Helping Users to Assess Content Online:

Integrating Behavioural Tools through an Adaptive Randomised Controlled Trial

Technical Paper

Economics Discussion Paper Series – Issue 16

Published 24 February 2026

Contents

Section

Overview

Ofcom is the United Kingdom's independent communications regulator. We regulate fixed and mobile telecoms, TV and radio broadcasting, post, the radio spectrum used by wireless devices, and online safety. Under the Communications Act 2003, Ofcom holds statutory duties to promote and to carry out research into media literacy. We aim to promote public understanding of online content and support users in navigating digital information environments. Our work is grounded in fundamental rights, particularly the right to freedom of expression under Article 10 of the European Convention of Human Rights. The Online Safety Act (OSA) bolstered our media literacy duties, including by adding a requirement for Ofcom to take steps to improve users' ability to assess the reliability, accuracy and authenticity of content, and ultimately to understand and reduce their exposure to the phenomena of disinformation and misinformation.[1] In practice, this involves enabling users of online services to critically assess the content they encounter online.

The experimental study set out in this Technical Paper is specifically targeted at exploring the efficacy of various interventions aimed at enabling users to assess the reliability, veracity and authenticity of online content they see, thereby supporting users in making more informed judgements about it. The research described in this paper tests a new methodology (Adaptive Randomised Control Trials) for exploring this area.

The study does not offer a broad assessment of platform interventions or their overall effectiveness. Rather, it presents findings from a controlled experiment that tested a range of behavioural tools — including educational quizzes, prompts and content labels — to understand how they might support users in identifying potentially misleading content. The insights generated here are intended to inform best practice and contribute to the wider evidence base on media literacy design, in line with our strategic commitment to supporting public resilience to online harms.

The study was designed to evaluate the effectiveness of various interventions aimed at enhancing users' ability to assess reliability, veracity and authenticity. Their ability to assess was tested based on their ability to identify content that had been independently verified as false. For brevity, we will refer to this content going forward in this report as "false content".

Adaptive Randomised Control Trials (ARCTs) employ statistical algorithms to update the experimental design in real time based on interim results. By adjusting the randomisation throughout the experiment, ARCTs enable the testing of a larger number of interventions and more efficient identification of the best interventions, compared to standard Randomised Control Trials (RCTs). In this study, 6,000 participants representative of the United Kingdom (UK) population were shown a simulated social media feed with some true and some false posts which were independently verified by fact checking organisations.[2] Participants were asked to assess the

[1] Online Safety Act, 2023. Chapter 8: Media Literacy. This includes a requirement to help users understand the nature and impact of disinformation and misinformation, reduce their and others' exposure to it and learn how to critically assess sources including manipulated content. While there is no agreed definition of 'misinformation' or 'disinformation', we use those terms because they are used in the statute.

[2] In particular for the posts containing false content, we used content that had been confirmed false by independent fact-checkers such as Full Fact and The Ferret. For the true posts, we collated content from reputable sources like the Bank of England, ONS, Science.org, and the BBC, as well as from fact-checkers like Snopes.

veracity of the information contained in each post. While all participants were shown the same set of posts, a different version of the feed was shown depending on the intervention.

Our interventions employed (1) user-focused tools, including an educational quiz where participants learned to detect potentially false content through evaluating posts and receiving instant feedback (a technique often referred to as “inoculation”) and a reminder prompt reinforcing lessons halfway through the feed; and (2) post-focused tools or informational labels including fact-checker labels, AI verification labels and crowdsourced notes. The interventions were tested individually as well as in combination.

Based on our experiment, we found:

- Combinations of user-focused tools and labels were most effective at improving participants’ ability to identify false content.
- When tested in isolation, user-focused tools generally outperformed labels.
- Interventions did not significantly reduce trust in verifiably true information.
- Crowdsourced notes were marginally less effective than AI and fact-checker labels.

While these findings provide novel evidence on the effect of different types of media-literacy tools and their complementarities, they should be interpreted in light of the study’s limitations. First, the experiment was conducted in a controlled, simulated environment, which cannot reproduce many key features of real-world online platforms and therefore limits the external validity of the results. Second, the findings depend heavily on the specific posts shown to participants, meaning results could vary if different or AI-generated false content were used. Finally, the effect of the interventions may differ in live platform contexts or over longer time horizons. Therefore, the findings should be viewed as indicative evidence in a controlled setting rather than direct estimates of real-world impact.

1. Introduction and Related Literature

Online news sources have become a primary source of information in recent years. According to Ofcom's News Consumption Report (2025), 71% of UK adults use online sources for news, with 51% specifically turning to social media. Among the 16-24 age group, these figures rise to 80% and 75%, respectively.[3]

As online platforms become more central to how people access information, concerns about the quality and reliability of content have intensified. False content has shown to be prevalent on these platforms.[4] Grinberg et al. (2019) reported that false content accounted for 6% of all news consumption during the 2016 US presidential election.[5] More recently, Ofcom's 2025 Online Nation report found that nearly four in ten adults had reported seeing or experiencing misinformation online in the previous four weeks, making misinformation the most prevalent potential harm identified in the survey.[6] Academic research further supports this trend: Broda and Strömbäck (2024) conducted systematic review of 1,200 studies on the prevalence, dissemination, and detection of misinformation and confirmed that misinformation has become significantly more widespread in the past decade, particularly following events like Brexit and the 2016 U.S. election.[7] [8]

A growing body of research highlights the role of social media in amplifying the spread of false content. Allcott and Gentzkow (2017) found that false news spreads extensively through these platforms. While social media referrals accounted for 11% of traffic on mainstream news sites, they made up a significantly larger share (42%) for "fake news websites".[9] Moreover, Vosoughi et al. (2018) demonstrated that false news spread significantly farther, faster, deeper, and more broadly online than the true news across all categories, with political information showing the most pronounced effects.[10]

---

[3] Ofcom, 2025, News Consumption in the UK.

[4] It is important to be clear from the outset that misinformation is a subjective term. Reported incidence of encountering misinformation is only part of the story, and this survey asks a range of other questions to understand the nature of people's beliefs and attitudes towards the news information they consume. There are a range of ways that misinformation can be characterised, which makes it important to understand the wider context of people's attitudes and knowledge. Perceptions of misinformation include and are not limited to: provision of empirically false information; provision of information that someone does not agree with; provision of information that does not fit with someone's prior knowledge of, or existing beliefs about, a subject – which can result in true information being reported as false, and vice versa; and something that a public figure has said and is being reported on by a news platform or service – with such reporting either identifying the statement as misleading, or providing it as if it were accurate.

[5] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019), Fake News on Twitter During the 2016 U.S. Presidential Election.

[6] Ofcom, 2025, Online Nation 2025 report.

[7] Broda, E., and Strömbäck, J. (2024), Misinformation, Disinformation, and Fake News: Lessons from an Interdisciplinary, Systematic Literature Review.

[8] We reference literature from the US, particularly studies centred on the US presidential election context, as they represent some of the most robust and comprehensive investigations into the prevalence and impact of misinformation online. While there is limited evidence available from the UK context, it is important to acknowledge that findings from the US may not fully translate to the UK.

[9] Allcott, H., and Gentzkow, M. (2017), Social Media and Fake News in the 2016 Election.

[10] Vosoughi, S., Roy D., and Aral, S. (2018), The spread of true and false news online.

