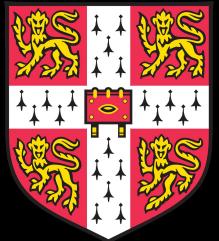


Fact-checking as a conversation

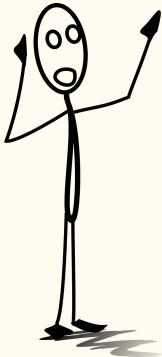
Andreas Vlachos

<http://andreasvlachos.github.io/>



What do fact-checkers do?

The United Kingdom has ten times Italy's number of immigrants.

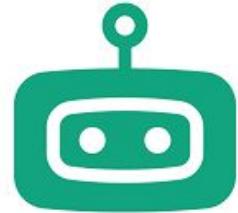


Country/ Immigration	Italy	UK
2014	4.92M	5.05M
2015	5.01M	5.42M
2016	5.03M	5.64M

FALSE: We find no data to support this claim. The UK does not have "ten times Italy's number of immigrants".

Automated fact-checking

The United Kingdom has ten times Italy's number of immigrants.



CHAT GPT



Country/ Immigration	Italy	UK
2014	4.92M	5.05M
2015	5.01M	5.42M
2016	5.03M	5.64M

FALSE: We find no data to support this claim. The UK does not have "ten times Italy's number of immigrants".

What do we want from automated fact-checking?

- Evidence!
 - Labels alone not conducive to fact-based discourse
 - Helps check the correctness of the fact-checks
- Learn with (relatively) little data
- Think about its intended uses; beware of the white hat bias!

(Vlachos and Riedel, 2014; Schlichtkrull et al. 2023)

New datasets needed

AI successes follow dataset availability (*Wissner-Gross, 2016*)

Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka “The Extended Book” (1991)	Negascout planning algorithm (1983)
2005	Google’s Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google’s GoogLeNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolution neural network algorithm (1989)
2015	Google’s Deepmind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		3 years	18 years

Fact Extraction and VERification (FEVER)

Claim:

The Rodney King riots took place in the most populous county in the USA.

SUPPORTED

Evidence:

[wiki/Los Angeles Riots]: The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los Angeles County]: Los Angeles County, officially the County of Los Angeles, **is the most populous county in the United States**.

- 185K claims verified on Wikipedia (*Thorne et al., 2018*)
- Evidence must be correct for verdicts to be correct

Verdict justification?

Retrieved evidence is a baseline. However, we also want to know:

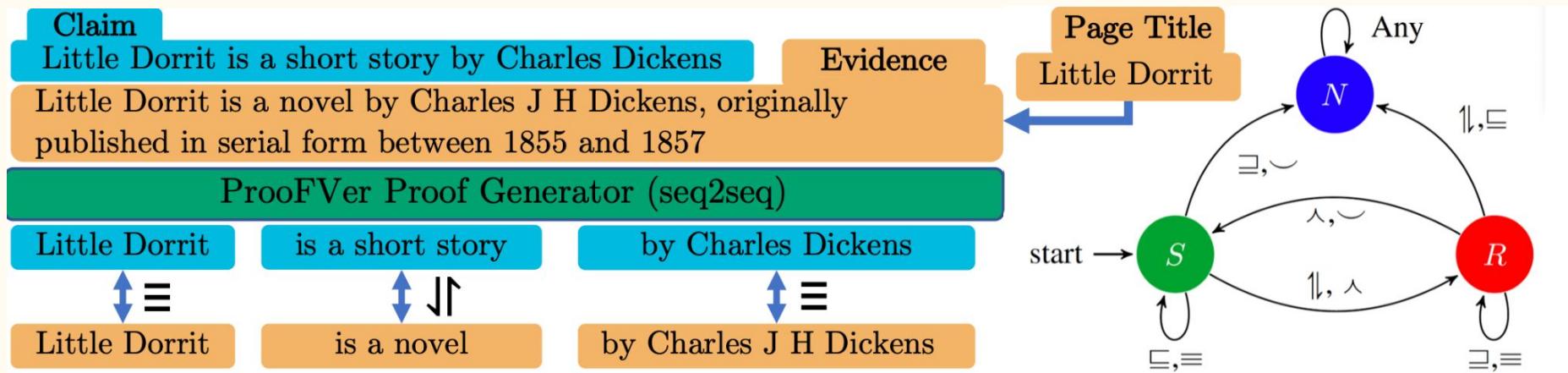
- *How* was the evidence used in the reaching the verdict?
- What were the assumptions/commonsense used?
- What was the reasoning process?

