



Reasoning in groups, with human and artificial agents

Tom Stafford
20th May 2026

tomstafford.github.io
[@tomstafford](https://twitter.com/tomstafford)

These slides:
bit.ly/tomstafford

Reasoning in groups, with human and artificial agents

In the right circumstances groups can use deliberation to outperform the ability of each individual group member. This talk will review work done in collaboration with Prof. Vlachos and other colleagues in the Department which looks at when, and how, the benefits of group deliberation can manifest, revealing insights into both human psychology and the power of argument exchange. Ultimately the ambition is to design artificial dialogue agents which can positively contribute to, and so enhance, group discussions.



tomstafford.github.io

Cognitive Science

Metascience

Research on Research

Reproducibility

Experiment design

Methods

Cognitive Science

Metascience

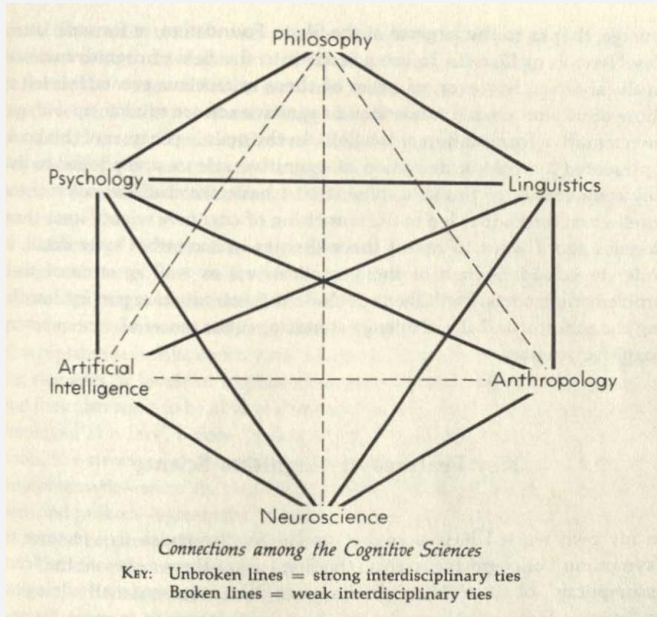
Research on Research

Reproducibility

Experiment design

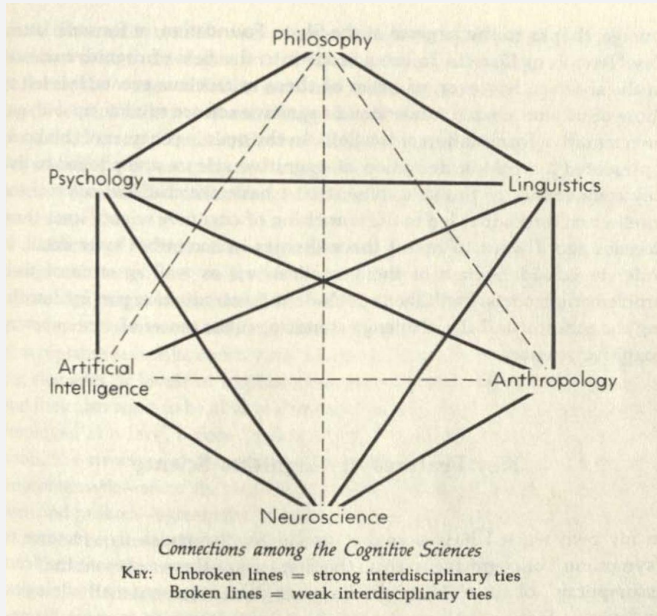
Methods

Cognitive Science



Gardner, H. (1987). *The mind's new science: A history of the cognitive revolution*. Basic books

Cognitive Science



Gardner, H. (1987). *The mind's new science: A history of the cognitive revolution*. Basic books

Interdisciplinarity

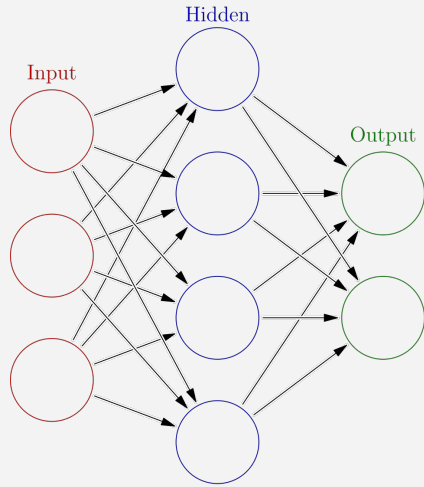
Formal models

Theory building

Object heterodoxy

“The interdisciplinary study of the mind, whether in animal, machines or other systems”

Neurons



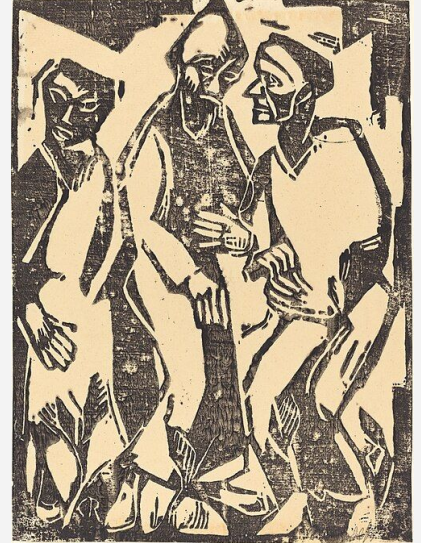
By Glosser.ca, CC
BY-SA 3.0, [Wikimedia
Commons](#)

Minds



Le Penseur, Musée
Rodin, Paris. Image:
Tom Stafford,
2023-08-15

Groups



Christian Rohlfs,
Deliberation (Beratung),
1913, woodcut. National
Gallery of Art, CC0, via
Wikimedia Commons

TODAY: Group decision making and dialogue



Andreas Vlachos



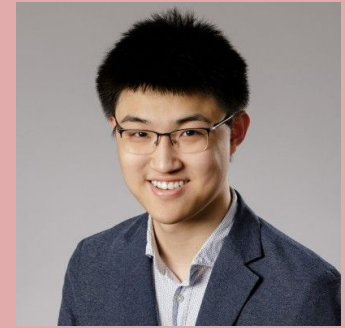
Georgi Karadzhov



Zahra
Arjmandi Lari



Christine de Kock



Xiaochen Zhu



Jacopo Amidei



Paul Piwek



Youmna Farag



Lotty Brand



Svetlana
Stoyanchev



Dongwon Lee



Jooyoung Lee

The psychology of reasoning

Foolproof

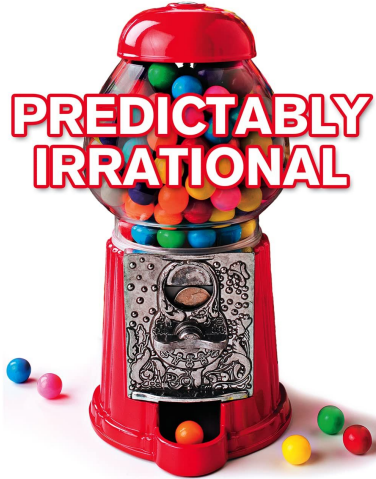
Why We Fall for Misinformation
and How to Build Immunity



Sander van der Linden

'An insightful and forensic examination'
Marianna Spring, BBC's Disinformation and
Social Media Correspondent

DAN ARIELY



The Hidden Forces that Shape Our Decisions

'A lifetime's worth of wisdom'
Steven D. Levitt, co-author of *Freakonomics*

The International
Bestseller

Thinking,
Fast and Slow



Daniel Kahneman
Winner of the Nobel Prize



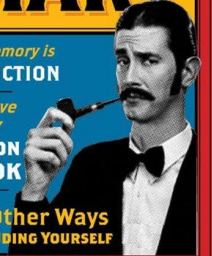
DAVID MCRANEY

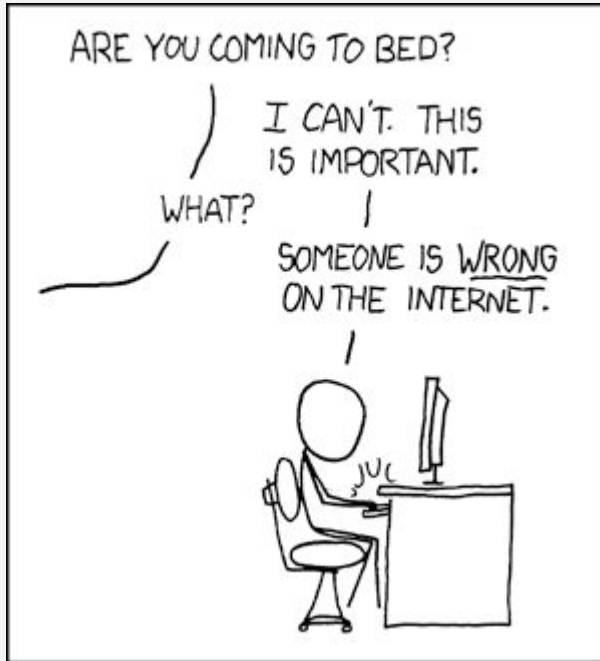
**YOU
ARE NOT SO
SMART**

*Why Your Memory is
MOSTLY FICTION*

*Why You Have
Too Many
FRIENDS ON
FACEBOOK*

*AND 46 Other Ways
YOU'RE DELUDING YOURSELF*





People are
wrong on the
internet

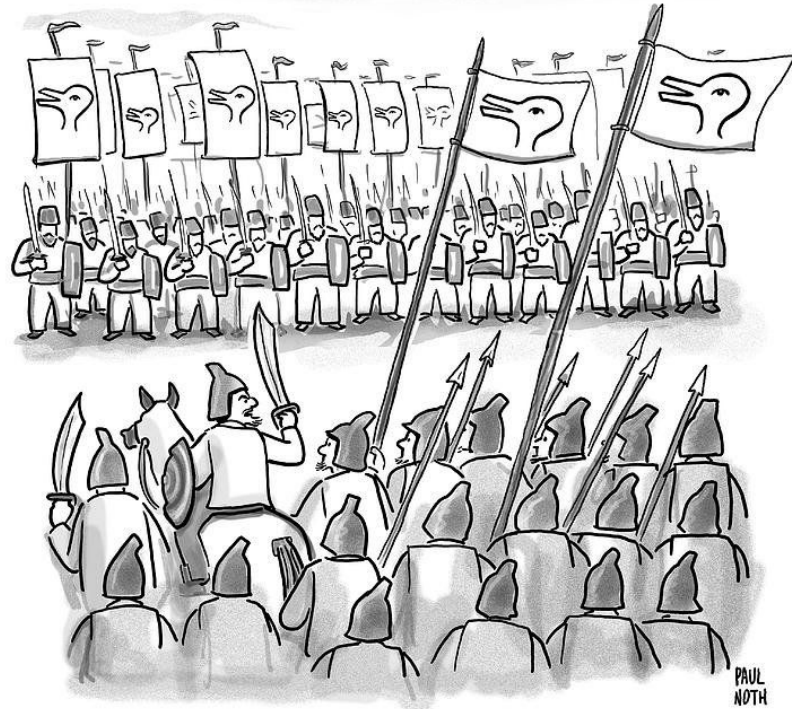
Image: "Duty Calls", xkcd,
[CC-BY-NC 2.5](https://creativecommons.org/licenses/by-nc/2.5/)
<https://xkcd.com/386/>

These slides:
bit.ly/tomstafford

Polarisation

“There can be no peace until they renounce their Rabbit God and accept our Duck God.”