Given the prevalence of false content online and the role of social media, understanding its societal impacts has become an important question. In a global survey conducted by Adobe involving over 6,000 participants from multiple countries - including more than 2,000 from the UK - 81% of UK respondents identified misinformation as one of the greatest threats to society. Additionally, 78% expressed concern that misinformation and deepfakes could influence upcoming elections and undermine the democratic processes.[11] Allcott and Gentzkow (2017) also found that 15% of survey respondents recalled seeing "fake stories" during the 2016 US presidential election, and about half of them reported believing those stories. Cantarella et al. (2023) showed that exposure to fake news affected election outcomes in Italy.[12] [13] Beyond the political sphere, research has focussed on the extent to which misinformation has also influenced health-related decisions. Pennycook et al. (2018, 2020) discusses that whether being priorly exposed to misleading COVID-19 content on social media might impact people's willingness to receive COVID-19 vaccinations.[14] [15]

Recent research in behavioural and social sciences has focused on developing interventions that strengthen users' ability to correctly assess the reliability of information. According to the OECD report on misinformation, behaviourally informed interventions have been shown to be effective, scalable tools for tackling the spread of misinformation and they can enhance system-level policies aimed at empowering users.[16] These interventions include debunking false claims (Lewandowsky et al., 2020), enhancing digital media literacy (Guess et al, 2020), building resilience through pre-emptive inoculation (Basol et al., 2020; Roozenbeek et al., 2022) introducing design features that slow the sharing of misinformation (Fazio, 2020), prompting users to focus on accuracy (Pennycook et al., 2021), and highlighting the trustworthiness of content (Clayton et al., 2020).[17]

Ofcom recently published an evidence review on the drivers and mitigations of misinformation which informed the selection of interventions tested in this research.[18] Along with the evidence review, Ofcom commissioned qualitative research to explore which approaches resonate with different groups when encountering misinformation.[19] Participants in this research study supported

---

[11] Adobe, 2024, Media Alert: Adobe Study Reveals High Concern Over Misinformation and Potential to Impact Elections.

[12] Cantarella, M., Fraccaroli, N., and Volpe R. (2023), Does fake news affect voting behaviour?.

[13] Although some researchers argue that fake news had a limited short-term impact on voting behaviour during the 2016 and 2020 US elections (such as Allcott and Gentzkow, 2017 and Guess et al, 2023), its long-term effects and contribution to political polarisation remain pressing concerns. Polarisation tends to build over time as diverse beliefs and biases accumulate, making it unrealistic to expect single short-term interventions to significantly counteract its effects.

[14] Pennycook, G., Cannon, T. D., and Rand D. G. (2018), Prior exposure increases perceived accuracy of fake news.

[15] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. (2020), Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention.

[16] OECD, 2022, Misinformation and disinformation.

[17] Lewandowsky, S., Ecker, U. K. H., and Cook, J. (2020), Beyond misinformation: Understanding and coping with the "post-truth" era. Guess, A. M., Nagler, J., and Tucker, J. (2020), Exposure to untrustworthy websites in the 2016 US election. Basol, M., Roozenbeek, J., and van der Linden, S. (2020), Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. Roozenbeek, J., Maertens, R., McClanahan, W. P., and van der Linden, S. (2022), Disentangling interventions to counter misinformation: Experimental evidence for the efficacy of accuracy nudges and inoculation. Fazio, L. K. (2020), Pausing to consider why a headline is true or false can help reduce the sharing of false news. Pennycook, G., and Rand, D. G. (2021), The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings. Clayton, K., Blair, S., Busam, J. A., Sorensen, G., and Nyhan, B. (2020), Real Solutions for Fake News? Measuring the Effectiveness of General, News-Specific, and Source-Specific Misinformation Interventions.

[18] Ofcom, 2025, Misinformation and Disinformation: Literature Review

[19] Ofcom, 2025, Co-creating ways to navigate and mitigate against mis and disinformation.

the proposal that helping people navigate misinformation requires a multi-channel approach - including person-to-person engagement, community spaces and social media campaigns- to effectively reach diverse audiences.

Under the broader theme of empowering users to critically assess online content, several interventions have proven to be effective. Strategies such as prompting users to consider accuracy, rating headlines and providing educational checklists enhance awareness and judgement (Pennycook et al., 2021).[20] Crowdsourced flagging and inoculation techniques (Basol et al., 2020), further build resilience, although reliance on warning labels alone may reduce users' ability to detect misinformation when such labels are absent. [21] Fact-checking has also been shown to reduce belief in misinformation over time (Porter and Wood, 2021), limit its spread through design interventions (Guriev et al., 2023), and increase engagement with fact-checked news when presented by politically news sources that do not align with their own views (Chopra et al., 2022).[22]

Additionally, our methodology builds on recent academic applications of adaptive experiments such as Offer-Westort et al. (2021), who evaluate interventions aimed at combatting the spread of COVID-19 on social media in sub-Saharan Africa.[23] Another notable example is Caria et al. (2024), who test job search assistance policies for refugees in Jordan.[24] Similarly to these studies, we use a tailored modification of so-called multi-armed bandit algorithms in the randomisation of participants. Our work also leverages recently developed statistical methods to address the technical challenges of inference with adaptively collected data.[25]

Introduction to our research

We carried out the research described in this paper as a means of testing a new methodology to explore how users' ability to assess the reliability, veracity and authenticity of online content might be improved. Insights from this research will inform the evidence-based recommendations and guidance Ofcom provides to platforms that might help build public resilience to online information threats. This experimental study contributes to the evidence base in two important ways.

The study provides comprehensive, robust, UK-specific insights into how different interventions can help people assess reliability, veracity and authenticity of content online.

Previous research has typically tested two or three interventions at a time, often using varied outcome measures, experimental settings, and samples. This variability made it difficult to compare findings across studies and to identify the most effective treatment. By evaluating a wide range of interventions within a single, unified experimental framework, our study overcomes these

[20] Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., and Rand, D. G. (2021), Shifting attention to accuracy can reduce misinformation online.
[21] Basol, M., Roozenbeek, J., and van der Linden, S. (2020), Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News.
[22] Porter, E., and Wood, T. J. (2021), The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. Guriev, S., Henry, E., Marquis, T., and Zhuravskaya, E. (2023), Curtailing False News, Amplifying Truth. Chopra, F., Haaland, I., and Roth, C. (2022), Do people demand fact-checked news? Evidence from US Democrats.
[23] Offer-Westort, M., Rosenzweig, L.R. and Athey, S. (2024), Battling the coronavirus 'infodemic' among social media users in Kenya and Nigeria.
[24] Caria, A. S., Gordon, G., Kasy, M., Quinn, S., Shami, S. O. and Teytelboym, A. (2024), An Adaptive Targeted Field Experiment: Job Search Assistance for Refugees in Jordan
[25] For user-friendly introduction to adaptive experiments, see: Practitioner's Guide: Designing Adaptive Experiments, Golub Capital Social Impact Lab, Stanford GSE (2021).

limitations and delivers a clearer, evidence-based understanding of which strategies work best within a controlled environment.

To make the study relevant to the UK, we created a mock social media feed with UK-specific content and recruited a representative UK sample. Most previous research is focused on the US; this study helps fill a gap in the literature by providing insights that can help inform UK policy and practice.

The study explores how interventions work in combination.

This study also tested how interventions interact when combined, an area that has received limited attention in prior research. This was made possible using an ARCT methodology. Unlike standard RCTs that allocate participants evenly across trial arms, ARCTs adjust allocations based on observed performance in real-time, sending more participants to more promising interventions. This efficient use of experimental resources allows us to identify the most effective among a larger set of treatments without requiring infeasibly large samples. We were therefore able to test combinations of interventions in a resource effective way to identify potential complementarities between them.

2. Trial Design and Methodology

Sample and Experimental Design

In March 2025, 6,000 adults (over 18 years) representative of the UK population took part in the experiment, recruited thorough the platform Prolific which uses an integrated algorithm designed to ensure samples are representative of the UK population.[26]

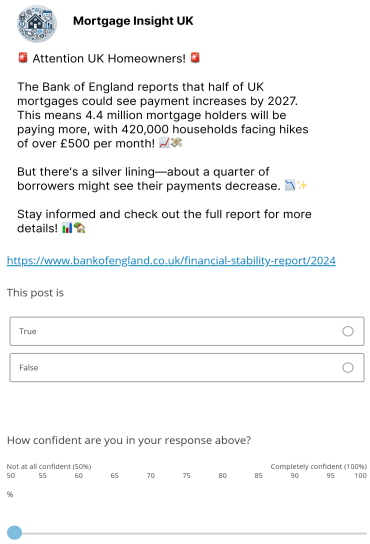
Participants were presented with 15 social media posts, shown in random order. Out of 15 posts, 10 posts contained true information ("true posts"), while 5 included false content ("false posts").[27] The posts covered a range of topics including economy/finance, health, migration, politics, and science.[28]

After each post, participants were asked:

- Question 1: To choose whether the post was true or false (binary choice)
- Question 2: "How confident are you in your response above?" rated on a scale between 50 (not at all confident) to 100 (completely confident).

