Advised or Paid Way to get it right. The contribution of fact-checking tips and monetary incentives to spotting scientific disinformation

Folco Panizza[1*],Piero Ronzani[1], Simone Mattavelli[2],Tiffany Morisseau[3,4,5], Carlo Martini[1,6],Matteo Motterlini[1]

1 Centre for Applied and Experimental Epistemology, Vita-Salute San Raffaele University, Via Borromeo, 41, 20811 Cesano Maderno (MB), Italy.
2 Department of Psychology, Bicocca University, Piazza dell'Ateneo Nuovo, 1, 20126 Milano, Italy.
3 Université de Paris, Laboratoire de Psychologie et d'Ergonomie Appliquées, Boulogne-Billancourt, France
4 Laboratoire de Psychologie et d'Ergonomie Appliquées, Université Gustave Eiffel, Versailles, France
5 Strane Innovation, Gif-sur-Yvette, France
6 TINT – Centre for Philosophy of Social Science, Department of Political and Economic Studies, University of Helsinki, P.O. Box 24 (Unioninkatu 40A), 00014 Helsinki, Finland.

* corresponding author: panizza.folco@hsr.it

Abstract

Disinformation about science can impose enormous economic and public health burdens. Several types of interventions have been proposed to prevent the proliferation of false information online, where most of the spreading takes place. A recently proposed strategy to help online users recognise false content is to follow the techniques of professional fact checkers, such as looking for information on other websites (lateral reading) and looking beyond the first results suggested by search engines (click restraint). In two preregistered online experiments (N = 5387), we simulated a social-media environment and set-out two interventions, one in the form of a pop-up meant to advise participants to follow such techniques, the other based on monetary incentive. In Experiment 1, we compared these interventions to a control condition. In Experiment 2 another condition was added to test the joint impact of the pop-up and the monetary incentive. We measured participants' ability to identify whether presented information was scientifically valid or invalid. Results revealed that while monetary incentives were overall more effective in increasing accuracy, the pop-up contributed when the post originated from an unknown source (and participants could rely less on prior information). Additional analysis on participants' search style based on both self-report responses and objectively measured behaviour revealed that the pop-up increased the use of fact-checking strategies, and that these in turn increased accuracy. Study 2 also clarified that the pop-up and the incentive did not interfere with each other, but rather acted complementarily, suggesting that attention and literacy interventions can be designed in synergy.

Introduction

The massive circulation of inaccurate scientific information can have nefarious societal consequences. Successful misconceptions influence the public debate on decisions regarding the effectiveness of a vaccine, the adoption of solutions mitigating climate change, or the cost of a social policy. The sharing of false information is easily fuelled by political or social motivations that disregard the best scientific evidence on the matter. It is indeed tempting to share information on social media without verifying its truthfulness, simply because the mere act of sharing allows us to exhibit our position on a given topic and to justify the validity of such a position. This phenomenon is amplified in crisis situations when scarce information is accompanied by multiple and contrasting rumours (also called infodemic) that could serve different views. People's propensity to accept scientifically dubious information can thus become a crucial problem for both democracy and public welfare.

There are structural challenges to fighting the spread of false information on social media. One key issue is that companies often perceive a trade-off between engaging users and combating viral but fake content, to the point of favouring the former over the latter [1]. Curtailment is made even more difficult when there is a deliberate intent behind the dissemination, what researchers refer to as disinformation. For example, at the peak of the coronavirus infodemic, only 16% of fact-checked disinformation was labelled as such by Facebook's algorithms, partly because content creators were able to simply repost content with minor changes, thus escaping detection [2]. It is therefore essential that, in combination with a systematic change in policy, users themselves are empowered against malicious or false content. User-based resilience needs to be part of a toolkit to fight disinformation: for instance, among the pillars of infodemic management, Eysenbach [3] lists eHealth Literacy, science literacy capacity, and critical thinking ability to fact-check information. Fighting science-related disinformation is harder than contrasting other forms of disinformation (e.g. political) because in the former case the lines between expertise and pseudoexpertise are blurred, and incompetent or otherwise biased sources pose as expert sources on topics like epidemiology or climate change.

Research on countering disinformation has developed substantially over the last decade, bringing a wealth of different approaches [4–8]. These include debunking, the systematic correction of false claims after they have been seen or heard [9, 10], pre-bunking, preventive measures before exposure to disinformation [5, 11], nudging, interventions affecting users' choices without limiting their freedom of choice [12], and boosting, the empowering of users by fostering existing competences or instilling new ones [12]. All of the above approaches have proven to be useful in a social media context, not least by adopting ingenious and innovative adaptations of classical paradigms. Debunking has been extensively studied, with several experiments focusing on the source [13–16] and the timing [17] of fact checking. Research has also explored whether evaluations about the quality of contents and sources can be delegated to the so-called wisdom of crowds, with encouraging results [18–20]. Studies on pre-bunkning have largely focused on the concept of inoculation [5, 21], namely exposing users to disinformation strategies in order to ease their recognition in future settings. Inoculation has demonstrated pronounced and lasting effects when introduced through games [22–25]. Nudging was also tested by showing warning labels for unchecked or false claims [26–29], but also by priming users to pay attention to the accuracy of content they might be willing to share [30–32] (however see [33] for a critique of this approach). Finally, boosting was tested by presenting users with a list of news/media literacy tips or guidelines on how to evaluate information on-line [34–38], producing some remarkable results and some non-significant ones.

A promising example of media literacy intervention has been carried out by researchers interested in understanding how fact checkers search for information about

unknown but institutional-looking sources [39]. Researchers catalogued fact checkers' strategies and distilled a series of questions to evaluate content, a set of skills that was named Civic Online Reasoning [40, 41]. Two are the most prominent strategies adopted by fact checkers. One is lateral reading, namely leaving a website and opening new tabs along a horizontal axis in order to use the resources of the Internet to learn more about a site and its claims. The other is click restraint, that is, skipping the first search results of a browser search to avoid biases created by results-ranking algorithms. These strategies seem particularly fit when a content has unknown origins that are hardly identifiable or that appear legitimate on the surface, a feature that has been associated with content creators spreading scientific disinformation and disinformation [42]. Detecting scientific disinformation often requires specific expertise to evaluate the content and cross-check sources. Under such conditions, assessing the truthfulness of information becomes tricky.

In the absence of expertise and content knowledge, users can rely on a number of external cues to infer whether information presented as scientific is reliable [43]. Unlike fake news, scientific disinformation relies on background knowledge about the expertise of the source. We can check, for example, whether a piece of information is agreed upon by the scientific community, or whether a source is a genuine expert one or a pseudo-expert one, and these elements can give us important cues as to whether a source conveys scientific, rather than pseudo-scientific, information. Lateral reading and click restraint can thus be used when scientific disinformation is deceptively sophisticated and difficult to detect. Indeed, training on Civic Online Reasoning has proven very effective in countering disinformation among high school and college students [44–46], as well as elderly citizens [47]. Despite extensive research on Civic Online Reasoning, so far little attention has been paid to the application of these techniques on social media. It is therefore unclear how effective presenting these strategies on a social network can actually be.

Critical thinking strategies might not be the only potentially effective tools in evaluating scientific (dis)information. For instance users might not be sufficiently motivated to evaluate the truthfulness of the content they see. Many users might share news simply because they come from a source they trust or like, or because those news align with their values, without paying much attention to trust. The spread of scientific disinformation then is not only related to false beliefs, but also to motivated behavior, paired with strong personal identities and values. In order to better exploit the benefits of critical thinking tools, it is therefore also important to identify the respective effects of being aware of truth-motivated strategies; i.e., being motivated to know the truth about a given topic. It may be that people, while being somehow familiar with fact-checking and civic online reasoning techniques, are only eager to apply them when identifying the truthfulness of the information is reinforced by specific incentives.

One way to test the effect of motivation then is the use of monetary incentives. In other words, does paying participants for their being accurate increase their accuracy in the evaluation of content? The idea behind this intervention is that money increases motivation, and thus the attention paid to otherwise ignored cues about the accuracy of content. Supporting this view, a study conducted with a sample of Mechanical Turk workers on comparable settings showed that monetary incentives are the main driver for people to spend time on the platform and, even in the face of small average earnings, aspects such as immediate payment play an important role in workers' motivation [48].

Monetary incentives have been proven to be a cost effective tool to modify behavior in domains such as health and human development [49], where often an early boost in motivation promotes the adoption of cheap preventive behaviours, avoiding this way costly consequences [50]. From a psychological perspective, the use of incentives builds on the attention-based account of disinformation spread. This account posits that

certain features of social networks favour the dissemination of interesting and unexpected content at the expense of accuracy [4, 51]. Recent research in this field has found both laboratory and field evidence that accuracy of content is often overlooked and that simple cues reminding participants to evaluate the accuracy of content they share has an impact in terms of the proportion of fake/true news shared [30, 32, 52, 53]. Increasing accuracy through incentives is not an entirely novel idea in social media either, as shown in a recent initiative promoted by Twitter [54]. Although these premises indicate that this type of intervention can be very effective, it is not a given that economic incentives will have a positive effect on scientific content evaluation. In an experimental setting in particular, social media content is subject to higher scrutiny than when users scroll through their news feed [30]. It is therefore possible that additional incentives may not further increase participants' accuracy.

The aim of the present study was to test and compare the effectiveness of Civic Online Reasoning techniques and monetary incentives in contributing to the recognition of science-related content on social media. We conducted two pre-registered experiments where participants observed and interacted with one out of several Facebook posts that linked to an article presenting science-themed information. Participants were free to conduct further research on external websites in order to form a more accurate idea of the scientific validity of the post. Once satisfied with the information they gathered, participants rated how scientifically valid the claims contained in the post were. To test for the usefulness of Civic Online Reasoning techniques, we designed a pop-up that preceded the post presenting the lateral reading and click restraint strategies 1. The use of a pop-up ensured that participants processed the content before observing the post, an approach that has also been adopted in previous research [52]. A pop-up could be easily adapted in a social media setting as regular reminders with the necessary precautions to avoid the reduction of their salience with time [55, 56]. To test the effect of monetary incentives instead, we doubled the participation fee (equivalent to an average +£8.40/hour) if participants guessed correctly the validity of the post they were evaluating.

Fig 1. Screenshot of the pop-up presented to participants.

Experiment 1

In Experiment 1, we tested separately the efficacy of pop-up and monetary incentives, and compared their effects to a control condition with no interventions. To assess that the effect of the interventions is effective over the widest possible range of contexts, we used a set of 9 different Facebook posts varying in various properties, such as the scientific topic, the source reputation, and its level of factual reporting. The original pre-registration of this experiment can be retrieved from osf.io/gsu9j.

Materials and methods

Participants

We recruited 2700 U.K. residents through the online platform prolific.co on 11 March 2021 (for a rationale of sample size, see S1 Methods). All participants gave their informed consent for participating in the experiment. Average age was 36 ($SD = 13.5$, 8 not specified), 60.7% of participants were female, (39.1% male, 0.2% other), and 55.6% had a Bachelor's degree or higher. Although recruitment explicitly specified that the experiment was supported only on computers or laptops, 316 participants (11.7%) completed the experiment on a mobile device. As our hypotheses were based on the assumption that search would happen on a computer, both stimuli and measures were not designed for mobile use. We therefore had to exclude these participants from the analyses. Analyses were thus conducted on 2384 participants.

Design

We conducted the experiment on Qualtrics and lab.js [57]. During the experiment, participants observed and were able to interact with one out of several Facebook-like posts (Fig 2 shows three examples; click here for an interactive example from Experiment 2). Participants' task was to rate the scientific validity of the statements reported in the title, subtitle, and caption of the post ("how scientifically valid would you rate the information contained in the post?"; 6-point likert scale from (1) "definitely non-valid" to (6) "definitely valid"). Researchers rated independently the scientific validity of the posts' content in terms of valid/invalid according to pre-specified criteria (see S3 Methods). Participants could take as much time as they wanted in giving their rating. Crucially, participants were also explicitly told that they were allowed to leave the study page before evaluating the post. After the rating, participants completed a questionnaire and were paid £0.70 for their time. Median completion time of the experiment was 5 minutes.

