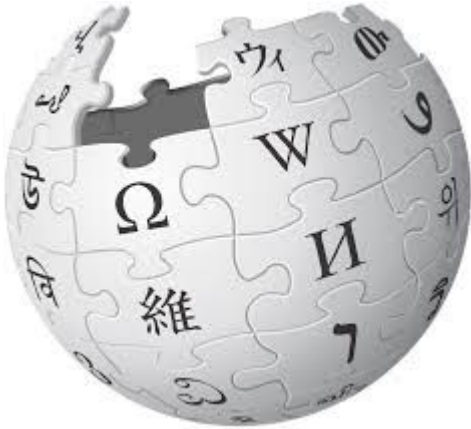


Towards constructive conversations

Prof. Andreas Vlachos,
Dept. of Computer Science and Technology,
Dinesh Dhamija Fellow, Fitzwilliam College



Do you know these two logos?



Wiki**TRIBUNE**



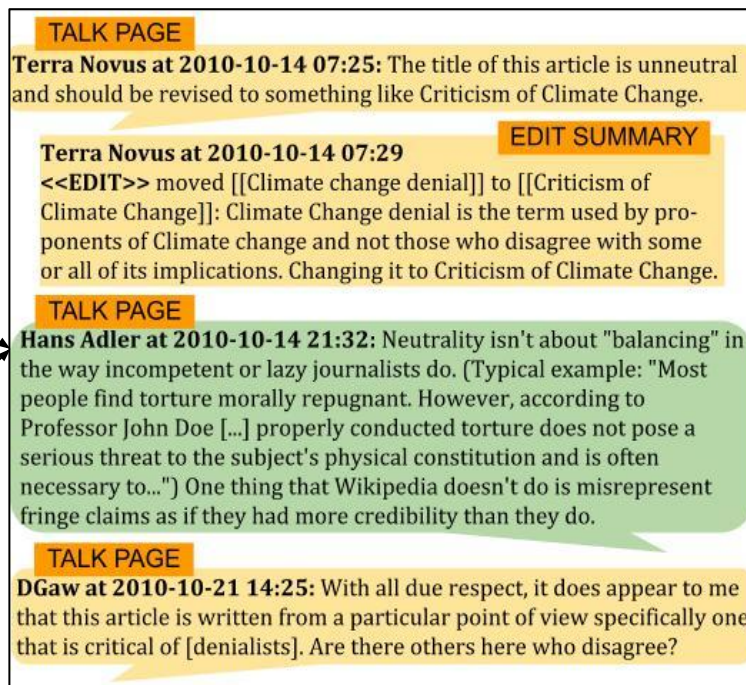
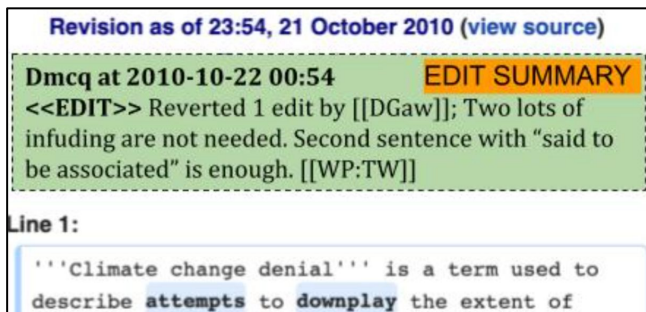
- Wikipedia: most successful large-scale online conversation
- Success not straightforward to replicate
- What can we learn from it?

WikiDisputes (*De Kock and Vlachos, 2021*)

A corpus of 7 425 disagreements on Wikipedia Talk pages



WikiConv: A Corpus of the Complete Conversational History of a Large Online Collaborative Community *Hua et al., 2018*





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

Shortcuts

[WP:DRN](#)

[WP:DR/N](#)

