Est. 1841 | YORK ST JOHN UNIVERSITY

Eggleston, Adam, Cook, Richard and Over, Harriet (2022) The influence of fake news on face-trait learning. PLOS ONE, 17 (12). e0278671.

Downloaded from: https://ray.yorksj.ac.uk/id/eprint/10450/

The version presented here may differ from the published version or version of record. If you intend to cite from the work you are advised to consult the publisher's version: https://doi.org/10.1371/journal.pone.0278671

Research at York St John (RaY) is an institutional repository. It supports the principles of open access by making the research outputs of the University available in digital form. Copyright of the items stored in RaY reside with the authors and/or other copyright owners. Users may access full text items free of charge, and may download a copy for private study or non-commercial research. For further reuse terms, see licence terms governing individual outputs. Institutional Repositories Policy Statement

RaY

Research at the University of York St John

For more information please contact RaY at ray@yorksj.ac.uk

PLOS ONE

RESEARCH ARTICLE

The influence of fake news on face-trait learning

Adam Eggleston[1]*, **Richard Cook**[1,2], **Harriet Over**[1]

1 Department of Psychology, University of York, York, United Kingdom, **2** The School of Psychology, University of Leeds, Leeds, United Kingdom

* adam.eggleston@york.ac.uk

Check for updates

OPEN ACCESS

Citation: Eggleston A, Cook R, Over H (2022) The influence of fake news on face-trait learning. PLoS ONE 17(12): e0278671. https://doi.org/10.1371/journal.pone.0278671

Editor: Pierluigi Vellucci, Roma Tre University: Universita degli Studi Roma Tre, ITALY

Received: November 15, 2021

Accepted: November 21, 2022

Published: December 21, 2022

Copyright: © 2022 Eggleston et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data and supplementary files are available at https://osf.io/t7c25/?view_only=b9a45cb252fa468eb897b637b6178c1a.

Funding: This research was supported by the European Research Council under the European Union's Horizon 2020 Programme, grant number ERC-STG- 755719 awarded to HO and grant number ERC-2016-StG-715824 awarded to RC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Humans spontaneously attribute a wide range of traits to conspecifics based on their facial appearance. Unsurprisingly, previous findings indicate that this 'person evaluation' is affected by information provided about the target's past actions and behaviours. Strikingly, many news items shared on social media sites (e.g., Twitter) describe the actions of individuals who are often shown in accompanying images. This kind of material closely resembles that encountered by participants in previous studies of face-trait learning. We therefore sought to determine whether Twitter posts that pair facial images with favourable and unfavourable biographical information also modulate subsequent trait evaluation of the people depicted. We also assessed whether the effects of this information-valence manipulation were attenuated by the presence of the "disputed tag", introduced by Twitter as a means to combat the influence of fake-news. Across two preregistered experiments, we found that fictional tweets that paired facial images with details of the person's positive or negative actions affected the extent to which readers subsequently judged the faces depicted to be trustworthy. When the rating phase followed immediately after the study phase, the presence of the disputed tag attenuated the effect of the behavioural information (Experiment 1: N = 128; M_{age} = 34.06; 89 female, 36 male, 3 non-binary; 116 White British). However, when the rating phase was conducted after a 10-minute delay, the presence of the disputed tag had no significant effect (Experiment 2: N = 128; M_{age} = 29.12; 78 female, 44 male, 4 non-binary, 2 prefer not to say; 110 White British). Our findings suggest that disputed tags may have relatively little impact on the long-term face-trait learning that occurs via social media. As such, fake news stories may have considerable potential to shape users' person evaluation.

Introduction

Adults spontaneously attribute a wide range of traits to others based on their facial appearance. For example, making judgments about their apparent trustworthiness, honesty, competence, intelligence, and likeability [1, 2]. These first impressions are thought to load on at least two principal dimensions–often referred to as trustworthiness or valence and dominance [1, 3] and are formed quickly, sometimes within 100 milliseconds of meeting another person [4–6].

Competing interests: All authors have declared that no competing interests exist.

Although individuals within a culture often form similar first impressions, judgments from appearance cues alone have little or no basis in reality, rarely reflecting the actual traits of the individuals being judged [7–10]. Nevertheless, these judgements exert a powerful influence over behaviour and decisions: Individuals judged competent are more likely to be elected to public office [2, 4], while individuals judged trustworthy are more likely to receive lenient sentences in criminal justice situations [11].