Experimental conditions. Participants were randomly assigned to one of three experimental conditions: control, incentive, and pop-up. In the control condition, participants completed the task as described above. In the incentive condition, participants were doubled their participation fee if their rating matched that given by the experimenters. Unbeknownst to participants, the correctness of the answer depended only on whether the answer was valid or invalid, and not on the extremity of the answer (e.g. having answered 4 instead of 5), even though we selected unambiguously valid or invalid content. In the pop-up condition, presentation of the post was preceded by a pop-up (Fig 1) presenting a list of civic online reasoning techniques (e.g., lateral reading, click restraint) as tips to verify the information in the post.

Stimuli. Each participant observed one out of nine possible Facebook posts (Fig 2; see S1 File for a full list). Posts varied in terms of: (i) scientific validity of the content (i.e., six valid and three invalid posts, either with verified or debunked information; S3 Methods); (ii) topic (i.e., three on climate change, three on the coronavirus pandemic, three on health and nutrition); (iii) factual reporting of the source, based on ratings

from mediabiasfactcheck.com (i.e., three high/very high versus six low/very low); (iv) source reputation, as measured in a screening survey (S4 Methods; three categories: trusted (2 posts), distrusted (4), unknown source (3)). Posts were balanced to have three posts for each topic, one from a source with high factual reporting displaying valid information, one from a source with low factual reporting displaying valid information, and one from a source with low factual reporting displaying invalid information.

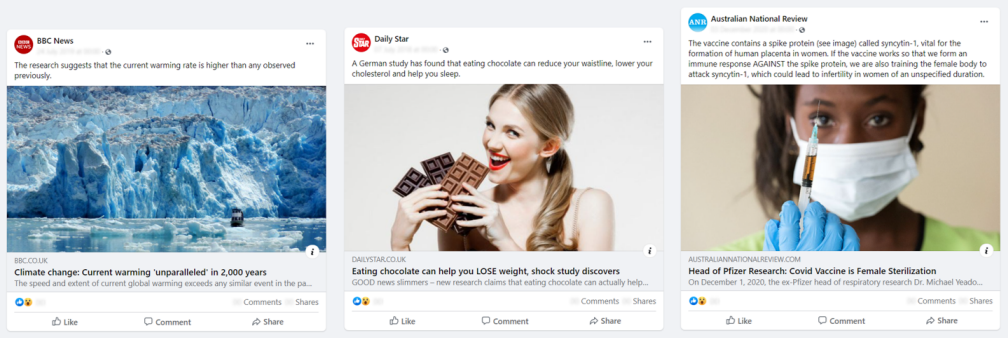

Fig 2. Examples of the stimuli presented, varying in topic (Climate Change, Health and Nutrition, COVID-19), factual reporting (high, low, low), scientific validity (high, low, low), and source reputation (trusted, untrusted, unknown source).

We standardised emoji reactions across all posts to control for their influence. In addition, post date, number of reactions and shares were blurred. The rest of the post was instead accessible to the participant, who could click on different links to access the source Facebook page, the original article, and the Wikipedia page (if present). Text and image were taken from the article. Captions were short statements of a scientific nature, i.e. facts or events pertaining to some scientific mechanism.

Measures

Accuracy. We computed two measures of accuracy–correct guessing and accuracy score. Correct guessing refers to a dichotomous variable that tracks whether participant gave a 'valid' (vs. 'invalid') rating when the post content was actually scientifically valid (vs. invalid). Accuracy score instead is a standardised measure ranging from zero to one, with 0 indicating an incorrect "1" or "6" validity rating, 0.2 indicating an incorrect "2" or "5" rating, 0.4 an incorrect "3" or "4" rating, 0.6 a correct "3" or "4" rating, 0.8 a correct "2" or "5" rating, and 1 a correct "1" or "6" rating. Accuracy score allows to distinguish validity evaluations that are associated with different behaviours: for instance, not all participants would be willing to share content that they rated as 4 in terms of scientific validity. In addition, accuracy score is statistically more powerful than correct guessing as it includes more possible responses [58]. We thus considered accuracy score as our main index.

Search behaviour. During the evaluation of the post, we tracked participants' behaviour on the study page. We measured the time spent both inside and outside the page, and a series of dummy variables tracking whether participants had clicked on any of the links present (e.g., Facebook page, article page, Wikipedia page). Based on these calculations we were able to estimate participants' response times and search behaviour.

Civic Online Reasoning. After having rated the scientific validity of the post, participants completed a questionnaire investigating those factors that could have influenced their choice. In order to test our hypotheses, we asked participants whether they engaged in lateral reading and click restraint. Participant were said to have used lateral reading if they reported having searched for information outside the study page

(yes/no question), and if they specifically searched on a search engine among other destinations (multiple selection question). Participants were said to have used click restraint if they further reported looking beyond the first results suggested by the search engine (multiple choice question). Critically, questions were formulated in such a way as to avoid any expectation as to which answer to select, and thus reduce the influence of the experimenter.

Control measures. In addition to measures of accuracy and civic online reasoning, we included a series of control measures for our analyses (S5 Methods). Other questions included self-report measures of confidence in the validity rating, plausibility of the post content, subjective relevance of obtaining accurate information about the post, familiarity with the source, perceived trustworthiness of the source, subjective knowledge of the topic, trust in scientists, conspiratorial beliefs, and a scientific literacy test. In addition to responses in the questionnaire, we obtained information about participants from the recruiting platform, such as their level of education, socio-economic status, social media use, and belief in climate change.

Analyses

Statistical tests were conducted using base R [59]. We adopted the standard 5% significance level to test against the null hypotheses. All tests were two-tailed unless otherwise specified. Post-hoc tests and multiple comparisons were corrected using the Benjamini-Hochberg procedure. Non-parametric statistics were log-transformed for conciseness. For probability differences, the lower boundary indicates the 2.5% quantile of the effect of the target variable starting from the 2.5% quantile of the baseline probability estimate, whereas the upper boundary indicates the 97.5% quantile of the effect of the target variable starting from the 97.5% quantile of the baseline probability estimate. Given the small number of stimuli ($N < 10$), we do not cluster errors by Facebook post in our regression analyses. The use of random effects yields however comparable results in magnitude and statistical significance unless otherwise reported.

Results

Participant randomisation was balanced across conditions (Chi squared test, $\chi^2(2) = .016$, $p = .99$). Median time to evaluate the Facebook post was 33 seconds in the control condition (incentive condition: 45 seconds; pop-up condition: 35 seconds; minimum overall time: 2 seconds, maximum overall time: 40 minutes). In the pop-up condition, participants spent an additional median time of 11 seconds on the pop-up. On a scale from 1 to 6 (3.5 response at chance level), average accuracy score in the control condition was 4.35 ($SD = 1.20$; incentive condition 4.48, $SD = 1.32$; pop-up condition 4.35, $SD = 1.19$). In the control condition, 78.2% of participants correctly guessed the scientific validity of the post (incentive condition: 80.1%; pop-up condition: 78.1%).

Effect of interventions

To test the effect of our interventions on accuracy, we adopted two tests, one for the accuracy scores, and one for correct guessing (original preregistered analyses are presented in S1 Analyses). Since accuracy scores were clearly non-normally distributed (Shapiro-Wilk test, all $p < .001$), we used an ordinal logistic regression in place of the linear regression to test the effect of condition on accuracy scores. Results showed a significant effect of incentive ($\beta = .293$ [$.092, .494$], $z = 3.225$, $p = .003$) and a lack of significance for the pop-up ($\beta = -.009$ [$-.207, .188$], $z = -0.103$, $p = .918$). According to the model, the probability of giving a "definitely valid" ("definitely invalid") correct

response increases by 4.4% [1.5%,8.2%] in the incentive condition compared to the control condition.

Technique adoption

To compare the adoption of Civic Online Reasoning techniques between experimental conditions (pre-registered hypothesis 2) we used a logistic regression with technique use (adoption of both lateral reading and click restraint) as predicted variable and experimental condition as predictor. Results revealed that both incentive and pop-up increased technique adoption (Fig 3; incentive: $\beta = 1.042$ [.527, 1.556], $z = 4.728$, $p < .001$; pop-up: $\beta = 1.556$ [1.065, 2.046], $z = 7.405$, $p < .001$), but that the increase was markedly higher with the presence of the pop-up than with monetary incentives ($\beta = .514$ [.157, .871], $z = 3.362$, $p < .001$).

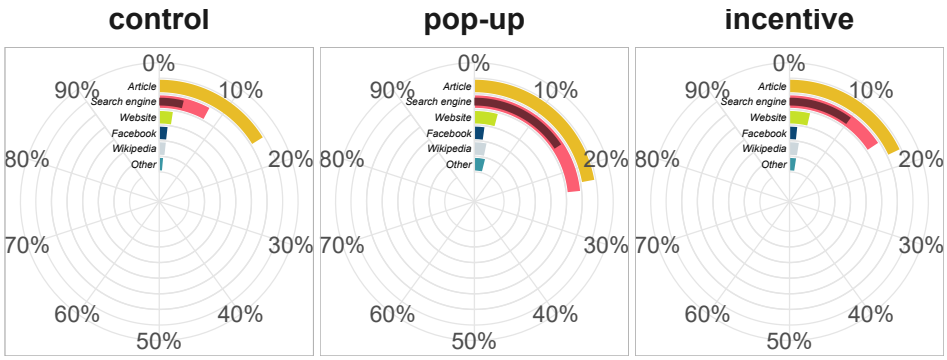

Fig 3. Race chart of self-report external search behaviour. Bars indicate the proportion of participants in each experimental condition reporting to have searched in either category of websites. Lateral reading is identified with the proportion of participants searching information on a search engine (light red), whereas click restraint is the subset of these participants who reported not stopping at the first algorithmically-ranked results of the search (dark red).

Since our measure of technique use is based on self-reporting, responses might have been biased by external expectations. We therefore checked whether participants who reported the use of techniques actually left the study by tracking their behaviour on the post's web page. According to our measures, 80% of these participants left the study in the control condition, compared to 87% in the pop-up and 90% in the incentive conditions. This result, if anything, suggests that our interventions did not increase the rate of false reporting. Moreover, even after accounting for false reports, results did not differ (incentive: $\beta = 1.156$ [.594, 1.719], $z = 4.791$, $p < .001$; pop-up: $\beta = 1.626$ [1.087, 2.166], $z = 7.024$, $p < .001$; pop-up > incentive: $\beta = .467$ [.095, .845], $z = 2.920$, $p = .004$; see sections S2 Analyses and S6 Analyses for an in-depth exploration of participants' search behaviour).

Did the use of lateral reading and click restraint actually improve post evaluation? And did the use of techniques mediate the effect of our interventions? To test our first question, we ran an ordinal logistic regression with accuracy score as predicted variable, and a standard logistic regression with correct guessing as predicted variable, both tests including adoption of techniques as the sole predictor. Results showed that accuracy score improved significantly if a participant reported using Civic Online Reasoning techniques ($\beta = .526$ [.274, .778], $z = 4.090$, $p < .001$). According to the model, the use of Civic Online Reasoning Techniques increased the probability of giving a "definitely valid" ("definitely invalid") correct response by 8.8% [4.0%,14.7%]. This result however was not confirmed by the standard logistic regression on correct guessing, which instead

found no significant effect of technique adoption ($\beta = .219$ $[-.121, .580]$, $z = 1.228$, $p = .220$).

Based on these results, we proceeded to test whether pop-up and incentives had some mediated impact on accuracy score through technique adoption. To test mediation we used the R package MarginalMediation [60]. Technique adoption was found to moderate the effect of both incentive and pop-up on accuracy score (incentive: unstandardised $\beta = .004$ $[.001, .006]$, $z = 4.728$, $p < .001$; pop-up: unstandardised $\beta = .007$ $[.003, .012]$, $z = 7.405$, $p < .001$), suggesting that both interventions affected indirectly accuracy scores.

