Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Psychology

Review

The wilful rejection of psychological and behavioural interventions

Jon Roozenbeek[1,a], David J. Young[1,2] and Jens Koed Madsen[2,a]

Psychology and behavioural science play a key role in the development, testing, and implementation of interventions aimed at addressing societal challenges. Some of these interventions have been impactful in shaping policy decisions, but their successful real-world implementation is beset by challenges, including a large number of people who might benefit from an intervention choosing to ignore it. However, there is almost no research on why people wilfully reject participating in an intervention: they notice it, consider participation, and decide against it. Addressing this knowledge gap is of critical importance for improving intervention uptake. Drawing on the literature on wilful ignorance, we propose a Bayesian model of the wilful rejection of psychological and behavioural interventions. People's prior beliefs about the relevance of an intervention, its effectiveness, and the goals and reliability of the intervention's source, strongly inform the probability of people wilfully rejecting an intervention when they come across it. Based on this model, we argue that people may downgrade their perceptions of the source's reliability if they perceive the intervention itself to be inefficacious, and that using intervention sources with high perceived reliability among target audiences is key to optimising intervention uptake.

Addresses
[1] Department of Psychology, University of Cambridge, United Kingdom
[2] Department of Psychological and Behavioural Science, London School of Economics, United Kingdom

Corresponding authors: Roozenbeek, Jon (jjr51@cam.ac.uk); Madsen, Jens Koed (j.madsen2@lse.ac.uk)
[a] Equal contribution.

Current Opinion in Psychology 2025, **66:**102138

This review comes from a themed issue on Willful Ignorance

Edited by Shaul Salvi, Susann Fiedler, Fiona Tho Pesch and Anna Baumert

For complete overview about the section, refer XXXXX

Available online 11 August 2025

https://doi.org/10.1016/j.copsyc.2025.102138

Psychology and behavioural science play an instrumental role in the development, testing, and implementation of interventions aimed that address keystone societal challenges. These include mis- and disinformation [1], mental health [2], political polarisation [3], and the COVID-19 pandemic response [4]. This effort has yielded substantial successes. For example, in the misinformation space, tech platforms such as Google and Meta have implemented "prebunking" and behavioural nudges as part of their online harms policies (see Ref. [5] for a review). More broadly, the United Kingdom and other national governments have long relied on behavioural insights for policymaking and evaluation [6].

However, the success of psychological and behavioural interventions has not been unequivocal. Some behavioural "nudge" interventions yield small or even negligible effects when implemented in real-world environments [7], intervention effects can decay rapidly without "booster shots" [8], and some counter-misinformation interventions have become increasingly politicised, particularly in the United States, limiting their effectiveness [9,10]. Moreover, interventions may backfire — for example, introducing a seemingly reasonable communal limit on fishing quotas may spark internal competition, which initiates a "race for fish" and biomass depletion [11]. It is therefore necessary to consider why some interventions, but not others, succeed at scale.

Implementation science and wilful ignorance

Researchers have argued that while lab-based interventions research may demonstrate improved performance on key outcome measures [1], the path towards intervention *effectiveness* (i.e., measurable real-world impact) is beset by several challenges. These include low intervention uptake, reduced impact over longer time periods, and the need to motivate people to take part [12—14]. Within this realm, psychology and behavioural science have generally neglected to integrate insights from implementation science, or the study of how innovations go from lab-based efficacy to adoption into routine usage [15,16].

A key open question in this domain is why individuals choose to participate in, or reject, an intervention: what makes a person "nudgeable" [17]? Why would someone choose to play a media literacy game [18], or watch a video about positive intergroup contact [3]? Landry and Halperin [12] argue that individuals often lack motivation to participate in interventions. However, understanding the *wilful* rejection of interventions is a critical and understudied research domain. By wilful rejection, we mean a situation where a person comes across an intervention, considers participating in it, and makes a conscious decision not to do so. Lewandowsky and Oberauer [19] argue that the cognitive mechanisms underlying the related concept of the wilful rejection of scientific findings "are found regardless of political orientation", and that education and literacy interventions do not mitigate this rejection, but rather "increase the polarisation of opinions along partisan lines".