- cartoon by
[@PaulNoth](#)



*“There can be no peace until they renounce their
Rabbit God and accept our Duck God.”*

Wikipedia: a Real Utopia

How to disagree well: Investigating the dispute tactics used on Wikipedia

Christine de Kock

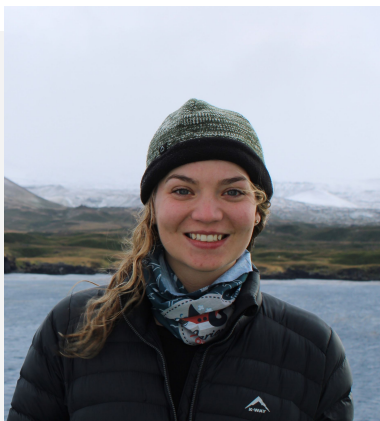
Department of Computer
Science and Technology
University of Cambridge
cd700@cam.ac.uk

Tom Stafford

Department of Psychology
University of Sheffield
t.stafford@sheffield.ac.uk

Andreas Vlachos

Department of Computer
Science and Technology
University of Cambridge
av308@cam.ac.uk



How to disagree well: Investigating the dispute tactics used on Wikipedia, Christine De Kock, Tom Stafford, Andreas Vlachos, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing

<https://arxiv.org/pdf/2212.08353>

These slides:
bit.ly/tomstafford



WIKIPEDIA
The Free Encyclopedia

[Main page](#)

[Contents](#)

[Current events](#)

[Random article](#)

[About Wikipedia](#)

[Contact us](#)

[Donate](#)

[Contribute](#)

[Help](#)

[Learn to edit](#)

[Community portal](#)

[Recent changes](#)

[Upload file](#)

[Tools](#)

[What links here](#)

[Related changes](#)

[Special pages](#)

[Permanent link](#)

[Page information](#)

[Cite this page](#)

[Wikidata item](#)

[Print/export](#)

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Article

[Talk](#)

Read

[Edit](#)

[View history](#)

Search Wikipedia

Fenugreek

From Wikipedia, the free encyclopedia

"Vendayam" and "Vendhayam" redirect here. For the 2011 film, see Vengayam.

Fenugreek (/ˈfɛnjʊɡriːk/; *Trigonella foenum-graecum*) is an annual plant in the family [Fabaceae](#), with leaves consisting of three small [obovate to oblong](#) leaflets. It is cultivated worldwide as a semiarid crop. Its seeds and leaves are common ingredients in dishes from the [Indian subcontinent](#), and have been used as a culinary ingredient since ancient times.^[2] Its use as a food ingredient in small quantities is safe.^{[3][4]}

Although sold as a [dietary supplement](#),^[5] there is no [clinical evidence](#) that fenugreek has therapeutic properties.^{[4][5]} Commonly used in [traditional medicine](#), fenugreek can increase the risk of serious [adverse effects](#), including [allergic reactions](#).^{[4][5]}

Contents [[hide](#)]

- [History](#)
- [Etymology](#)

Fenugreek



Scientific classification

Kingdom: [Plantae](#)

Contents [\[hide\]](#)

- 1 [Comment moved to talk page](#)
- 2 [spelling?](#)
- 3 [Lead sentence](#)
- 4 [Rosh HaShanah custom =](#)
- 5 [Exaggerated Health Claims](#)
- 6 [This is an article not an advertisement](#)
- 7 [Un sourced Material](#)
- 8 [Taxonomy](#)
- 9 [Maybe yes or maybe not?](#)
- 10 [unsourced language section](#)
- 11 [a herb/an herb](#)
- 12 [A question about sources](#)
- 13 [Traditional Medicine](#)
- 14 [A herb/an herb](#)
- 15 [To \(hopefully\) avoid future disputes such as the above \(and the one further up on this page\)](#)
- 16 [Breastfeeding](#)
- 17 [Contradictory statements](#)
- 18 [Serious Medical Side Effects](#)
- 19 [So what's wrong with this randomized, double-blind, placebo-controlled, n=154, 90-day trial?](#)
- 20 [Perspiration smell](#)
- 21 [Fenugreek seed tea as a decongestant](#)
- 22 ['Uses' Section - Information without Citation](#)
- 23 [Lactation stimulant? No](#)
- 24 [Staple food in Galilee?](#)

*The following discussion is closed. **Please do not modify it.** Subsequent comments should be made on the appropriate discussion page. No further edits should be made to this discussion.*

User:Porphyro -- 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 piece of the MOS in place, articles would be beset by endless edit wars over style issues that would then generate endless RfCs or other forms of DR that would become a time sink across the encyclopedia. Please get consensus to change i; this is not something you can do unilaterally, much less across several articles. Jytdog (talk) 17:09, 15 March 2017 (UTC) [reply]

Hi User:Jytdog, 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. Additionally, in the above discussion, it appears to me that what consensus there was, was formed on the side of "a herb". Porphyro (talk) 20:25, 15 March 2017 (UTC) [reply]

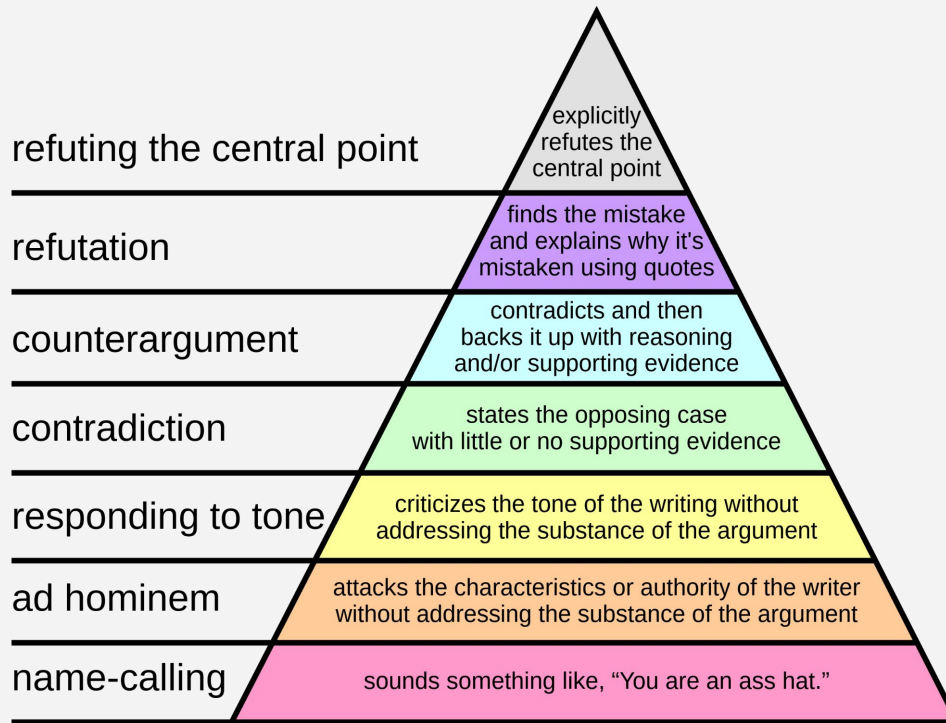
it is what it is and it already was "an herb" when the previous editor (now indeffed) starting messing with it. just leave it alone already. Jytdog (talk) 21:20, 15 March 2017 (UTC) [reply]

Be that as it may, per policy it should be "a herb" on this page, and your advice to "leave it alone" could just as easily be applied to yourself. 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. Porphyro (talk) 00:00, 16 March 2017 (UTC) [reply]

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. Jytdog (talk) 00:44, 16 March 2017 (UTC) [reply]

RESOLVED

Consensus reached per WP:ENGVAR and specifically MOS:RETAIN, that as "a herb" was the original wording, it is what should be used in the article.



Paul Graham. 2008. *How to disagree*.

<http://www.paulgraham.com/disagree.html>

Also see https://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution

These slides:
bit.ly/tomstafford

Rebuttal	Coordination
<p>DH7: Refuting the central point</p> <p>DH6: Refutation</p> <p>DH5: Counterargument</p> <p>DH4: Repeated argument/ Stating your stance</p> <p>DH3: Policing the discussion</p> <p>DH2: Attempted derailings / off-topic comments</p> <p>DH1: Attacks to the credibility of the person or the argument</p> <p>DH0: Name calling, insults and hostility</p>	<p>Coordinating edits</p> <p>Contextualisation</p> <p>Asking questions</p> <p>Providing clarification</p> <p>Suggesting a compromise</p> <p>Conceding / recanting</p> <p>Bailing out</p> <p>“I don’t know.”</p>

Fenugreek: A herb / an herb

Coordination

Rebuttal

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.

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.

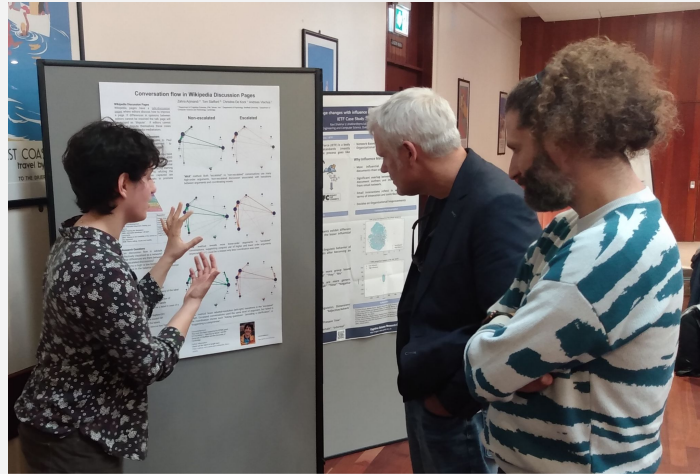
Coordinating edits
DH3: Policing the discussion

Wikitalk data

213 disagreements (3,865 utterances) from Wikipedia Talk page, 50% escalated / 50% resolved

- weak correlation with argument level and resolution
- For 535 users with >1 utterance, the median difference between the highest and lowest rebuttal level employed is 4
- 60.7% of personal attacks are in escalated conversations

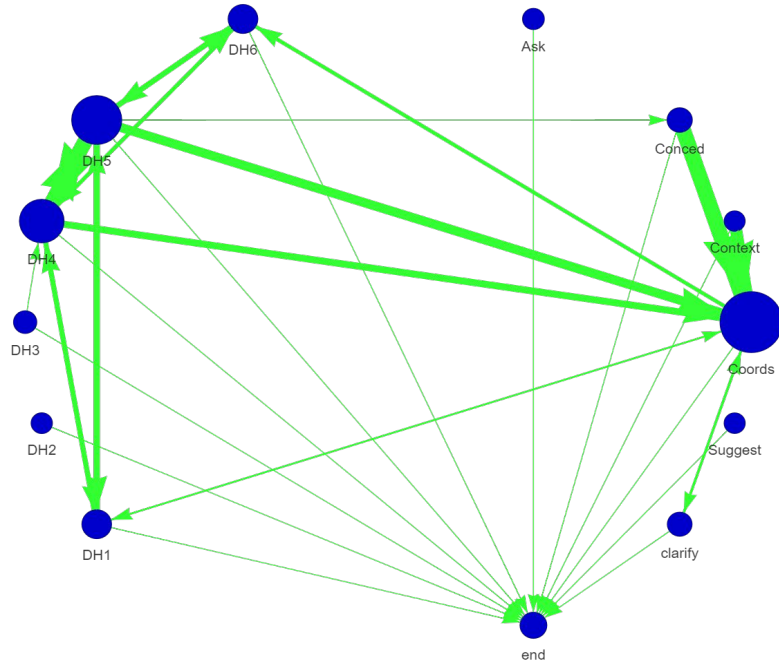
Wikitalk visualisations by Zahra Arjmandi Lari



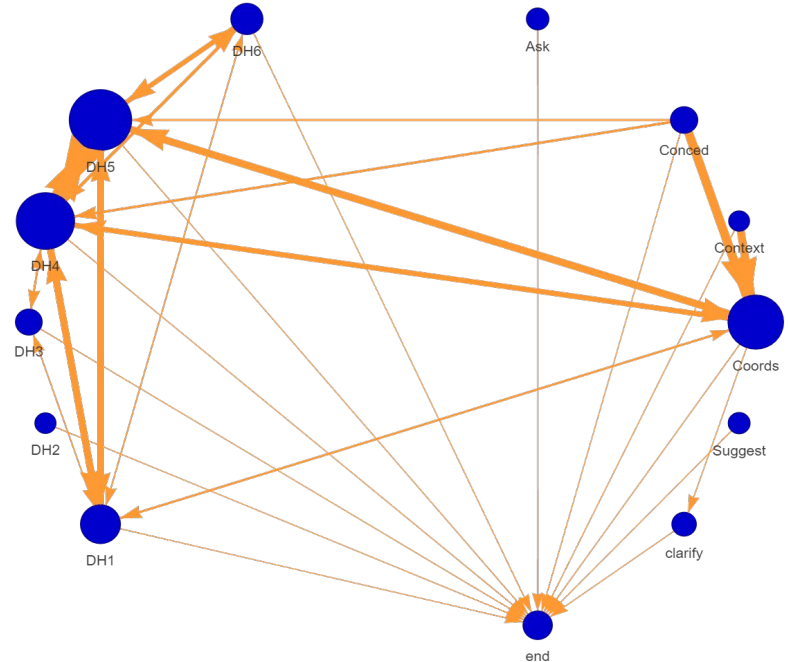
Full report

https://github.com/zahra-arjm/wikipedia_discussions

Non-escalated



Escalated



Both “escalated” vs “non-escalated” conversations use many high-order arguments. Non-escalated discussion associated with transitions between arguments and coordinating moves