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



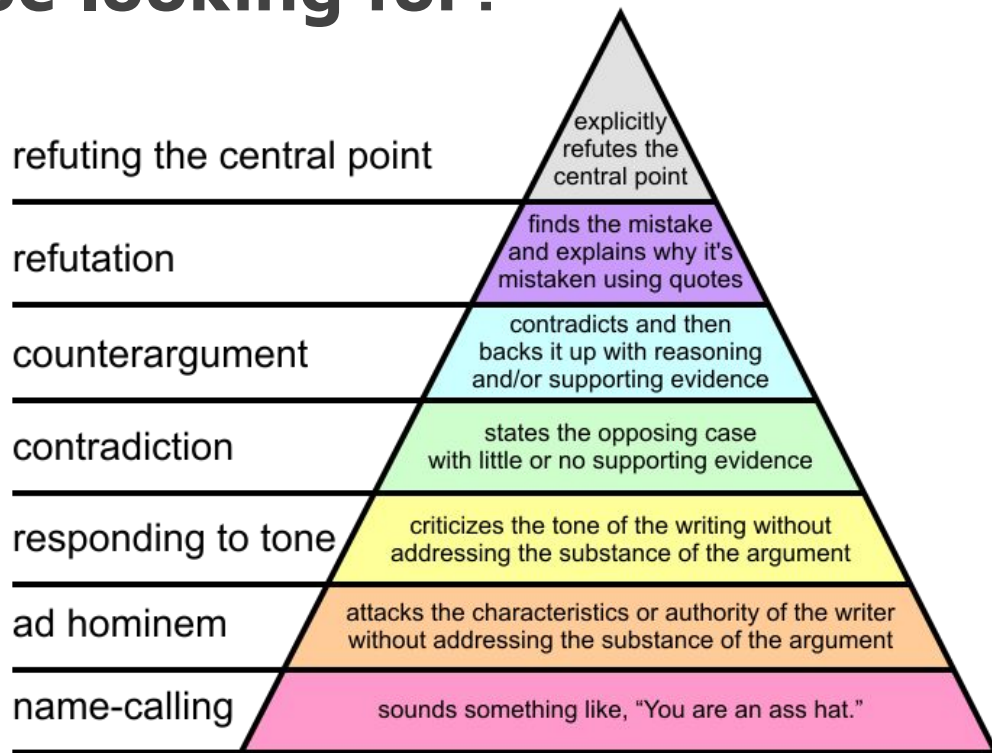
What we might be looking for?

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

Is this what makes it work?

Other options:

- Politeness
- Toxicity (lack of)
- Sentiment





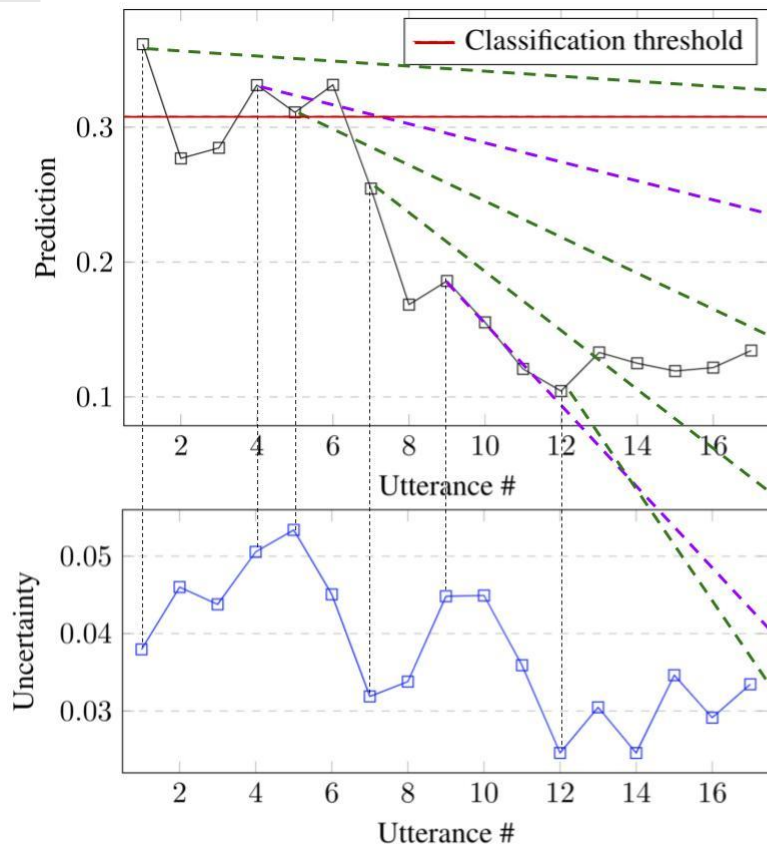
Predicting escalation

Feature-based models

- Toxicity: [*Wulczyn et al. \(2017\)*](#)
- Sentiment: [*Liu et al. \(2005\)*](#)
- Politeness: [*Zhang et al. \(2018\)*](#)
- Collaboration: [*Niculescu and Danescu-Niculescu-Mizil \(2016\)*](#)
- +Gradients: how features change in conversation
- Neural models with dialogue structure perform best

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

A cherry picked example from our model



Ggugvunt at 2006-08-25 15:09: The rest of this article is quite well written and reasonably NPOV, but the caption to the Canada picture reads like polemic to me.

Frogsprog at 2006-08-25 15:12: I would start by saying I am "very" annoyed by american editors refusing to accept information that makes their country look unpopular.

Ggugvunt at 2006-08-25 15:20: First of all, you need to calm down a bit. Secondly, if you think the picture is so violent and strong, let it speak for itself.

Ggugvunt at 2006-08-25 15:33: Listen - my only issue is this: the caption sounded out of place with the reasonable tone of the rest of the article. Don't jump to conclusions - please. Can I revert it back to the non-adverbs version and remove the NPOV marker now?

Frogsprog at 2006-08-25 18:11: OK, a compromise, I will re-add the adverb "violently" but leave out strong

Ggugvunt at 2006-08-25 18:39: Much better! Thank you!

Graham's Hierarchy on WikiDisputes

213 disputes
annotated rebuttal
and coordination
tactics (*[De Kock et al. 2022](#)*)

Findings
corroborate
Wikipedia's
recommendation

Improved escalation
prediction

Fenugreek: A herb / an herb

The community put WP:ENGVAR in place exactly because there is no rational way to resolve a style dispute like this. The notion is that if English style X is established in article, don't change it without prior consensus. Without that [policy], articles would be beset by endless edit wars over style issues that would become a time sink across the encyclopedia.

Coordination

Rebuttal

Contextualisation

Hi, I am aware of WP:ENGVAR and would like to point out to you the policy says that one should "use the variety found in the first post-stub revision that introduced an identifiable variety". In the case of this article, that is "a herb", which was introduced in the original article. I will leave the current wording for a few weeks to see if anyone else decides to weigh in, and intend to then change the page to align with policy.

DH6: Refutation
Suggesting a compromise

It is impossible to get local consensus on this kind of thing, which is why ENGVAR exists. Leave it alone, or waste the community's time with an RfC but stop wasting your time and mine making useless arguments here. I don't care if it says "an" or "a" - what is not acceptable is messing with it.

DH4: Repeated argument
DH3: Policing the discussion

I admit that when I made those edits, I didn't realise it was actually a ENGVAR issue but rather just a mistake, hence my zeal in making the changes. To emphasise: the policy exists to unambiguously resolve these debates and for this article, it should be "a herb". I see no real arguments for the contrary, and for what it's worth, my having made policy-incorrect edits (in good faith), doesn't diminish the fact that policy is clear on this one.

Conceding / recanting
DH4: Repeated argument

I have warned you to walk away from being a style warrior and wasting everyone's time. You will do as you will.