Following this line of argument, we posit that the wilful rejection of interventions is not a function of people's background (e.g., their partisan affiliation or education level), but is instead grounded in their subjective degrees of belief about the reliability of the source of the intervention, its perceived effectiveness, and the probability that the issue that the intervention addresses constitutes a problem that requires a solution [20]. These considerations are related to wilful ignorance, which occurs when people deliberately avoid information, for example when avoiding learning about negative consequences of one's actions [21]. The literature on wilful ignorance identifies identity projection and self-image preservation as being strongly linked to motivating psychological factors around the emergence of trust perceptions. That is, akin to belief bias, if a person provides a statement that challenges my identity, my immediate response may be to distrust them [22]. Here, we propose that the wilful rejection of interventions is analogous to identity-preserving wilful ignorance, with both referring to the deliberate avoidance of information based on subjective prior beliefs. However, the two concepts are not synonymous, as wilful rejection does not require ignorance about an intervention per se. Rather, we propose a process by which people may reject an intervention, even when having full information about it. This process is also related to, but distinct from, psychological reactance, which occurs when individuals perceive that interventions limit their freedom of choice [23]: instead, wilful rejection can occur under many conditions, even when perceived freedom of choice is not affected.

A Bayesian lens

We argue that the wilful rejection of interventions in the real world can be explained from a Bayesian perspective, which departs from people's subjective degrees of belief, ranging between 0 and 1. Grounded in Bayes' theorem [24], this approach describes how people should update their beliefs about the world given evidence (or *testimony*) from sources. Relevant beliefs about a given intervention include the belief that the problem it addresses is real (and important), that the person's own behaviour contributes to the issue, and that the intervention is benign (e.g., it doesn't seek to "brainwash" them). Degrees of belief are subjective, meaning that people can differ in their perception of a problem due to their exposure to prior evidence, their information sampling, or how they perceive sources' reliability or trustworthiness. When a person is offered the opportunity to take part, they receive testimony from the source of the intervention that they should take part, and they may also receive persuasive messages designed to target related beliefs concerning the reality and relevance of the problem that an intervention seeks to address. But how much the person changes their beliefs in response to these messages will be influenced by factors related to the intervention itself, the problem it addresses, or the intervention's source (see Refs. [25,26]). For example, a person may choose to remain ignorant of the contents of a proposed intervention promoting vaccination if they believe the disease is not dangerous, if it is not found in their geographical location, or if they believe the intervention is likely to be ineffective or even counterproductive. Moreover, if messages attempting to encourage participation come from sources their receivers do not consider reliable — perhaps, for vaccine sceptics, the government or pharmaceutical companies (see Ref. [27]) — these messages are likely to prompt wilful intervention rejection.

Furthermore, encountering a source promoting an intervention which appears to be irrelevant or nefarious in its intentions should also lead people to downgrade their view of the source's reliability [24]. In other words, promoting an intervention ineffectively could negatively affect not only the intervention's success but the reputation of its source, making further attempts at intervention more difficult. This suggests a bi-directional element to wilful rejection: interventions from unreliable sources can be expected to have little or no uptake, and interventions perceived as absurd or unnecessary should further downgrade the perceived reliability of the source.

Modelling wilful intervention rejection

We propose a Bayesian computational model to formalise these propositions, following others who have applied Bayesian models to study contemporary socio-political problems like polarisation [28], conspiracy theories [29], and the spread of false information [30]. To do so, we adapt a previous Bayesian Network model [25] which contends that how a person updates their beliefs in response to a source's testimony about a given hypothesis (e.g. "you should play this online game that helps you spot misinformation better") is conditioned on their pre-existing belief about the source's reliability. We model the person's beliefs and source reliability perceptions using Beta distributions. Beta distributions are typically used to model binomial processes, where there is a probability of the process resulting in a "success", i.e., guessing a coin flip correctly, and a probability of resulting "failure" (guessing a coin flip incorrectly). The shape of the distribution is controlled by two parameters, α and β, which equate to the number of previously experienced "successes" and "failures", respectively. When further successes and failures are encountered, the shape of the distribution can be updated by simply adding the number of successes to α and failures to β. The fraction $\alpha/(\alpha + \beta)$ indicates a person's current best guess estimate of the probability of success, and the sum $\alpha + \beta$ indicates the "precision" of the distribution, which is equivalent to person's confidence in the accuracy of their estimate.