Response times

As we expected monetary incentives to increase motivation, we tested whether response times (a common proxy for increased deliberation and attention) were affected by our interventions. We compared participants' evaluation time of the post across conditions by way of a Kruskal-Wallis rank sum test. The test was significant ($\chi^2(2) = 67.63$, $p < .001$), thus we conducted post hoc comparisons. All comparisons were significant, with participants in the incentive condition taking significantly more time than control ($\log(V) = 8.02$, $p < .001$) and pop-up ($\log(V) = 5.54$, $p < .001$) participants, and pop-up participants taking more time than control ($\log(V) = 2.41$, $p = .016$).

We tested whether longer evaluation times predicted higher accuracy scores by means of an ordinal logistic regression with log-transformed evaluation time as predictor and accuracy score as predicted variable. Results revealed a significant and positive association ($\beta = .182$ $[.095, .268]$, $z = 4.12$, $p < .001$). The result was confirmed also for correct guessing (logistic regression, $\beta = .242$ $[.120, .366]$, $z = 3.87$, $p < .001$).

We additionally looked at how much time participants spent outside the study page when they left without clicking any link (a proxy of lateral reading). The Kruskal-Wallis test was again significant ($\chi^2(2) = 13.482$, $p = .001$): of those participants who performed such external searches, control participants spent less time outside the page than participants in both the incentive ($\log(V) = 2.85$, $p = .006$) and the pop-up conditions ($\log(V) = 3.58$, $p = .001$), whereas we found no significant difference between incentive and pop-up ($\log(V) = .92$, $p = .360$).

Source reputation

Civic Online Reasoning techniques were originally designed for helping to evaluate content from seemingly legitimate but unknown websites [39]. We thus analysed differences in our interventions based on the recognisability and perceived trustworthiness of the posts' sources. The importance of a source's perceived trustworthiness was exemplified by two posts covering the same scientific article, one from BBC News (a source trusted by most participants), and another one from the Daily Mail (a source barely trusted by most participants). Despite the posts covered the same content and presented similar wording, participants' evaluation of the two posts differed considerably: average accuracy score was 4.7 for the BBC piece ($SD = 1.05$) and 4.05 for the Daily Mail piece ($SD = 1.08$; ordinal regression: $\beta = 1.255$ $[.926, 1.584]$, $z = 7.470$, $p < .001$), and the proportion of correct guesses was 90.7% and 77.3%, respectively (logistic regression: $\beta = 1.059$ $[.568, 1.576]$, $z = 4.132$, $p < .001$).

Perhaps not surprisingly, we observed that, in the pop-up condition, adoption of lateral reading and click restraint was strongly linked with source type (Chi squared test with technique adoption and source category as variables, $\chi^2(2) = 15.407$, $p < .001$): when the source was trusted, only 6.7% of participants used these techniques, whereas the proportion was 20% when the source was unknown. We then tested differences of the interventions by source type in accuracy scores and correct guessing.

Likelihood-ratio tests confirmed the importance of this variable for both analyses ($p < .001$), however family-wise corrected contrasts revealed only one significant result, the effect of incentive on accuracy scores for unknown sources ($\beta = .558$ [.114, 1.001], $z = 3.445$, $p = .005$; Fig 4; see S4 Analyses for results about the uncorrected contrasts).

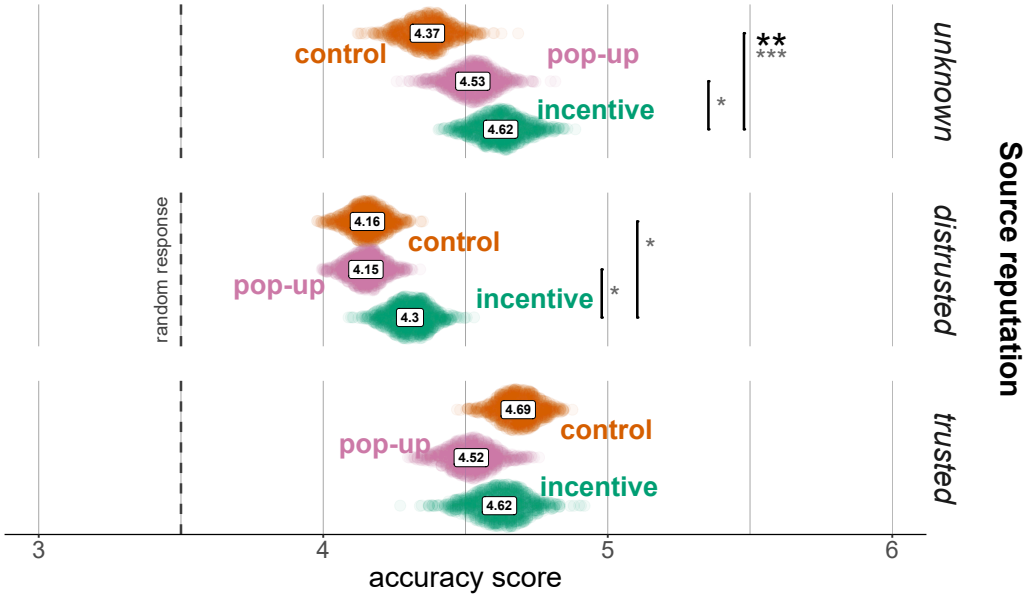

Fig 4. Bootstrap estimates of the average accuracy score by experimental condition and source reputation (Min. 1, Max. 6, random response: 3.5). Asterisks refer to significance of contrasts in the ordinal logistic regression. Black: family-wise corrected contrasts; dark grey: uncorrected contrasts. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Discussion

Results from Experiment 1 suggest that paying participants to be accurate does increase the accuracy score but not the proportion of participants correctly guessing the scientific validity of the posts. Compared to control, participants with an incentive gave more extreme answers, reported engaging in Civic Online Reasoning techniques more often (and did leave the page more often), spent more time in searching information outside the study page, and took longer to evaluate the post (even compared to pop-up participants). These results support the idea that monetary incentives affect accuracy, possibly by increasing motivation and attention in the task, although this hypothesis would need further testing.

By contrast, the presence of the pop-up seemed not to affect directly any indicator of accuracy. In spite of that, participants in the pop-up condition reported more lateral reading and click restraint, as well as the frequency of searches outside the study page. In turn, this increment of Civic Online Reasoning techniques (up to +13.5% when source is unknown) seems to mediate a small but significant increase in accuracy scores (marginal mediation analysis), suggesting an indirect effect of the pop-up. An effect of pop-up is possibly seen in posts produced by unknown sources, where correct guessing (but not accuracy scores) is slightly higher in the pop-up condition than in control (S4 Analyses).

These results suggest that monetary incentives might have more consistent effects over the presentation of Civic Online Reasoning techniques. At the same time, we observe considerable variability in participants' behavior depending on specific features of the posts. For instance, source reputation seems to have a remarkable effect on the adoption of Civic Online Reasoning techniques, which were (foreseeably) overlooked by almost all participants when looking at posts from generally trusted sources.

One potential takeaway from these findings is that some initial biases might affect the rate at which participants look for information outside the content provided (e.g. familiarity and opinion about the source), as well as in the way they look for such information. To explore this possibility, we designed a second experiment in which we tried to reduce the presence of initial biases by presenting posts from generally unknown sources. In addition, we included a fourth condition where we test the combination of monetary incentives and Civic Online Reasoning techniques, to explore whether and how the two interact.

Experiment 2

In line with evidence in the literature, we expected an increased impact of our interventions in a context where participants could rely on less prior information. We thus conducted a second experiment that was statistically powered to test for this possibility. In the Experiment 2 we replicated the format of the first one, with two main modifications: 1) we ran a pre-screening survey to identify lesser-known sources of information and only used those sources as the basis for the Facebook posts the participants were asked to evaluate; 2) we added an experimental condition that included both incentive and pop-up interventions, to test the interaction between the two. we advanced the idea that the two intervention strategies might trigger distinct behavioral outcomes (i.e., increased time spent on the task and use of Civic Online Reasoning). If this is the case, then combining the two interventions should produce even stronger effects on accuracy. The original pre-registration of this experiment can be retrieved from osf.io/w9vfb.

Materials and methods

Participants

3004 U.K. residents were recruited through the online platform prolific.co on 24 May 2021 (for a rationale of sample size, see S2 Methods). All participants gave their informed consent for participating in the experiment. Average age was 36 ($SD = 13.2$, 6 not specified), 63.1% of participants were female, (36.7% male, 0.2% other), and 59.4% had a Bachelor's degree or higher. Per our pre-registered criteria, we excluded one participant who was not a resident in the United Kingdom. Analyses were thus conducted on 3003 participants.

Design

The second experiment was a replication of the first one, with the major difference that sources of the Facebook posts were unknown to most participants. In addition, we included a fourth condition where we gave participants a monetary incentive and also showed them the pop-up with the Civic Online Reasoning techniques. Thus, the experiment had a between-subjects design with 2 factors, pop-up (present, absent) and monetary incentive (present, absent). Median completion time of the experiment was 5 minutes.

Stimuli

Participants observed one out of 6 posts that vared in terms of: the scientific validity of the content, i.e. the validity of the scientific statements in the title, subtitle, and caption of the post; the topic (climate change, coronavirus pandemic, and health and nutrition); factual reporting of the source, based on ratings from mediabiasfactcheck.com (3 high/very high versus 3 low/very low). All posts came from sources relatively unknown to participants, as measured in a preliminary survey and confirmed by participants' familiarity ratings. There were two distinct posts for each topic, one from a source with high factual reporting displaying valid information, one from a source with low factual reporting displaying invalid information.

Some titles, subtitles and captions of the posts included references to governmental or academic institutions. To prevent that these references could affect the evaluation of the content, we slightly rephrased some sentences to remove this information. In addition, we corrected also grammatical mistakes in the text that could have given away the reliability of the source.

Results

Participant randomisation was balanced across conditions (Chi squared test, $\chi^2(1) = .409$, $p = .52$); average N per post, per condition was 125, minimum 106, maximum 146. Median time to evaluate the Facebook post was 33 seconds in the control condition, 48 seconds in the incentive condition, 34 seconds in the pop-up condition, and 58 seconds in the incentive + pop-up (minimum overall time: 2.5 seconds, maximum overall time: 22 minutes). When the pop-up was present, participants spent an additional median time of 11 seconds on the pop-up. On a scale from 1 to 6 (3.5 response at chance level), average accuracy score in the control condition was 3.96 ($SD = 1.33$; incentive condition: 4.20, $SD = 1.41$; pop-up condition: 4.07, $SD = 1.33$; incentive + pop-up: 4.29, $SD = 1.44$; Fig 5). In the control condition, 64.6% of participants correctly guessed the scientific validity of the post (incentive condition: 71.2%; pop-up condition: 66.2%; incentive + pop-up: 72.9%). Overall performance was generally lower than in Experiment 1, most likely due to the use of relatively unknown news sources that forces participants not to rely on source knowledge to evaluate content.

Effect of interventions

To test the individual and combined effects of pop-up tips and monetary incentives we conducted two tests, one for each accuracy index. For accuracy scores, We used two ordinal logistic regression models, one with pop-up, monetary incentive as predictors, and another regression including the same variables and the interaction between pop-up and incentive as an additional predictor. For correct guessing, we compared two logistic regressions, one with correct guessing as dependent variable and pop-up, monetary incentive as predictors, and another regression including the same variables and the interaction between pop-up and incentive as an additional predictor. For both the indices, we then adopted the model fitting data best according to a likelihood-ratio test. Perhaps surprisingly, model comparison favoured models without the interaction term (accuracy score: $\chi^2(1) = .032$, $p = .858$; correct guessing: $\chi^2(1) = .007$, $p = .931$); we thus tested the effect of incentives and pop-up assuming that they are (approximately) orthogonal. Results revealed a significant effect of incentive on both accuracy scores ($\beta = .350$ [.194, .505], $z = 5.371$, $p < .001$) and correct guessing ($\beta = .313$ [.124, .501], $z = 3.954$, $p < .001$), and a significant effect of pop-up on accuracy scores ($\beta = .137$

$[-.018, .292]$, $z = 2.115$, $p = .034$)[1], but not on correct guessing ($\beta = .076$ $[-.018, .292]$, $z = 0.966$, $p = .334$). In addition, we found that the combination of the two interventions significantly increased both accuracy indices compared to control (accuracy score: $\beta = .487$ $[.268, .705]$, $z = 5.315$, $p < .001$; correct guessing: $\beta = .389$ $[.123, .654]$, $z = 3.496$, $p < .001$), and that the contribution of incentive was greater than the contribution of pop-up (accuracy score: $\beta = .213$ $[-.007, .432]$, $z = 2.307$, $p = .028$; correct guessing: $\beta = .2362$ $[-.032, .504]$, $z = 2.103$, $p = .047$). According to the ordinal logistic regression model, the combination of the two interventions led to a 10.4% [5.4%,14.2%] increase in correct guessing, and a 6.9% [2.8%,12.4%] increase in "definitely" correct responses compared to control.