An example of a true post, as shown to participants, is given in Figure 1.

Figure 1: Example Post



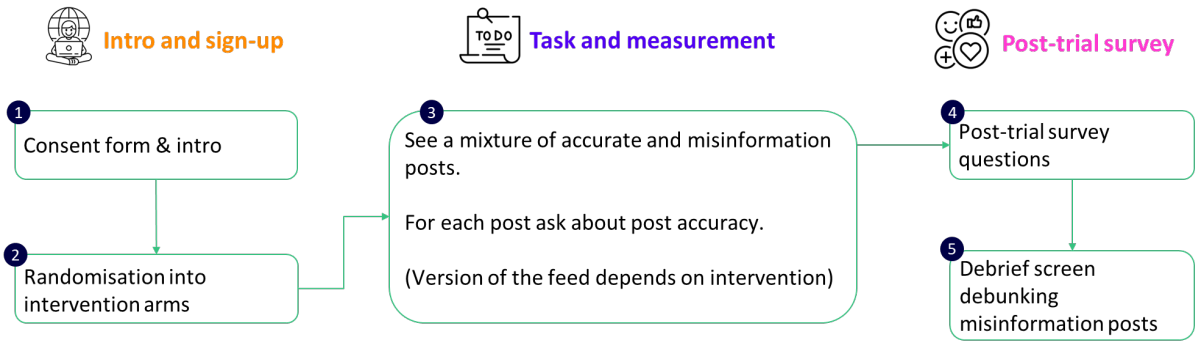
[26] Prolific recruits participants by connecting researchers with a diverse pool of eligible individuals who are willing to participate in studies. Those who sign up are screened for eligibility and invited to take part in studies based on criteria specified by researchers.

[27] For the false posts, we used content that had been confirmed false by independent fact-checkers such as Full Fact and The Ferret. The posts featured randomly generated common usernames and AI-generated profile pictures. The text was kept close to the original but not word-for-word to mitigate privacy risks. For the true posts, we collated content from reputable sources like the Bank of England, ONS, Science.org, and the BBC, as well as from fact-checkers like Snopes.

[28] The topics have been chosen to replicate a representative UK news diet based on Ofcom's previous research (Ofcom, 2021, Understanding online false information in the UK).

After they had seen all the posts participants were asked questions about their media consumption habits, social media attitudes and usage, and demographic characteristics including age, gender, ethnicity, education, and political leanings. At the end of the study, participants were shown which posts were false, given links to trustworthy sources, and offered extra information to help them learn more.[29] See Figure 2 for experimental flow.

Figure 2: User Journey



Participants were paid for taking part in the experiment (conditional of completion of task) and received a bonus for each post that they correctly identified as true or false.[30] As outlined in our pre-registry, we excluded from analysis participants who failed both attention checks embedded in the journey, did not complete the task or completed it exceptionally fast (less than 3 minutes) or slow (50 minutes or more).

Interventions and Trial Arms

The experiment considered two main types of interventions aimed at helping people identify false content online: user-focused tools and post-focused tools.[31]

User-focused tools:

- **Educational Quiz/Inoculation**: Before viewing the simulated social media feed, participants engaged in a short quiz designed to train them to detect false content. They judged whether the post was true or false and got instant feedback, along with tips on how to spot false content (see Figure 3).
- **Educational Quiz + Reminder Prompt**: In addition to the Educational Quiz, participants also saw a quick reminder of the key lessons from the quiz halfway through the feed (see Figure 4).

[29] This included referring participants to the independent fact-checker websites that were used to collate false content for this experiment.
[30] The payment for participation was £1.20, while the bonus for each post correctly identified as true or false was £0.06.
[31] We used an existing classification of interventions against online misinformation that have been shown to be effective in the literature (Kozyreva et al, 2024, Toolbox of individual-level interventions against online misinformation).

Figure 3: Example Educational Quiz

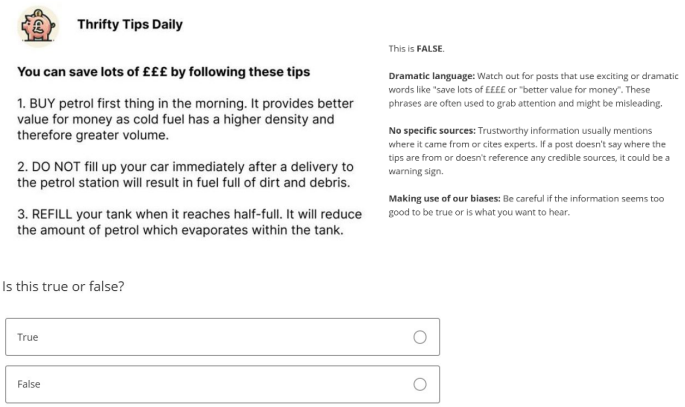
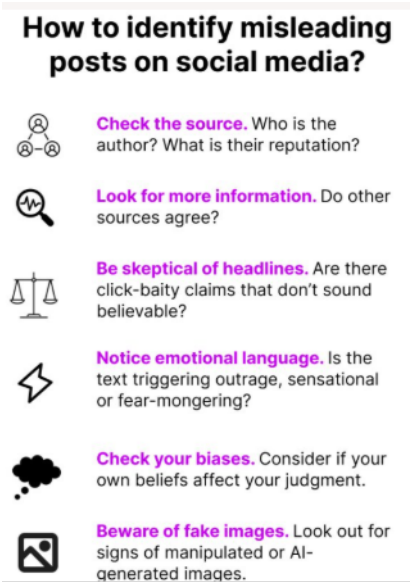


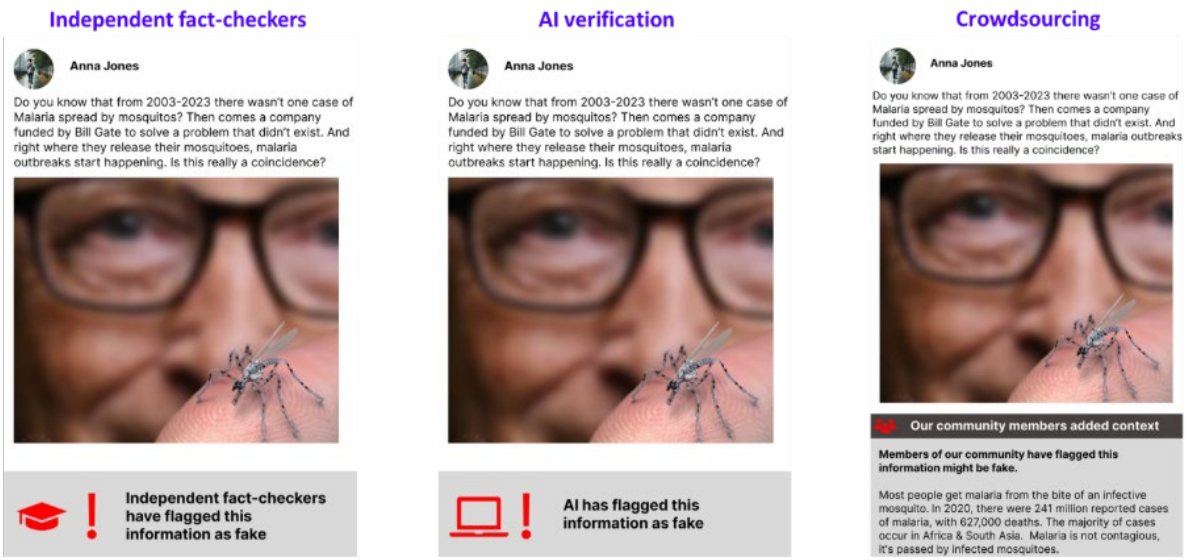
Figure 4: Example Reminder Prompt



Post-focused tools (Labels)

- **Fact-Checker Label**: Posts containing false content were flagged with the message "Independent fact-checkers have flagged this information as fake."
- **AI Verification Label**: Posts containing false content were flagged with the message "AI has flagged this information as fake.". This label indicates that an AI system has analysed the content and assessed it to be false or misleading.
- **Crowdsourced Notes**: Posts containing false content were flagged with the message "Our community members added context. Members of our community have flagged this information might be fake", followed by additional explanatory/de-bunking information provided by community members (see Figure 5 for examples of all three types of intervention).