We can then model a person's belief in a **hypothesis (HYP)** with a Beta distribution, as the weight of previously-encountered pieces of evidence in favour of the hypothesis can be counted as the number of "successes", and the weight of previous evidence against as the number of "failures"; $\alpha/(\alpha + \beta)$ therefore gives their estimate of the probability that the hypothesis is true, and $\alpha + \beta$ their confidence in this estimate (or their "precision"). Similarly, we can model a person's perception of the **reliability (REL)** of a source using a Beta distribution, as previously-encountered cases of the source providing reliable **testimony (TES)**, can be thought of as "successes", and previous cases of incorrect testimony as "failures"; see Figure 1.

The calculation of the posterior degree of belief in the hypothesis *and* the reliability of the source of the intervention is captured formally by the following set of equations:

$$
\begin{array}{ll}
 & \sim \mathrm{Beta}(\alpha_H, \beta_H) \\
\text{Prior Belief:} & \alpha_H = p(HYP) \times Precision_h \\
 & \beta_H = (1 - p(HYP)) \times Precision_h \\
 & \sim \mathrm{Beta}(\alpha_R, \beta_R) \\
\text{Prior Reliability:} & \alpha_R = p(REL) \times Precision_r \\
 & \beta_R = (1 - p(REL)) \times Precision_r \\
\text{Posterior Belief:} & \begin{array}{l} \mathit{if}\ \text{TES} = \text{TRUE}, \sim \mathrm{Beta}(\alpha_H + \alpha_R, \beta_H + \beta_R) \\ \mathit{if}\ \text{TES} = \text{FALSE}, \sim \mathrm{Beta}(\alpha_H + \beta_R, \beta_H + \alpha_R) \end{array} \\
\text{Posterior Reliability:} & \begin{array}{l} \mathit{if}\ \text{TES} = \text{TRUE}, \sim \mathrm{Beta}(\alpha_R + \alpha_H, \beta_R + \beta_H) \\ \mathit{if}\ \text{TES} = \text{FALSE}, \sim \mathrm{Beta}(\alpha_R + \beta_H, \beta_R + \alpha_H) \end{array}
\end{array}
$$

Figure 2 shows the results of simulations where agents (i.e., individuals targeted with an intervention) receive a message (AKA "testimony") from a source that a particular hypothesis is true (e.g., "you should