These slides:
bit.ly/tomstafford

Wason & interactivist accounts

J'accuse...moi



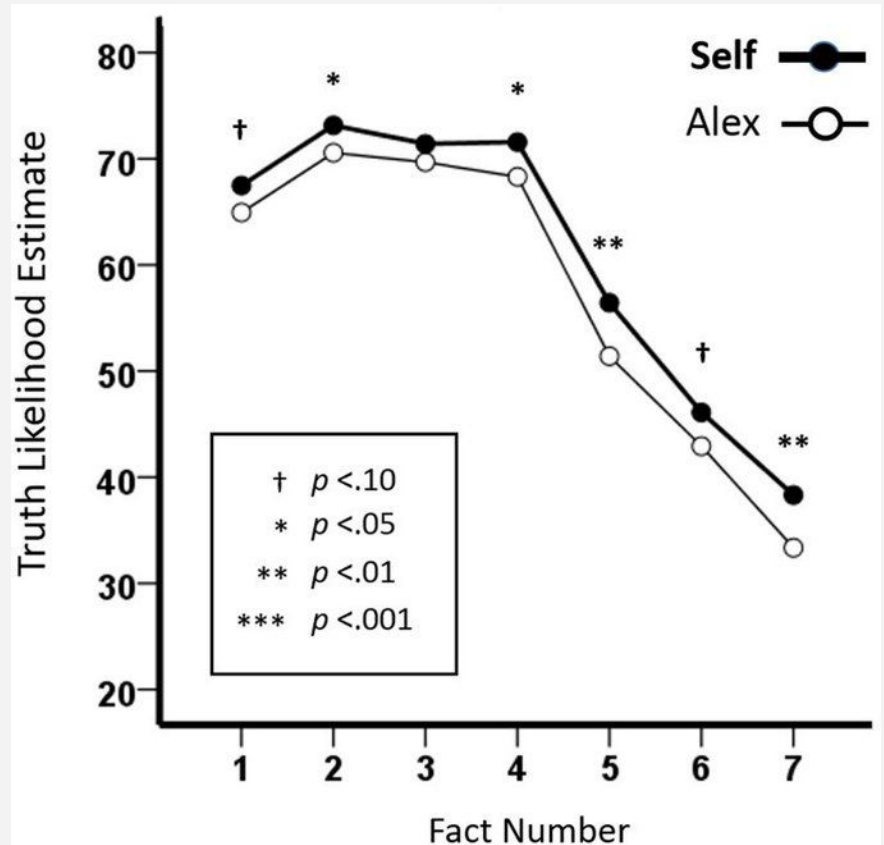
Christensen-Szalanski, J. J. J., & Beach, L. R. (1984). The citation bias: Fad and fashion in the judgment and decision literature. *American Psychologist*, 39(1), 75–78. <https://doi.org/10.1037/0003-066X.39.1.75>

Gregg et al (2017): bias (self vs other) massively smaller than the response to evidence (x-axis)

Product of the way experiments are designed to allow isolation of effects of bias (Stafford, 2014).

Gregg, A. P., Mahadevan, N., & Sedikides, C. (2017). The SPOT effect: People spontaneously prefer their own theories. *Quarterly Journal of Experimental Psychology*, 70(6), 996-1010.

Stafford, T. (2014). [The perspectival shift: how experiments on unconscious processing don't justify the claims made for them](#). *Frontiers in Psychology*, 5, 1067.



Wason's Selection Task



Which cards must you turn to test the rule :
“All cards with vowels on one side have an
even number on the other” ?

Wason's Selection Task : Lure



Which cards must you turn to test the rule :
“All cards with vowels on one side have an
even number on the other” ?

Wason's Selection Task : Easy Mode

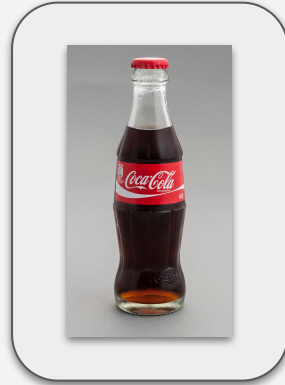


32

16

Whose drink must you check to test the rule :
“Everyone drinking alcohol must be 18” ?

Wason's Selection Task : Easy Mode



32

16

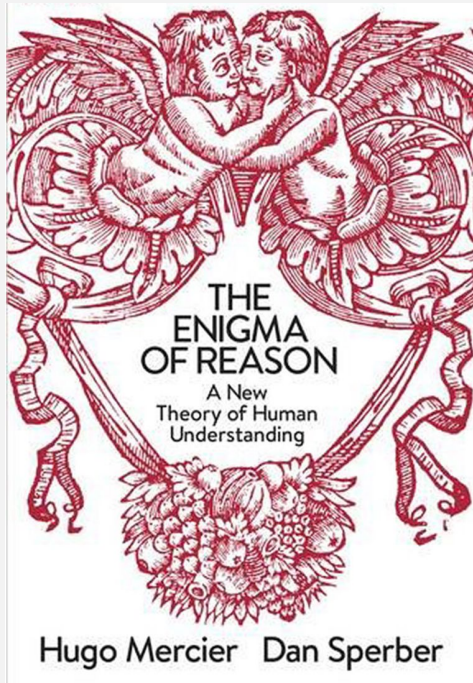
Whose drink must you check to test the rule :
“Everyone drinking alcohol must be 18” ?

Wason's Selection Task : Correct



Which cards must you turn to test the rule :
“All cards with vowels on one side have an
even number on the other” ?

Interactionist accounts of reason



Reasoning can be successful because, not despite, our biases

Mercier, H. & Sperber, D. (2018).
The enigma of reason: A New
Theory of Human Understanding.
Harvard University Press.

These slides:
bit.ly/tomstafford

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.



x

x

Zebra: What answers did everyone put?



Beaver: I put U and 2



Duck: I selected only U



Zebra: Hmm... Do we actually need 2?



Duck: If we turn 2, there may be a consonant, but that's fine



Zebra: Surely we don't need 7 right



Beaver: Why not 7 though?



Zebra: Actually, I am wrong. There may be a vowel behind 7



Duck: Do we all agree on U and 7?



[A lens on the magic of deliberation](#)

Karadzhov, G. M., Vlachos, A., & Stafford, T. (2024). The effect of diversity on group decision-making. Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0).

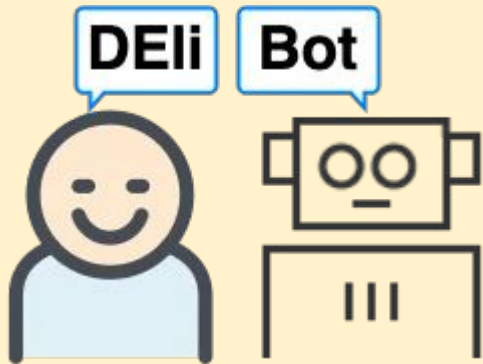
<https://escholarship.org/uc/item/6685n22z>
. arXiv preprint [arXiv:2402.01427](https://arxiv.org/abs/2402.01427).

Also:

<https://tomstafford.substack.com/p/a-lens-on-the-magic-of-deliberation>

Bee	yeah sorry
Bee	i get it now haha sorry
Giraffe	and that wouldn't break the rule
Bee	oooh hang on
Bee	yeah no i agree
Giraffe	but if 5 was a vowel then it would
Giraffe	so i think 5 needs testing
Bee	so 5 does
Bee	yep
Beaver	But if there is a vowel on the other side then it would prove it, and by not including it were missing evidence
Giraffe	to check if it is or not
Bee	i think as long as its not disproved i think its ok
Giraffe	do we all agree that 5 needs testing or
Bee	yeah
Beaver	Yeah
Beaver	So what do we think overall then?
Bee	ok so 5 defo
Bee	and e?
Beaver	Yeah
Giraffe	and E needs testing to check its an even number on other side?
Giraffe	i think 5 and e
Bee	we happy with 5 and e
Giraffe	yes
Beaver	Sure
Bee	cool
Giraffe	ive submitted that

A delibot!



delibot.xyz

DEliBots - Deliberation Enhancing Bots

Chatbots: focus on the goal of the dialogue

The dialogue is the goal

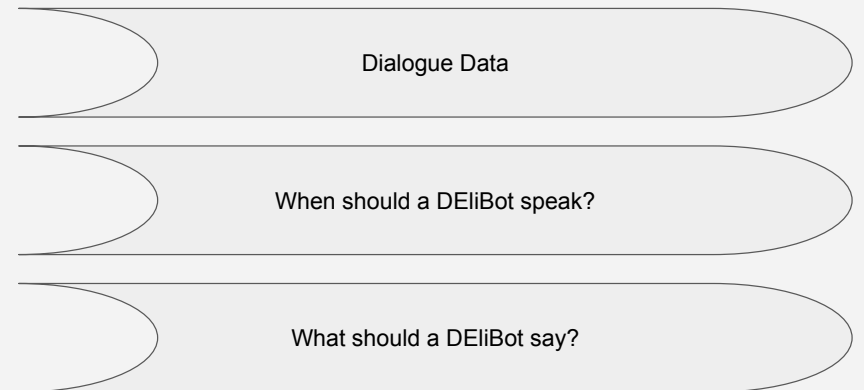
Complete a task for the person

Provide a framework that enables people to contribute towards a common goal

Recruit domain knowledge

Task agnostic

Karadzhov, G. (2023). *DEliBots : Deliberation Enhancing Bots*. PhD Thesis. <https://doi.org/10.17863/CAM.109182>



Karadzhov, G., Stafford, T., & Vlachos, A. (2021). [DeliData: A dataset for deliberation in multi-party problem solving](#). *arXiv preprint arXiv:2108.05271*.

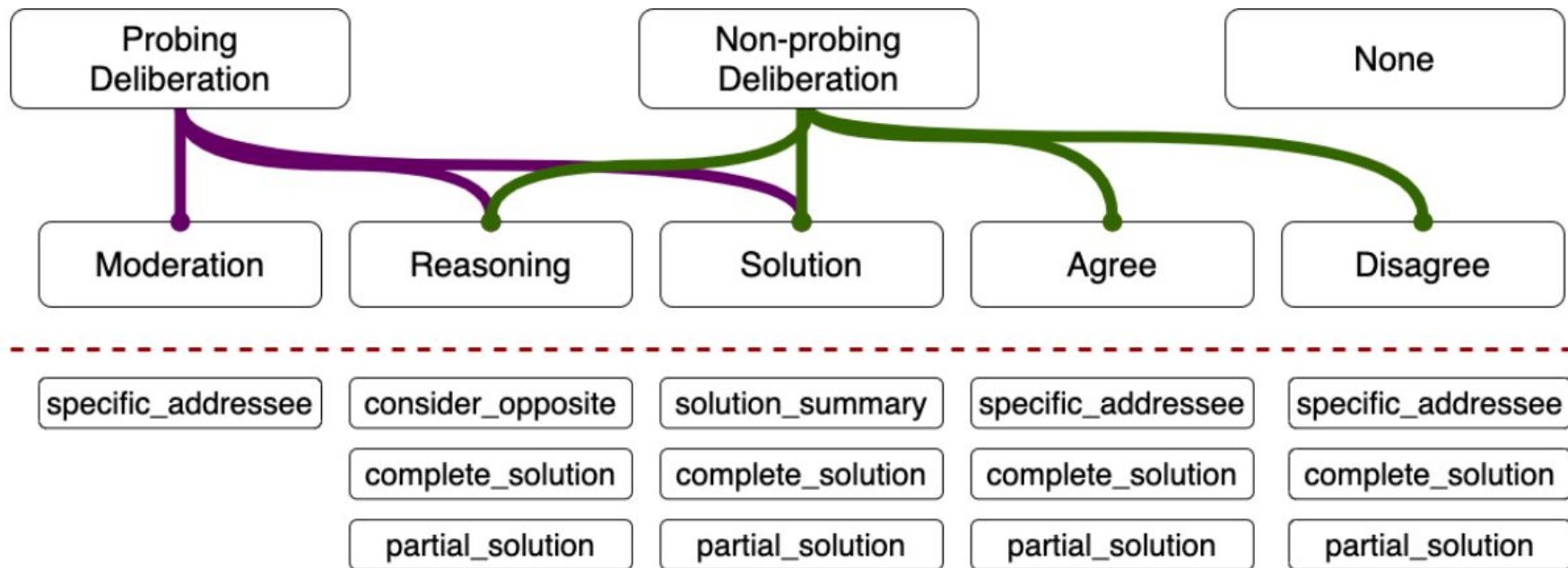
In 64% of these conversations, the group members are able to find a better solution than they had identified individually.