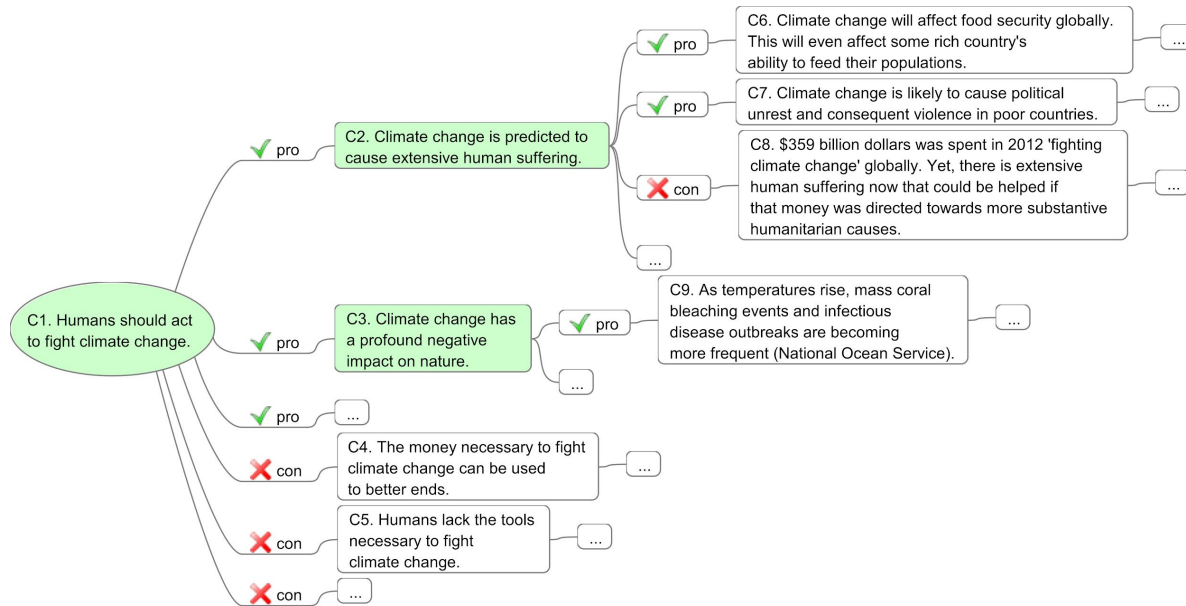
DH1: Ad hominem attack

No one further has weighed in on this and so I am making the change in accordance with policy, as I have done on each of the herb-related pages. Each of these articles is now in accordance with WP:ENGVAR. Please do not edit it without an RFC or DR. We are now within the spirit and letter of policy on each of these pages and I hope we can draw a line under this ridiculous matter.

Coordinating edits
DH3: Policing the discussion

How do we encourage open minds?