Figure 1

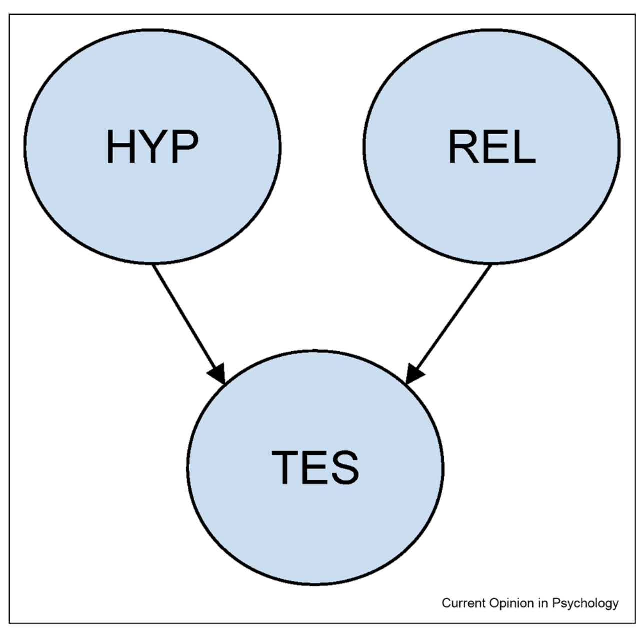
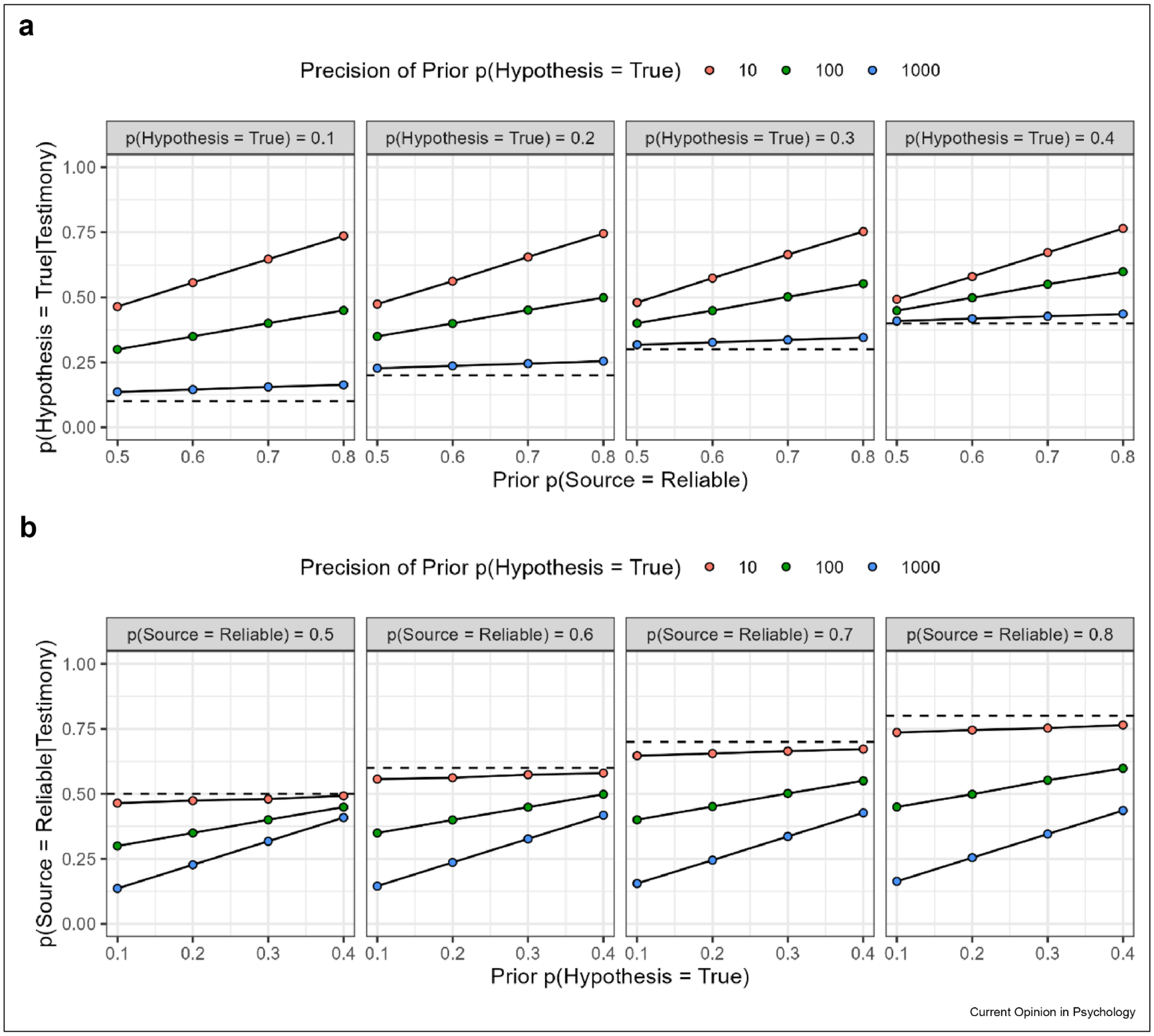


Diagram of the Bayesian network model.