Annotated according to deliberative purpose

Thanks to the Isaac Newton Trust and Cambridge University Press - without their support this would not be possible

	Pilot	Mturk	Total
Number of Dialogues	18	482	500
Total Participants	53	1526	1579
Total number of utterances	705	13298	14003
AVG utterances	39.2	27.6	28
AVG utterance length	8.19	8.62	8.59
AVG unique tokens	78.1	67.6	68
AVG number of participants	2.94	3.17	3.16
Solo Performance (fine-grained)	0.59	0.59	0.59
Group performance (fine-grained)	0.81	0.71	0.72
Solo Performance (coarse-grained)	0.19	0.11	0.11
Group performance (coarse-grained)	0.57	0.32	0.33
AVG group agreement	0.92	0.83	0.83

Annotation Schema



Utterances associated with positive change

MODERATION

So , lets go one by one I think is the simplest way to reason this out

which should we go with

What did everybody put ?

so what is your thinking

What do you guys think is the answer ?

REASONING

doesn't showing rule is correct verify it ?










But , what if three has a vowel on the other side too ?

Why do you think that <MENTION> ?

So why would you flip the other 3 ?

we need to check te rule letter-number , why not the middle 2 ?

Annotated Example

	DEliBot: What answers did everyone put?	Probing	Moderation	
	Beaver: I put U and 2	NPD	Solution	complete_solution
	Duck: I selected only U	NPD	Solution	complete_solution
	DEliBot: Hmm... Do we actually need 2?	Probing	Reasoning	
	Duck: If we turn 2, there may be a consonant, but that's fine	NPD	Reasoning	
	Zebra: Surely we don't need 7 right	NPD	Solution	partial_solution
	DEliBot: Why not 7 though?	Probing	Reasoning	consider_opposite
	Zebra: Actually, I am wrong. There may be a vowel behind 7	NPD	Reasoning	
	DEliBot: Do we all agree on U and 7?	Probing	Solution	solution_summary

DeliBot with Wason -> works!

	Control	DEliBot 1.0	DEliBot 2.0
Number of utterances (by participants)	29.64	42.52 (33.52)	35.5 (31.2)
AVG utterance length	6.5	6.4	7.3
Proportion DEliBot utterances	0	0.24	0.16
Solo Performance	0.605	0.59	0.574
Group Performance	0.634	0.645	0.655
Performance Gain	0.029	0.06	0.082

How helpful was the DElibot for the discussion? (in terms of aiding collaboration and argumentation)



Collaborative Evaluation of Deepfake Text with Deliberation-Enhancing Dialogue Systems

Jooyoung Lee¹, Xiaochen Zhu², Georgi Karadzhov²
 Tom Stafford³, Andreas Vlachos², Dongwon Lee¹

¹The Pennsylvania State University, USA, {jfl5838, dongwon}@psu.

²University of Cambridge, United Kingdom, {xz479, gmk34, av308}@cam.ac.uk

³University of Sheffield, United Kingdom, t.stafford@sheffield.ac.uk

ICWSM 2026!

Session 6: Synthetic Veracity,
 08:30 AM on May 28th.

Task - Article 1

Step 1: Select the AI Machine-generated Paragraph

Please read the article with the title and first three paragraphs, where 1 (out of the 3) paragraph was generated by an AI machine and the other two were written by humans.

Please choose which one paragraph was generated by AI machine.

Title: "Feds charge woman allegedly heard during Capitol riot saying she was looking for Pelosi 'to shoot her in the friggin' brain'"

Select **The first three paragraphs**

Paragraph 1 "Feds charge woman allegedly heard during Capitol riot saying she was looking for Pelosi 'to shoot her in the friggin' brain'"

Paragraph 2 The woman, Dawn Bancroft, was charged along with Diana Santos-Smith for violent entry on Capitol grounds, remaining in a restricted area and disorderly conduct in a restricted building.

Paragraph 3 In an affidavit, investigators cited a selfie video they say was taken by Bancroft. Investigators claim she is heard saying, "We broke into the Capitol. ... We got inside, we did our part."

Step 2: Reasons to explain your choice.

To explain why the paragraphs are AI machine-generated, here is a summary of their drawbacks. Please check all explanations that satisfy the reason(s) for your choice below.

- Grammatical issues
- Repetition
- Lacks common sense
- Contains logical errors/fallacies
- Contradicts previous sentences
- Lack of creativity or boring to read
- Writing is erratic (i.e., does not have a good flow)

If other, please provide explanation below.

User A I think the answer is paragraph 1 because of grammatical issues. What do you think?

User B I chose paragraph 2 because of logical flaws.

User A Hm I see. I still think reporters would not make that type of grammar mistakes.

DeLiBot It seems like we have two conflicting arguments. What do you think User C?

Type ...

Setting	Mean Accuracy	<i>p</i>
Individual vs. Group (w/o DeepFakeDeLiBot)	45.83% vs. 54.76% (8.93% ↑)	0.0004
Individual vs. Group (w. DeepFakeDeLiBot)	48.86% vs. 57.43% (8.67% ↑)	0.0013
Group (w/o DeepFakeDeLiBot) vs. Group (w. DeepFakeDeLiBot)	54.76% vs. 57.43% (2.67% ↑)	0.482

Table 2: T-test results for individual detection performance.

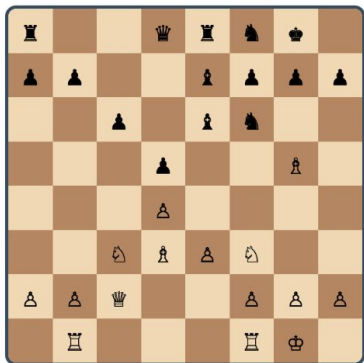
A Task description (left panel)

B Chat room (right panel)

DeliChess: A Multi-party Dialogue Dataset for Deliberation in Chess Puzzle Solving - Xiaochen Zhu, Georgi Karadzhov, Tom Stafford, Andreas Vlachos

Chess Move Analysis Workflow

Collaborative decision-making system for chess position evaluation



Move Options

1. Pawn b2 → b4
2. Bishop d3 → h7 (takes, check)
3. Knight f3 → e5
4. Pawn e3 → e4
5. Bishop g5 → f6 (takes)

Team Discussion

Koala

yeah, i did option 3 as well

Koala

either that one or option 4

Chipmunk

But with 5, if you sacrifice the bishop, then you're in with the other bishop

Emu (You)

So we should go for option 3 then? I think its the best

|Type your message as Emu...

Statistic	Value
Dialogues collected	107
Avg. messages	71.65 ± 32.76
Avg. time elapsed (min)	18.80 ± 9.12

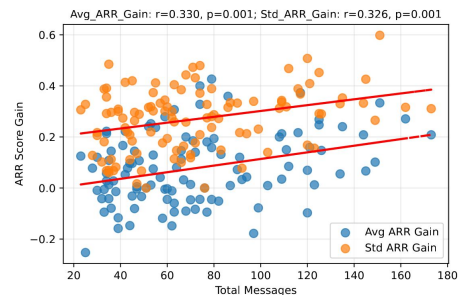
Table 1: Summary statistics of the dialogue dataset.

Puzzle Type	Messages	Time (min)
Endgame	24.35 ± 18.03	6.23 ± 5.82
Positional	26.17 ± 18.93	7.23 ± 5.91
Tactical	21.72 ± 13.31	5.34 ± 5.84

Table 2: Averaged statistics across puzzle types.

- **Endgame:** Precise play when few pieces remain, typically focusing on pawn promotion, king activity, zugzwang, or key theoretical positions.
- **Positional:** No immediate material gain or mating attack; it's about improving the position gradually to exploit opponent's weaknesses.
- **Tactical:** Spot immediate sequences that win material or deliver checkmate, through forks, pins, skewers, discovered attacks, etc.

Correlation: Dialogue Length vs ARR Performance Gains



Onward collaboration

Conformity in Large Language Models

Xiaochen Zhu*[🏰] Caiqi Zhang*[🏰] Tom Stafford[🏰] Nigel Collier[🏰] Andreas Vlachos[🏰]

[🏰]University of Cambridge [🏰]University of Sheffield

{xz479, cz391, nhc30, av308}@cam.ac.uk

t.stafford@sheffield.ac.uk

[ACL 2025 Main](#)

Demystifying Multi-Agent Debate: The Role of Confidence and Diversity

Xiaochen Zhu*[🏰] Caiqi Zhang*[🏰] Yizhou Chi[🏰] Tom Stafford[🏰] Nigel Collier[🏰] Andreas Vlachos[🏰]

[🏰]University of Cambridge [🏰]University of Sheffield

{xz479, cz391, yc697, nhc30, av308}@cam.ac.uk

t.stafford@sheffield.ac.uk

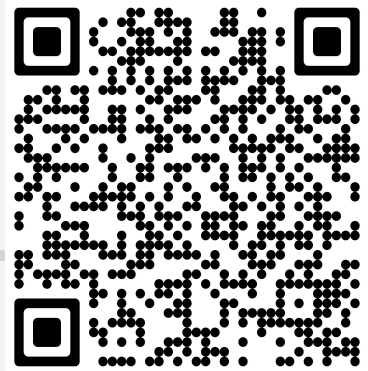
[ACL 2026 Findings](#)

These slides:
bit.ly/tomstafford

To conclude

- The promise of cognitive science is alive! CS and PSY still have a lot to learn from each other
- Human psychology (including our biases) may have evolved to suit dialogic reasoning
- Group decision-making can outperform individual decision-making, when reasons are exchanged
- Dialogue models are well suited to engage and support group function

These slides: tomstafford.github.io



RESERVE SLIDES

FOLLOW

(not for show)

Opening Up Minds

Brand, C.O., Brady, D. and Stafford, T. (2025), The Ideological Turing Test: A Behavioral Measure of Open-Mindedness and Perspective-Taking. *Cognitive Science*, 49: e70126.

<https://doi.org/10.1111/cogs.70126>

Farag, Youmna; Brand, Charlotte; Amidei, Jacopo; Piwek, Paul; Stafford, Tom; Stoyanchev, Svetlana and Vlachos, Andreas (2022). Opening up Minds with Argumentative Dialogues. In: Findings of EMNLP (Empirical Methods in Natural Language Processing), 7-11 Dec 2022, Abu Dhabi, United Arab Emirates.

Opening Up Minds:
engaging dialogue
generated from
argument maps



PI: Paul Piwek

http://mcs.open.ac.uk/pp2464/Opening_Up_Minds

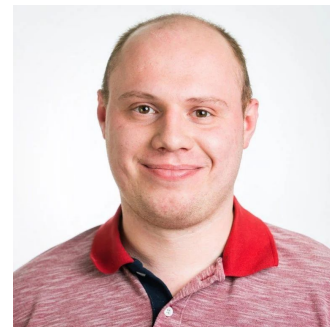


**Engineering and
Physical Sciences
Research Council**

In progress: DEliBots!