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

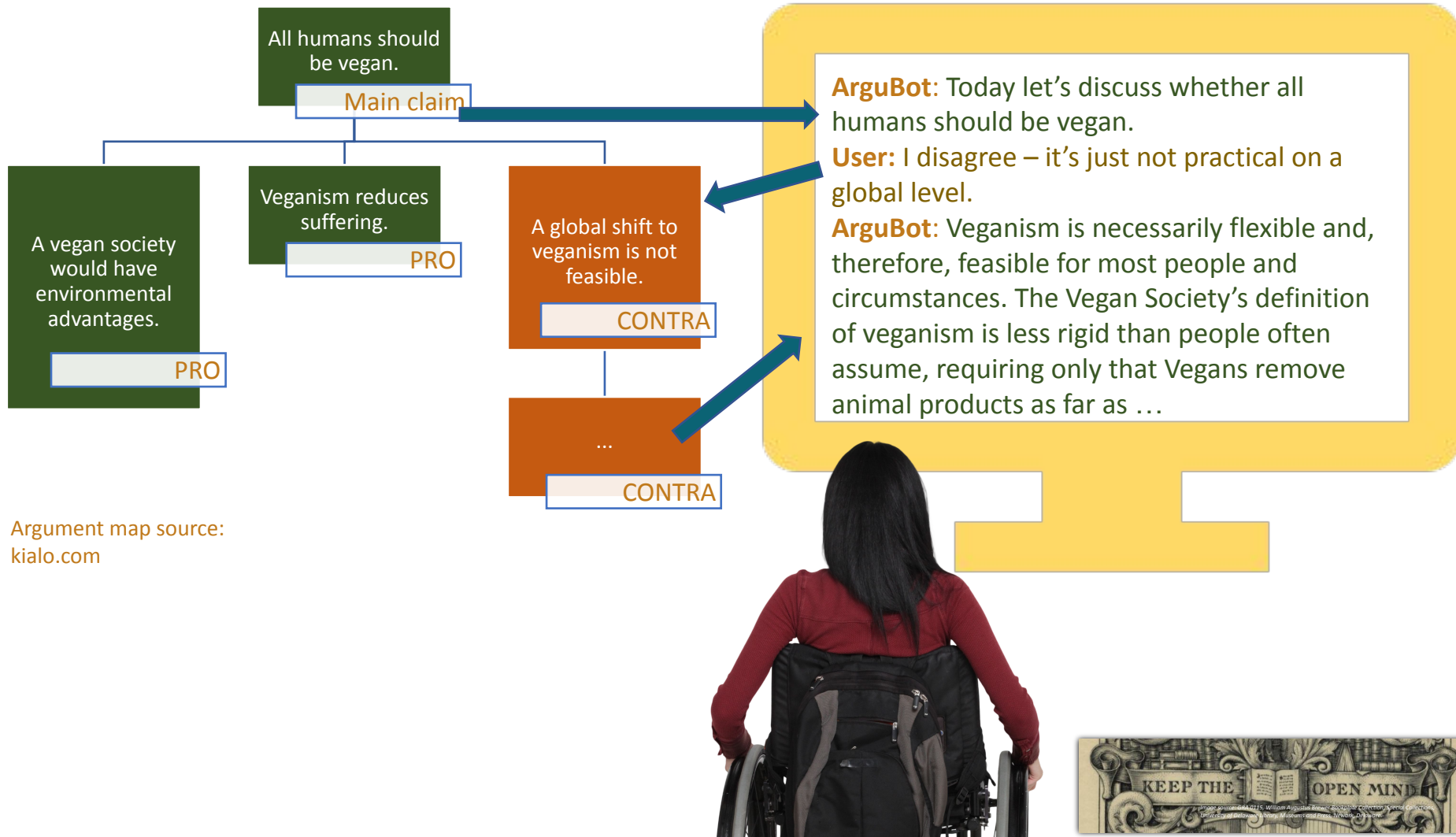


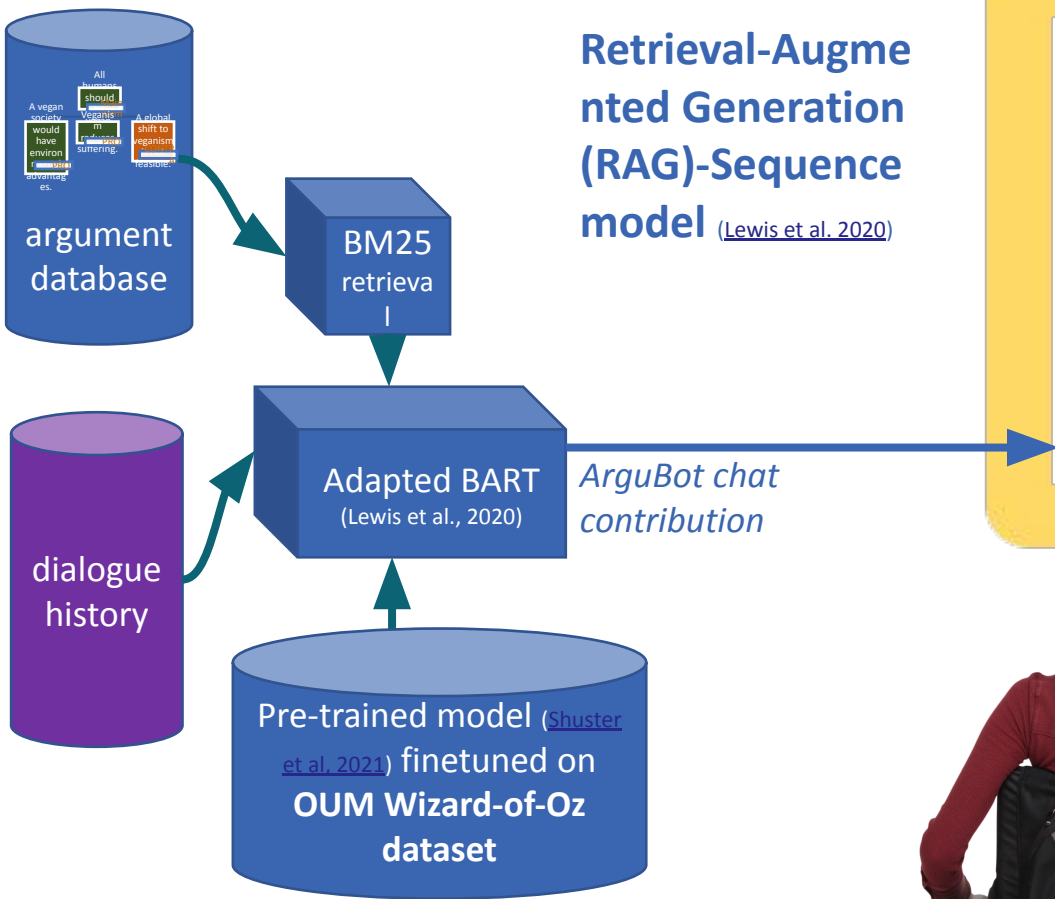
ArguBot: Today let's discuss whether all humans should be vegan.