Figure 2



Panel a: Simulations show that agents who reason using the model are more likely to agree with a message telling them they should partake in an intervention (p(Hypothesis = True|Testimony)) when they have a higher prior belief that they should partake (p(Hypothesis = True)), lower precision for that belief (i.e., higher-order confidence that their prior estimate is correct), and a higher belief that the source is reliable (p(Source = Reliable). **Panel b:** The same simulations show that when agents have relatively low priorbelief that they should take part in an intervention (all p(Hypothesis = True) are below 0.5), they revise their estimate of the source's reliability downwards after receiving the message (p(Source = Reliable|Testimony). Their posterior belief in the source's reliability is lower when the prior for this belief (p(Source = Reliable)) is lower to begin with, their prior estimate of the hypothesis being true (p(Hypothesis = True)) is lower, and their prior precision for this belief is higher. *Note*: For the *R* script to create the figure, see our OSF page: https://osf.io/2j4y9/.

play this online game that helps you spot misinformation better" as mentioned above), and thus decide whether to take part in an intervention. Agents differ in three respects: their initial "prior" estimate of the probability that the hypothesis is correct before receiving the message ($\alpha_H/(\alpha_H + \beta_H)$, AKA p (Hypothesis = True), with levels 0.1, 0.2, 0.3, and 0.4); their initial *precision* of this prior estimate, reflecting their confidence in the estimate ($\alpha_H + \beta_H$, with levels 10, 100, and 1000); and their initial "prior" estimate of the source's *reliability* ($\alpha_R/(\alpha_R + \beta_R)$, AKA p (Source = Reliable), with levels 0.5, 0.6, 0.7, 0.8). The precision of their prior for the source's reliability is held at 100 for all agents. In all cases, every agent receives a message from a source which conflicts with their prior – they think it is likelier that they shouldn't participate than that they should to begin with, and the source's message is that they should participate.

We take two measurements from each agent after they update in response to the message. One is their posterior estimate of the hypothesis being true, p(Hypothesis = True|Testimony). A higher p(Hypothesis = True|Testimony) should make an agent more likely to partake in the intervention; as Figure 2a shows, agents are likelier to participate when they have a higher prior belief in the hypothesis being true, have lower precision for that prior belief, and ascribe higher prior reliability to the intervention's source. The second measure is their posterior estimate of the source's reliability, p(Source = Reliable|Testimony). A lower p(Source = Reliability|Testimony) means the source has lower reliability in the eyes of the agent; as Figure 2b shows, perceived source reliability should be lower *after* receiving a message from that source, the lower the agent's prior belief in the source's reliability, the lower their prior belief in the hypothesis being true, and the higher their precision for that prior belief.

As an example of how this model might play out in real life, Johnson and Madsen [29] showed that people's willingness to watch an "inoculation" video about misinformation depends strongly on people's trust in the intervention's source, with a high-trust source like Harvard University yielding higher uptake intentions than low-trust sources such as the Russian government.[b] In other words, compared to a low-trust source, an intervention proposed by a high-trust source was found to be less likely to be wilfully rejected. Importantly, however, Johnson and Madsen [31] also showed that even the source with the highest level of trust (Harvard University) failed to outperform the no-source condition in terms of uptake potential. This indicates that population-level prior beliefs about the organisations in charge of intervention implementation are often more negative than positive (as no sources were better than having no source at all), which explains low intervention uptake of many types of interventions (e.g., see Ref. [32]). Practically, however, having no intervention source is often not feasible. One could hide this information, but this brings about ethical issues and risks backlash. Another option is that the source is *unknown* to the intervention recipient, for example when small organisations (or companies) build and implement interventions. However, this too may be impractical due to budget constraints and other limitations around scalability, especially where interventions are concerned that must be implemented at large scale to work, for example interventions implemented on social media platforms that require cooperation from platform curators.

More broadly, our above distinction between wilful ignorance and wilful rejection points to a wider discussion concerning the nature of subjective probabilities. Bayesian accounts, such as the one we propose here, are grounded in people's subjective degrees of belief. However, it is not always clear where these probability estimates come from [33]. It is possible that identity-preserving motivations may trigger subjective perceptions of the strength of an argument, the reliability of a source, or the relevance of an intervention. Then, given such a subjective perception, a Bayesian account should describe how people accept or reject an intervention. Future research may explore the relationship between motivational functions (e.g., wilful ignorance of information to preserve identity or emotional reactance to unpleasant information) and process functions (e.g., wilful rejection of an intervention despite having full information).