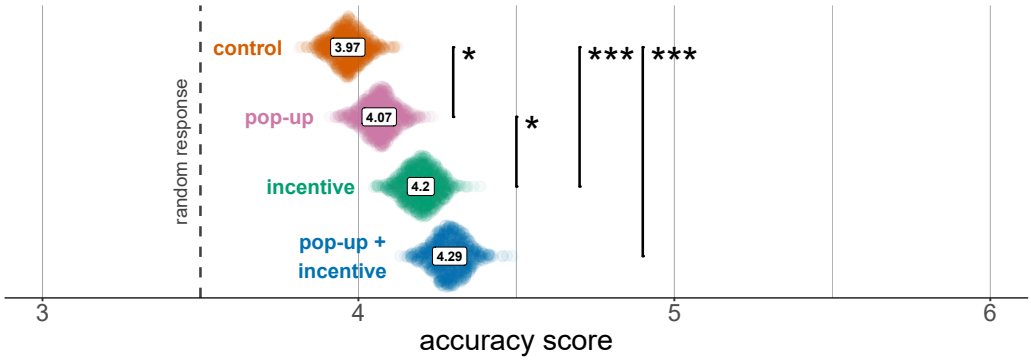

Fig 5. Bootstrap estimates of the average accuracy score by experimental condition (Min. 1, Max. 6, random response: 3.5). Asterisks refer to significance of contrasts in the ordinal logistic regression. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Technique adoption

We tested whether technique adoption was influenced by either interventions following a similar procedure to our test for correct guessing (comparison of two logistic regressions with/without interaction). likelihood-ratio tests again favoured the model without interaction ($\chi^2(1) = .245$, $p = .621$). Model contrasts revealed several significant differences (Fig 6): both incentive ($\beta = .725$ $[.471, .978]$, $z = 6.829$, $p < .001$) and pop-up ($\beta = 1.191$ $[.926, 1.455]$, $z = 10.736$, $p < .001$) increased significantly the use of Civic Online Reasoning techniques, but pop-up effect was significantly stronger than the effect of the incentive ($\beta = .466$ $[.106, .826]$, $z = 3.093$, $p = .002$). In addition, the combined effect of pop-up and incentive was also significant ($\beta = 1.915$ $[1.542, 2.288]$, $z = 12.263$, $p < .001$), leading to an estimated 16.5% [8.6%,26.0%] increase in technique use compared to control.

To test the robustness of these findings, we checked as in Experiment 1 the rate of false reporting (i.e., participants who said they used fact-checking techniques while they did not even leave the study page). False reporting was 22.2% in the control condition, 16% in the pop-up condition, 15.3% in the incentive condition, and 12.8% in the condition with both interventions. The results did not differ after accounting for false reporting (pop-up: $\beta = 1.210$ $[.924, 1.496]$, $z = 10.094$, $p < .001$; incentive: $\beta = .761$ $[.488, 1.033]$, $z = 6.669$, $p < .001$; pop-up¿incentive: $\beta = .449$ $[.061, .838]$, $z = 2.759$, $p = .006$; pop-up + incentive: $\beta = 1.971$ $[1.570, 2.372]$, $z = 11.729$, $p < .001$; see S8 Analyses for an exploration of participants' search behaviour).

[1]Mixed-effects regression with errors clustered by post: $p = .052$.

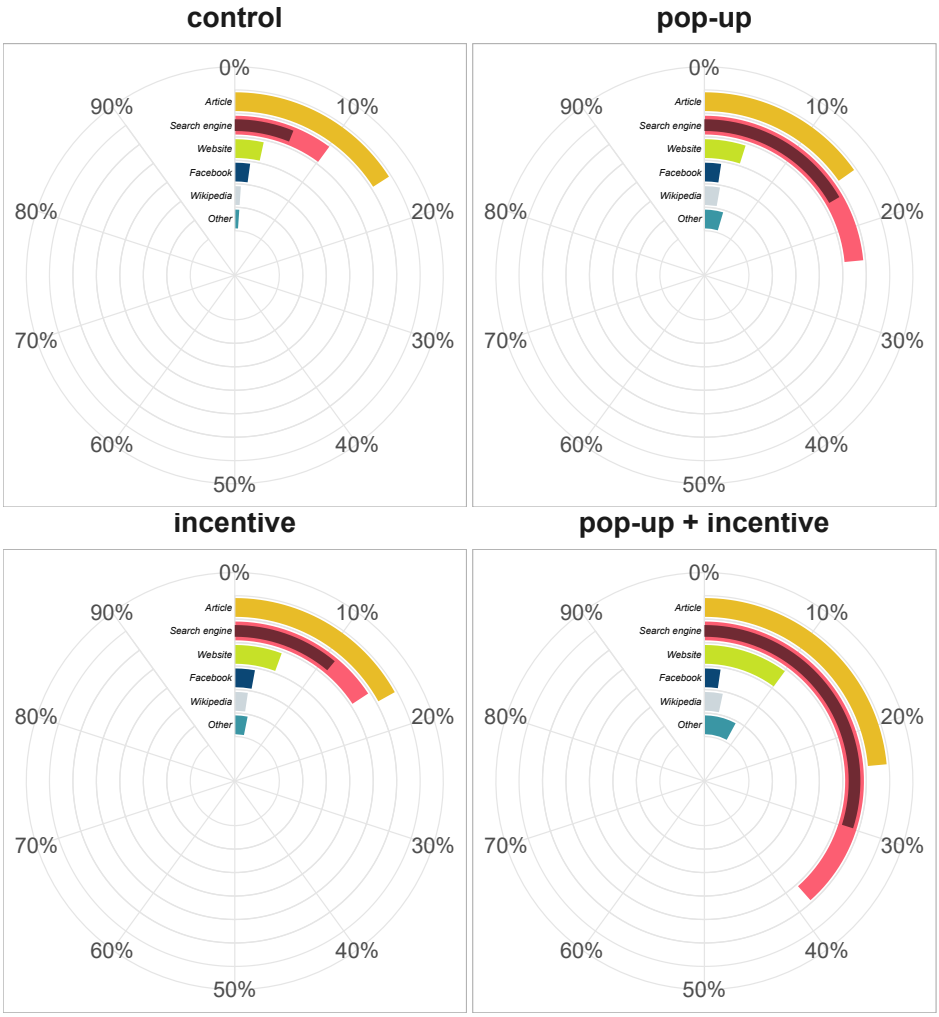

Fig 6. Race chart of self-report external search behaviour. Bars indicate the proportion of participants in each experimental condition reporting to have searched in either category of websites. Lateral reading is identified with the proportion of participants searching information on a search engine (light red), whereas click restraint is the subset of these participants who reported not stopping at the first algorithmically-ranked results of the search (dark red).

To test whether participants who adopted civic online reasoning techniques performed better in the task we run two tests, one for each accuracy index. For accuracy scores, since accuracy scores were non-normally distributed (Shapiro-Wilk test, all $p < .001$) we used a ordinal logistic regression model, with accuracy score as dependent variable and adoption of techniques as a dummy predictor variable. For correct guessing, we used a logistic regression, with correct guessing as dependent variable and adoption of techniques as a dummy predictor variable. According to the models, participants adopting Civic Online Reasoning techniques were more accurate in terms of both accuracy score ($\beta = .591$ [.414, .767], $z = 6.560$, $p < .001$) and correct guessing ($\beta = .506$ [.281, .738], $z = 4.345$, $p < .001$). According to the ordinal regression model, technique adoption increased the probability of giving a "definitely valid" ("definitely invalid") correct response increases by 9.5% [5.9%,13.7%].

We also tested whether the use of Civic Online Reasoning techniques mediated the effect of the interventions with two marginal mediation analyses on accuracy score and

correct guessing. Technique adoption was found to moderate the effect of both incentive and pop-up on accuracy score (incentive: unstandardised $\beta = .007$ [.003, .010], $z = 6.829$, $p < .001$; pop-up: unstandardised $\beta = .011$ [.006, .015], $z = 10.736$, $p < .001$) and correct guessing (incentive: unstandardised $\beta = .008$ [.004, .013], $z = 6.829$, $p < .001$; pop-up: unstandardised $\beta = .014$ [.007, .021], $z = 10.736$, $p < .001$).

Response times.

We compared participants' evaluation time of the post across conditions using linear regressions with rank-transformed time as dependent variable and pop-up and incentives as predictors, with and without interaction. Again, model comparison favoured the model without interaction ($F(1) = 1.104$, $p = .293$). All contrasts were significant: both incentives ($\beta = 370$ [297, 443], $t(2928) = 12.127$, $p < .001$) and pop-up ($\beta = 61$ [−12, 134], $t(2928) = 2.011$, $p = .044$) increased evaluation times, however incentives did so to a greater extent ($\beta = 309$ [205, 413], $t(2928) = 7.105$, $p < .001$). Also, the combination of incentives and pop-up led to higher evaluation times than control ($\beta = 431$ [329, 534], $t(2928) = 10.070$, $p < .001$). We tested whether longer evaluation times were associated with higher accuracy scores by means of an ordinal logistic regression with log-transformed evaluation time as predictor and accuracy score as predicted variable. Results revealed a significant and positive association association ($\beta = .152$ [.081, .223], $z = 4.22$, $p < .001$). The result was confirmed also for correct guessing (logistic regression, $\beta = .204$ [.117, .292], $z = 4.56$, $p < .001$). We also compared the duration of non-click external searches across conditions with the same procedure as total evaluation times, again finding no interaction between interventions ($F(1) = 0.1746$, $p = .676$). Results showed a significant effect of incentive ($\beta = 52$ [15, 90], $t(726) = 3.355$, $p = .001$), pop-up ($\beta = 80$ [43, 116], $t(726) = 5.170$, $p < .001$), and their combination ($\beta = 132$ [80, 184], $t(726) = 6.100$, $p < .001$), but found no significant difference between the interventions ($\beta = 27$ [−26, 80], $t(726) = 1.217$, $p = .224$).

Discussion

Results from Experiment 2 confirmed the effectiveness of monetary incentives on accuracy, and presented evidence in favour of the potential usefulness of fact-checking tips when the post's source is unknown. Monetary incentives increased both accuracy scores and correct guessing, the rate of (self-reported) Civic Online Reasoning techniques, as well as the frequency and duration of non-link searches outside the study page. Participants offered with a monetary incentive spent more time evaluating the post than those who were not. Lastly, incentives seemed to increase the sharing intentions of valid information compared to control (S11 Analyses).

Contrary to the Experiment 1, the pop-up intervention seems to increase accuracy scores, but not correct guessing. We observed that the presence of the pop-up dramatically increased technique adoption (even compared to the presence of incentives) and the rate of non-link external searches, which in turn were linked to an increase in both measures of accuracy. Marginal mediation analyses confirm an indirect effect of pop-up on accuracy measures via an increase of search outside the post page.

In this experiment, we also tested the interaction between incentive and pop-up. Model comparison showed no interaction between the two interventions, suggesting that pop-up and monetary incentives contributed separately to the increase in accuracy. We additionally observe that monetary incentives increased participants' time spent on reading the pop-up: median time is 12.3 seconds with incentive compared to 9.6 when incentive is absent (S7 Analyses). Despite this increase in reading times, our statistical tests do not detect an increased pop-up effects by any other metric.