- Using dialogue agents to expose people to arguments for other positions and facilitate group deliberation

- [Collaborative Evaluation of Deepfake Text with Deliberation-Enhancing Dialogue Systems](#). J Lee, X Zhu, G Karadzhov, T Stafford, A Vlachos, D Lee. arXiv preprint arXiv:2503.04945
- [The effect of diversity on group decision-making](#). G Karadzhov, A Vlachos, T Stafford. Cognitive Science Conference 2024
- [DeliData: A dataset for deliberation in multi-party problem solving](#). G Karadzhov, T Stafford, A Vlachos. CSCW'2023
- [What makes you change your mind? An empirical investigation in online group decision-making conversations](#)
- G Karadzhov, T Stafford, A Vlachos. SIGDIAL'2022 Special Interest Group on Dialogue and Discourse



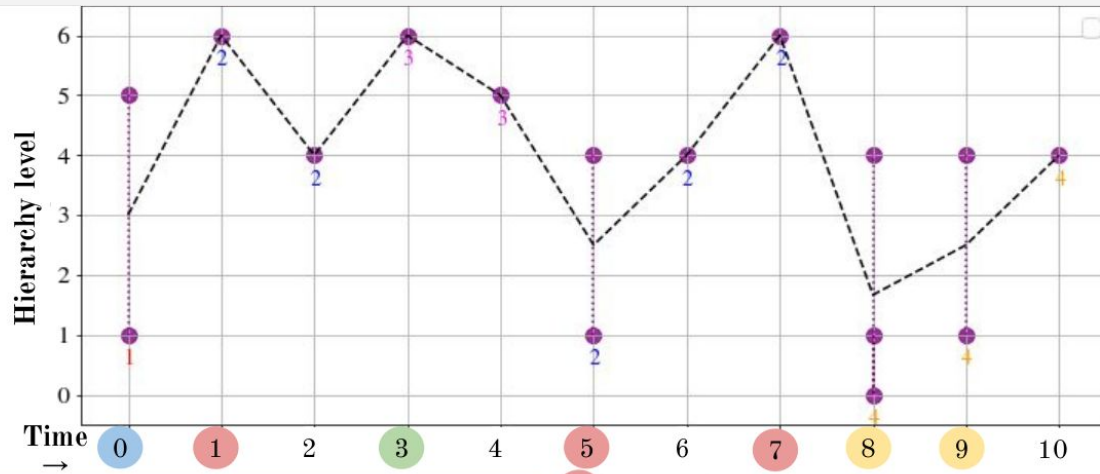
delibot.xyz

“Reasonable People” newsletter

<https://tomstafford.substack.com>



More on
Wason



0 Indian tamils are not stateless, it has to be the most absurd thing i have ever come across, it's very biased propaganada. Tamil Nadu is where majority of Tamils live and we don't identify anything other than Indian and Tamil. Tamil is one of many official languages of India and also has classical language status.

1 Tamil as a whole are a stateless nation. Tamil claim to be a nation but there is no sovereign Tamil state. If Tamils consider themselves a nation and somewhere in the world exists a movement for a Tamil state, then Tamils as whole are a stateless nation. To claim that Sri Lankan Tamils are stateless nation and India Tamils not, makes no sense, because both are the same ethnic group and so also the same nation."

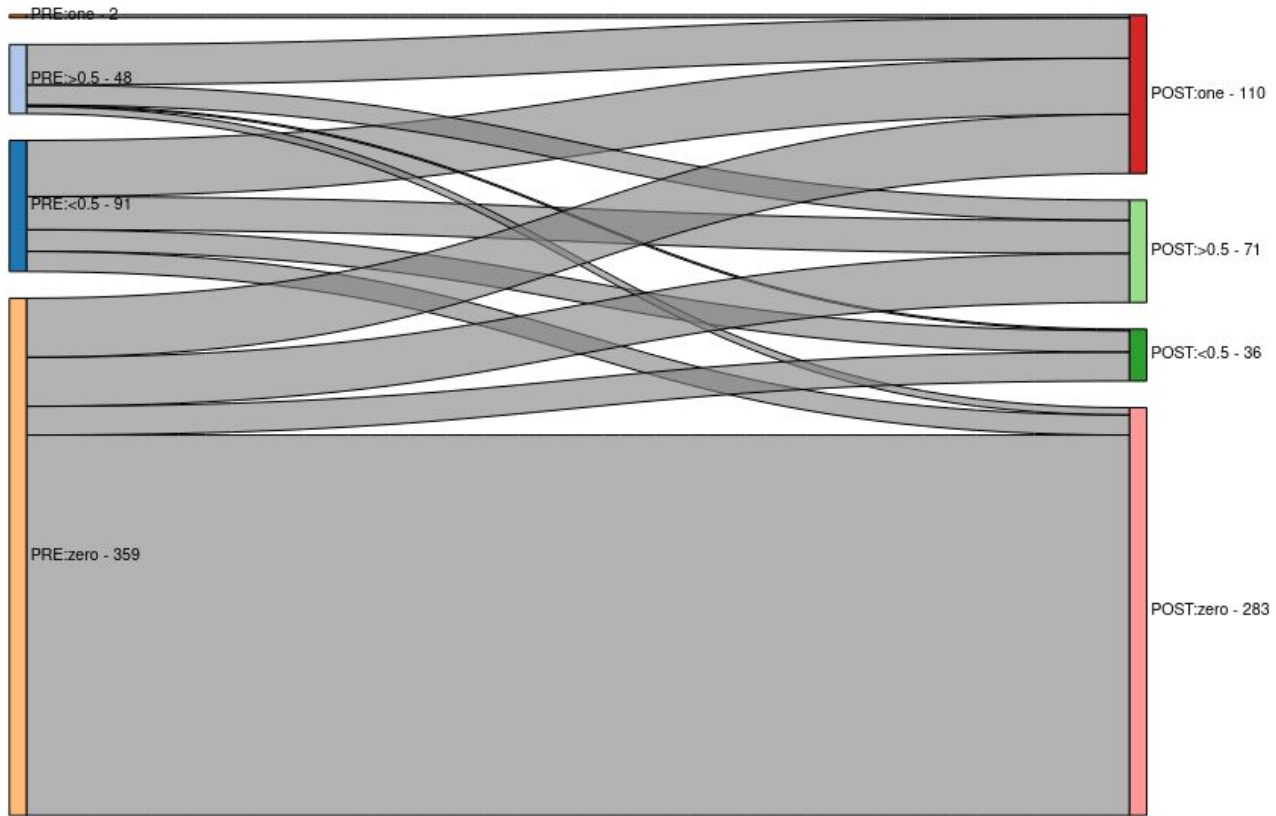
3 This concept of stateless nation only applies to Sri Lankan Tamils who are demanding Tamil Eelam NOT Indian Tamils. Some political movements in Tamil Nadu may support Eelam nation but that does not mean we want to leave India. There is difference between Indian Tamils and Sri lankan-Tamils when it comes to politics, we are fully absorbed into Indian identity and society and have always been. Same can't be said about Sri Lankan-Tamils and Sinhalese in modern times."

5 We have not even finished our discussion and you're already imposing your POV on Tamil people. Probably you have a ulterior motive to divide Tamils and to deny the Tamil nationalism. The entire Tamil population belongs to a nation and this Tamil nation has no sovereign state, which makes Tamils to the largest stateless nation in the world.

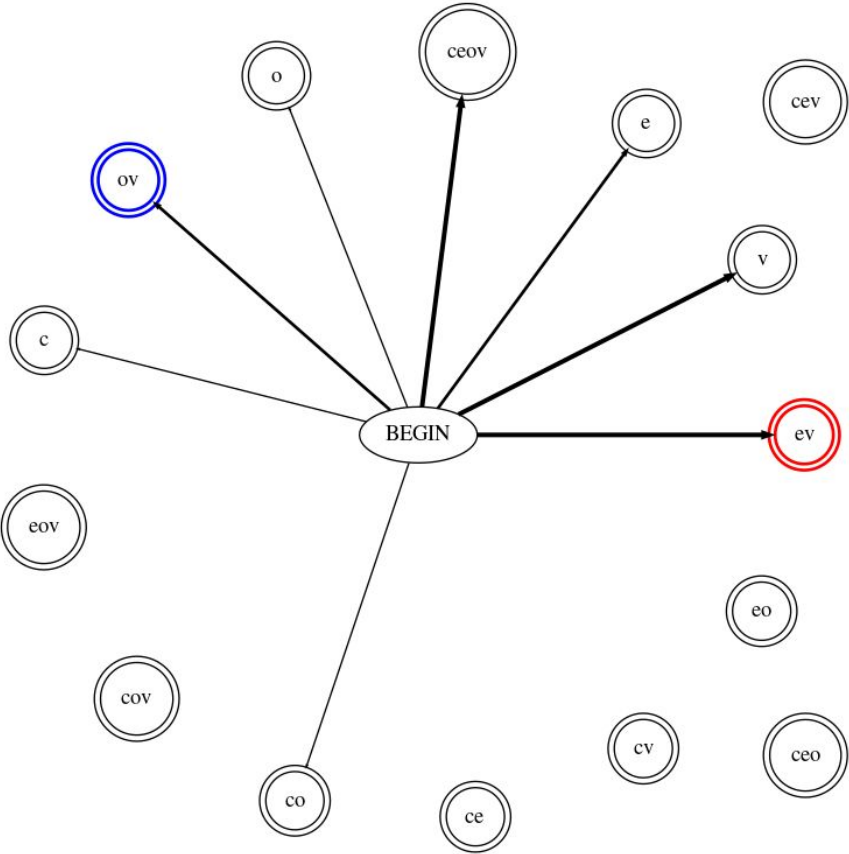
7 Tamil secession movement even existed in India in the past. Indian government had added a legislation that outlawed anyone wanting independence from India and so the Tamil secession movement in India got weaker. You can not deny this because it's history. A famous quote by Tamil poet Kannadasan about the Tamils as stateless nation. Historical Dictionary of the Tamils (2007), p. 319.

8 Only person who is imposing anything here is YOU with your "77 million tamils being stateless nation" are you insane? It's obviously propaganada! You'll be laughed at your face if you asked anyone from Tamil Nadu about wanting separate nation for Indian Tamils or to join Eelam.

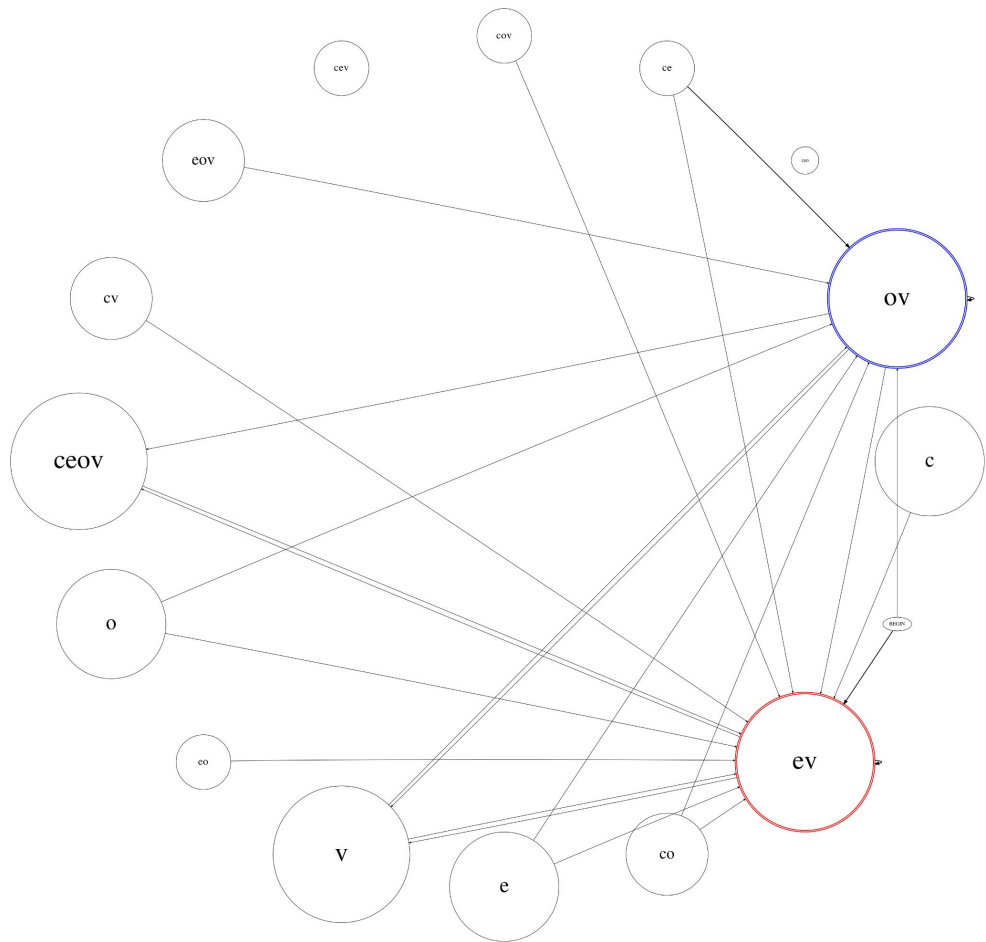
9 Again, just because some Indian parties have solidarity with Sri Lankan Tamils for Eelam state does NOT mean Indian Tamils want to secede from Indian union, even you know that.