Common approaches:

- Highlight(attention)-based, e.g. *Popat et al. (2018)*, but not clear if attention is an explanation indeed
- (Evidence) Summarization, e.g. *Kotonya and Toni (2020)*, *Atanasova et al. (2020)*, but does not correspond to reasoning
- *Faithfulness* is lacking in both

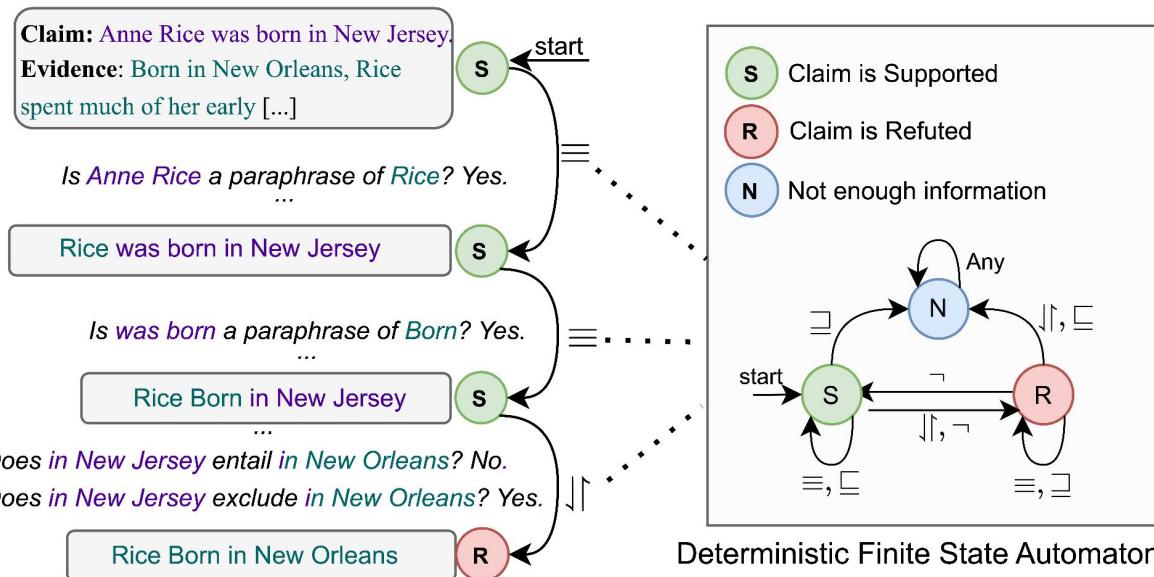
Proof System for Fact Verification (ProoFVer)

When comparing the claim with the evidence, we generate the proof directly and infer the verdict from it (*Krishna et al., 2022*)



The six operators are from Natural Logic (Angeli and Manning, 2014) indicating negation, equivalence, alternation, etc.

Question-answering for proof generation



Aly et al. (2023) turned the natural logic operator assignment to question answering; using cross-lingual LLMs and 64 training instances and run it on Danish FEVER!

Real-world claims?

FEVER and similar datasets contain purpose-made claims derived from Wikipedia. Facilitates dataset creation, but:

- evidence limited to Wikipedia
- claims very different from those tackled by fact-checkers

Datasets with real-world claims exist, but evidence:

- is superficially annotated (e.g. search engine results)
- contains pages created after the claim (including fact-checks!)

From Wikipedia to WWW: AVeriTec

Claim: The USA has succeeded in reducing greenhouse emissions in previous years.

Date: 2020.11.2

Q1: What were the total gross U.S. greenhouse gas emissions in 2007?

A1: In 2007, total gross U.S. greenhouse gas emissions were 7,371 MMT.

Q2: When did greenhouse gas emissions drop in the USA?

A2: In 2017, total gross U.S. greenhouse gas emissions were 6,472.3 MMT, or million metric tons, carbon dioxide.

Q3: Did the total gross U.S. greenhouse gas emissions rise after 2017?

A3: Yes. After 3 years of decline, US CO2 emissions rose sharply last year. Based on preliminary power generation, natural gas, and oil consumption data, we estimate emissions increased by 3.4% in 2018.

Verdict: Conflicting Evidence/Cherrypicking.

Justification: It is true they did reduce emissions however they have now increased again. It is unknown exactly what years are being referred to.