Unsurprisingly, this 'person evaluation' is influenced by information about the target's past actions and behaviours [12–15]. Indeed, even a single behavioural statement is sufficient to influence perceptions of another person's trustworthiness. For example, faces that have been paired with a positive behaviour (e.g., "Gave his balloon to a child who had let hers go") are judged more trustworthy than those that have been paired with a negative behaviour (e.g., "Stole money and jewellery from the relatives he was living with") [13]. Interestingly, in line with the Associative-propositional evaluation model, learning of this sort generalises to novel targets with a similar appearance [16–20]. Attempts to modify person perception have yielded mixed results. Learned associations between appearance and character can be difficult to override as new experiences do not necessarily cancel out old associations [21–25]. Due to the ease with which face-trait learning takes place, the potential for it to transfer to novel targets, and its seeming resistance to counter training, it is crucial to understand the real-world implications of face-trait learning.

Social media sites may be one source of face-trait learning. Sites such as Twitter are becoming increasingly popular sources of news and current affairs [26, 27]. The success of these platforms is in part attributable to the ease with which users can share news items that affect themselves, their friends, and family. Depending on the platform and the size of the user's network, shared items can be viewed by thousands of others in a matter of minutes. Many of the news items shared on social media sites describe the actions and behaviours of individuals. Frequently, these news stories are accompanied by a picture of the person's face. This format closely approximates that employed in lab-based studies of face-trait learning [12–15]. Given the results of these studies, it is likely that social media posts pairing facial images and behavioural information will affect the evaluation of the individuals depicted by users who encounter this content.

Given the reach of social media and the speed with which items can be shared across a network, posts that pair faces and trait relevant biographical information potentially exert a powerful influence on the public's perception of the individuals depicted. Alarmingly, however, much of the information shared online is misleading or even inaccurate [28]. Indeed, fake news appears to spread faster and further online than stories verified to be true [29]. There is a clear danger, therefore, that people will be unfairly evaluated because of misleading information shared on social media. Moreover, this possibility could be exploited as an instrument of propaganda, used to tarnish the perception of a political rival or improve perception of a favoured candidate [30].

The proliferation of fake news is a concern for several reasons. For example, fake news is thought to hinder the success of public health programmes [31, 32] and undermine democratic elections [33–35]. In response, social media platforms are attempting to mitigate the effects of fake news via several means including the prioritisation of news from sources judged trustworthy [36], the incorporation of web plug-ins that quickly highlight sites known to spread fake news [37], and the removal of unreliable accounts [36, 38]. Similarly–and of particular relevance to the present investigation–sites including Twitter have started adding a "disputed tag" to flag news stories that may be inaccurate or unreliable.

Several studies have shown that participants are able to rationally incorporate information about the credibility of sources when deciding what to believe [39, 40]. For example, the

presence of a disputed tag significantly decreases participants' perception of an article's believability and accuracy [41–43] even up to a week later [44]. However, recent research suggests that repeated exposure to a fake news story increases its believability even when a disputed tag is present [45, 46]. Furthermore, work by Baum and colleagues [14, 47] suggests that trait-relevant information influences person evaluation irrespective of source credibility. For example, when asked to evaluate faces that had been paired with favourable or unfavourable news headlines, participants' judgements were based on the valence of the headline even when the source was distrusted [14]. Similalry, faces paired with negative biographical details (e.g., "he bullied his apprentice") were judged untrustworthy even when the information was qualified by the addition of "allegedly" [47].

The present investigation had two aims: First, we sought to determine whether Twitter posts that pair facial images with favourable and unfavourable biographical information modulate person evaluation in a manner consistent with previous empirical findings [12–15]. In line with these previous findings, we predicted that targets displayed with a negative headline would subsequently be judged as less trustworthy than targets displayed with a positive headline. Our second aim was to establish whether the effects of the information-valence manipulation were attenuated by the presence of the "disputed tag". Some previous research has shown that participants are able to utilise information on source credibility in their decision making [39, 40], given the ease with which face-trait learning occurs [4] and previous work on the robustness of this learning [14, 47], it is also possible that a target's trustworthiness will be influenced by a tweet's valence even when the information within it is marked as disputed.

