost of living scraped from the Scottish and UK governments' websites
- **Research questions:**
 1. What can we observe about the number and length of articles on the cost of living in the last 3 years? How do the data sets compare?
 2. Can we see a difference in the wording about the cost of living between the data sets?
- **Topics covered:** keywords in context, term frequency, term correlation; topic modelling; sentiment categorisation





THE UNIVERSITY *of* EDINBURGH
Centre for Data, Culture & Society



A blurred photograph of a person sitting at a desk, viewed from the side and back. They appear to be working on a laptop. The background is a warm, reddish-orange color.
TIME FOR R



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



COFFEE BEAK

**WE ARE GOING TO RESTART AT
15:30**



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



BYOD

SESSION 2

Wildlife Management and Muirburn (Scotland) Bill

- The Scottish Parliament committees run inquiries on specific topics including legislation
- The Parliament uses the digital engagement tool Citizen's Space
- This provides structured data based on how the survey was set up
- This inquiry received over 5,000 responses
- The responses are published and publicly available



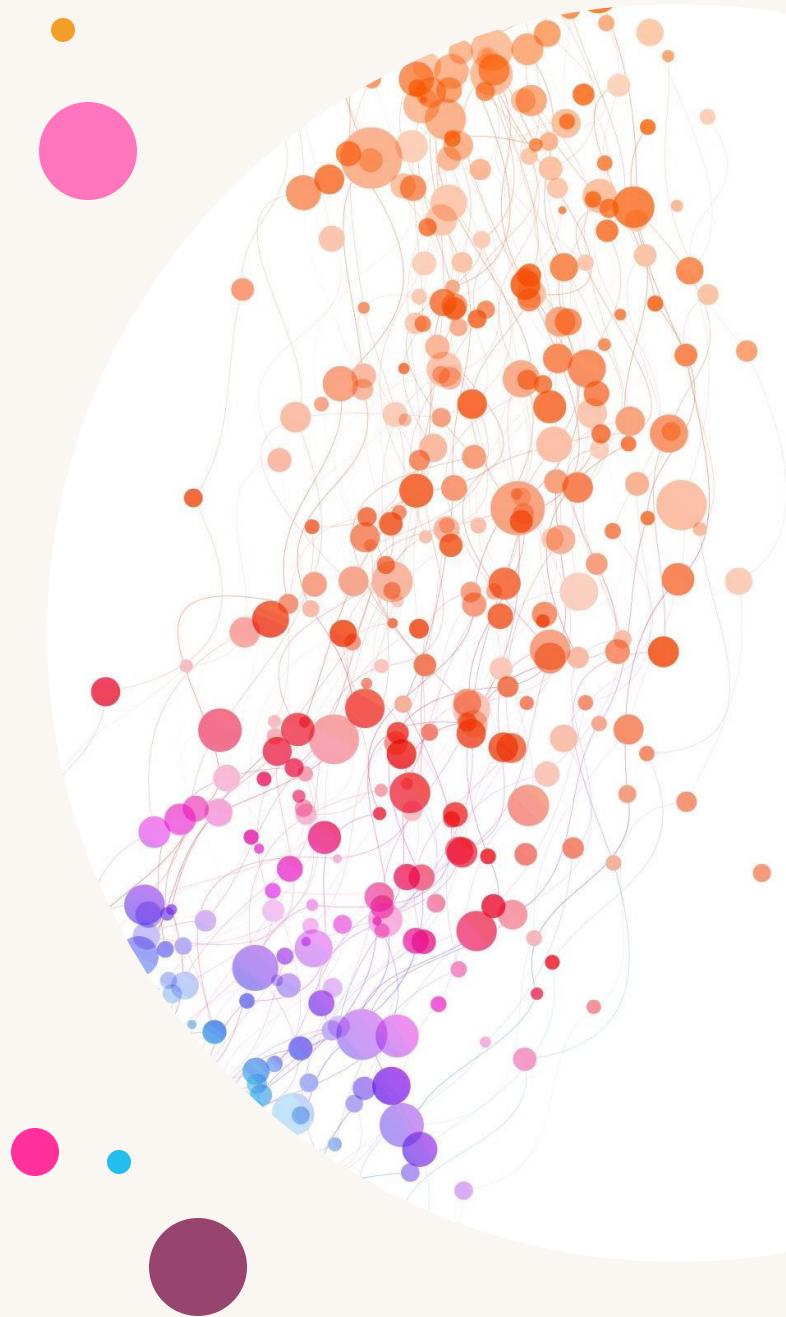
CLINICO-EPIDEMIOLOGICAL PROFILES OF PATIENTS WITH COVID-19 DURING SECOND WAVE OF PANDEMIC IN BANGLADESH