General Discussion

In this research, we studied whether presenting fact-checking tips and monetary incentives increases the correct evaluation of science-themed Facebook posts. In two experiments, participants rated the scientific validity of the content of one out of several posts, with some participants receiving a monetary reward when they responded correctly and other participants being shown a pop-up window (superimposed on the Facebook post itself) that contained a list of fact-checking techniques proposed in the literature (Civic Online Reasoning). Results showed that monetary incentives work as an accuracy booster. Moreover, data on search times and extremity of validity ratings corroborated the hypothesis that incentives operate by increasing motivation and, subsequently, attention on the content and other features of the post. This effect is particularly remarkable given the strong benchmark against which it was compared: in fact, participants in the control condition were already primed for accuracy [30], and are therefore likely to exert a greater degree of attention than when routinely browsing social media. The effectiveness of the pop-up as a way of introducing participants to fact-checking techniques received support in cases where the source of the post was relatively unknown, i.e. when participants could rely on low prior information to evaluate posts. Furthermore, given that the presence of the pop-up significantly increases the adoption of Civic Online Reasoning techniques, and that the use of these techniques is, in turn, a strong predictor of participants' performance on the task, marginal mediation analyses support the hypothesis that the pop-up may have an indirect positive effect on performance.

One of the original aims of this study was to establish whether incentives and techniques could be compared in their effectiveness in improving the evaluation of scientific content, even when not directly accessible without technical expertise. In this respect, our results suggest that the presence of the pop-up has less impact on subsequent evaluation than monetary incentives. We suspect that the effectiveness of fact-checking advice may be hampered by several factors. A first explanation is that the adoption of the techniques might not have been effective enough to avoid the influence of previous beliefs about the content or of the search style. For example, if participants considered a content to be plausible in the first place, they might have selectively ignored conflicting information even when it was clearly present in the search results (i.e. confirmation bias [61]); similarly, if a participant relied primarily on certain sources of information, consulting these sources might have steered the interpretation in the wrong direction. It is unclear however how such biases might have meaningfully reduced the effectiveness of the pop-up but not of the monetary incentives. A second possibility is that participants failed to follow click restraint recommendations and did not search deep enough to find relevant information and instead relied on unreliable sources favoured by ranking algorithms. Lastly, the reduced impact of the pop-up may derive from its brevity: Civic Online Reasoning techniques have in fact been tested so far after being taught in extensive courses. It is therefore possible that simply presenting a condensed set of tips on the best techniques is not enough to fully understand and master them. This possibility is in line with similar unsuccessful previous interventions presenting news literacy tips [35, 55, 62]. Thus, true ability to recognise pseudo-scientific information might only come from a minimal mastery of critical-thinking skills, which cannot be achieved by simply adding a snippet of information to a post, in the form of a pop-up. Testing whether critical-thinking skills learned in the appropriate context can boost people's capacity to spot pseudo-scientific information is however problematic, due to the subjective nature of critical thinking courses and their instructors in a virtual or physical classroom setting.

Despite the asymmetric contribution of monetary incentives and fact-checking techniques, our results also indicate that the interventions may work in a

complementary way. In particular, Experiment 2 shows that these two interventions do not appear to interact with each other. This result, which was replicated by testing different variables of interest, suggests that the working mechanisms of the interventions are largely orthogonal, and thus can be combined to achieve an even stronger evaluation performance by participants.

Our results on incentives are in line with an attention-based account of information processing on social media; that is, increased deliberation is sufficient to decrease belief in false content [4]. Our results add to the literature of attention-based interventions by showing how monetary incentives can additionally modulate motivation and attention and increase performance.

These promising results were not self-evident, as several experiments have cautioned against the universal effectiveness of monetary incentives as a behavioural driver. [63–65]. In fact, under some circumstances incentives decrease rather than increase motivation [66]. One crucial aspect lies in incentives' calibration, as it has been proven that if the effect of incentives on performance is non-monotonic and too small incentives are often counterproductive [66]. Moreover, when explicit incentives seek to modify behaviour in areas such as education, environmental actions, and the formation of healthy habits, a conflict arises between the direct extrinsic effect of incentives and how these incentives may crowd out intrinsic motivations. Seeking accuracy in judging news is certainly driven by the intrinsic motivations of individuals. In all likelihood, however, these intrinsic motivations do not conflict with monetary incentives. Seeking accuracy, unlike deliberately adopting ecological behaviour or going on a diet, is a largely automatic process.

Another concern was that motivation and attention might not have been sufficient for content that is hardly accessible to non-experts. The effectiveness of incentives is then even more remarkable when considering that participants were asked to evaluate information based on scientific and technical reports, and thus had to rely external knowledge and intuition when claims and data were not immediately available.

Compared to work on Civic Online Reasoning [39], our study finds correlational and causal evidence supporting the importance of lateral reading and click restraint as predictors of accurate information, especially (as initially intended) when the information about the source is scarce. Notably, this is the first reported evidence of a general population intervention in a social media context, extending the evidence for its applicability. We note however that the connection between our intervention (the pop-up) and technique use is only indirect, as participants were free to ignore recommendations. Stronger evidence for the efficacy of Civic Online Reasoning techniques could come from within-subject studies that could limit selectively the use of the techniques to assess their direct impact on users' behaviour.

Our results also partly support literature on media and news literacy [34]. Previous successful attempts at using fact-checking tips relied on presenting participants with some of the Facebook guidelines for evaluating information [36, 37]. Critically, these tips acted by reducing post engagement (liking, commenting, sharing) and perceived accuracy of headlines by hyper-partisan and fake-news sources. Given that our results highlight the effectiveness of fact-checking tips when participants are less familiar with the source, we suspect that the use of such tips is inversely associated to the knowledge and reputation of the source: that is, the more the source is well-known and widely respected, the less participants will rely on guidelines and recommendations. This interpretation goes against previous studies in the literature claiming that source information has little impact on the accuracy judgement of social media content [67–69]. Although we did not directly test for the presence/absence of source information, we did find that familiarity with and trust in a source largely affected the search style and evaluation of the content, suggesting that providing this information to participants had

a meaningful effect on their validity evaluations. One way to reconcile these apparently antithetical conclusions is by considering the relative capability of participants to assess the plausibility of information: source knowledge can be a viable heuristic when information is harder to evaluate. Indeed, we suspect that in our experiment information about the source was often easier to assess than the plausibility of the content itself. In addition, compared to previous experiments, participants could open the original article of the post to confirm that it had actually been produced by the source and not fabricated, a factor that probably increased reliance on the source. These considerations and our findings are not sufficient to ascertain whether and under what circumstances reliance on the source is beneficial or detrimental; however, we argue that source information is important in many situations [70, 71].

Our study does not come without limitations. Possibly the most critical issue is the limited number of stimuli that were used across experiments (15), which did not allow us to properly control for many features that could impact the evaluation of the posts. Even though we cannot exclude confounding variables and biases in the selection of stimuli, we tried as much as possible to follow a standardised procedure with pre-defined criteria in order to exclude stimuli that could be considered problematic. Moreover, even though most of the literature and the present study have focused on standardised stimuli reporting content from news sources, we recognize that scientific (dis)information comes in several formats that also depend on the topic, the audience, and the strategy of the creator. We decided to exclude other types of formats (e.g. videos or screenshots) to try to minimise the differences in experience between users, we think however that future research should explore more in depth the impact of varying media on the impact of disinformation spread and on possible counteracting interventions. Lastly, the study explored the effectiveness of interventions when using a computer, as the very concept of lateral reading is based on browsing horizontally through internet tabs on a computer. Although nothing precludes the use of such techniques on other devices such as a mobile phone or tablet, the user interface is often not optimised to search for different contents at the same time, making their use more cumbersome. This is particularly problematic considering that social media are predominantly accessed through mobile devices. A promising direction in the fight to disinformation will be to study the influence of the device and UI in the ability of users to access high-quality information. Further studies should also investigate how much easiness of accessing information from within a specific app could prompt users to fact-check what they see. For example, many apps allow to check information on the internet via an internal browser without leaving the app itself.

Conclusion

This study set out to assess the relative effectiveness of monetary incentives and fact-checking tips in recognising the scientific validity of social media content. We found strong evidence that incentivising participants increases accuracy evaluations; we also found evidence that fact-checking tips increase accuracy evaluation when the source of the information is unknown. These results suggest a promising role of attention and search strategies, and open the way to the test of multiple approaches in synergy to achieve the most effective results.

Supporting information

S1 File. Facebook posts, Experiment 1. The spreadsheet including the list of Facebook posts for Experiment 1.

S2 File. Facebook posts, Experiment 2. The spreadsheet including the list of Facebook posts for Experiment 2.

S1 Methods. Sample size estimation, Experiment 1. Based on related findings in the literature [36], we expected a small effect size (Cohen's $d \approx .15 - .20$). Assuming no differences across the posts used as stimuli, and hence computing the sample size based on the main effect of a one-way ANOVA with three levels (experimental condition) yielded a minimum sample size of 1269 participants assuming $\alpha = 5\%$ and power $(1 - \beta) = 90\%$. Aside from the main contrast, we expected also to analyse the impact of secondary variables such as the topic of the post or trustworthiness of the source. For this reason, we planned to recruit the maximum number of participants possible given our budget constraints.

S2 Methods. Sample size estimation, Experiment 2. Our target sample size was 3000 participants. We based our sample size estimation on the main effect of pop-up on one of the two accuracy indices, correct guessing (analysis: logistic regression [72]). Estimate of this effect was based on the analyses of the first experiment (8% increase in correct guesses compared to control). To compute this effect size, we filtered observations from the first experiment based on two criteria: the source of the post had to be unknown to most participants, and participants had to have completed the task on a computer. Power $1 - \beta$ was set to 95% and significance α was set to 5%. Results yielded a sample size of $n = 733$ per condition. We thus decided to recruit 750 participants per condition, total $N = 3000$.

We further simulated achieved power for pre-registered hypotheses 3, 4 and 5, for both accuracy indices (correct guessing and accuracy score). Simulations were based on $N = 3000$, $\alpha = 5\%$, and effects sizes estimated from the first experiment.For correct guessing (test: logistic regression), achieved power is 96% for hypothesis 3 (pop-up main effect), and 88% for hypothesis 4 (incentive main effect). Combined effect of pop-up and incentive (hypothesis 5) depends on whether the two interventions interact. Therefore, we simulate different scenarios exploring the effect of interaction on power. Results reveal that to achieve at least 95% power for this contrast, the interaction effect should not be less than -4% (effect: change in the proportion of correct guesses). For accuracy scores (test: ordinal logistic regression), achieved power is 51% for hypothesis 3 (pop-up main effect), and $\approx 100\%$ for hypothesis 4 (incentive main effect). Combined effect of pop-up and incentive (hypothesis 5) depends on whether the two interventions interact. Therefore, we simulate different scenarios exploring the effect of interaction on power. Results reveal that to achieve at least 95% power for this contrast, the interaction effect should not be less than -0.25 (effect: change in log odds).

S3 Methods. Scoring of scientific validity. Sources of scientific information usually comply with standards approved by the community to guarantee that the information provided is obtained using rigorous methods and goes through several quality checks. In order for a content to be considered scientifically valid it had to satisfy the following requirements:

- the original research could be found in a peer-reviewed publication;
- authors of the research had a track record certifying their expertise in their field of competence;
- research was not falsified by concomitant research in the field;
- there was no potential conflict of interest, or alternately the content had been independently evaluated by a source with no conflicts of interest;

- the media article represented accurately data and claims of the original research.

S4 Methods. Source familiarity and trustworthiness. Since we suspected that assessing familiarity and perceived trustworthiness of the source could be affected by the observation of the Facebook post, we ran two separate surveys with independent raters to categorise and select the Facebook posts (first survey: $N = 100$, mean age $M = 26.5$, $SD = 7.8$, 2 not specified; 71 female, 1 not specified; second survey: $N = 100$, mean age $M = 33.2$, $SD = 12.4$; 68 female, 2 not specified). Raters were recruited on the online platform prolific.co and had to assess the familiarity and trustworthiness of several sources using a questionnaire taken from a previous study [20] (Fig 7). To categorise sources based on the raters' responses, we ran an expectation maximisation model-based clustering algorithm using the McLust package in R [73]. Results revealed four clusters, one collecting known, trustworthy sources ($N = 4$; e.g., National Geographic and BBC), one known, untrustworthy sources ($N = 5$; e.g., Daily Mail and Daily Star), one unknown sources ($N = 21$; e.g., Duluth News Tribune and the American Enterprise Institute), and a last one including sources with mixed recognition ($N = 7$, e.g., the Washington Times and Live Science).