Contextualised claims (from ClaimReview) with **metadata**: resolved references, dates, speaker, etc

Human-written **questions** inspired by the fact-checking article, with **answers** and sources manually retrieved from the Web (Google + archive.org)

Four-way classification:

- **supported**
- **refuted**
- **not enough evidence**
- **conflicting evidence/cherry-picking**

Double-checked to ensure evidence sufficiency, without reference to the fact-checking article

Justification: explain how the evidence combines to give the label

Fact-checking approach

- **Search:** top 30 pages using Google search with the claim initially and questions generated with in-context training of BLOOM open-source language model
- **Evidence selection:** top 100 paragraphs similar to the claim, generate question for each paragraph, classify for relevance
- **Veracity prediction:** label the pairing of each question-answer with the claim as support/refute/irrelevant and then:
 - If mixed: conflicting evidence/cherry picking
 - If only supported/refuted: supported/refuted
 - Otherwise: not enough evidence

Results

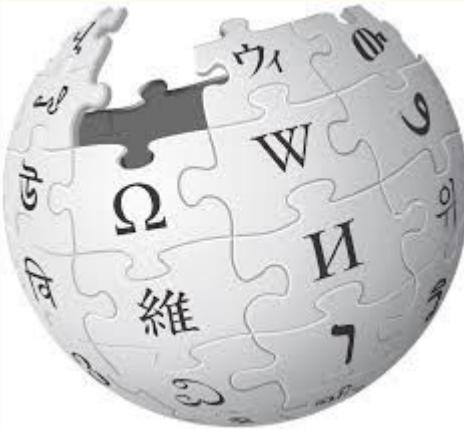
Model	Questions	Questions + Answers	Veracity w. Evidence
No Search	0.19	0.11	0.02
Gold Evidence	1.00	1.00	0.49
AVeriTec	0.26	0.21	0.15
ChatGPT	0.29	0.16	0.10

- Evidence retrieval is the main challenge for the systems
 - Using the claim as a search query is insufficient
 - Evaluation of the correctness of evidence is very challenging
- ChatGPT asks good questions!
 - But provides wrong answers as it doesn't retrieve them from sources
 - Even worse, hallucinates evidence plausibly
- Veracity accuracy is conditional on evidence correctness

Future work

- FEVER7 and AVeriTeC shared task at EMNLP!
- Evidence trustworthiness taken into account?
- Claim detection/prioritization
- Work with human fact-checkers
- Other inputs:
 - Images, video
 - Languages beyond English