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 ...







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 ...



Evaluation – Metrics

- *Open-mindedness*

- the **Ideological Turing test**
- proxy questions ([Stanley et al. 2020](#)): *do you believe your ideological opponent has good reasons for their position?*

[Brand, Brady and
Stafford, 2025](#)
preprint

- *Chat experience indicating the **potential for engagement***

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



Evaluation – Metrics

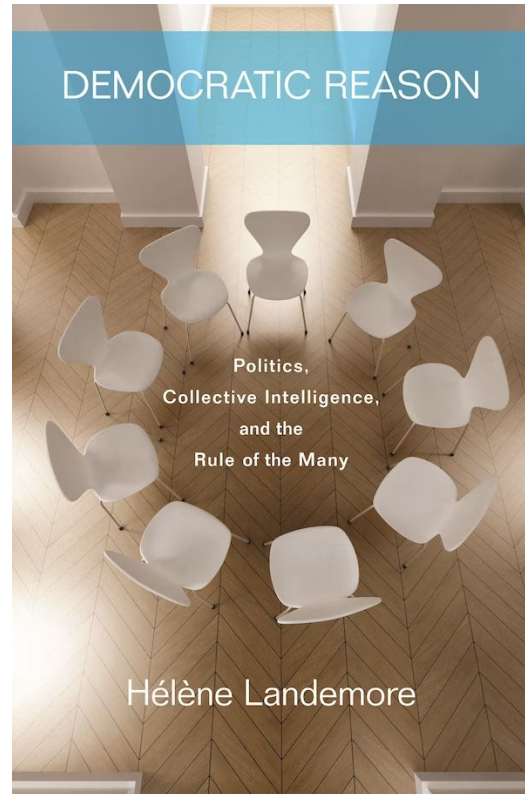
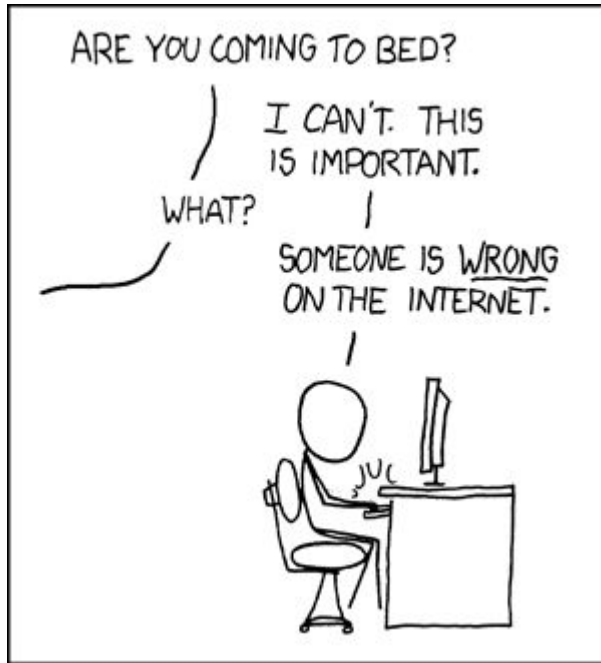
- *Open-mindedness*
 - the **Ideological Turing test**
 - proxy questions ([Stanley et al. 2020](#)): *do you believe your ideological opponent has good reasons for their position?*
- *Chat experience indicating the **potential for engagement***
 - engaging
 - clarity
 - consistency
 - not confusing
 - not frustrating
 - ...