Experiment 1

In Experiment 1, we presented participants with displays pairing target faces with positively or negatively valanced headlines that either appeared with or without a disputed tag. During this study phase, we asked participants how believable they found each headline. This served as a manipulation check to ensure that participants were attending to and encoding the disputed tag. In the subsequent test phase, we presented the target faces in isolation (i.e., in the absence of the tweet context and the news headline) and asked participants to judge their perceived trustworthiness. After the crucial test phase, we also measured whether participants could recall whether the headline associated with each face was positive or negative, disputed or non-disputed.

Method

Participants. Protocols were approved by the University of York's Psychology Ethics Committee, participants gave informed written consent before taking part. One-hundred-and-twenty-eight participants completed the experiment based on a power analysis using MorePower 6.0.4 that found a minimum *N* of 126 would be necessary to detect interactions with a medium effect size (partial eta squared .06) with an alpha of .05 and power of .8 (M_{age} = 34.06, SD_{age} = 12.93; 89 female, 36 male, 3 non-binary). All participants were recruited via the online platform Prolific (www.prolific.co). A further 42 were tested, but replaced having reported that they did not notice the disputed tag. Of the 128 participants in the final sample, all reported English as their first language and 126 resided in the UK (2 selected 'prefer not to say'). Of these 128 participants, 116 identified as White British, 2 as "Other" White background (not-specified), 4 as Indian, 2 as Chinese, 1 as Irish, 1 as Caribbean, and 1 as Pakistani. All participants received a small honorarium (£2.50) for their participation.

Materials. In a pre-test, we asked 30 participants, recruited via Prolific, to rate 60 headlines (30 positive, 30 negative) on positivity and on 3 items related to credibility (authenticity,

believability, accuracy). All scales ran from -50 to +50. Based on these scores, the 12 most credible headlines of each valence were chosen for use in the current experiment. All positive headlines had a positivity rating above +25, while all negative headlines had a positivity rating of less than -25.

The target stimuli used in this experiment were photographs of AI generated faces taken from the openly available Academic Dataset by Generated Photos (https://generated.photos/datasets). This database was chosen as it provided a large number of high quality and constrained images of forward-facing faces featuring a range of ethnicities, genders and ages. A subset of 100 (50 male, 50 female) adult faces were chosen from the database based on their neutral expression and forward-facing head position.

In a separate pre-test, we recruited two groups of 60 participants to rate the 100 faces on measures of trustworthiness and attractiveness (both scales: -50 to +50) as well as age (scale: under-18 to over-60). Thirty participants rated the male faces, 30 participants rated the female faces. Based on the mean ratings for each measure we created four sets of 6 faces (3 male, 3 female) with each set closely matched on all three measures.

Combinations of facial images and headlines were transformed into tweets using an online fake tweet generator (www.tweetgen.com) with or without a disputed tag (Fig 1). A full list of the headlines, stimuli and all ratings can be found in "Supplementary materials" at the Open Science Framework: (https://osf.io/t7c25/?view_only=8e6ca4f1249c4579a0e49e1ce5f4d747).

Design and counterbalancing. During the study phase, participants observed fictious news tweets that were positive and negative disputed and not disputed in a 2 × 2 factorial design. In total, participants viewed 24 tweets, six for each of the four factorial combinations. Each tweet was presented twice across two blocks, once in block-1 and once in block-2, yielding 48 study trials. The order in which all tweets were presented within each block was randomised. The allocation of face set (1, 2, 3, or 4) to factorial combination (positive disputed, positive non-disputed, negative disputed, negative non-disputed) was perfectly counterbalanced across the sample.

During the test phase, all participants completed the same rating procedure, in which the 24 faces were presented in a randomised order. Each face was rated once, yielding 24 test trials. In the subsequent recall phase, the order of two final questions used to assess participants' memory for headline valence and disputed status was counterbalanced.

Procedure. At the start of the study phase, participants were informed that they would be shown "a series of tweeted news stories" and that "each tweet contains a picture of the person referred to in the headline." They were asked to rate how believable they considered the tweet to be using a sliding scale that appeared below the statement: "I think the tweet I just read is believable". The scale ranged from -50 ("Describes the tweet very poorly") to +50 ("Describes the tweet very well"). Participants were unable to see the numerical value they selected on the slider. Each tweet was visible for 3 secs before the question and rating scale appeared. The tweet remained on screen until participants had entered their rating and clicked to proceed to the next trial.