S5 Methods. Post-rating questionnaire. After rating the post's scientific validity, the participant completed a questionnaire. Below is the full list of questions asked:

- Confidence in rating: "How confident are you in your response?"; 6-point likert scale from (1) "don't know" to (6) "absolutely certain"
- Sharing intention (Experiment 2): Would you consider sharing this story online (e.g., through social networks or messaging apps)?; Yes/no
- Sharing behaviour (Experiment 2): Approximately how many news articles, memes, opinion pieces, etc. have you shared in the last week?; numeric free-text response
- Source familiarity: "Did you know [name of source] before the experiment?"; Yes/no
- Source trustworthiness: "How much do you trust [name of source]?"; 5-point likert scale from (1) "not at all" to (5) "entirely"
- Content plausibility (Experiment 1): "How plausible do you find the content of the post?"; 6-point likert scale from (1) "totally implausible" to (6) "totally plausible"
- Content plausibility (Experiment 2): "Please respond as if you did not read the Facebook post: does it sound plausible to you that [statement based on the post]?"; 6-point likert scale from (1) "Totally implausible" to (6) "totally plausible"
- External search: "While you were evaluating the Facebook post, did you look for information outside the study page?"; Yes/No
- if No: "Why not?"; randomised: It did not occur to me to do it; I had enough information already; I thought I would lose the experiment; I thought it was not allowed; I thought it was not possible; Other (free text entry).
- if Yes: "Where did you look for information? (select all that apply)"; randomised: The article's web page; Wikipedia; Other web pages from the article's website; Search engine (e.g., Google); Facebook. (lateral reading = search engine is selected)

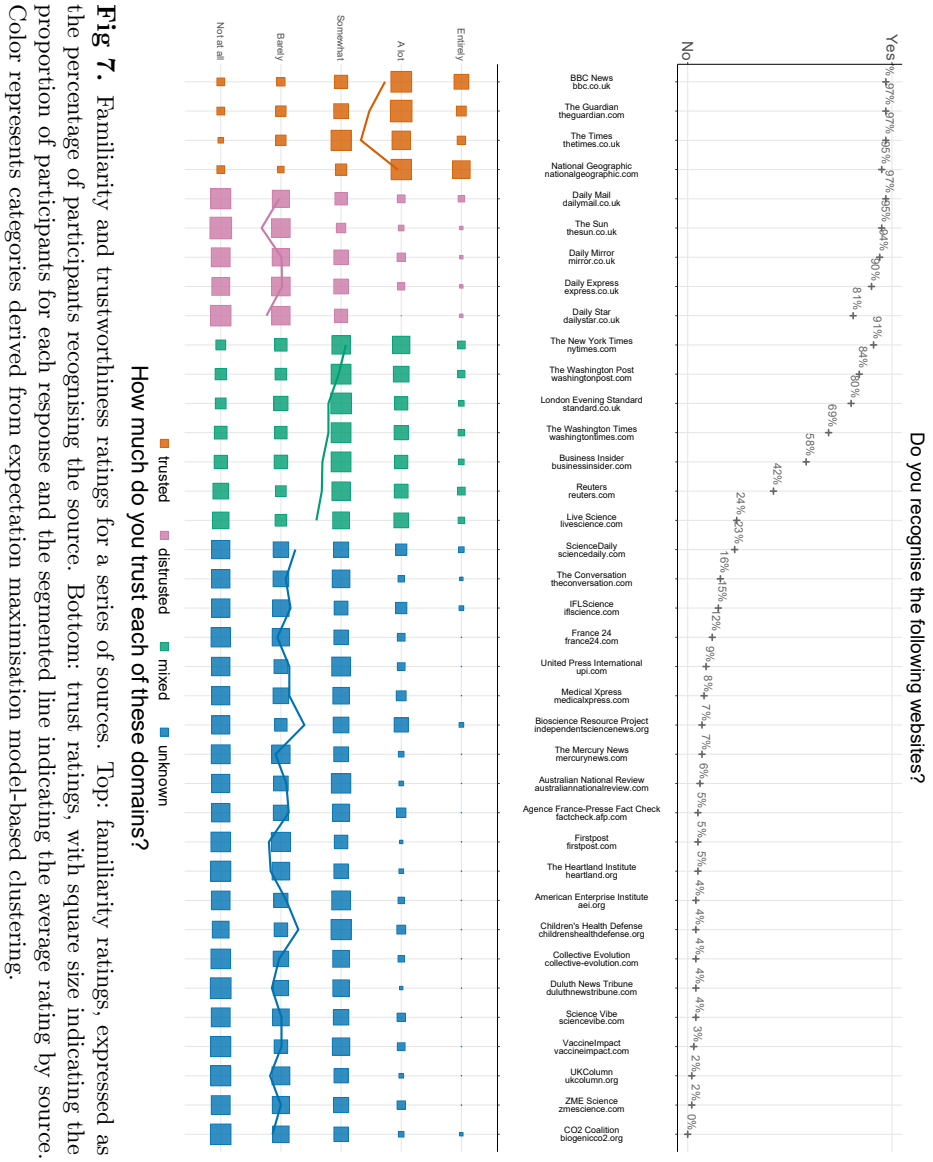

Fig 7. Familiarity and trustworthiness ratings for a series of sources. Top: familiarity ratings, expressed as the percentage of participants recognising the source. Bottom: trust ratings, with square size indicating the proportion of participants for each response and the segmented line indicating the average rating by source. Color represents categories derived from expectation maximisation model-based clustering.

- if Search engine is selected: "While you were looking at the search browser results, what links did you open?"; options: The first search results suggested.; The subsequent search results.; Both the first and subsequent search results.; I did not open any search results. (click restraint = either "The subsequent search results." or "Both the first and subsequent search results" is selected)
- Subjective knowledge of the topic [74] (study 2): "How much do you know about [topic]?"; 6-point likert scale from (1) "nothing at all" to (6) "a great deal"
- Relevance of obtaining accurate information: "We are considering compiling a comprehensive summary of the scientific discussion behind the content of the post. If so, would you be interested in receiving it by private message on your prolific account?"; Yes/No
- Trust in scientists: "In general, how much do you trust scientists to do what is right?"; 6-point likert scale from (1) "not at all" to (6) "A lot" (adapted from the

Edelman Trust Barometer Yearly online survey)

- Conspiracy ideation trait [14]: 4, 5-point likert scales combined into a mean index
- Scientific literacy [74]: 15 true/false questions

S6 Methods. Supplementary measures in Experiment 2. Measures of Experiment 2 were identical to those administered in Experiment 1, with three exceptions.

Scientific validity. In Experiment 1, all the scale points used to measure scientific validity were labelled with an adjective (e.g., 4 corresponded to "possibly valid"). We removed intermediate labels and left only the ones for 1 and 6 ("definitely invalid/valid"). We removed these labels to make sure that adjectives could not influence the evaluation in the conditions with incentives, where the participants were asked to give a response that matched the ratings of the experimenters.

Plausibility. We changed one control measure, plausibility, to reflect more specifically on the content of the post than on its general appearance. We thus singled out on claim from the post and asked participants if it sounded plausible, *disregarding the information they had gathered during the task*. The content of a source should sound plausible to a participant if their background information is in agreement with the content itself, so measuring plausibility in this allows us to make inferences about a participant's background beliefs regarding the post they were given.

Sharing behaviour. As an additional exploratory measure we also asked participants' intention to share the post. This question is widely adopted in the literature (see for instance [29]). We also asked participants to estimate their weekly amount of sharing on social media, since this rate could affect the intention to share.

S1 Analyses. Original pre-registered analyses (Experiment 1). We tested differences in accuracy scores using a linear probabilistic model with accuracy score as predicted variable, and experimental condition as predictor. Contrasts revealed a small but significant impact of incentive on accuracy score ($\beta = .0264$ $[-.003, .055]$, $t(2381) = 2.133$, $p = .0495$[2]), but not of the pop-up ($\beta = -.0005$ $[-.0296, .0285]$, $t(2381) = -.042$, $p = .9667$); we also found that accuracy scores were higher in the incentive condition than in the pop-up condition ($\beta = .0269$ $[-.002, .056]$, $t(2381) = 2.177$, $p = .0495$). To test correct guessing, we used a logistic[3] regression with the guess of participant (i.e., "valid" or "invalid") as dependent variable and actual validity of the post content, experimental condition, and their interaction as predictors. Neither experimental condition nor its interaction with post validity yielded significant results (all $p > .119$), thus we could not reject the null hypothesis that there is no difference in terms of correct guessing between conditions.

We additionally tested whether results differ when excluding participants who either failed attention checks, encountered technical issues with the display of the Facebook post, or who did not close the pop-up (and therefore could not observe the post). Tests were robust to all these exploratory exclusions.

S2 Analyses. Differences in recorded search behaviour (Experiment 1). We tracked participants' search behaviour on the post page as an additional proxy of technique use. Since the page did not include a link to a search engine, we tracked

[2]Mixed-effects regression with errors clustered by post: $p = .052$.

[3]Original preregistered analyses proposed the use of a probit regression instead of a logistic regression. The two regressions yield the same results, but since we adopt ordinal logistic regressions for non-parametric analyses, we choose to report the results of the logistic regression for ease of comparison across tests.

whether participants in each condition did leave the post page without clicking any link. Results confirm that more participants in the incentive and pop-up conditions left the page than participants in the control condition (Fig 8, light red bar; incentive: $\beta = .6754$ $[.3759, .9748]$, $z = 5.282$, $p < .001$; pop-up: $\beta = .5226$ $[.2182, .8270]$, $z = 4.021$, $p < .001$), however the difference between the two interventions was not significant ($\beta = -.1528$ $[-.4258, .1203]$, $z = -1.310$, $p = .190$).

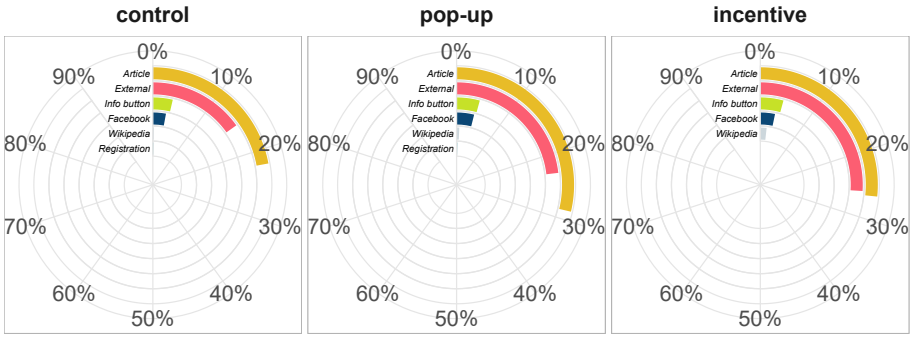

Fig 8. Race chart of recorded search behaviour, Experiment 1. Bars indicate the proportion of participants in each experimental condition that clicked on one of the available links, on Facebook's info button, or left the page without clicking any links (named "external" searches). The links available to participants led to the original article, the source's Facebook page, a site containing the source's domain registration, and a Wikipedia page where one existed.

S3 Analyses. Response extremity and confidence (Experiment 1). We tested whether our itnerventions affected the extremity (i.e. 4 versus 5 versus 6) and confidence of ratings. To measure extremity of responses we looked at the three levels of evaluation regardless of their correctness. We ran an ordered logistic regression with extremity of response as predicted variable and experimental condition as predictor. Participants in the incentive condition gave more extreme responses than participants in the control ($\beta = .4425$ $[.2351, .6499]$, $z = 5.000$, $p < .001$) and pop-up ($\beta = .4847$ $[.27709, .6924]$, $z = 5.471$, $p < .001$) conditions, whereas we found no effect of pop-up over control ($\beta = -.0422$ $[-.2480, .1636]$, $z = -.481$, $p = .692$). We also compared confidence ratings between conditions, but found no statistically significant difference between conditions (ordinal logistic regression, all $p > .05$).