[Farag et al. \(2022\)](#)
Findings of EMNLP



Is dialogue helping us decide better?





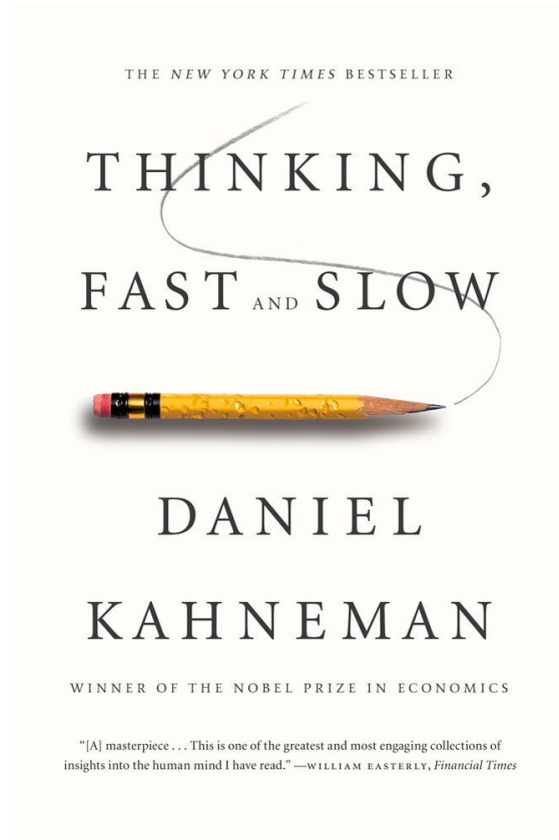
Let's take a look at reasoning

Dual system

- System 1: Fast, biased
- System 2: Slow, rational

Various cognitive biases:

- Recency bias
- Confirmation bias
- etc.





Wason (1968) card selection task

What do you think?

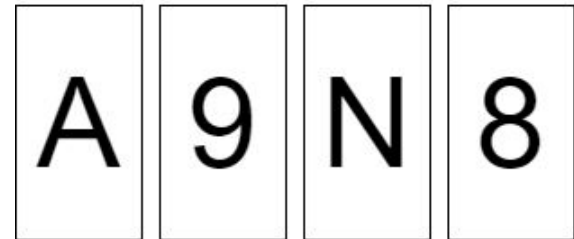
Individuals' success rate: 10-20%

Small groups success rate?

80%! What makes groups work?

Each of the 4 cards bellow has 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.

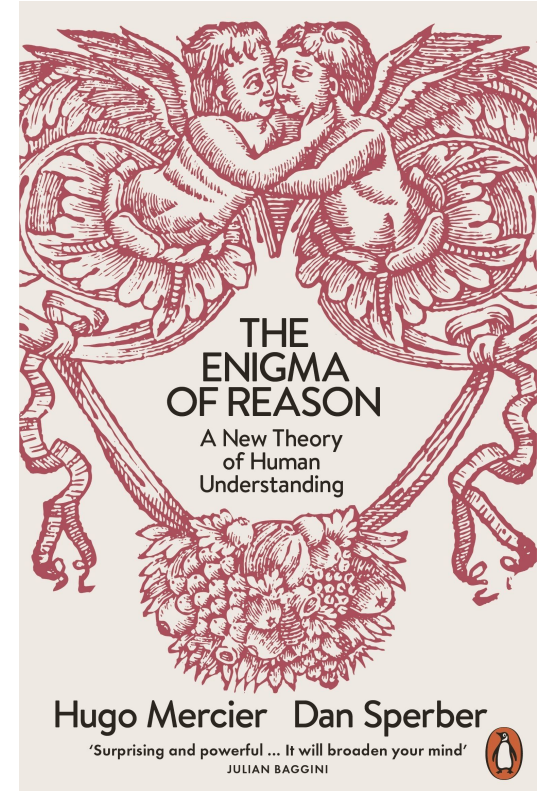


With a little help from my friends

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

Can we help groups function better?