During the test phase, target faces (without headlines or tweet context) appeared one at a time in the centre of the screen. Underneath the facial image was a question prompt: "How trustworthy do you think this person is?" Participants recorded their judgements using a scale that ranged from -50 ("Not at all Trustworthy") to +50 ("Extremely Trustworthy"). Participants were unable to see the numerical value they selected on the slider. Faces were visible until a response was recorded.

In the penultimate recall phase, participants again viewed each of the target faces, one at a time. For each face, they were asked two questions in a counterbalanced order: The first was whether or not the face was paired with a positive or a negative headline. Participants

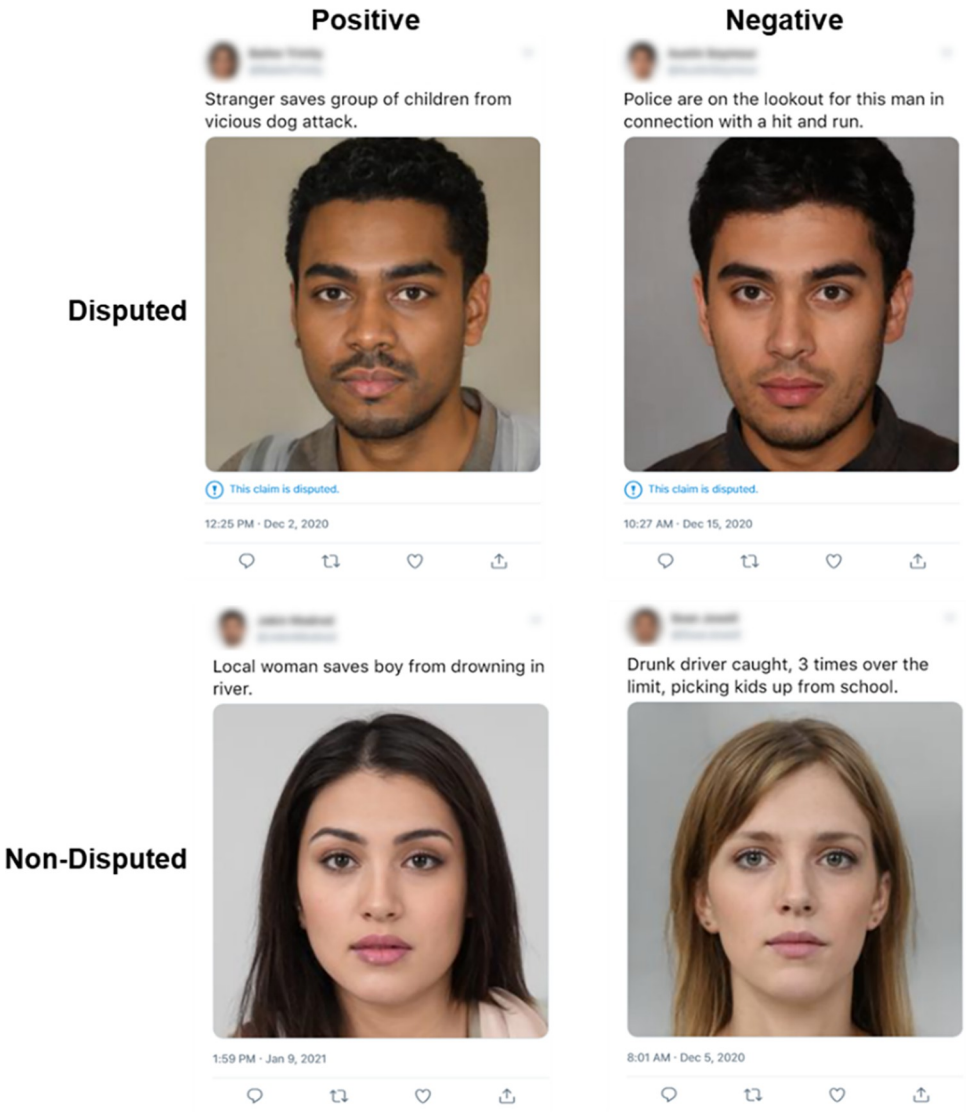

Fig 1. Example tweet stimuli. Stimuli encountered by participants during the study phase. Face stimuli taken from Generate Photos (https://generated.photos/datasets) and tweets created in (www.tweetgen.com).

https://doi.org/10.1371/journal.pone.0278671.g001

responded to the statement "The headline paired with this face was positive" using a scale ranging from -50 ("Not at all Confident") to +50 ("Extremely Confident"). The second was whether or not the associated headline had a disputed tag present. Participants responded to the statement "The headline paired with this face was disputed" using a scale ranging from -50 ("Not at all Confident") to +50 ("Extremely Confident").