To conclude, we argue that using intervention sources with high perceived reliability among target audiences is key to optimising the uptake of psychological and behavioural interventions. Prior to intervention design and implementation, we therefore suggest studying whether target communities believe the intervention serves a purpose for themselves (i.e., is it perceived to be potentially effective), and which sources they trust to implement an intervention in a way they believe is reasonable. Our proposed model further suggests that people who believe the intervention is irrelevant or that the problem addressed by the intervention is non-existent should downgrade their perception of reliability of sources that propose interventions. We therefore argue that backfire effects that are the topic of debate in the intervention research space [34,35] may take place at the *source* level, rather than at the level of individual-level behavioural or psychological outcome measures. This would require empirical work to test.

Credit author statement

Note: Jon Roozenbeek and Jens Koed Madsen contributed equally to this work.

Jon Roozenbeek: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing.

David J. Young: Conceptualization; Data curation; Formal analysis; Methodology; Software; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing.

Jens Koed Madsen: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Supervision; Validation; Visualization; Roles/Writing – original draft; Writing – review & editing.

[b] Source reliability was assessed in a pilot study where participants were asked to rate 17 possible intervention sources according to their trustworthiness and effectiveness.

Declaration of competing interest

No funding was required for this project. The authors declare no competing interests.

References

References of particular interest have been highlighted as:

* of special interest
** of outstanding interest

1. Fazio L, Rand DG, Lewandowsky S, Susmann M, Berinsky AJ, Guess AM, Kendeou P, Lyons B, Miller JM, Newman E, Pennycook G, Swire-Thompson B: **Combating misinformation: a megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines**. In *PsyArXiv preprints*; 2024, https://doi.org/10.31234/osf.io/uyjha.

2. Greene Barker T, O'Higgins A, Fonagy P, Gardner F: **A systematic review and meta-analysis of the effectiveness of co-designed, in-person, mental health interventions for reducing anxiety and depression symptoms**. *J Affect Disord* 2024, **350**:955–973, https://doi.org/10.1016/j.jad.2023.12.080.

3. Voelkel JG, Stagnaro MN, Chu JY, Pink SL, Mernyk JS, Redekopp C, Ghezae I, Cashman M, Adjodah D, Allen LG, Allis LV, Baleria G, Ballantyne N, Van Bavel JJ, Blunden H, Braley A, Bryan CJ, Celniker JB, Cikara M, … Willer R: **Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity**. *Science* 2024, **386**, https://doi.org/10.1126/science.adh4764.

4. Van Bavel JJ, Baicker K, Boggio PS, Capraro V, Cichocka A, Cikara M, Crockett MJ, Crum AJ, Douglas KM, Druckman JN, Drury J, Dube O, Ellemers N, Finkel EJ, Fowler JH, Gelfand M, Han S, Haslam SA, Jetten J, … Willer R: **Using social and behavioural science to support COVID-19 pandemic response**. *Nat Hum Behav* 2020, **4**:460–471, https://doi.org/10.1038/s41562-020-0884-z.

5. Roozenbeek J, Culloty E, Suiter J: **Countering misinformation: evidence, knowledge gaps, and implications of current interventions**. *Eur Psychol* 2023, **28**:189–205, https://doi.org/10.1027/1016-9040/a000492.

6. Mertens S, Herberz M, Hahnel UJJ, Brosch T: **The effectiveness of nudging: a meta-analysis of choice architecture interventions across behavioral domains**. *Proc Natl Acad Sci* 2022, **119**, https://doi.org/10.1073/pnas.2107346118.

7. DellaVigna S, Linos E: **RCTs to scale: comprehensive evidence from two nudge units**. *Econometrica* 2022, **90**:81–116, https://doi.org/10.3982/ECTA18709.

8. Maertens R, Roozenbeek J, Simons JS, *et al.*: **Psychological booster shots targeting memory increase long-term resistance against misinformation**. *Nat Commun* 2025, **16**:2062. https://doi-org.ezp.lib.cam.ac.uk/10.1038/s41467-025-57205-x.

9. Druckman JN: **The crisis of politicization within and beyond science**. *Nat Hum Behav* 2017, **1**:615–617, https://doi.org/10.1038/s41562-017-0183-5.

10. Rekker R: **The nature and origins of political polarization over science**. *Publ Understand Sci* 2021, **30**:352–368, https://doi.org/10.1177/0963662521989193.