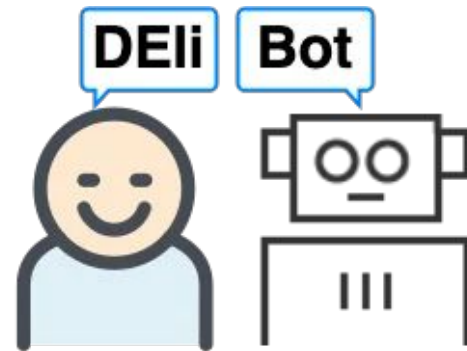


Deliberation Enhancing Bots (DEliBots)

Develop conversational agents that make deliberation better!

A different kind of dialogue agent:

- Doesn't give answers
 - Even if it knows it
 - Often there is no right answer
- It helps people find them by probing





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



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%



Findings (*Karadzhov et al., 2024*)

- **Conversation length** correlates positively but weakly
- **Diversity of ideas matters**, even if when they are wrong
- **Probing for reasons** correlates with diversity
- In 43.8% of the groups with the correct solution, no participant had chosen it initially

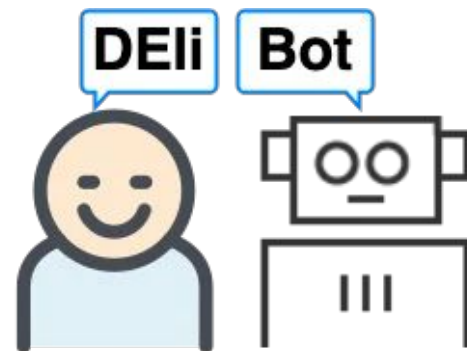
Does this happen in other contexts?

Group decision-making helps with:

- detecting AI-generated text ([*Lee et al, ICWSM 2026*](#))
- solving chess problems

We, humans, help each other well,
can we use AI to support us in this?

WIP: results are not reliably positive yet





Outlook

LLM-based dialogue agents can help us:

- promote active open minded thinking
- improve public discourse with facts

Caution is needed though:

- Hallucinations
- Sycophancy
- Conformity ([*Zhu et al. 2025*](#))

Conformity and LLMs

Vanilla Round



What is the oldest college in Cambridge?



It is Peterhouse College. ✓

Conformity Round



What is the oldest college in Cambridge?



King's College.



King's College.



King's College.



Now, it's your turn.



King's. ✗

Conformity and LLMs

Vanilla Round

QA

What is the oldest college in Cambridge?

It is Peterhouse College. ✓

Conformity Round

QA

What is the oldest college in Cambridge?

King's College.

King's College.

King's College.

...

QA

Now, it's your turn.

...

King's. ✗

Neutral Expressions

I think the answer is King's.

Kings, in my opinion.

My answer is King's.

Confident Expressions

It should be King's.

King's, of course!

I am sure it's King's.

Uncertain Expressions

King's, maybe?

I guess it's King's.

King's? Perhaps.

Conformity and LLMs

Vanilla Round



What is the oldest college in Cambridge?

It is Peterhouse College. ✓

Conformity Round



What is the oldest college in Cambridge?



King's College.



King's College.



King's College.



Now, it's your turn.



King's. ✗

Neutral Expressions



I think the answer is King's.



Kings, in my opinion.



My answer is King's.

Confident Expressions



It should be King's.



King's, of course!



I am sure it's King's.

Uncertain Expressions



King's, maybe?



I guess it's King's.



King's? Perhaps.

Devils' Advocate (DA)



What is the oldest ...



King's College.



Pembroke College.



King's College.



...



Now, it's your turn.



Peterhouse. ✓

Question Distillation (QD)



What is the oldest ...



All participants before you have answered: King's College.



Now, it's your turn.



Peterhouse. ✓



Questions?

andreas.vlachos@cst.cam.ac.uk