Finally, participants were asked to complete a short demographic questionnaire recording gender, ethnicity, age, twitter use and whether or not they had noticed that some of the tweets contained a disputed tag. After participants had completed all parts of the questionnaire they were thanked and debriefed. The experiment was conducted using Gorilla Experiment Builder (https://gorilla.sc). Participants' reported twitter use and examples of all instruction and response screens can be found in Supplementary materials at the Open Science Framework (https://osf.io/t7c25/?view_only=8e6ca4f1249c4579a0e49e1ce5f4d747).

Results

The data for this experiment, as well as our pre-registered hypotheses and analysis plans, are available open access at Open Science Framework (https://osf.io/t7c25/?view_only=8e6ca4f1249c4579a0e49e1ce5f4d747).

Believability ratings. The believability ratings obtained during the study phase were subjected to ANOVA with Valence (positive, negative), Credibility (disputed, non-disputed) and Presentation (first, second) as within-subjects factors. As expected, there was a significant main effect of Credibility [$F(1,127) = 130.80$, $p < .001$, $\eta_p^2 = .34$], whereby non-disputed tweets ($M = 18.37$) were rated as more believable than disputed tweets ($M = -7.57$). There was also a significant main effect of Valence [$F(1,127) = 103.63$, $p < .001$, $\eta_p^2 = .10$], whereby positive tweets ($M = 12.30$) were rated as more believable than negative tweets ($M = -1.51$). The analysis revealed a significant Credibility × Presentation interaction [$F(1,127) = 11.90$, $p < .001$, $\eta_p^2 = .002$]. However, the believability of disputed tweets at presentation one ($M = -6.49$) and presentation two ($M = -8.66$) did not differ significantly ($p = .158$). Similarly, the believability of non-disputed tweets at presentation one ($M = 17.33$) and presentation two ($M = 19.41$) did not differ significantly ($p = .200$). There was no main effect of Presentation [$F(1,127) = .004$, $p = .951$, $\eta_p^2 = < .001$], no Valence × Credibility interaction [$F(1,127) = 1.55$, $p = .216$, $\eta_p^2 = .xx$], no Valence × Presentation interaction [$F(1,127) = 1.16$, $p = .284$, $\eta_p^2 = < .001$], and no Valence × Credibility × Presentation interaction [$F(1,127) = 1.08$, $p = .300$, $\eta_p^2 = < .001$].

Trustworthiness ratings. The trustworthiness ratings obtained during the test phase were subjected to ANOVA with Valance (positive, negative) and Credibility (disputed, non-disputed) as within-subjects factors (Fig 2). The analysis revealed a significant main effect of