Figure 5: Example Post-Focused Interventions



The experiment tested interventions from the two types -both individually and in combinations- alongside a control group that did not receive any intervention (see Table 1 for a list of all trial arms).

Table 1: Trial Arms

	Educational quiz	Reminder prompt	Label
T1 Control	No	No	No
T2 Educational q.	Yes	No	No
T3 Educational q.+ Reminder prompt	Yes	Yes	No
T4 AI Label	No	No	AI Label
T5 Fact-checker label	No	No	Fact-checker label
T6 Crowdsourced notes	No	No	Crowdsourced notes
T7 Educational q. + AI label	Yes	No	AI Label
T8 Educational q. + Fact-checker label	Yes	No	Fact-checker label
T9 Educational q. + Crowdsourced notes	Yes	No	Crowdsourced notes
T10 Educational q. + AI label+ Reminder Prompt	Yes	Yes	AI Label
T11 Educational q. + Fact-checker label + Reminder Prompt	Yes	Yes	Fact-checker label
T12 Educational q. + Crowdsourced notes + Reminder Prompt	Yes	Yes	Crowdsourced notes

Primary and secondary outcomes

We measure the impact of treatments in terms of a primary outcome that combines a participant's assessment of the veracity of posts and the confidence levels attached to those assessments.

Let X_i^p be a binary variable taking value 1 if participant i correctly identifies whether the post p contains false content, 0 otherwise. Z_i^p is the confidence level attached to X_i^p, ranging between 50 and 100. We compute the a*ccuracy score* of participant i' assessment of post p as

$$Y_i^p = X_i^p \cdot Z_i^p + (1 - X_i^p) \cdot \left(100 - Z_i^p\right),$$

which ranges between 0 (maximum confidence in the incorrect answer) and 100 (maximum confidence in the correct answer).

The final target outcome Y_i (*Average Accuracy Score*) for participant i is then computed as the average over the 15 posts:

$$Y_i = \frac{1}{15}\sum_{p=1}^{15} Y_i^p.$$

We have chosen this as the primary outcome for its comprehensive measurement of participants' ability to assess false content. In fact, it allows to detect causal shifts in beliefs induced by the treatments even when these do not translate into a switch from an incorrect to a correct assessment of veracity of a post (or vice versa)—for example, in cases where participants find a post easy to correctly assess in the absence of any intervention. In such instances, changes in confidence will still be captured by the primary outcome to reflect causal effects of interventions on belief strength.[32]

To isolate the ability of the treatments to change assessments of veracity from incorrect to correct (or vice versa), we also consider the percentage of correct answers ("*Correct Identification Rate*"):

$$X_i = \frac{1}{15}\sum_{p=1}^{15} X_i^p.$$

For both the accuracy score and correct identification rate, we report disaggregated measures with averages are computed separately for true and false posts.[33]

Randomisation and Adaptive Allocation

The trial utilised an adaptive randomisation approach, conducted over five rounds.[34] Over the course of Rounds 1–4, each trial phase recruited 1,050 participants. In the first round, participants were evenly distributed across twelve trial arms. Assignment probabilities were then updated at the end of each round based on interim results for the target outcome Y_i and used for randomisation in the subsequent round. In Round 5, the most effective intervention was determined using the data gathered from previous rounds, and 1,800 participants were randomly assigned between the top-performing intervention (based on posterior probabilities) and the control arm with equal probability. This allowed us to study the performance of the best treatment against the control with sufficiently large sample size and circumvent issues related to selective inference. See Figure 6 for the summary of the randomisation process.

Throughout the experiment, assignment probabilities were updated based on Exploration Sampling (Kasy and Sautmann, 2021), a variation of Thompson Sampling geared towards identification of the best arm.[35] The Exploration Sampling algorithm was augmented with probability floors, ensuring that each treatment is sampled with a minimum probability of 0.02 in each round. Probability floors allow to circumvent well-known issues with inference in adaptive experiments (see, e.g., Bibaut and Kallus, 2025) and invoke asymptotic validity of standard normal confidence intervals for sample means.[36]

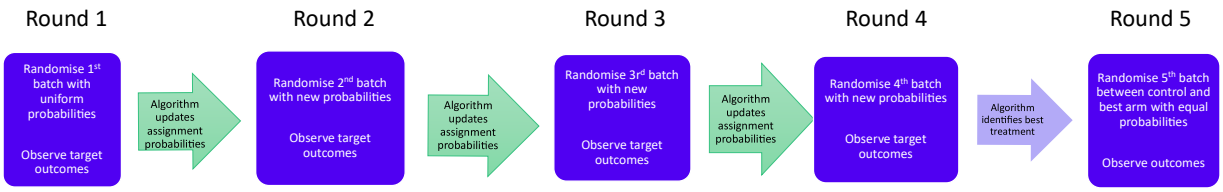
Given that our target outcome Y_i is continuous, the algorithm uses posterior distributions based on a normal model for Y_i with normal-inverse-gamma independent priors. Posterior probabilities are computed via simulation from the posterior distributions of the mean parameter for Y_i for each treatment arm.

Further details on the randomisation algorithm and the computation of posterior probabilities (including the choice of prior parameters) are given in the Appendix.

---

[32] Because the reported confidence levels were not used to determine the financial bonus for participation, using the average accuracy score as primary measure partly mitigates concerns that the measured effects might be driven by monetary incentives, thereby supporting the external validity of the analysis.
[33] See Findings section for details of the analysis on primary and secondary outcomes.
[34] Each round was conducted on consecutive days and at the same time of day.
[35] Kasy, M. and Sautmann, A. (2021), Adaptive Treatment Assignment in Experiments for Policy Choice.
[36] Bibaut, A. and Kallus, N. (2025), Demistifying Inference after Adaptive Experiments.

Figure 6: Structure of adaptive randomisation



3. Findings

In this section, we present the result of the experiment based on sample means and 95% confidence intervals. Throughout, reported p-values refer to the effect of treatments against the Control arm, corrected.[37]

The ranking of treatments at the end of Round 4 is shown in Table 2 (see Figure A.1 in the Appendix for the evolution of assignment and posterior probabilities). Quiz + Prompt + Fact-checkers label (Treatment 11) was the best performing treatment in terms of the target outcome, with associated posterior probability just under 50%. When tested against the Control in Round 5 of the experiment, it resulted in a 4.86-point increase in average accuracy score (95% confidence interval: [4.04, 5.69]) from a baseline of 65.48 for the Control.

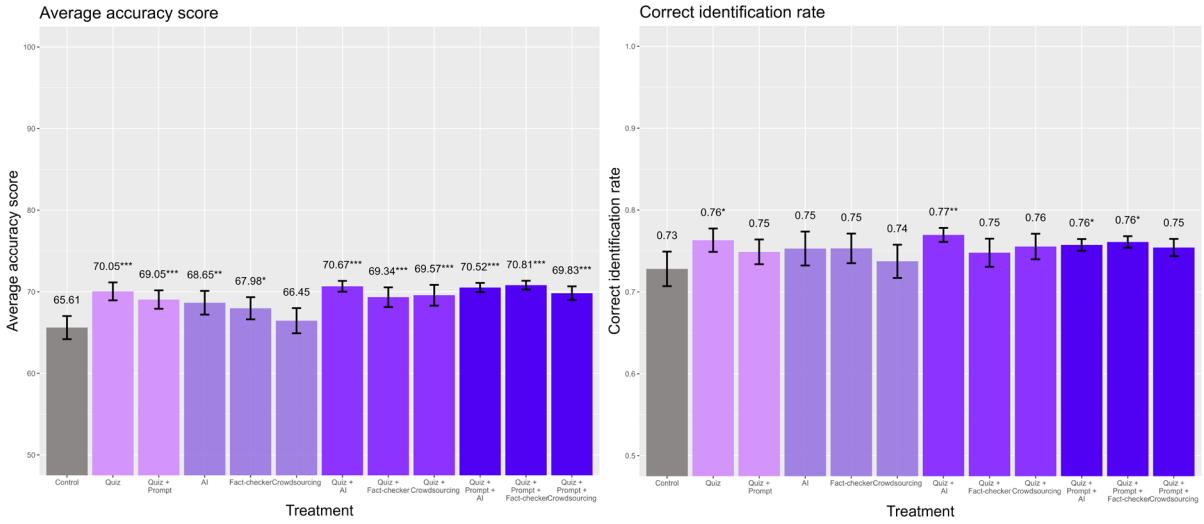
Table 2: Final ranking of trial arms (Rounds 1-4)