S4 Analyses. Uncorrected contrasts for source reputation (Experiment 1). Given the smaller power for the exploratory analysis on source reputation, we looked at uncorrected contrasts. These tests suggested that, for unknown sources, accuracy scores were higher in the incentive condition than in the pop-up condition ($\beta = .3655$ $[.0843, .8153]$, $z = 2.226$, $p_{\text{uncorr}} = .026$), and that correct guessing was higher in the pop-up condition than in the control condition ($\beta = .45177$ $[-.1323, 1.0358]$, $z = 2.118$, $p_{\text{uncorr}} = .034$); for generally distrusted sources, participants in the incentive condition had higher accuracy scores than control ($\beta = .2807$ $[-.0905, .6519]$, $z = 2.071$, $p = .038$) and than pop-up participants ($\beta = .2788$ $[-.0845, .6420]$, $z = 2.102$, $p = .036$); lastly, for generally trusted sources, correct guessing was lower in the pop-up condition than in the control condition ($\beta = -.8168$ $[-1.7956, .1621]$, $z = -2.285$, $p = .022$). This last counter-intuitive result may suggest that providing Civic Online Reasoning techniques when the source is known might actually backfire. However, this interpretation should be taken with caution, since all trusted sources in the experiment were in fact presenting valid information, and thus we cannot exclude the influence of post validity (see S5 Analyses).

S5 Analyses. Effect of post type on accuracy (Experiment 1). Here we tested for any potential post differences in terms of scientific validity and scientific topic. When testing for differences across valid and invalid posts, likelihood-ratio tests confirmed the importance of this variable for accuracy scores ($chi^2(3) = 92.331$, $p < .001$) but not for correct guessing ($chi^2(3) = 5.479$, $p < .140$); we thus tested only for differences in accuracy scores. Contrasts revealed a significant effect of incentives when posts contained valid information: accuracy scores were higher in the incentive condition than in the control ($\beta = .3582$ [.07329, .6431], $z = 3.268$, $p = .003$) and pop-up conditions ($\beta = .3713$ [.0913, .6514], $z = 3.447$, $p = .003$). Uncorrected contrasts did not reveal any other significant result. A possible interpretation of these findings is that there was a bias in the task favouring the interpretation of the posts' content as scientifically invalid, and that the increase in time and attention produced by the incentives mitigated this bias. We do not however have the data to confirm or dis-confirm this conclusion. We also note that posts from trusted sources were all presenting valid content, and this could play a potential confound.

We then tested for differences between posts by scientific topic. Scientific topic had to have a significant effect on both accuracy scores and correct guessing (likelihood-ratio test, all $p < .001$). Contrasts reveal a significant effect of incentive on accuracy scores for posts about the COVID-19 pandemic (against control: $\beta = .4016$ [−.0426, .8459], $z = 2.476$, $p = .040$; against pop-up: $\beta = .4556$ [.0176, .8936], $z = 2.849$, $p = .020$) and climate change (against pop-up: $\beta = .4505$ [.0175, .8835], $z = 2.850$, $p = .020$). Uncorrected contrasts did not reveal any other significant result.

S6 Analyses. Search behaviour and post evaluation (Experiment 1). As an exploratory analysis, we tested what type of behaviour on the post page predicted higher accuracy scores and correct guessing in the task. We tracked whether participants clicked on the links on the post's web page (Facebook page; original article; Facebook's info button; who.is, a website tracking information about the source domain; source's Wikipedia page, when existing), or if they left the page without clicking any links. We ran an ordinal logistic regression for accuracy score and a logistic regression for correct guessing, with predictors a series dummy variables indicating whether the participant performed each behaviour or not. Results revealed that leaving the page without clicking any link was a significant predictor both for accuracy scores ($\beta = .4500$ [.2649, .6352], $z = 4.760$, $p < .001$) and correct guessing ($\beta = .4273$ [.1552, .7100], $z = 3.022$, $p = .003$). In addition, participants who opened the original article were more likely to correctly guess the validity of the post ($\beta = .4137$ [.1600, .6754], $z = 3.149$, $p = .002$).

As a confirmatory test, we ran an expectation maximisation model-based clustering algorithm to categorise participants based on their tracked behaviour on the page. Specifically, we fed the algorithm with participants' total search time (either reading the info window related to the post or searching outside the page), and the proportion of time for each activity. Cluster analyses revealed four clusters of behaviours, plus a fifth group including participants who never left the study page. Results reveal that, for both accuracy scores and correct guessing, two clusters of participants performed better than those who did not leave the study page: those who predominantly searched without clicking links (accuracy score: $\beta = .5876$ [.3932, .7820], $z = 5.920$, $p < .001$; correct guessing: $\beta = .6338$ [.3493, .9317], $z = 4.272$, $p < .001$), and those who searched predominantly via the link to the article (accuracy score: $\beta = .3130$ [.1203, .5056], $z = 3.180$, $p = .003$; correct guessing: $\beta = .4761$ [.1969, .7671], $z = 3.277$, $p = .002$). We speculate (also based on comments in the post-experimental questionnaire) that participants searching on the original article used this exploration to confirm whether the post content was not fabricated, and thus rely more directly on their opinion of the

source; we do not have results confirming this hypothesis.

S7 Analyses. Additional pre-registered analyses (Experiment 2).

Main effect of pop-up on adoption of techniques. To test whether the presence of the pop-up increases the adoption of civic online reasoning techniques, we used a chi squared test comparing the proportion of participants reporting to adopt the fact checking techniques (lateral reading and click restraint, dichotomous variable) when pop-up was present versus absent. Proportions were indeed significantly different ($\chi^2(1) = 122.66$, $p < .001$), with 23.6% of participants adopting lateral reading and click restraint when the pop-up was present compared to 8.7% when the pop-up was absent.

Effect of incentive on pop-up reading times. To test whether the monetary incentive increases attention towards the pop-up, we used a t-test (or an equivalent non-parametric alternative) to compare the reading times of the pop-up between participants who did or did not receive a monetary incentive. Given that reading times (and their log-transformation) were not normally distributed (Shapiro-Wilk test, all $p < .025$), we adopted a Wilcoxon rank-sum test. The test was significant ($\log(W) = 5.32$, $p < .001$), with median reading times being 2.1 [1.5,2.8] seconds longer when the incentive was present.

S8 Analyses. Differences in recorded search behaviour (Experiment 2). We checked how many participants in each condition left the post page without clicking any link, a proxy of technique use. Likelihood-ratio test again suggested no interaction between incentive and pop-up ($\chi^2(1) = .678$, $p = .410$). Results confirmed the significant effect of both incentive ($\beta = .5239$ [.3159, .7320], $z = 6.010$, $p < .001$) and pop-up ($\beta = .3901$ [.1841, .5961], $z = 4.519$, $p < .001$), but did not find any significant difference in strength between the two interventions ($\beta = .1339$ [−.1579, .4256], $z = 1.095$, $p = .274$; Fig 9).

S9 Analyses. Exclusion criteria (Experiment 2). We tested whether results differed when excluding participants who reported being familiar with the source, or who were not regular Facebook users. Pre-registered results did not differ with one exception: when controlling for source familiarity, the contrast comparing the strength of intervention between incentive and pop-up was no more significant (accuracy score: $\beta = .1645$ [−.0625, .3916], $z = 1.730$, $p = .084$; correct guessing: $\beta = .2124$ [−.0647, .4895], $z = 1.830$, $p = .090$).

S10 Analyses. Response extremity and confidence (Experiment 2). We measured differences in extremity of responses and confidence ratings across conditions as in Experiment 1. Both analyses favoured the model without interaction (likelihood-ratio tests, all $p > .05$). Contrasts for response extremity revealed that incentives increased the ratio of extreme answers ($\beta = .4963$ [.3328, .6597], $z = 7.245$, $p < .001$) whereas pop-up did not ($\beta = .1191$ [−.0434, .2815], $z = 1.749$, $p = .080$). Contrasts for confidence ratings revealed that only when incentive and pop-up were combined confidence ratings were significantly higher than control ($\beta = .2830$ [.0617, .5043], $z = 3.053$, $p = .009$).

S11 Analyses. Sharing behaviour (Experiment 2). In Experiment 2, after the rating of the post, we asked participants about their willingness to share it. We tested the effect of incentives and pop-up on sharing behaviour. We ran two logistic regressions, one with sharing intention as predicted variable, and incentive, pop-up, scientific validity, the interaction between incentive and scientific validity, and the interaction between pop-up and scientific validity as predictors, and a second regression

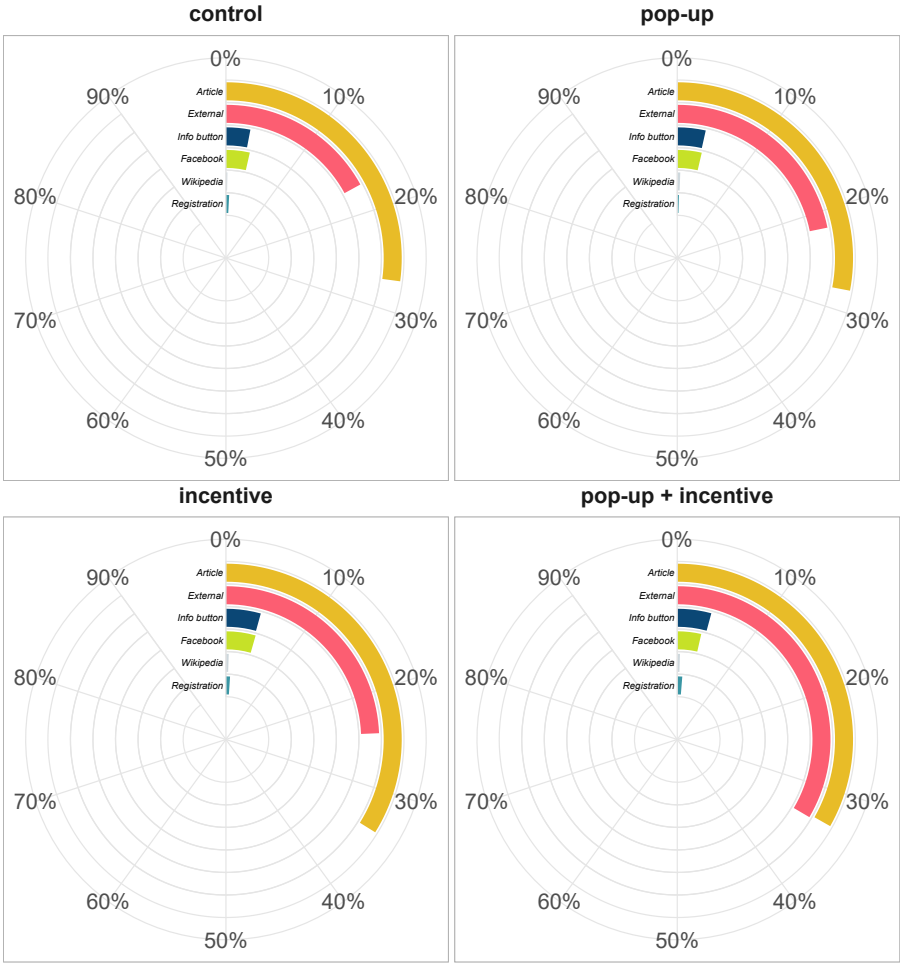

Fig 9. Race chart of recorded search behaviour, Experiment 2.

identical to the first one with the additional interaction between incentive and pop-up. Both regressions included also a variable controlling for the self-report number of weekly shares of posts on social media. Given that one participant reported sharing an implausibly large number of posts (50000; S6 Methods) we excluded this participant from this analysis. Comparison of the two models favoured the model without interaction between the two interventions ($\chi^2(1) = .072$, $p = .788$). Analyses revealed only an increase in sharing intention when the post was valid and participants received a monetary incentive ($\beta = .7311$ $[.3856, 1.0765]$, $z = 5.392$, $p < .001$). One possible interpretation of this increase is that the task (assessing the scientific validity of a post) increases scepticism towards the content of the post, and that incentives counteract this scepticism by prompting people to investigate further.