Fact checking as a conversation



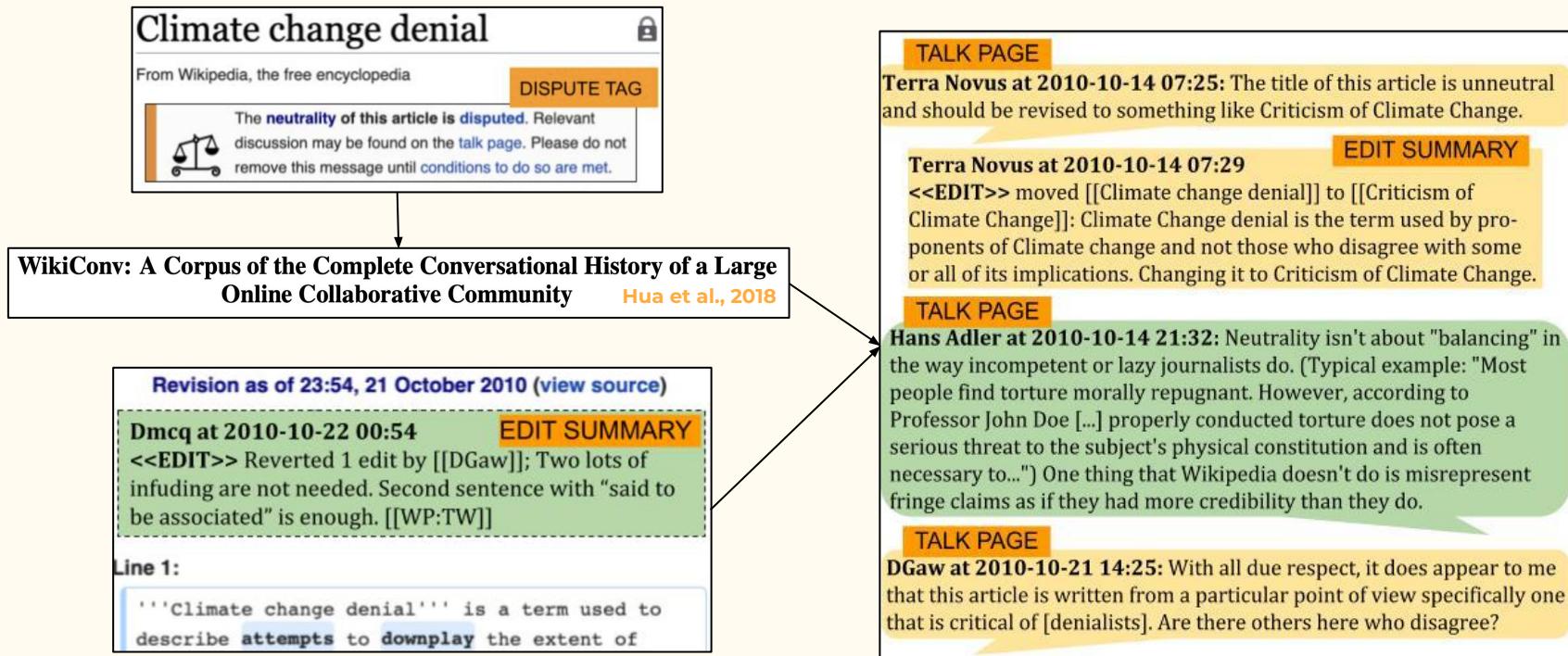
WikiTRIBUNE



- Wikipedia: most successful large-scale online conversation
- Success not straightforward to replicate
- How can we make it happen again?

WikiDisputes (De Kock and Vlachos, 2021)

- A corpus of 7 425 disagreements on Wikipedia Talk pages



Predicting escalation

Welcome to the dispute resolution noticeboard (DRN)

This is an *informal* place to resolve small **content disputes** as part of [dispute resolution](#). It may also be used as a tool to direct certain discussions to more appropriate forums, such as [requests for comment](#), or other noticeboards. You can ask a question on the [talk page](#). This is an early stop for most disputes on Wikipedia. You are [not required to participate](#), however, the case filer must participate in all aspects of the dispute or the matter will be considered failed. Any editor may volunteer! Click this button  to add your name! You don't need to volunteer to help. Please feel free to comment below on any case. **Be civil and remember; Maintain Wikipedia policy: it is usually a misuse of a talk page to continue to argue any point that has not met policy requirements.** "Editors must take particular care adding *information about living persons* to any Wikipedia page. This may also apply to some [groups](#).

Noticeboards should not be a substitute for talk pages. Editors are expected to have had extensive discussion on a talk page (not just through edit summaries) to work out the issues before coming to DRN.

Do you need assistance? Would you like to help?

[Request dispute resolution](#) [Become a volunteer](#)

- +
- ## Escalation labels:
- 201 Escalated
 - 7224 Not escalated*
- *sub-sampled to correct for length imbalance

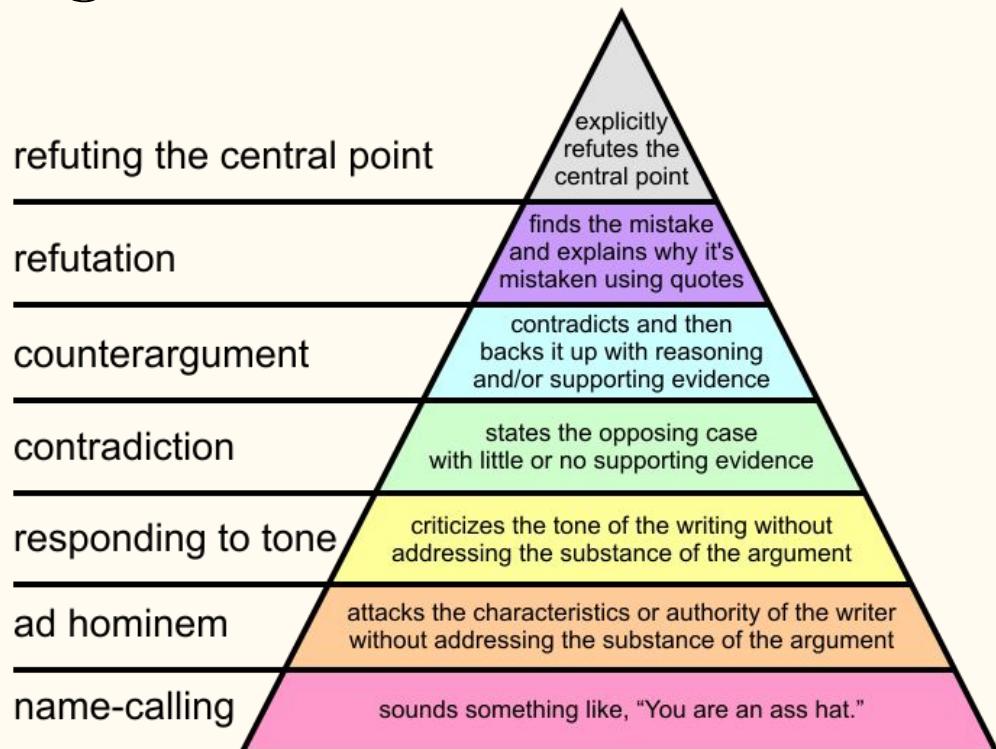
What we might be looking for?