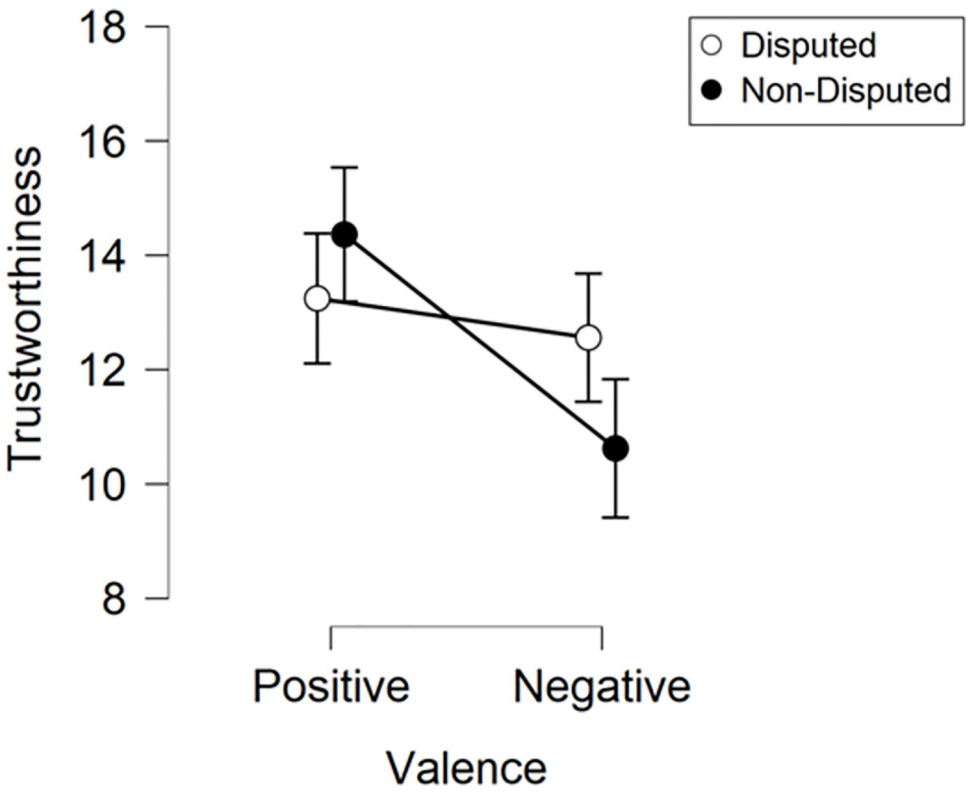

Fig 2. Experiment 1 ratings. Trustworthiness ratings obtained during the test phase of Experiment 1. Error bars depict ±SEM.

https://doi.org/10.1371/journal.pone.0278671.g002

Valence [$F(1,127) = 11.40$, $p < .001$, $\eta_p^2 = .08$], whereby targets associated with positive tweets ($M = 13.80$) were rated as more trustworthy than targets associated with negative tweets ($M = 11.59$). There was no main effect of Credibility [$F(1,127) = 0.98$, $p = .325$, $\eta_p^2 = .01$]. However, there was a significant Valence × Credibility interaction [$F(1,127) = 6.34$, $p = .013$, $\eta_p^2 = .05$]. Follow up t-tests, with a Bonferroni-adjusted alpha level of .025 (.05/2), were conducted to explore this interaction. Results showed that trustworthiness ratings of faces paired with positive disputed ($M = 13.25$) and negative disputed headlines ($M = 12.56$) did not differ significantly [$t(127) = 0.89$, $p = .378$, d = 0.08]. However, the trustworthiness ratings of faces paired with positive non-disputed headlines ($M = 14.36$) and negative non-disputed headlines ($M = 10.62$) did differ significantly [$t(127) = 3.75$, $p < .001$, d = 0.33]. This suggests that the modulation of facial trustworthiness induced by headline valence is attenuated by the presence of disputed tag.

Recall of headline valence. Next, we sought to assess participants' recall of the valence of the headline associated with each target face. Participants' confidence ratings were subjected to ANOVA with Valence (positive, negative) and Credibility (disputed, non-disputed) as within-subjects factors. The analysis revealed a significant main effect of Valence [$F(1,127) = 80.91$, $p < .001$, $\eta_p^2 = .39$], whereby participants were more confident that targets were paired with positive headlines when they were paired with a positive headline ($M = 2.09$), than when paired with a negative headline ($M = -8.33$). This suggests participants had some recollection of headline valence. A significant effect of Credibility was also found [$F(1,127) = 5.90$, $p = .017$, $\eta_p^2 = .04$], with participants more confident that the headlines were positive if they were non-disputed ($M = -2.01$), than if they were disputed ($M = -4.23$). We observed no Valence × Credibility interaction [$F(1,127) = 0.92$, $p = .340$, $\eta_p^2 = .01$].

Recall of headline credibility. Finally, we sought to assess participants' recall of the credibility of the headline associated with each target face. Participants' confidence ratings were subjected to ANOVA with Valence (positive, negative) and Credibility (disputed, non-disputed) as within-subjects factors. The analysis revealed a significant main effect of Credibility [$F(1,127) = 7.78$, $p = .006$, $\eta_p^2 = .06$], with participants being more confident that targets were presented with a disputed headline when the headline was disputed ($M = -8.42$) than when the headline was not disputed ($M = -11.01$). This suggests that participants had some recollection of the headline credibility. We observed no main effect of Valence [$F(1,127) = 2.41$, $p = .123$, $\eta_p^2 = .02$], and no Valence × Credibility interaction [$F(1,127) = 0.26$, $p = .612$, $\eta_p^2 = .002$].