Ranking	Trial arm	Posterior probability	Sample size	Sample mean for Y_i
1	Quiz + Prompt + Fact-checker	0.4979	980	70.80 (0.28)
2	Quiz + AI	0.2869	632	70.70 (0.34)
3	Quiz + Prompt + AI	0.1299	828	70.50 (0.29)
4	Quiz	0.0547	228	70.00 (0.56)
5	Quiz + Crowdsourcing	0.0161	172	69.60 (0.65)
6	Quiz + Prompt + Crowdsourcing	0.008	387	69.80 (0.43)
7	Quiz + Fact-checker	0.0046	142	69.30 (0.72)
8	AI	0.001	136	68.70 (0.74)
9	Quiz + Prompt	0.0008	205	69.00 (0.58)
10	Fact-checker	0	158	68.00 (0.69)
11	Crowdsourcing	0	149	66.50 (0.79)
12	Control	0	141	65.62 (0.68)

[37] The effective sample sizes after applying sample exclusion restrictions were 4,158 for Rounds 1-4 and 1,763 for Round 5, out of a total of 4200 and 1800 participants, respectively.

Figure 7 presents final results for all the treatments considered in Rounds 1-4 (see Tables A.1-A.2 in the Appendix for full reporting of results). Treatments combining user-focused interventions and labels display very similar and statistically significant positive effects on average accuracy score (in the range of 3.73-5.20 increases), with comparable performance by Educational Quiz alone (Treatment 2). By contrast, treatments involving labels alone provided more modest gains, with Crowdsourced Notes notably failing to deliver statistically significant positive effects at conventional levels. These insights are corroborated by Figure 8, where results are reported after consolidating treatments by treatment type and label type, respectively. Combinations of user-focused interventions and labels generally performed best, with the former accounting for much of the improvement. Crowdsourced Notes performed worse than AI and Fact-checker labels across all treatments that included Post-focused interventions, although only to a marginal extent.[38] Results for the correct identification rate (Figure 7, right panel) show overall positive effects that fail to achieve statistical significance for most treatments, with Educational Quiz + AI (Treatment 7) delivering the largest increase of 0.042 to the 0.728 correct identification rate observed in the Control arm.

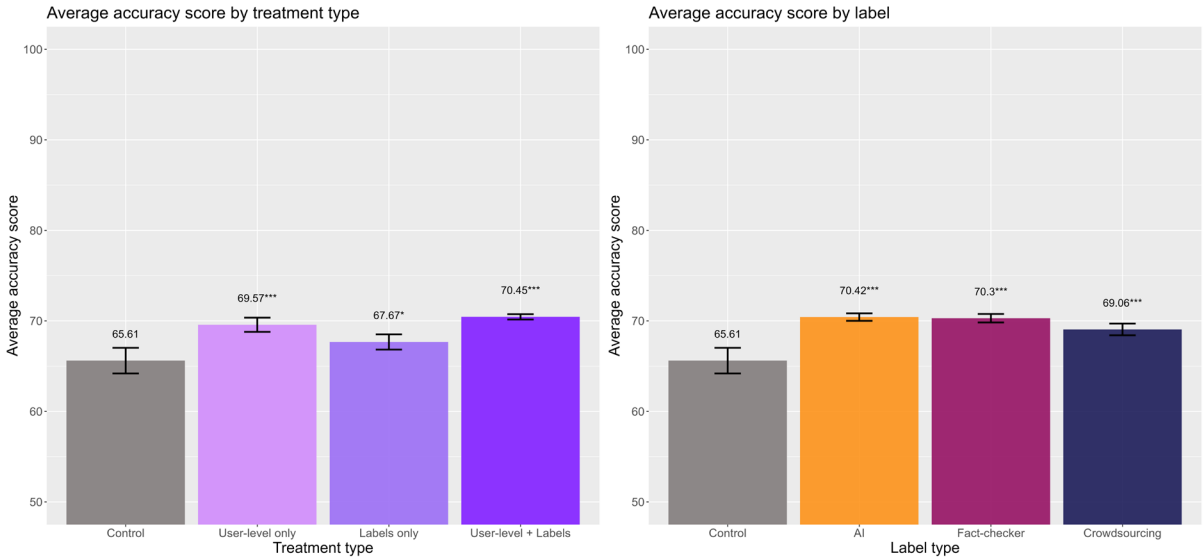
Figure 7: Results for average accuracy score (left) and correct identification rate (right). (Rounds 1-4)



*p-values (vs. control) corrected for 11 multiple comparisons (Benjamini-Hochberg): *** <0.01, ** <0.05, * <0.10*

[38] We have also computed estimates and confidence intervals for the primary outcome using the leave-future-out adaptively-weighted AIPW estimator with variance-stabilising weights in Hadad et al. (2021, link). The results are presented in the Appendix (Table A.2) and are very close to those obtained with sample means. This reassures us that sample sizes in Rounds 1-4 were sufficiently large to ensure estimates with negligible bias coming from adaptivity.

Figure 8: Results by treatment type (left) and label type (right). (Rounds 1-4)