Wikipedia's guidelines for dispute resolution follow Graham's argument hierarchy

Is this what makes it work?

Other options:

- Politeness
- Toxicity (lack of)
- Sentiment



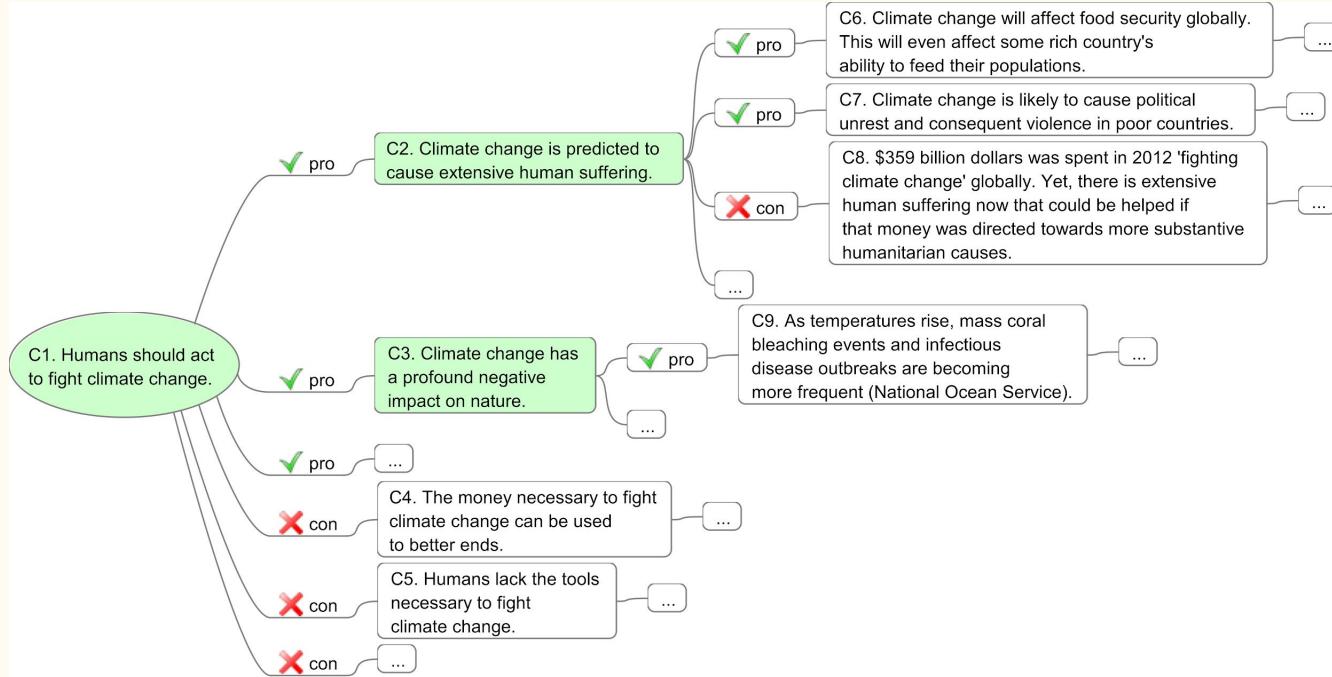
Predicting escalation

- **Toxicity**: *Wulczyn et al. (2017)*
- **Sentiment**: *Liu et al. (2005)*
- **Politeness**: *Zhang et al. (2018)*
- **Collaboration**: *Niculae and Danescu- Niculescu-Mizil (2016)*
- **+Gradients**: how features change in conversation
- Neural models with dialogue structure perform best
- Improved further with Graham hierarchy (De Kock et al. 2023)