11. Arton A, Carrella E, Madsen JK, Bailey RM: **Triggering the tragedy: the simulated effects of alternative Fisher goals on marine fisheries and fisheries policy**. *Ecol Complex* 2024, **57**, 101070.

12. ** Landry AP, Halperin E: **Intergroup psychological interventions: the motivational challenge**. *Am Psychol* 2025, **80**: 206–219, https://doi.org/10.1037/amp0001289.

13. * Brashier NM: **Fighting misinformation among the most vulnerable users**. *Curr Opin Psychol* 2024, **57**, 101813, https://doi.org/10.1016/j.copsyc.2024.101813.

14. * Roozenbeek J, Remshard M, Kyrychenko Y: **Beyond the headlines: on the efficacy and effectiveness of misinformation interventions**. *Adv Psychol* 2024, **2**, https://doi.org/10.56296/aip00019.

15. Bauer MS, Kirchner J: **Implementation science: what is it and why should I care?** *Psychiatry Res* 2020, **283**, 112376, https://doi.org/10.1016/j.psychres.2019.04.025.

16. Eccles MP, Mittman BS: **Welcome to implementation science**. *Implement Sci* 2006, **1**:1, https://doi.org/10.1186/1748-5908-1-1.

17. de Ridder D, Kroese F, van Gestel L: **Nudgeability: mapping conditions of susceptibility to nudge influence**. *Perspect Psychol Sci* 2021, **17**:346–359, https://doi.org/10.1177/1745691621995183.

18. Hopkins KL, Lepage C, Cook W, Thomson A, Abeyesekera S, Knobler S, Boehman N, Thompson B, Waiswa P, Ssanyu JN, Kabwijamu L, Wamalwa B, Aura C, Rukundo JC, Cook J: **Co-designing a mobile-based game to improve misinformation resistance and vaccine knowledge in Uganda, Kenya, and Rwanda**. *J Health Commun* 2023, **28**:49–60, https://doi.org/10.1080/10810730.2023.2231377.

19. Lewandowsky S, Oberauer K: **Motivated rejection of science**. *Curr Dir Psychol Sci* 2016, **25**:217–222, https://doi.org/10.1177/0963721416654436.

20. Madsen JK, de-Wit L, Ayton P, Brick C, de-Moliere L, Groom CJ: **Behavioral science should start by assuming people are reasonable**. *Trends Cognit Sci* 2024, **28**:583–585.

21. Vu L, Soraperra I, Leib M, van der Weele J, Shalvi S: **Ignorance by choice: a meta-analytic review of the underlying motives of willful ignorance and its consequences**. *Psychol Bull* 2023, **149**:611.

22. Reher K, Götz M, Gross J: **Blind spots: the (social) factors of willful ignorance**. *Curr Opin Psychol* 2025, 102100.

23. Bilfinger L, Brummernhenrich B, Jucks R: **The effects of fear appeals on reactance in climate change communication**. *J Exp Soc Psychol* 2024, **115**, 104666, https://doi.org/10.1016/j.jesp.2024.104666.

24. Oaksford M, Chater N: *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford University Press; 2007.

25. Bovens L, Hartmann S: *Bayesian epistemology*. Oxford: Oxford University Press; 2003.

26. Collins PJ, Hahn U, von Gerber Y, Olsson EJ: **The Bi-directional relationship between source characteristics and message content**. *Front Psychol* 2018, **9**, https://doi.org/10.3389/fpsyg.2018.00018.

27. Lindholt MF, Jørgensen F, Bor A, Petersen MB: **Public acceptance of COVID-19 vaccines: cross-national evidence on levels and individual-level predictors using observational data**. *BMJ Open* 2021, **11**, e048172.

28. * Young DJ, Madsen JK, de-Wit LH: **Belief polarization can be caused by disagreements over source independence: computational modelling, experimental evidence, and applicability to real-world politics**. *Cognition* 2025, **259**, 106126, https://doi.org/10.1016/j.cognition.2025.106126.