Uncorrected contrasts also suggested that such increase was significantly stronger than any potential increase due to the pop-up ($\beta = .4821$ $[-.0582, 1.0224]$, $z = 2.273$, $p_{\text{uncorr}} = .023$). Moreover, pop-up appeared to slightly reduce the number of shared when the content was not valid, both compared to control ($\beta = -.3606$ $[-.7862, .0650]$, $z = -2.159$, $p_{\text{uncorr}} = .031$) and to the effect of incentive ($\beta = -.5416$ $[-1.1933, .1100]$, $z = -2.117$, $p_{\text{uncorr}} = .034$). Whereas these results indicate an influence of our interventions towards sharing behaviour, it is important to keep in mind that despite the ecological validity of this task, participants in all conditions were asked to evaluate

the validity of the content which they were seeing, which could in turn influence any subsequent sharing intention [30, 32, 75].

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870883. The information and opinions are those of the authors and do not necessarily reflect the opinion of the European Commission. We would like to thank Torbjørn Gundersen, Philipp Lorenz-Spreen, David J. Grüning, and the members of the Prosocial Design Network for their insightful comments and advice.

References

1. Roose K, Isaac M, Frenkel S. Facebook Struggles to Balance Civility and Growth; 2020. https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html.
2. AVAAZ. Facebook's Algorithm: A Major Threat to Public Health; 2020. https://secure.avaaz.org/campaign/en/facebook_threat_health/.
3. Eysenbach G, et al. How to fight an infodemic: the four pillars of infodemic management. Journal of medical Internet research. 2020;22(6):e21820.
4. Pennycook G, Rand DG. The psychology of fake news. Trends in cognitive sciences. 2021;.
5. Lewandowsky S, Van Der Linden S. Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology. 2021; p. 1–38.
6. Kozyreva A, Lewandowsky S, Hertwig R. Citizens versus the internet: Confronting digital challenges with cognitive tools. Psychological Science in the Public Interest. 2020;21(3):103–156.
7. Lorenz-Spreen P, Lewandowsky S, Sunstein CR, Hertwig R. How behavioural sciences can promote truth, autonomy and democratic discourse online. Nature human behaviour. 2020; p. 1–8.
8. Lewandowsky S, Ecker UK, Cook J. Beyond misinformation: Understanding and coping with the "post-truth" era. Journal of applied research in memory and cognition. 2017;6(4):353–369.
9. Lewandowsky S, Ecker UK, Seifert CM, Schwarz N, Cook J. Misinformation and its correction: Continued influence and successful debiasing. Psychological science in the public interest. 2012;13(3):106–131.
10. Lewandowsky S, Cook J, Ecker U, Albarracín D, Amazeen MA, Kendeou P, et al.. The Debunking Handbook 2020; 2020.
11. Cook J, Lewandowsky S, Ecker UK. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. PloS one. 2017;12(5):e0175799.
12. Hertwig R, Grüne-Yanoff T. Nudging and boosting: Steering or empowering good decisions. Perspectives on Psychological Science. 2017;12(6):973–986.

13. Walter N, Brooks JJ, Saucier CJ, Suresh S. Evaluating the impact of attempts to correct health misinformation on social media: a meta-analysis. Health Communication. 2020; p. 1–9.

14. Bode L, Vraga EK. See something, say something: Correction of global health misinformation on social media. Health communication. 2018;33(9):1131–1140.

15. Bode L, Vraga EK. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. Journal of Communication. 2015;65(4):619–638.

16. Colliander J. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. Computers in Human Behavior. 2019;97:202–215.

17. Brashier NM, Pennycook G, Berinsky AJ, Rand DG. Timing matters when correcting fake news. Proceedings of the National Academy of Sciences. 2021;118(5).

18. Allen J, Arechar AA, Pennycook G, Rand DG. Scaling up fact-checking using the wisdom of crowds. Science Advances. 2021;7:1–10.

19. Allen J, Arechar AA, Rand DG, Pennycook G. Crowdsourced Fact-Checking: A Scalable Way to Fight Misinformation on Social Media; 2020.

20. Pennycook G, Rand DG. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences. 2019;116(7):2521–2526.

21. McGuire WJ. Inducing resistance to persuasion. Some Contemporary Approaches. In: Berkowitz L, editor. Advances in Experimental Social Psychology. vol. 1. Academic Press; 1964. p. 191–229. Available from: `https://www.sciencedirect.com/science/article/pii/S0065260108600520`.

22. Roozenbeek J, Van Der Linden S. The fake news game: actively inoculating against the risk of misinformation. Journal of Risk Research. 2019;22(5):570–580.

23. Roozenbeek J, van der Linden S, Nygren T. Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. Harvard Kennedy School Misinformation Review. 2020;1(2).

24. Roozenbeek J, van der Linden S. Breaking Harmony Square: A game that "inoculates" against political misinformation. The Harvard Kennedy School Misinformation Review. 2020;.

25. Cook J. Cranky Uncle Vs. Climate Change: How to Understand and Respond to Climate Science Deniers; 2020.

26. Clayton K, Blair S, Busam JA, Forstner S, Glance J, Green G, et al. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Political Behavior. 2019; p. 1–23.

27. Mena P. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. Policy & internet. 2020;12(2):165–183.

28. Gaozhao D. Flagging Fake News on Social Media: An Experimental Study of Media Consumers' Identification of Fake News. Available at SSRN 3669375. 2020;.

29. Pennycook G, Bear A, Collins ET, Rand DG. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. Management Science. 2020;.

30. Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG. Shifting attention to accuracy can reduce misinformation online. Nature. 2021; p. 1–6.

31. Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand D. Understanding and reducing the spread of misinformation online. Unpublished manuscript: https://psyarxiv com/3n9u8. 2019;.

32. Pennycook G, McPhetres J, Zhang Y, Lu JG, Rand DG. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological science. 2020;31(7):770–780.

33. Roozenbeek J, Freeman AL, van der Linden S. How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al.(2020). Psychological science. 2021; p. 09567976211024535.

34. Tully M, Maksl A, Ashley S, Vraga EK, Craft S. Defining and conceptualizing news literacy. Journalism. 2021; p. 14648849211005888.

35. Vraga EK, Bode L, Tully M. Creating news literacy messages to enhance expert corrections of misinformation on Twitter. Communication Research. 2020; p. 0093650219898094.

36. Guess AM, Lerner M, Lyons B, Montgomery JM, Nyhan B, Reifler J, et al. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. Proceedings of the National Academy of Sciences. 2020;117(27):15536–15545.

37. Lutzke L, Drummond C, Slovic P, Árvai J. Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. Global Environmental Change. 2019;58:101964.

38. Jones-Jang SM, Mortensen T, Liu J. Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. American Behavioral Scientist. 2019; p. 0002764219869406.

39. Wineburg S, McGrew S. Lateral reading: Reading less and learning more when evaluating digital information; 2017.

40. Breakstone J, Smith M, Wineburg S, Rapaport A, Carle J, Garland M, et al. Students' civic online reasoning: A national portrait. Educational Researcher. 2019; p. 0013189X211017495.

41. McGrew S, Ortega T, Breakstone J, Wineburg S. The Challenge That's Bigger than Fake News: Civic Reasoning in a Social Media Environment. American educator. 2017;41(3):4.

42. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, et al. The spreading of misinformation online. Proceedings of the National Academy of Sciences. 2016;113(3):554–559.

43. Martini C. Ad Hominem Arguments, Rhetoric, and Science Communication. Studies in Logic, Grammar and Rhetoric. 2018;55(1).

44. McGrew S, Breakstone J, Ortega T, Smith M, Wineburg S. Can students evaluate online sources? Learning from assessments of civic online reasoning. Theory & Research in Social Education. 2018;46(2):165–193.

45. McGrew S, Smith M, Breakstone J, Ortega T, Wineburg S. Improving university students' web savvy: An intervention study. British Journal of Educational Psychology. 2019;89(3):485–500.

46. McGrew S, Byrne VL. Who Is behind this? Preparing high school students to evaluate online content. Journal of Research on Technology in Education. 2020; p. 1–19.

47. Moore RC, Hancock JT. The Effects of Online Disinformation Detection Training for Older Adults; 2020.

48. Kaufmann N, Schulze T, Veit D. More than fun and money: Worker motivation in crowdsourcing-a study on Mechanical Turk. Working paper. 2011;.

49. Gneezy U, Meier S, Rey-Biel P. When and why incentives (don't) work to modify behavior. Journal of economic perspectives. 2011;25(4):191–210.

50. Rickard JA, Russell AM. Interest in Advance and Other Up-front Incentives. Graduate School of Management, University of Melbourne; 1986.

51. Pennycook G, Rand DG. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition. 2019;188:39–50.

52. Epstein Z, Berinsky AJ, Cole R, Gully A, Pennycook G, Rand DG. Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. Harvard Kennedy School Misinformation Review. 2021;.

53. Jahanbakhsh F, Zhang AX, Berinsky AJ, Pennycook G, Rand DG, Karger DR. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW1):1–42.

54. Crawford E. Introducing Tip Jar; 2021. `https://blog.twitter.com/en_us/topics/product/2021/introducing-tip-jar`.

55. Tully M, Vraga EK, Bode L. Designing and testing news literacy messages for social media. Mass Communication and Society. 2020;23(1):22–46.

56. Vraga EK, Tully M. Media literacy messages and hostile media perceptions: Processing of nonpartisan versus partisan political information. Mass Communication and Society. 2015;18(4):422–448.

57. Henninger F, Shevchenko Y, Mertens U, Kieslich PJ, Hilbig BE. lab. js: A free, open, online study builder. PsyArXiv. 2019;.

58. Taylor AB, West SG, Aiken LS. Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. Educational and psychological measurement. 2006;66(2):228–239.

59. R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: `https://www.R-project.org/`.

60. Barrett TS. MarginalMediation: Marginal Mediation; 2019. Available from: `https://CRAN.R-project.org/package=MarginalMediation`.

61. Nickerson RS. Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology. 1998;2(2):175–220.

62. Vraga E, Tully M, Bode L. Assessing the relative merits of news literacy and corrections in responding to misinformation on Twitter. New Media & Society. 2021; p. 1461444821998691.

63. Frey BS, Oberholzer-Gee F. The cost of price incentives: An empirical analysis of motivation crowding-out. The American economic review. 1997;87(4):746–755.

64. Fryer Jr RG. Financial incentives and student achievement: Evidence from randomized trials. The Quarterly Journal of Economics. 2011;126(4):1755–1798.

65. Chao M. Demotivating incentives and motivation crowding out in charitable giving. Proceedings of the National Academy of Sciences. 2017;114(28):7301–7306.

66. Gneezy U, Rustichini A. Pay enough or don't pay at all. The Quarterly journal of economics. 2000;115(3):791–810.

67. Dias N, Pennycook G, Rand DG. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. Harvard Kennedy School Misinformation Review. 2020;1(1).

68. Pennycook G, Rand DG. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of personality. 2020;88(2):185–200.

69. Tsang SJ. Motivated fake news perception: The impact of news sources and policy support on audiences' assessment of news fakeness. Journalism & Mass Communication Quarterly. 2020; p. 1077699020952129.

70. Kim A, Moravec PL, Dennis AR. Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. Journal of Management Information Systems. 2019;36(3):931–968.

71. Nadarevic L, Reber R, Helmecke AJ, Köse D. Perceived truth of statements and simulated social media postings: an experimental investigation of source credibility, repeated exposure, and presentation format. Cognitive Research: Principles and Implications. 2020;5(1):1–16.

72. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Statistics in medicine. 1998;17(14):1623–1634.

73. Fraley C, Raftery AE. MCLUST version 3: an R package for normal mixture modeling and model-based clustering. Washington Univ. Seattle Dept. of Satistics; 2006.

74. Fernbach PM, Light N, Scott SE, Inbar Y, Rozin P. Extreme opponents of genetically modified foods know the least but think they know the most. Nature Human Behaviour. 2019;3(3):251–256.

75. Pennycook G, Binnendyk J, Newton C, Rand D. A practical guide to doing behavioural research on fake news and misinformation; 2020.