Model	PR-AUC
Baselines	
Random	0.121
Bag-of-words	0.213
Feature-based models	
Toxicity	0.140
Sentiment	0.150
Politeness	0.232
+ <i>gradients</i>	0.275
Collaboration	0.261
+ <i>gradients</i>	0.269
Politeness and collaboration	0.255
+ <i>gradients</i>	0.281
Neural models	
Averaged embeddings	0.243
LSTM	0.263
HAN	0.373
+ <i>edit summaries</i>	0.400

How do we encourage Open minds? ArguBots

- Joint project with Open University, Sheffield and Toshiba
- Develop bots that help users engage with the “other side”

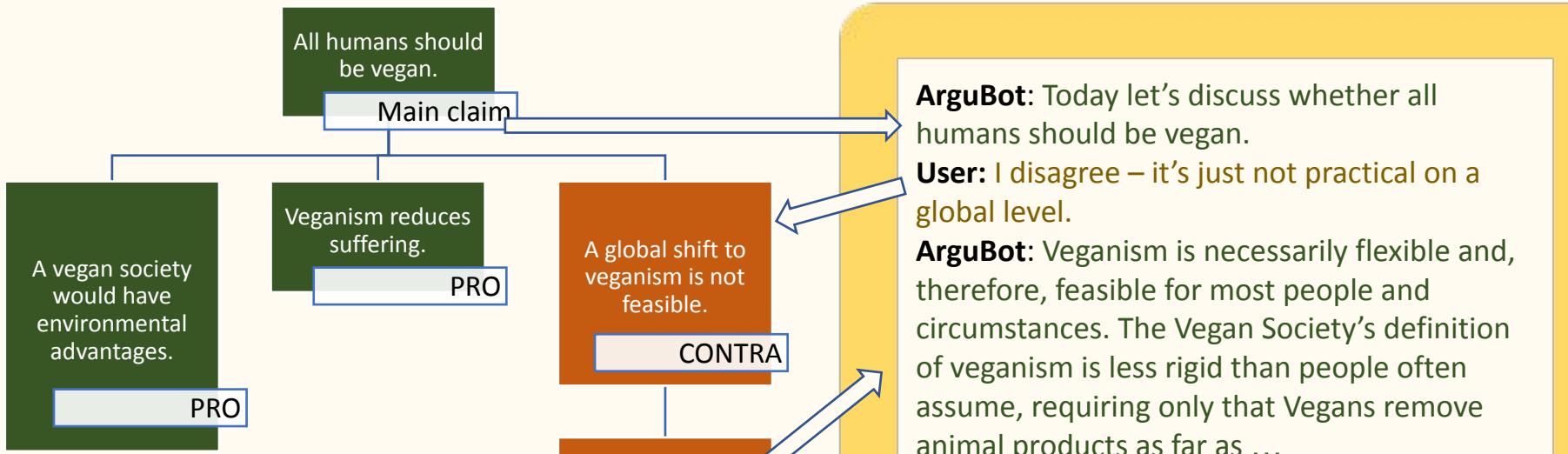


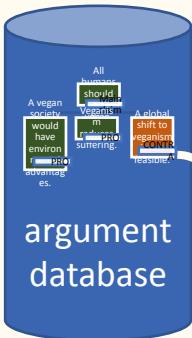
ArguBot: Today let's discuss whether all humans should be vegan.

User: I disagree – it's just not practical on a global level.

ArguBot: Veganism is necessarily flexible and, therefore, feasible for most people and circumstances. The Vegan Society's definition of veganism is less rigid than people often assume, requiring only that Vegans remove animal products as far as ...







argument database



BM25 retrieval



Adapted BART
(Lewis et al., 2020)

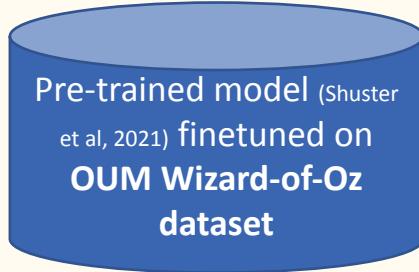
Retrieval-Augmented Generation (RAG)-Sequence model

(Lewis et al. 2020)

ArguBot chat contribution



dialogue history



Pre-trained model (Shuster et al, 2021) finetuned on
OU M Wizard-of-Oz dataset

Argument source:
kialo.com

ArguBot: Today let's discuss whether all humans should be vegan.

User: I disagree – it's just not practical on a global level.

ArguBot: Veganism is necessarily flexible and, therefore, feasible for most people and circumstances. The Vegan Society's definition of veganism is less rigid than people often assume, requiring only that Vegans remove animal products as far as ...