Odd + Vowel : "Target"



Even + Vowel : "The Lure"



Opening Up Minds

BBC Radio 4 programme 'The Moral Maze'

Well, the problem we always have with debt, which I would use rather than credit, is that ultimately there is interest. And the interest has to be paid back as well as the debt.

witness




chair



panellists



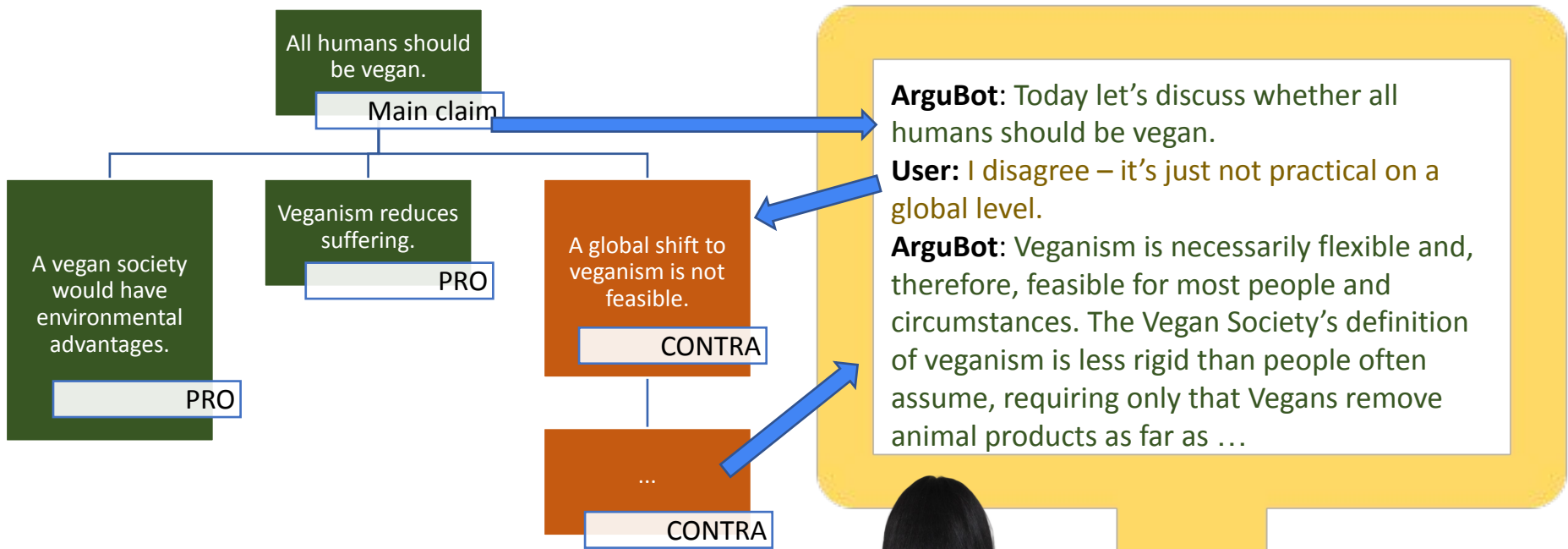
Well, it's interesting that you should imply that saving is morally good and can lead you to not be dependent, because actually sometimes taking out credit can also help you not be dependent. Is that not true?

A person with long black hair, wearing a red long-sleeved shirt, is seen from behind, sitting in a black wheelchair. They are looking at a large, stylized yellow monitor. The monitor has a white rectangular area in the center containing text. The monitor's base is a thick yellow bar with a central rectangular cutout.

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

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

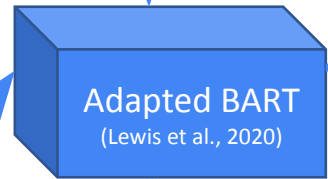
ArguBot: Veganism is necessarily flexible and, therefore, feasible for most people and circumstances. The Vegan Society's definition of veganism is less rigid than people often assume, requiring only that Vegans remove animal products as far as ...



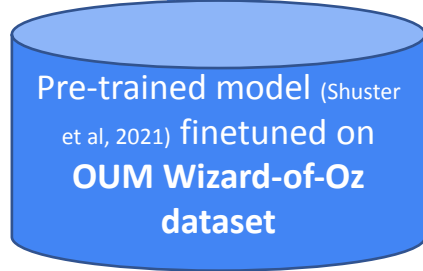
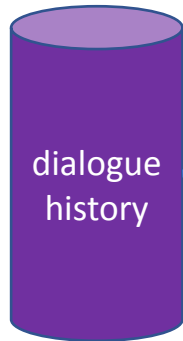
Argument map source:
kialo.com



Retrieval-Augmented Generation (RAG)-Sequence model (Lewis et al. 2020)



ArguBot chat contribution



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 ...



ROYAL SOCIETY OPEN SCIENCE

Open Access

Check for updates

View PDF

Tools Share

Cite this article

Section

Abstract

1. Introduction
 2. Method
 3. Results
 4. Discussion
- Ethics

Research articles

Using dialogues to increase positive attitudes towards COVID-19 vaccines in a vaccine-hesitant UK population

Charlotte O. Brand and Tom Stafford

Published: 12 October 2022 | <https://doi.org/10.1098/rsos.220366>

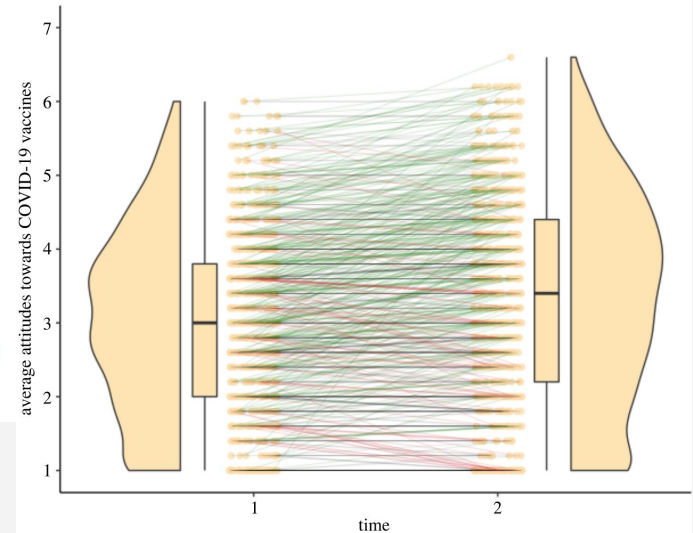
Review history

Abstract

Recently, Altay *et al.* (Altay *et al.* 2021. *J. Exp. Psychol.: Appl.* (doi:10.1037/xap0000400)) showed that 5 min of interaction with a chatbot led to increases in positive COVID-19 vaccination attitudes and intentions in a French population. Here we replicate this effect in a vaccine-hesitant, UK-based population. We attempt to isolate what made the chatbot condition effective by controlling the amount of information provided, the trustworthiness of the information and the level of interactivity. Like Altay *et al.*, our



attitudes before and after dialogue exposure



Which question are you most interested in?

Are vaccines dangerous in general?

Do we know enough about Covid-19 to develop a vaccine?

Could the virus mutate, reducing the effectiveness of the vaccine?

Can we trust vaccines from all countries?

Is Covid-19 dangerous for someone young like me?



<https://tomstafford.substack.com/p/vaccine-dialogues>

These slides:
<http://bit.ly/tom-talks>

Comments from vaccine hesitant participants after taking part in our study



Use these buttons to see comments including the following words, or to see all comments:

Government

Media

Trust

Big Pharma

Science

Research

Time

Health

Safety

Thanks

See All Comments

Search:

Comments

"Scientific consensus" is not to be trusted when so many are reliant on funding from pharmaceutical companies. They have much to lose if they do not follow the narrative.

From an individual perspective, I have no issues surrounding vaccinations. With an extensive travel history I have received more vaccinations than the average individual, however my concern regarding the Covid vaccination lies in just how much of a threat is Covid to the general population? I do not trust the Media narrative or figures hence it is a person's free will choice whether they should accept this vaccine or not. Much more research needs to be done into just how dangerous Covid 19 actually is.

I had one vaccine, felt obligated to do so and everyone I spoke to was having it. I had the Astra Zeneca vaccine, cheap one. Then all the news was on about blood clots and countries blocking the AZ vaccine as it wasn't safe. I then cancelled my 2nd vaccine appointment after that. The vaccine is untested, too many intelligent people are against it, the Government tells lies and too many of them for me to believe a word they say, to the point I will never vote in an election ever again. For such a killer virus going around there has been far too much normality and the vaccine rules and protocol continually change so why should people trust it. If it's a killer virus and everyone has to be vaccinated then why are people allowed to fly in and out of the country since we first knew about it. We don't even know if the vaccine stops you being a virus carrier, it's clearly an untested vaccine. Also since when does the Government care about old people, disabled, sick people but now they are the important ones at the top of the list to be vaccinated.

information about vaccines seems v biased towards having it hence the distrust i'm not anti vaccines - i just want to be fully informed and resent being referred to as an 'anti vaxxer' (you haven't! :-))

<https://lottybrand.shinyapps.io/vaccineComments/>

The (Ideological) Turing Test

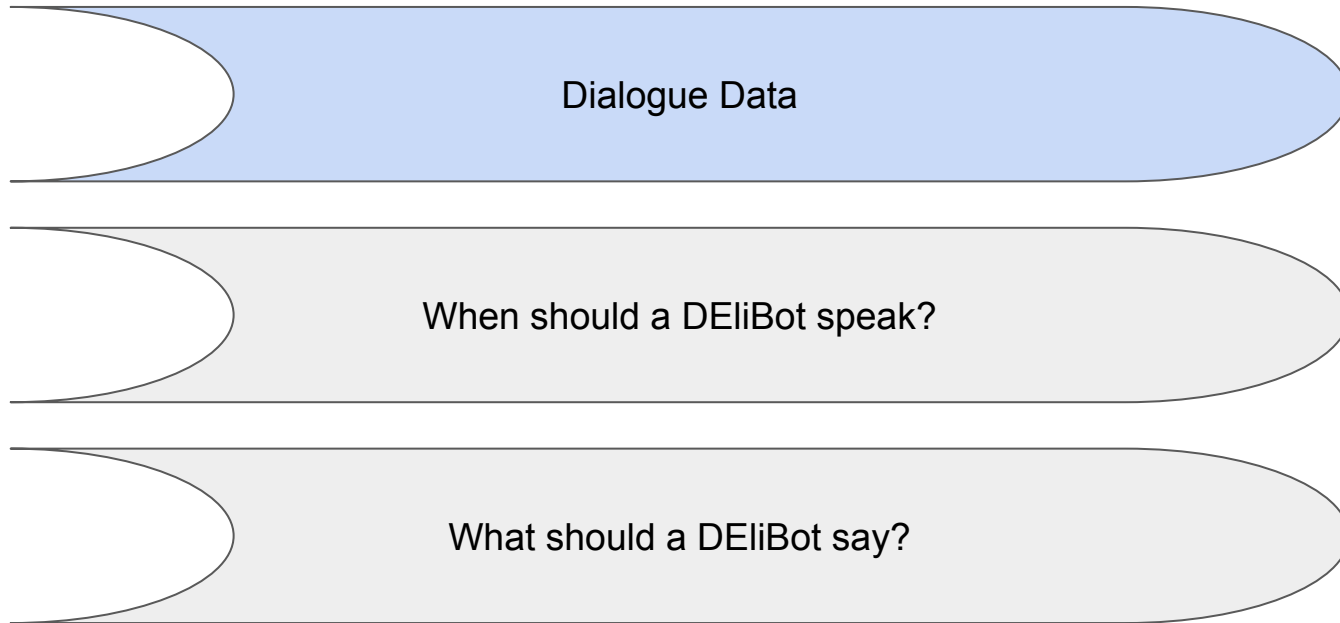
Originally, Bryan Caplan (2011),
after Alan Turing (1950).