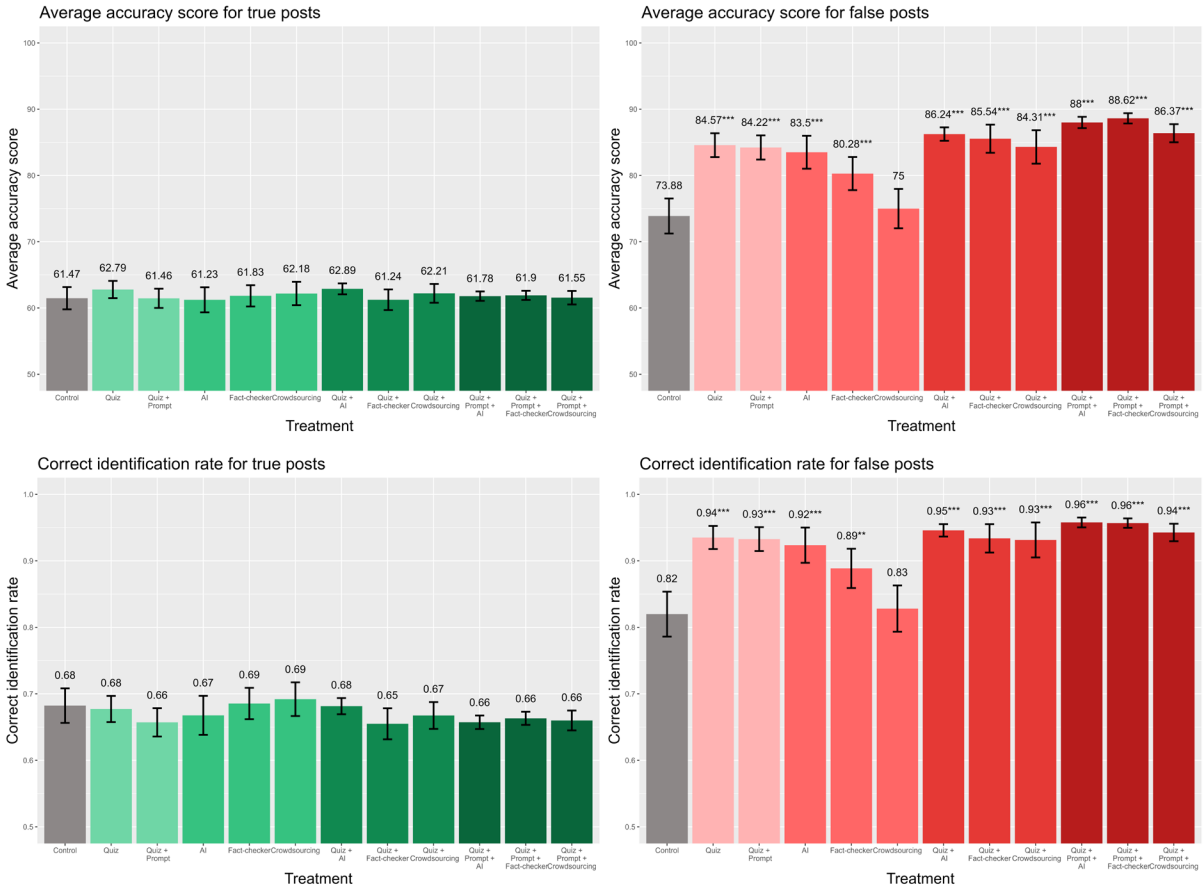


*p-values (vs. control) corrected for 4 multiple comparisons (Benjamini-Hochberg): *** <0.01, ** <0.05, * <0.10*
"User-level only" includes observations for Treatments 2 and 3; "Labels only" for treatments 4, 5 and 6; "User-level + labels" for treatments 7-12.
"AI" includes observations for treatments 4, 7 and 10; "Fact-checker" for treatments 5, 8 and 11; "Crowdsourcing" for treatments 6, 9 and 12.

Disaggregated analysis for true and false posts in Figure 9 reveals key insights into the drivers of the results. While consistent with the aggregated analysis in terms of relative treatment effectiveness, results for the false posts show substantially larger positive effects in terms of both accuracy score and correct identification rate. Differences in positive effects between treatments are also more pronounced for false posts, with increases in correct identification rate ranging from 0.008 for Crowdsourced Notes (Treatment 6) to 0.138 for Quiz + Prompt + Fact-Checker label (Treatment 11). In contrast, none of the treatments produced statistically significant changes in average accuracy scores relative to the Control for true posts. Small decreases in correct identification rates relative to the Control arm are observed for treatments that delivered the largest gains on false posts, suggesting that user-focused interventions may be partly operating through a general reduction in trust for all content. However, such decreases are not statistically significant for any of the treatments.

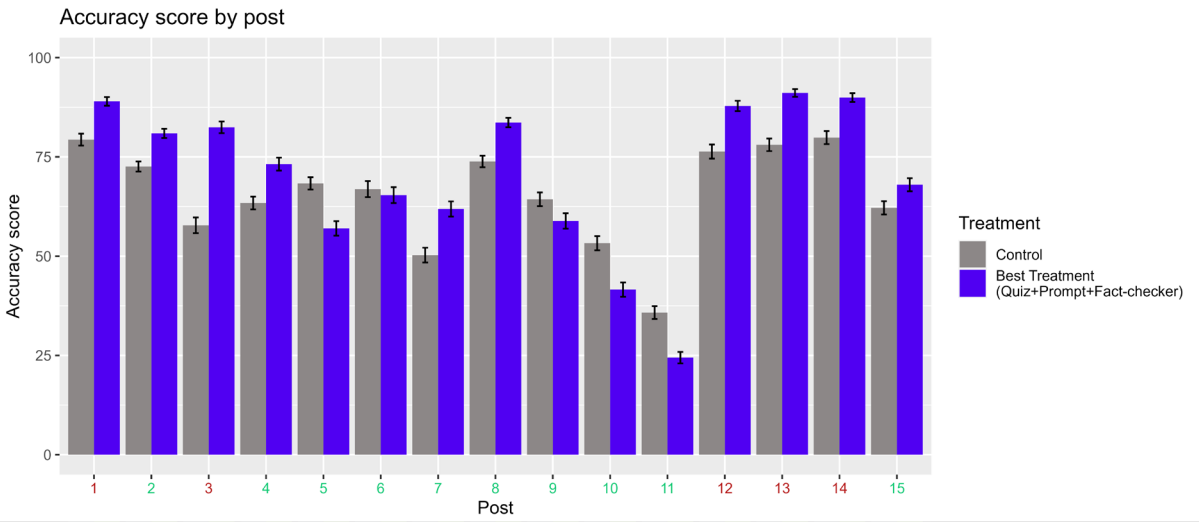
Finally, results for Round 5 reported in the Appendix closely resemble those obtained in Rounds 1-4 for both Control and Treatment 11 for all outcome measures. Analysis by individual post for Round 5 (Figure 10) further indicates that the effect of the best treatment is positive across all false posts. By contrast, there is considerable heterogeneity in effects for true posts, with statistically significant negative effects observed for some.

Figure 9: Results for true posts (left) and false posts (right). (Rounds 1-4)



*p-values (vs. control) corrected for 11 multiple comparisons (Benjamini-Hochberg): *** <0.01, ** <0.05, * <0.10*

Figure 10: Accuracy score by post: Best Treatment vs Control. (Round 5)



True posts (green): 2, 4, 5, 6, 7, 8, 9, 10, 11, 15. False posts (red): 1, 3, 12, 13, 14.

4. Discussion

This is the first study in which Ofcom has used an ARCT. This adds to the tools and techniques we use in our work, allowing for the efficient and cost-effective testing of numerous treatments, including combinations of interventions to assess their complementarity. Using a nationally representative UK sample, the study tested and delivered new experimental evidence on the effectiveness of various media literacy tools and offered valuable insights into how users respond to interventions designed to help them assess reliability, veracity and authenticity of content online.

First, our results show that user-focused tools and labels were most effective when used in combination, pointing to complementarities between these two types of interventions. This finding provides evidence that they may act through partly distinct behavioural channels, as user-focused interventions provide tools that encourage users to actively assess content, while labels are more likely to instigate heuristic judgements based on the perceived credibility of the underlying verification mechanism (Koch et al, 2023).[39] It also indicates that combining interventions did not give rise to adverse effects associated with 'over-warning' in the context of this controlled experiment. The ability to provide novel evidence on these complementarities reflects the value of the adaptive randomisation design, which allowed us to test combinations of interventions as additional treatments.

Another relevant insight is the comparatively stronger performance of user-focused interventions relative to labels. This is notable given that labels explicitly flagged false posts, whereas user-focused interventions did not provide direct indications of the veracity at the post level. This finding is consistent with existing literature suggesting that interventions which train users to recognise misleading content can produce robust improvements in discernment (Pennycook and Rand, 2019, 2021; Roozenbeek and van der Linden, 2019).[40] Evidence on the effectiveness of post-focused labelling is instead more mixed and appears to depend on design and context (Clayton et al., 2020; Oeldorf-Hirsch et al., 2020).[41] This study contributes some of the first direct comparative evidence on their relative effectiveness.[42]

Importantly, we also find no evidence that these interventions—whether used in isolation or in combination in this study—tangibly reduced accuracy in identifying genuine information. This finding helps assuage concerns that interventions might produce adverse effects by inducing generalised scepticism and thus reduce trust in accurate content (Pennycook and Rand, 2021).[43]

Finally, Crowdsourced Notes labels were slightly less effective than Fact-Checker and AI labels in improving participants' ability to identify false posts. This may reflect the greater perceived impartiality of verification mechanisms backed by established institutional or technological

[39] Koch, T. K., Frischlich, L., and Lermer, E. (2023), Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media.
[40] Roozenbeek, J., and van der Linden, S. (2019), The Bad News Game: Active Inoculation Against Online Misinformation.
[41] Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., and Boyle, M. P. (2020), The Ineffectiveness of Fact-Checking Labels on News Memes and Articles.
[42] Offer-Westort et al. (2024) also report a qualitatively similar pattern for the relative effectiveness of user- and post-focused interventions on sharing behaviour among social media users in Kenya and Nigeria.
[43] Pennycook, G., and Rand, D. (2021), The Psychology of Fake News.

infrastructure and is consistent with previous findings from the US context showing that expert-based signals tend to be more influential than peer-generated cues (Kim et al., 2022). [44] [45]

Several limitations should be noted when interpreting these results:

First, the effectiveness of any intervention may vary depending on the context in which it is studied. This experiment simulated a social media environment, which may not fully capture real-world user behaviour. On actual platforms, users' motivations and incentives may be different and algorithmic curation and personalisation may play a significant role. Additionally, although this study used a nationally representative sample from the UK population, the findings may not be generalisable to contexts with different demographic and cultural characteristics.