29. Botvinik-Nezer Rotem, Jones Matt, Wager Tor D: **A belief systems analysis of fraud beliefs following the 2020 US election**. *Nat Hum Behav* April 10, 2023, **7**:1106–1119, https://doi.org/10.1038/s41562-023-01570-4.

30. Hahn U, Merdes C, von Sydow M: **How good is your evidence and how would you know?** *Top Cogn Sci* 2018, **10**:660–678. https://onlinelibrary.wiley.com/doi/10.1111/tops.12374.

31. * Johnson A, Madsen JK: **Inoculation hesitancy: an exploration of challenges in scaling inoculation theory**. *R Soc Open Sci* 2024, **11**, https://doi.org/10.1098/rsos.231711.

32. Roozenbeek J, van der Linden S, Goldberg B, Rathje S, Lewandowsky S: **Psychological inoculation improves resilience against misinformation on social media**. *Sci Adv* 2022, **8**, https://doi.org/10.1126/sciadv.abo6254.

33. Madsen JK: **Time during time: multi-scalar temporal cognition**. In *Cognition beyond the brain: computation, interactivity*

and human artifice. Cham: Springer International Publishing; 2017:155–174.

34. Swire-Thompson B, DeGutis J, Lazer D: **Searching for the backfire effect: measurement and design considerations**. *Journal of Applied Research in Memory and Cognition* 2020, **9**: 286–299, https://doi.org/10.1016/j.jarmac.2020.06.006.

35. Wood T, Porter E: **The elusive backfire effect: mass attitudes' steadfast factual adherence**. *Polit Behav* 2019, **41**:135–163, https://doi.org/10.1007/s11109-018-9443-y.

Further information on references of particular interest

12. ** **Outstanding interest**: Landry and Halperin are among the first to explicitly seek to understand why people might choose to engage with interventions, They argue that the key challenge within this domain is motivation, i.e., ensuring that sufficient motivation exists to take part in the intervention. However, they do not explicitly look at the reasons why an intervention might be willfully rejected by individuals.

13. * **Special Interest:** Brashier argues that interventions that seek to counter susceptibility to or sharing of misinformation should be aimed at, and tested on, individuals who are most vulnerable (i.e., are most likely to engage in these types of behaviours). She further argues that testing for moderation (e.g., testing if an intervention "works" for both Liberals and Conservatives") is not enough: interventions should be designed with target audiences in mind.

14. * **Special Interest:** Roozenbeek, Remshard, and Kyrychenko argue that the field of misinformation research has focused excessively on efficacy (successful lab studies) and not enough on effectiveness (real-world implementation and impact). They identify six challenges that hamper this transition: (1) an overabundance of lab research and a lack of field studies; (2) the presence of testing effects, which impede intervention longevity and scalability; (3) modest effects for small fractions of relevant audiences; (4) a reliance on item evaluation tasks (e.g., rating a series of headlines as true or false) as the primary efficacy measure of interest; (5) low replicability in the Global South and a lack of audience-tailored interventions; and (6) an underappreciation of potential unintended consequences of intervention implementation.

20. * **Special Interest**: Madsen and colleagues argue that behavioural science and policymaking have too often assumed that people are fundamentally irrational creatures, prone to errors in judgment leading to suboptimal outcomes. Instead, they propose a series of empirical, philosophical, and theoretical arguments that suggest that human behaviour is indeed reasonable and rational. This, in turn, should lead to more optimal insights into how and why humans behave in certain ways and in certain situations.

28. * **Special Interest**: Young and colleagues show how Bayesian reasoning about inter-dependencies between information sources can cause belief polarization between completely rational people, presenting model-based simulations, experimental evidence, and analysis of political opinion data in support.

31. * **Special Interest**: Johnson and Madsen present the results of an experiment into the uptake potential of a hypothetical "inoculation" intervention (aimed at reducing susceptibility to misinformation), depending on the perceived reliability of the source of the intervention: high-reliability (e.g., Oxford University), low reliability (the Russian government) or a partisan source (e.g., the Republican or Democratic Party). They find not only that intervention uptake potential is strongly dependent on source credibility, but also that no source outperformed the "no source" condition, indicating that optimizing intervention uptake may be complicated, especially when implemented by sources that suffer from low credibility among certain groups in the population.