“We should “attempt to re-express your target’s position so clearly, vividly, and fairly that your target says, ‘Thanks, I wish I’d thought of putting it that way’,”

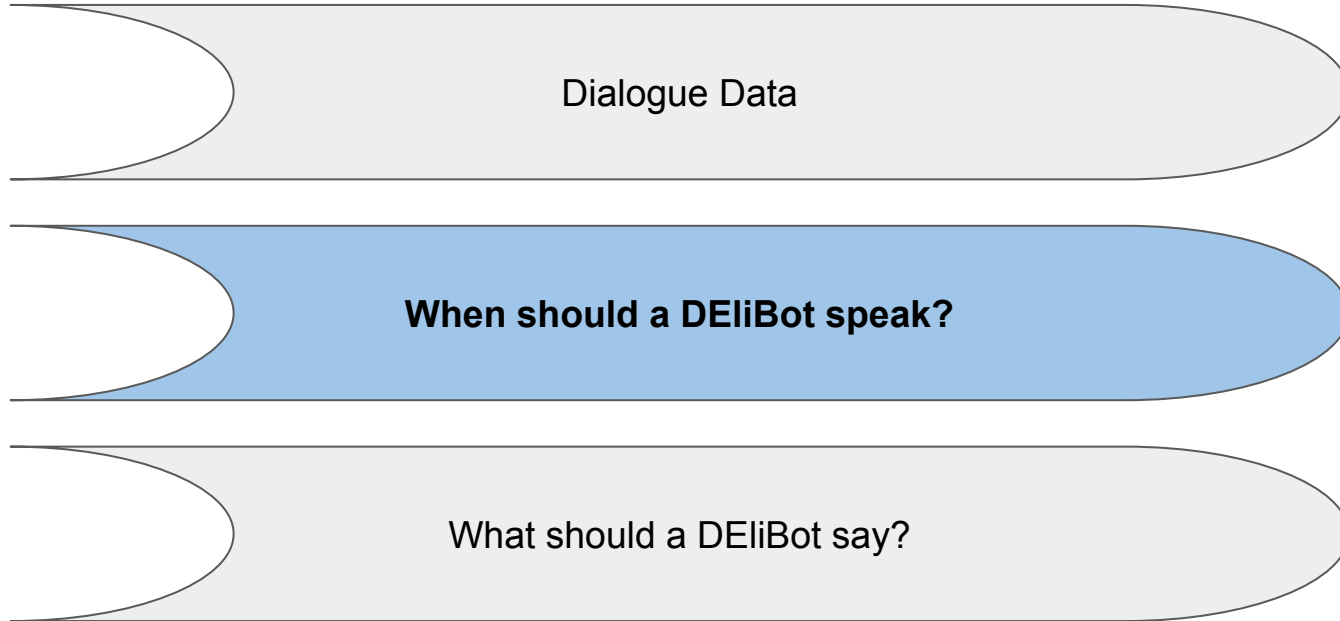
Dennett, D. C. (2013). Intuition pumps and other tools for thinking. WW Norton & Company.

Delibot!

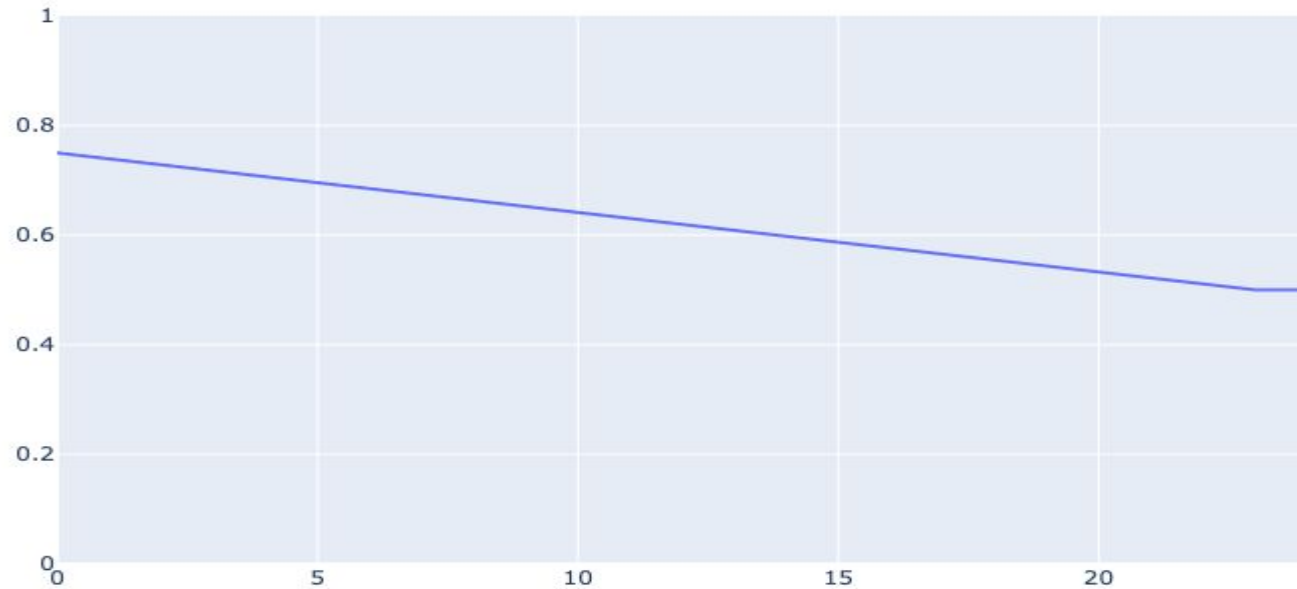
How to build a DEliBot?



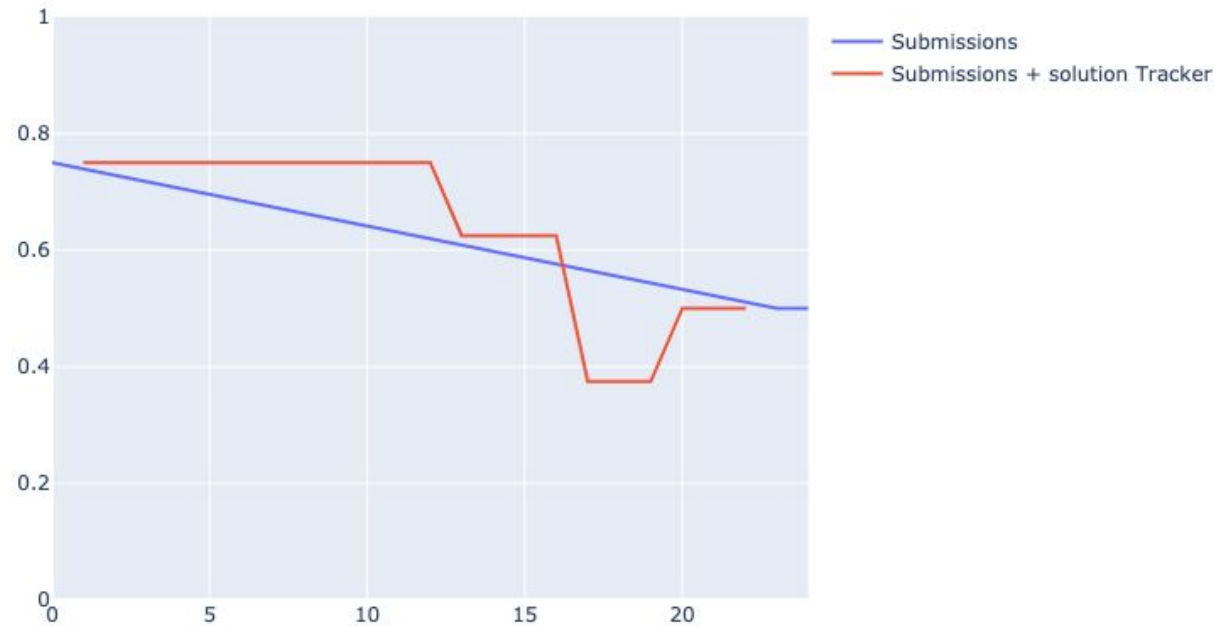
How to build a DEliBot?



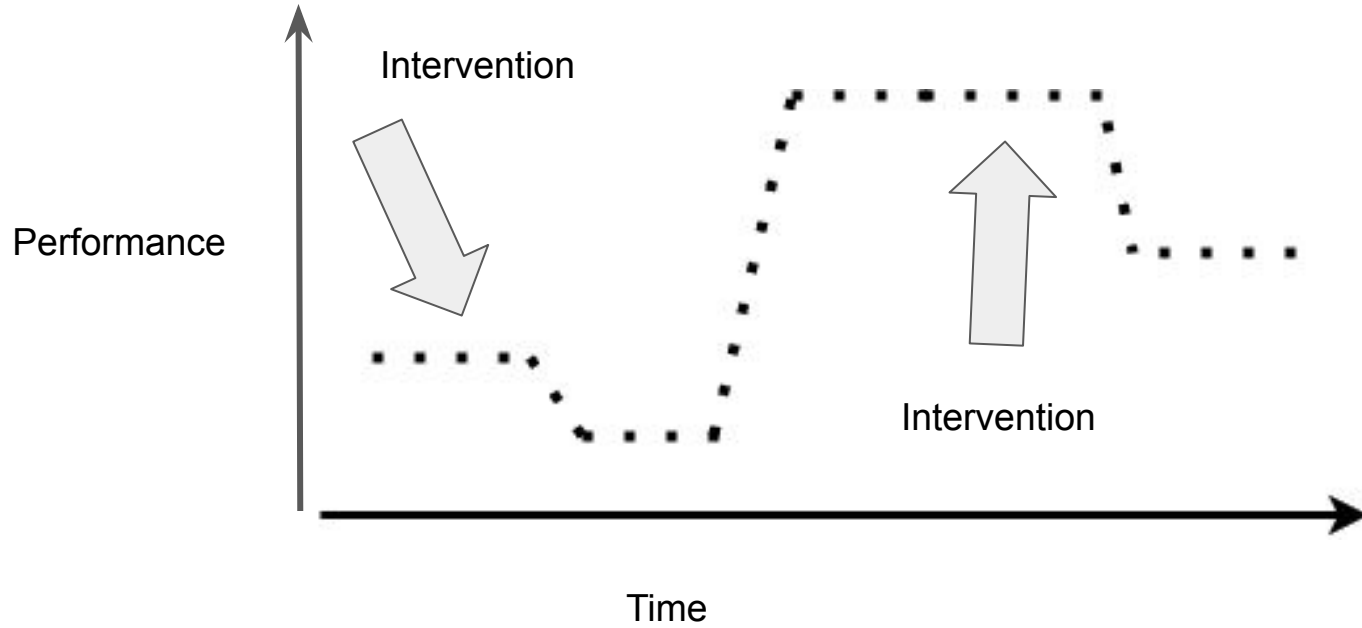
How did the conversation go?



How did the conversation go?



Addressing the timing of DEliBot participation



Can we predict when will someone change their mind?

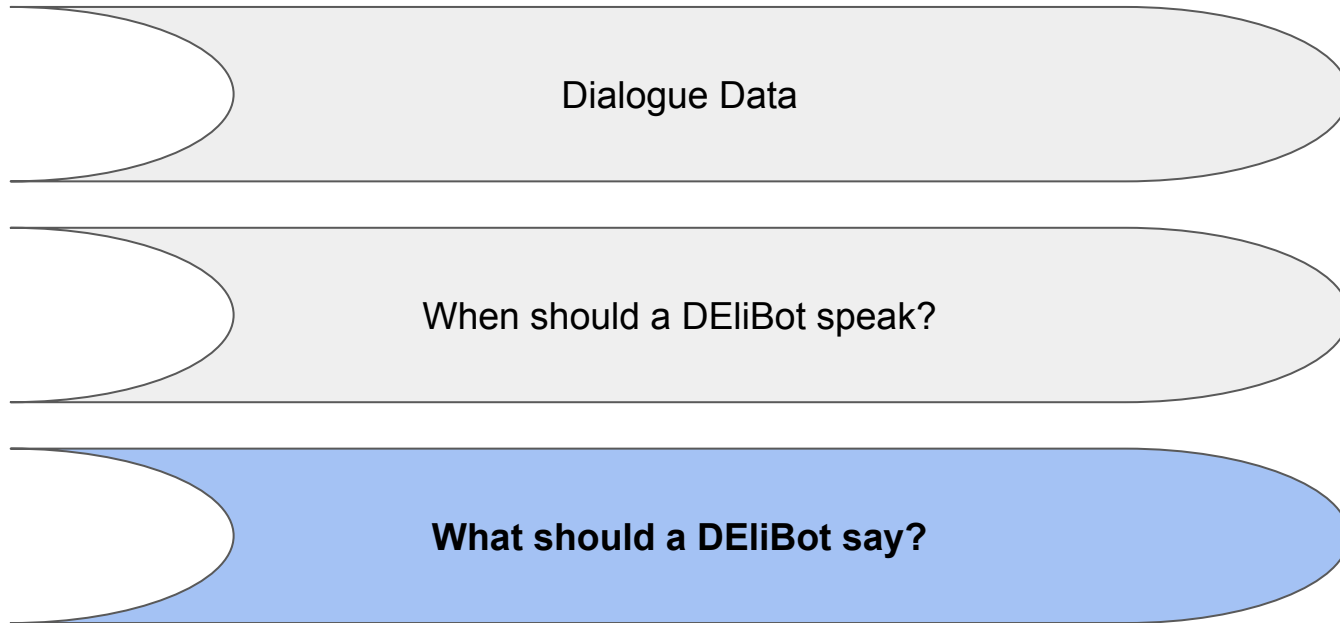
- Formalise it as a task of change-point detection (reliability engineering)
- Hard task, but not impossible