The study contains data related to:

- Patients' Demographic Profile
- Symptoms Severity
- Comorbidity Conditions
- Vaccination Status
- Diagnosis
- Outcome

The data is a transcript from one of the group discussions in a focus group study on digital media use and personal data management.

Text analysis generates the critical content and context in personal data management strategy in daily-based digital personalised services. It explores the type of devices, media platforms, individual data categories, their relations with trust and risk assessment in the decision-making process, and users' knowledge of personal data management.

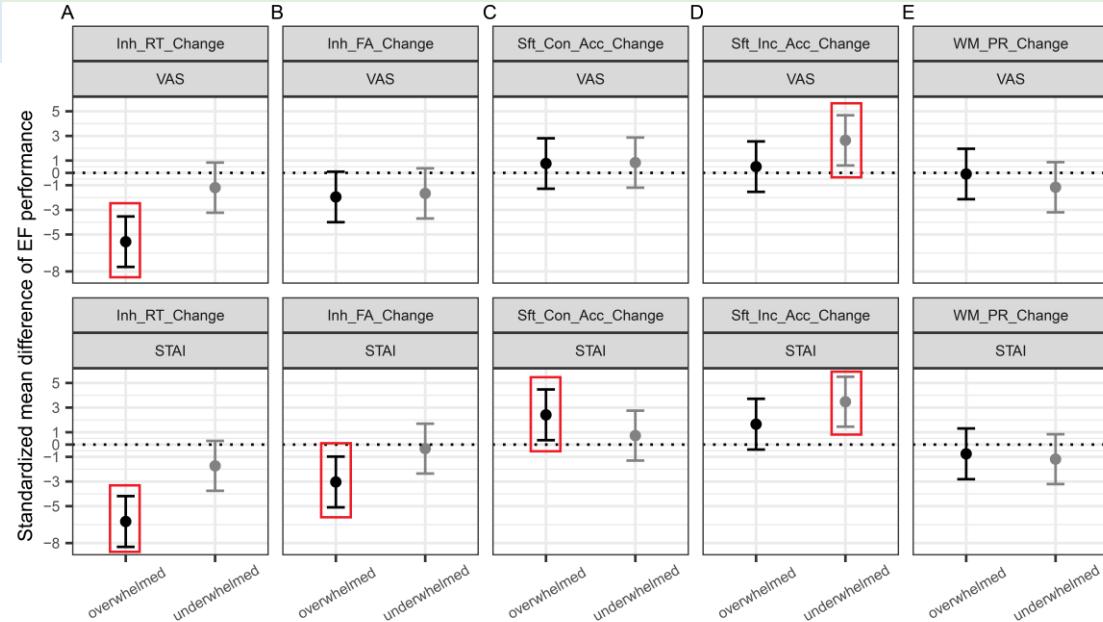


Effects of Stress on Cognition (Mental Health)

Background

While higher-order cognitive functions such as executive function skills are highly critical to daily functions, these skills do not operate optimally in all contexts and are particularly vulnerable to the effects of stress, an inherent aspect of being social-emotional individuals. A new cognitive game-based assessment was developed to measure cognition in the context of stress. Using data collected from adolescent players from across the US to investigate the impact of game-based stressors on cognitive function.

Key data points



adolescents who are under higher stress levels **first** will have better EF performance compared to those who are under lower levels.

Sample size: 80

Codebook link

<https://m87t2wrqpm.larksuite.com/docx/NtyMdi5wAoBcdoxB9gluQD0zsJe>



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society



PUB CRAWL

DOCTORS PUB

32 Forrest Road
Edinburgh EH1 2QN



www.ccds.ed.ac.uk