Experiment 2

The results of Experiment 1 indicate that, in the absence of a disputed tag, participants tended to judge the faces associated with positive headlines as more trustworthy than those associated with negative headlines. When a disputed tag was present, however, target faces associated with positive and negative disputed headlines did not differ significantly in their perceived trustworthiness. These results suggest that the disputed tag may be effective in mitigating the potentially harmful effects of fake news on person evaluation.

These results accord with previous reports that people use disputed tags when deciding what to believe, and disregard information from less credible sources [39, 40]. They are, however, somewhat inconsistent with previous evidence that trait-relevant information influences person evaluation irrespective of source credibility [14, 48]. In our second experiment, we therefore sought to examine whether the 'protective' effects of the disputed tag are short-lived. It is conceivable, for example, that participants might find it difficult to hold in their memory which tweets were disputed and which were not. If their ability to 'bind' the disputed tag to the correct face-headline pairings was tenuous, increasing the interval between the study and test phases might render the disputed tags less effective.

Method

Participants. Protocols were approved by the University of York's Psychology Ethics Committee, participants gave informed written consent before taking part. One-hundred-and-twenty-eight participants completed the experiment (M_{age} = 29.12, SD_{age} = 9.36; 78 female, 44 male, 4 non-binary, 2 prefer not to say), recruited via Prolific. A further 42 were tested, but replaced having reported that they did not notice the disputed tag. Of the 128 participants in the final sample, all reported English as their first language and 126 resided in the UK (one resided in Portugal and one in Ireland). Of these 128 participants, 110 identified as White British, 1 as Black British, 2 as White and Black Caribbean, 1 as White and Black African, 4 as White and Asian, 3 as Indian, 2 as Pakistani, 1 as Chinese, 1 as "other" Asian background (not specified), 1 as African, and 1 as Caribbean. All participants received a small honorarium (£3.50) for their participation.

Design and procedure. The design of Experiment 2 was almost identical to that of Experiment 1. The only change was the inclusion of a 10-minute delay between the initial study phase and the test phase. Participants completed a distractor task during this interval, in which they were asked, in turn, to list as many uses as possible for five objects (a wellington boot, blanket, brick, watering can and a paperclip). Participants were given two minutes per object with the trustworthiness rating phase presented directly after.

Results

The data for this experiment, and pre-registered hypotheses and analysis plans, are available open access at Open Science Framework (https://osf.io/t7c25/?view_only=8e6ca4f1249c4579a0e49e1ce5f4d747).

Believability ratings

The believability ratings obtained during the study phase were subjected to a within subjects ANOVA with Valence (positive, negative), Credibility (disputed, non-disputed), and Presentation (first, second) as within-subjects factors. The analysis revealed a significant main effect of Credibility [$F(1,127) = 107.53$, $p < .001$, $\eta_p^2 = .46$], whereby non-disputed tweets (M = 15.46) were rated as more believable than disputed tweets (M = -6.90). There was also a significant main effect of Valence [$F(1,127) = 69.82$, $p < .001$, $\eta_p^2 = .36$], whereby positive tweets (M = 9.28) were rated as more believable than negative tweets (M = -0.72). A significant Credibility × Presentation interaction was observed [$F(1,127) = 19.74$, $p < .001$, $\eta_p^2 = .13$]. Simple contrasts revealed that disputed tweets at presentation one (M = -5.49) were judged as more believable than disputed tweets at presentation two (-8.30) ($p < .001$). In contrast, non-disputed tweets at presentation one (M = 14.54) were seen as less believable than non-disputed tweets at presentation two (M = 16.38) ($p = .038$). A significant Valence × Presentation interaction was also found [$F(1,127) = 9.27$, $p = .003$, $\eta_p^2 = .07$]. Follow up t-tests, with a Bonferroni-adjusted alpha level of .025 (.05/2), were conducted to explore this interaction. Results revealed that positive tweets were judged more believable at presentation one (M = 10.13) than at presentation two (M = 8.43) [$t(127) = 2.83$, $p = .005$, d = 0.25]. The believability ratings of negative tweets at presentation one (M = -1.08) and presentation two (M = -0.35) did not differ significantly [$t(127) = -1.34$, $p = .182$, d = -0.12]. There was no main effect of Presentation [$F(1,127) = 1.40$, $p = .239$, $\eta_p^2 = .01$], no Valence × Credibility interaction [$F(1,127) = .21$, $p = .645$, $\eta_p^2 = .002$], and no Valence × Credibility × Presentation interaction [$F(1,127) = .12$, $p = .730$, $\eta_p^2 = < .001$].