[What makes you change your mind? An empirical investigation in online group decision-making conversations](#)

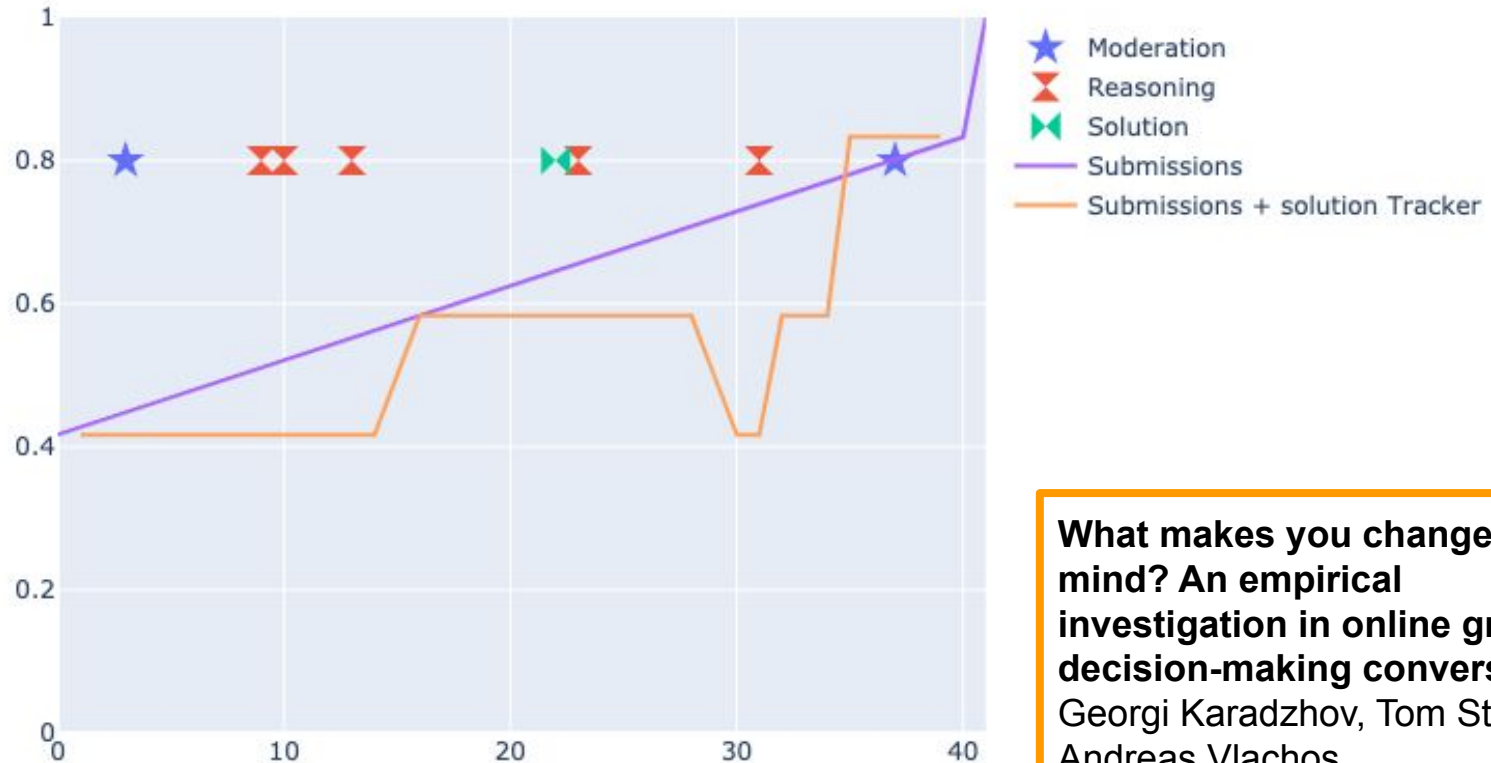
Georgi Karadzhov, Tom Stafford,
Andreas Vlachos. SIGDIAL'22

	Micro AUC	Cutoff P-R
BoW + Random Forest	0.19	0.21
Bayesian Online CP detection	0.18	0.22
BOCP + Positional Prior	0.21	0.26
Linguistic Model (Neural)	0.20	0.23
Learning to Rank	0.23	0.24
Learning to Rank + BOCP + Positional Prior	0.25	0.30

How to build a DEliBot?

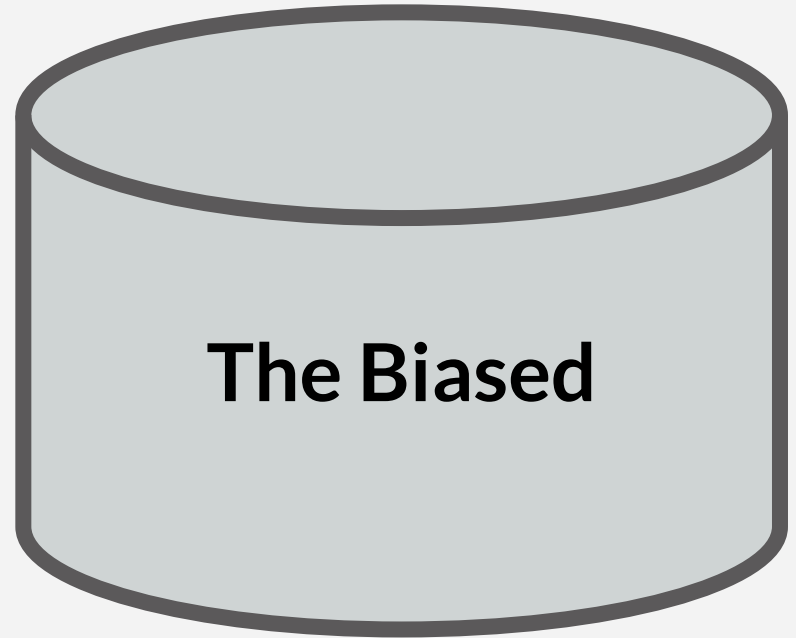
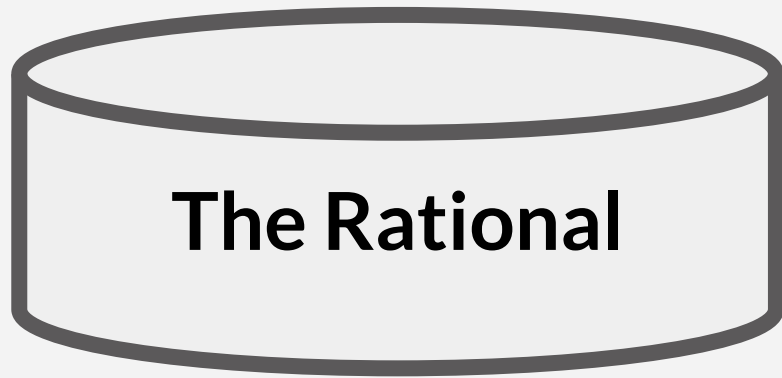


Finding good things to say

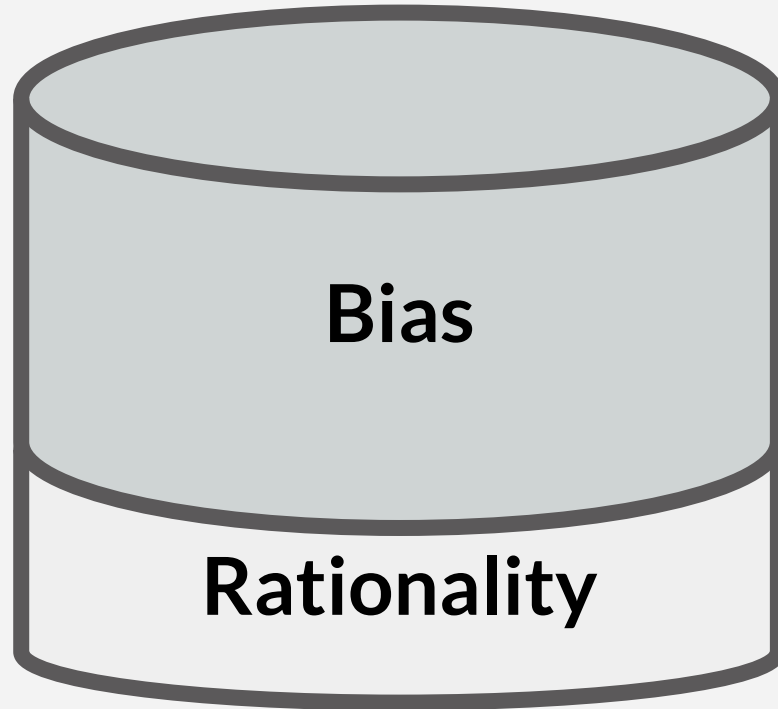


What makes you change your mind? An empirical investigation in online group decision-making conversations
Georgi Karadzhov, Tom Stafford, Andreas Vlachos

More on
human bias



W. Phillips Davison, "The Third-Person Effect in Communication,"
Public Opinion Quarterly 47, no. 1 (1983): 1–15





Rationality

Bias

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Every rational reason has a bias



Image: torbakhopper
"fire alarm bell, scott richard"

CC BY-ND 2.0

<https://www.flickr.com/photos/gazeronly/28520377874>

Signal detection theory

		You judge:	
Claim:		“True”	“False”
True		HIT!	Miss
False		False alarm	Correct Rejection

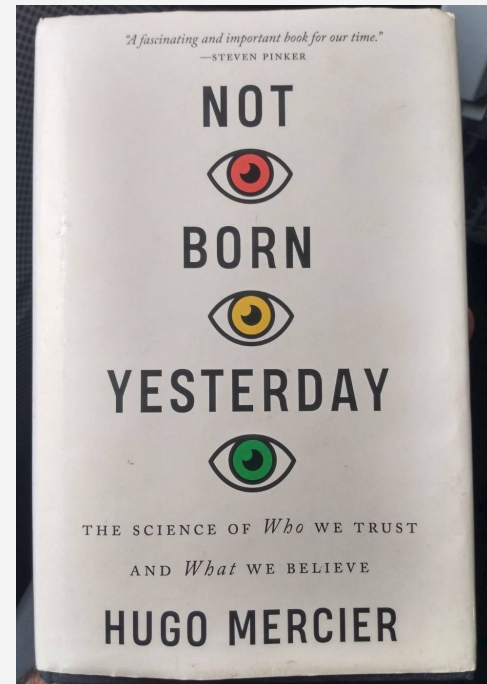
Most misinformation not seen, not believed or not acted upon

Polarisation is exaggerated

Echo chambers don't exist online

Backfire effect rare

Biased assimilation



Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media+ Society*, 9(1), 20563051221150412.

Coppock, A. (2022). *Persuasion in Parallel: How Information Changes Minds about Politics*. University of Chicago Press.

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, 118(15), e1912440117.

On

fact-checking

How we reason

Social is primary

dialogue as the natural interface to human reason

this springs from interactionist accounts of reason

following from this
reputation tracking

trust systems

adversariality

BUY A FUCKING SHOVEL



Well fuck. Here we are again – it's another snowma-fucking-ggeddon.

As if being trapped indoors for months at a time wasn't bad enough (fuck you COVID), your relatively blissful once-a-week trip to the supermarket (that previously served as a welcome escape from the monotony of home-fucking-schooling) now requires you to first excavate the stupid fucking car using a stupid fucking tea tray – and all because you didn't heed our previous warnings to [buy a proper fucking shovel](#)*.

To make matters worse, because you didn't buy the kids a [decent fucking sledge](#)* the last time, you've had to send them off with a fucking tea tray and a couple of plastic fucking bags, so they're probably going to hate you forever.

Sure, you could try having them delivered from Amazon, but it'll probably cost you twice the normal price and by the time it comes it'll be fucking Easter.

So yeah, 2021 has started off about as good as 2020 was, but all is not lost. Fill in the form below, and on June 15th 2021 when snow shovels are cheap and plentiful again we'll email you a reminder to **buy a fucking shovel**.

Act now to make sure next winter is a fucking success for once:

Buy a fucking shovel

It looks like you're trying to
reconsider my legacy.

Would you like help?

Yes

No



What is the space of effect?

Will Moy: “fact checking is a behavioural intervention”

No: fact checking is a reputational system intervention