Second, the results of this study crucially depend on the specific content shown to the participants. While our choice of posts was driven by realism and relevance to the UK context, the impact of interventions could change if different sets of content were to be used in the experiment. Moreover, both the nature and perception of information are context-dependent, limiting the extent to which our findings can generalise over time or across different information environments.

Third, the study only measured short-term impacts. This limits our ability to draw conclusions about long-term effects, especially since the effect of educational tools may attenuate over time.

Therefore, while our findings provide valuable insights, further research is needed to test interventions in more realistic settings and over longer horizons. Such evidence is crucial to establish which tools best allow users to identify the reliability of information and how they may be most effectively deployed.

[44] Kim, A., Moravec, P. L., and Dennis, A. R. (2019), Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings.

[45] The results may also reflect the different presentation of Crowdsourced notes in terms of wording and information provided, compared to Fact-checker and AI labels.

Appendix

Randomisation algorithm

The algorithm used to update assignment probabilities at the end of each round is based on Exploration Sampling (Kasy and Sautmann, 2021), a modification of Thompson Sampling geared towards identification of the best arm.

We augmented the algorithm by enforcing probability floors, ensuring that the probabilities of assignment $p^*_{t,k}$ for each arm $k = 1, \dots, 12$ at every round $t = 1, \dots, 4$ do not fall below $\underline{p}$=0.02.

For the Bayesian updating of probabilities, we use a normal likelihood with normal-inverse-gamma independent priors, initialised with prior parameters as follows:

- Prior means m_k=50 for all *k*,
- Prior count ν_k=1 for all *k*,
- Prior shape α_k=1 for all *k*,
- Prior scale β_k=1 for all *k*.

The first round of the experiment is carried out with equal assignment probabilities across arms. Assignment probabilities at each subsequent round are then updated based on data from all previous rounds using our algorithm, as described by the pseudo-code below.

Algorithm pseudo-code

Inputs:

- Outcomes Y_i and Treatments D_i for $i = 1 \dots N_t$, obtained up to round t
- Total number of treatments K
- Size of next wave n_{t+1}
- Probability floor $\underline{p}$
- Vector of tuning parameters:
  - **a.** Prior means m_k
  - **b.** Prior count ν_k
  - **c.** Prior shape α_k
  - **d.** Prior scale β_k
- Number of Monte Carlo draws R for computation of probabilities

Output:

- Vector of final assignment probabilities $p^*_{t+1,k}$.
- Vector of posterior probabilities $p_{t,k}$.

Additional notation:

- $N_{t,k} = \sum_i^{N_t} 1\{D_i = k\}$, the sample size for treatment k at round t

Algorithm flow

Compute parameters for posterior

For $k = 1, \dots, K$

Set $\overline{Y}_k \leftarrow \frac{\sum_i^{N_t} Y_i \cdot 1\{D_i = k\}}{N_{t,k}} \times 1\{N_{t,k} > 0\}$,

$s_k^2 \leftarrow \sum_{i=1}^{N_t} \left(Y_i - \overline{Y}_k\right)^2 \cdot 1\{D_i = k\}$,

$\hat{\rho}_k \leftarrow \frac{N_{t,k} \cdot \overline{Y}_k + \nu_k \cdot m_j}{N_{t,k} + \nu_k}$,

$\hat{\beta}_k \leftarrow \beta_k + \frac{s_k^2}{2} + \frac{N_{t,k} \cdot \nu_k \cdot \left(\overline{Y}_k - m_k\right)^2}{2 \cdot \left(N_{t,k} + \nu_k\right)}$

End

Obtain Thompson Sampling Probabilities by simulation

For $i = 1, \dots R$

For $k = 1, \dots, K$

Draw $\hat{\sigma}_{k,i}^2 \sim$ Inv-Gamma$\left(0.5 \cdot N_{t,k} + \alpha_k,\ \hat{\beta}_k\right)$

Set $\hat{\zeta}_{k,i}^2 = \frac{\hat{\sigma}_{k,i}^2}{N_{t,k} + \nu_k}$

Draw $\hat{\theta}_{k,i} \sim \mathcal{N}\left(\hat{\rho}_k, \hat{\zeta}_{k,i}^2\right)$

End

Set $D_i^* \leftarrow \text{argmax}_k \hat{\theta}_{k,i}$

End

For $k = 1, \dots, K$

Set $p_{t,k} \leftarrow \frac{\sum_{r=1}^{RR} 1\{D_i^* = k\}}{RR}$

End

Obtain Exploration Sampling probabilities

If $\max_k p_{t+1,k} < 1$

For $k = 1, \dots, K$

Set $\tilde{p}_{t,k} \leftarrow p_{t,k} \cdot \left(1 - p_{t,k}\right)$

End

Set $\tilde{p}_t \leftarrow \frac{\tilde{p}_t}{\sum_{k=1}^K \tilde{p}_{t,k}}$ *(this is intended as a vector operation)*

End

Enforce probability floors

Set $p^*_{t+1} = \vec{0}$

While $\min_k p^*_{t+1,k} < \underline{p}$

For $k = 1, \dots, K$

If $\tilde{p}_{t,k} \leq \underline{p}$

Set $\check{p}_{t,k} = max\{\tilde{p}_{t,k}, \underline{p}\}$

End

End

Set $\underline{S}^* = \sum_{k=1}^{K} \check{p}_{t,k} \cdot 1\{\tilde{p}_{t,k} \leq \underline{p}\}$

Set $\overline{S} = \sum_{k=1}^{K} \tilde{p}_{t,k} \cdot 1\{\tilde{p}_{t,k} > \underline{p}\}$

For $k = 1, \dots, K$

If $\tilde{p}_{t,k} > \underline{p}$

Set $\tilde{p}_{t,k} = \tilde{p}_{t,k} \cdot \frac{1-\underline{S}^*}{\overline{S}}$

End

End

$p^*_{t+1} = \tilde{p}_t$ *(this is intended as a vector operation)*

End

Additional results

Table A.1: Results for Average Accuracy Score (Rounds 1-4)

Trial arm	Average accuracy score (Hadad et al., 2021)	Average accuracy score	Average accuracy score: effect vs control	Accuracy score for true posts	Effect vs Control (True posts)	Accuracy score for false posts	Effect vs Control (False posts)
T1 Control	65.58 (0.86)	65.60 (0.72)	–	61.50 (0.86)	–	73.90 (1.35)	–
T2 Quiz	70.66 (0.67)	70.00 (0.56)	4.44*** (0.91)	62.80 (0.67)	1.31 (1.08)	84.60 (0.92)	10.69*** (1.63)
T3 Quiz + Prompt	68.80 (0.72)	69.00 (0.58)	3.44*** (0.92)	61.50 (0.74)	-0.02 (1.13)	84.20 (0.93)	10.34*** (1.63)
T4 AI Label	68.62 (0.68)	68.70 (0.74)	3.04** (1.03)	61.20 (0.97)	-0.24 (1.29)	83.50 (1.27)	9.62*** (1.84)
T5 Fact-checker label	67.94 (0.81)	68.00 (0.69)	2.37* (1.00)	61.80 (0.82)	0.36 (1.18)	80.30 (1.28)	6.41*** (1.85)
T6 Crowdsourcing	66.62 (0.78)	66.50 (0.79)	0.84 (1.07)	62.20 (0.90)	0.71 (1.24)	75.00 (1.51)	1.12 (2.02)
T7 Quiz + AI label	70.60 (0.36)	70.70 (0.34)	5.06*** (0.80)	62.90 (0.42)	1.41 (0.95)	86.20 (0.52)	12.36*** (1.44)
T8 Quiz + Fact-checker	69.47 (0.72)	69.30 (0.62)	3.73*** (0.95)	61.20 (0.80)	-0.24 (1.17)	85.50 (1.09)	11.66*** (1.73)
T9 Quiz + Crowdsourcing	69.77 (0.71)	69.60 (0.65)	3.97*** (0.97)	62.20 (0.72)	0.73 (1.12)	84.30 (1.29)	10.43*** (1.86)
T10 Quiz + Prompt + AI	70.53 (0.33)	70.50 (0.29)	4.91*** (0.78)	61.80 (0.36)	0.31 (0.93)	88.00 (0.43)	14.12*** (1.41)
T11 Quiz + Prompt + Fact-checker	70.84 (0.29)	70.80 (0.28)	5.20*** (0.77)	61.90 (0.35)	0.43 (0.93)	88.60 (0.40)	14.74*** (1.40)
T12 Quiz + Prompt + Crowdsourcing	69.74 (0.72)	69.80 (0.43)	4.22*** (0.84)	61.60 (0.52)	0.08 (1.00)	86.40 (0.69)	12.50*** (1.51)