Image source: 461715, William Augustus Brown, Education Collection, Special Collections, University of California, Santa Barbara, Manuscripts and Special Collections, University of California, Santa Barbara

Evaluation – Results for the Argubot

Open-mindedness

- the **Ideological Turing test** (Brand and Stafford, to appear)
- proxy questions (Stanley et al. 2020): *do you believe your ideological opponent has good reasons for their position?* 

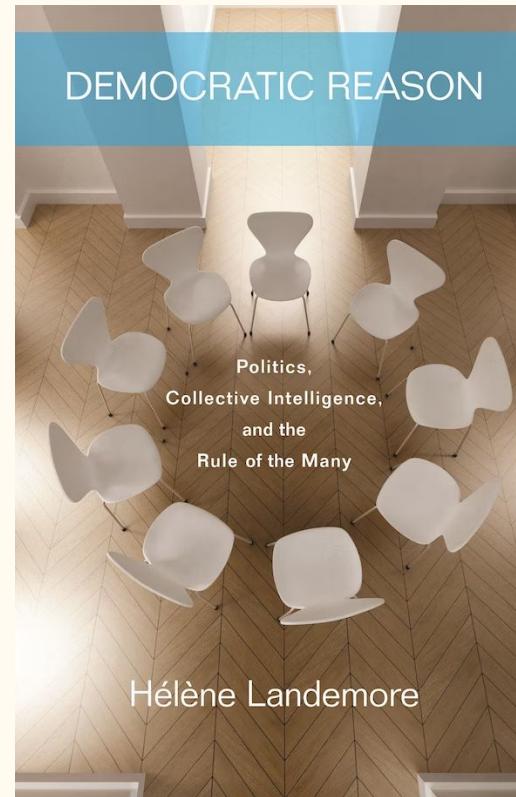
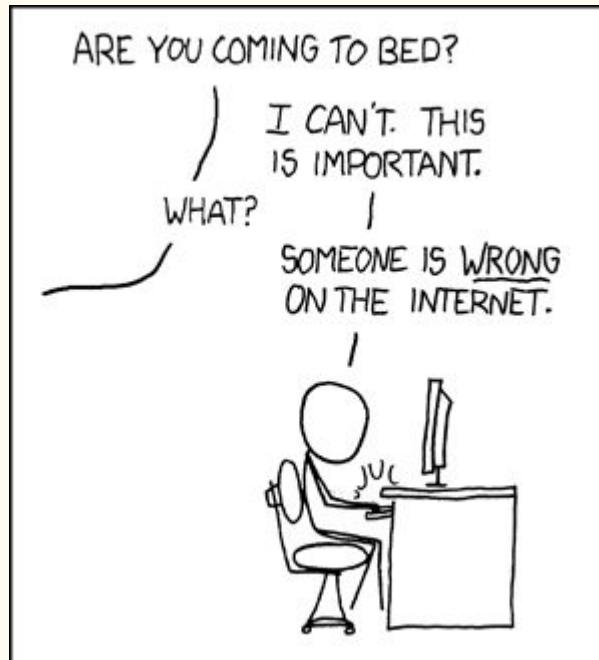
Chat experience indicating the potential for engagement

- engaging 
- clarity 
- consistency 
- not confusing 
- not frustrating 
- ... 

More in *Farag et al. (EMNLP 2022)*



Is dialogue helping us reach better decisions?



Wason (1968) selection task

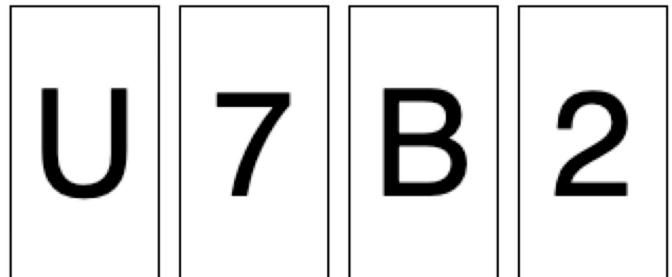
What do you think?

Individuals' success rate: 10-20%

Small groups success rate?

80%! What makes groups work?

Each of the 4 cards below has a letter on one side and a number on the other. Which card(s) do you need to turn to test the rule:
All cards with vowels on one side have an even number on the other.