*p-values for effects (vs. control) corrected for 11 multiple comparisons (Benjamini-Hochberg): *** <0.01, ** <0.05, * <0.10. Estimates computed via sample means, except for column "Average accuracy score (Hadad et al.)" where estimates and standard errors are computed using the leave-future-out adaptively-weighted AIPW estimator with variance-stabilising weights (see* Hadad, V., Hirshberg, D. A., Zhan, R., and Athey S. (2021), Confidence intervals for policy evaluation in adaptive experiments).

Table A.2: Results for Correct Identification Rate (Rounds 1-4)

Trial arm	Correct identification rate	Effect vs control	Correct identification rate for true posts	Effect vs Control (True posts)	Correct identification rate for false posts	Effect vs Control (False posts)
T1 Control	0.728 (0.011)	–	0.682 (0.013)	–	0.820 (0.017)	–
T2 Quiz	0.763 (0.007)	0.035* (0.000)	0.677 (0.010)	-0.005 (0.017)	0.935 (0.009)	0.115*** (0.019)
T3 Quiz + Prompt	0.749 (0.008)	0.021 (0.000)	0.657 (0.011)	-0.025 (0.017)	0.933 (0.009)	0.113*** (0.020)
T4 AI Label	0.753 (0.011)	0.025*** (0.000)	0.668 (0.015)	-0.015 (0.020)	0.923 (0.013)	0.104*** (0.022)
T5 Fact-checker label	0.753 (0.009)	0.025*** (0.000)	0.685 (0.012)	0.003 (0.018)	0.889 (0.015)	0.069** (0.023)
T6 Crowdsourcing	0.737 (0.010)	0.009 (0.000)	0.692 (0.013)	0.010 (0.018)	0.828 (0.018)	0.008 (0.025)
T7 Quiz + AI label	0.770 (0.004)	0.041** (0.000)	0.681 (0.006)	-0.001 (0.015)	0.946 (0.005)	0.126*** (0.018)
T8 Quiz + Fact-checker	0.748 (0.009)	0.020 (0.000)	0.655 (0.012)	-0.027 (0.018)	0.934 (0.011)	0.114*** (0.020)
T9 Quiz + Crowdsourcing	0.755 (0.008)	0.027 (0.000)	0.667 (0.010)	-0.015 (0.017)	0.931 (0.013)	0.112*** (0.022)
T10 Quiz + Prompt + AI	0.757 (0.004)	0.029* (0.000)	0.657 (0.005)	-0.025 (0.014)	0.958 (0.004)	0.138*** (0.018)
T11 Quiz + Prompt + Fact-checker	0.761 (0.004)	0.033* (0.000)	0.663 (0.005)	-0.019 (0.014)	0.957 (0.004)	0.137*** (0.018)
T12 Quiz + Prompt + Crowdsourcing	0.754 (0.005)	0.026 (0.000)	0.660 (0.008)	-0.022 (0.015)	0.943 (0.007)	0.123*** (0.019)

*p-values for effects (vs. control) corrected for 11 multiple comparisons (Benjamini-Hochberg): *** <0.01, ** <0.05, * <0.10. Estimates computed via sample means.*

Table A.3: Results for Average Accuracy Score: Best Treatment vs Control. (Round 5)

Trial arm	Average accuracy score	Average accuracy score: effect vs control	Accuracy score for true posts	Effect vs Control (True posts)	Accuracy score for false posts	Effect vs Control (False posts)
T1 Control	65.48 (0.30)	–	61.09 (0.37)	–	74.28 (0.55)	–
T11 Quiz + Prompt + Fact-checker	70.35 (0.29)	4.86*** (0.42)	61.49 (0.36)	0.40 (0.52)	88.06 (0.43)	13.79*** (0.70)

*p-values for effect (vs. control): *** <0.01, ** <0.05, * <0.10. Estimates computed via sample means.*

Table A.4: Results for Correct Identification Rate: Best Treatment vs Control. (Round 5)

Trial arm	Correct identification rate	Effect vs control	Correct identification rate for true posts	Effect vs Control (True posts)	Correct identification rate for false posts	Effect vs Control (False posts)
T1 Control	0.726 (0.004)	–	0.673 (0.006)	–	0.832 (0.007)	–
T11 Quiz + Prompt + Fact-checker	0.757 (0.004)	0.031*** (0.006)	0.659 (0.005)	-0.015 (0.008)	0.955 (0.004)	0.123*** (0.008)

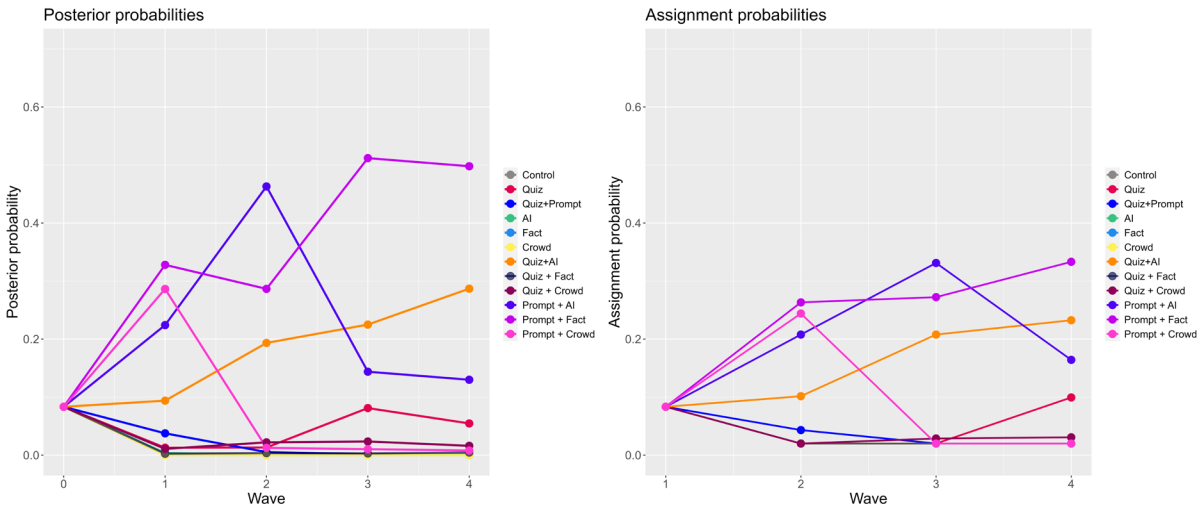
*p-values for effect (vs. control): *** <0.01, ** <0.05, * <0.10. Estimates computed via sample means.*

Additional tables and figures

Table A.5: Number of posts used by topic

	Educational Quiz	Main Feed
False Posts	**2**	**5**
Economy/ Finances	1	
Health		1
Migration		1
Political	1	3
True Posts	**1**	**10**
Economy/ Finances		2
Health	1	2
Migration		2
Political		3
Science		1
Grand Total	**3**	**15**

Figure A.1: Evolution of posterior probabilities (left) and assignment probabilities (right). (Rounds 1-4)



Posterior probabilities for each round reflect the likelihood that each treatment is the best in terms of the target outcome Y_i, based on data collected up to (and including) each round. Assignment probabilities for each round were computed by our algorithm based on data collected in previous rounds.