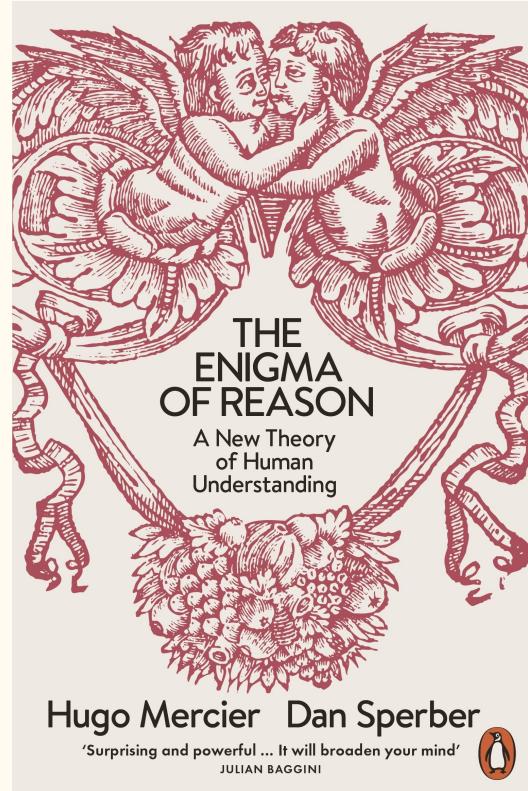


With a little help from my friends

Reasoning has evolved in the context of communication, not in isolation:

- arguments are made to help us justify ourselves and convince each other
- we are bad judges of our own arguments but good for the others
- Scientists are no different!

Can we help groups work better?



Deliberation Enhancing Bots (DEliBots)

- Develop conversational agents that make conversations better!
- A different kind of dialogue agent:
 - Unlike chatbots, they help users accomplish a task
 - Unlike task-oriented bots (e.g. restaurant booking), they don't know or give the answer

Data collection (Karadzhov et al. 2023)

- 500 groups, 2-5 persons (avg 3.16) (smaller group, fewer ideas)
- each group member submits responses at onboarding
- the group deliberates and members submit again
- no need for the group consensus but bonus for correct response

Onboarding success rate: 11%

Success rate after deliberation: 33%

In 43.8% of the groups with the correct solution, no participant had chosen it initially

Improving deliberation?

Ask questions/probes for:

- moderation
- solutions
- reasons

Hypothesis: probing for reasoning makes a difference

Beaver: What do you think?

moderation

Cat: I think A and 8

Duck: I thought A and 8 too, but we may be wrong

Cat: @Duck, well we need A for sure

Beaver: What if we don't turn 8 at all?

reason

Duck: Yes! We don't care what is behind the even numbers

Cat: This may be right, but we may need to check the odds

Duck: So, A and 9?

solution

Beaver: Yes

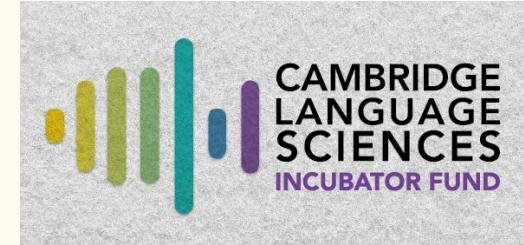
Analysis (Karadzov et al. 2024)

- Key correlations:
 - **Conversation length** correlates positively but weakly
 - **Diversity of ideas matters**, even if when they are wrong
 - **Probing for reasons** correlates with diversity
- We have now built a Delibot that improves group decision-making in Wason, publication forthcoming!

Next steps

- Real-world applications
 - Detecting AI-generated text
 - Chess problem solving
 - Peer-reviewing
- Data and more here: <https://www.delibot.xyz/delidata/>

Thanks to the funding agencies



Thanks to my collaborators

- Cambridge: **Rami Aly, Zhenyun Deng, Christine De Kock** (Melbourne), **Youmna Farag** (Toshiba), **Zhijiang Guo** (Huawei), **Georgi Karadzhov**, **Amrith Krishna** (Uniphore, Learn.AI), **Nedjma Ousidhoum** (Cardiff), **Michael Schlichtkrull** (QMUL), **Marek Strong**, **James Thorne** (KAIST AI)
- Elsewhere:
 - **Christos Christodoulopoulos** (Amazon)
 - **Oana Cocarascu** (Imperial College, King's College)
 - **Arpit Mittal** (Amazon, Meta)
 - **Sebastian Riedel** (UCL, Meta, Google)
 - **Lotty Brand, Tom Stafford** (Sheffield)
 - **Paul Piwek, Jacopo Amidei** (Open University)
 - **Svetlana Stoyanchev** (Toshiba Research)



Questions?

andreas.vlachos@cst.cam.ac.uk