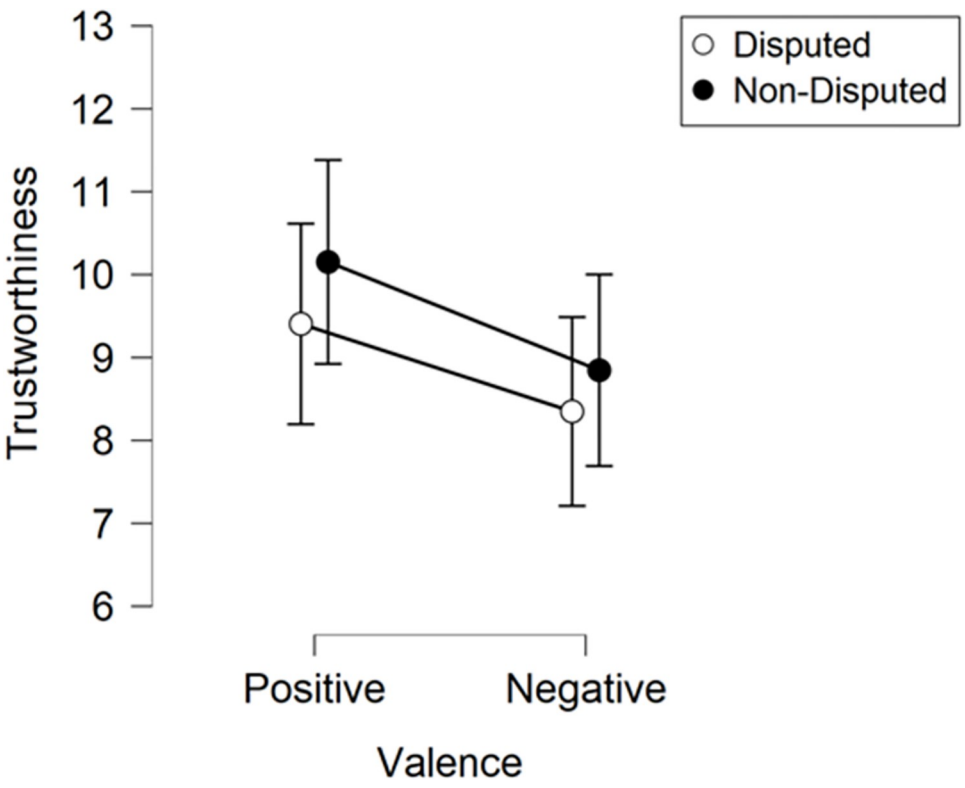

Fig 3. Experiment 2 ratings. Trustworthiness ratings obtained during the test phase of Experiment 2. Error bars depict ±SEM.

https://doi.org/10.1371/journal.pone.0278671.g003

Trustworthiness ratings

The trustworthiness ratings obtained during the test phase were subjected to ANOVA with Valence (positive, negative) and Credibility (disputed, non-disputed) as within-subjects factors (Fig 3). The analysis revealed a significant main effect of Valence [$F(1,127) = 5.09$, $p = .026$, $\eta_p^2 = .04$], whereby targets associated with positive headlines ($M = 9.78$) were rated as more trustworthy than targets associated with negative headlines ($M = 8.60$). We observed no main effect of Credibility [$F(1,127) = 1.17$, $p = .281$, $\eta_p^2 = .009$] and no Valence × Credibility interaction [$F(1,127) = .04$, $p = .840$, $\eta_p^2 = < .001$].

Recall of headline valence

Next, we sought to assess participants' recall of the valence of the headline associated with each target face. Participants' confidence ratings were subjected to ANOVA with Valence (positive, negative) and Credibility (disputed, non-disputed) as within-subjects factors. The analysis revealed a significant main effect of Valence [$F(1,127) = 85.83$, $p < .001$, $\eta_p^2 = .40$], whereby participants were more confident that targets were paired with positive headlines when they were paired with positive headlines ($M = 4.19$) than when they were paired with negative headlines ($M = -6.40$). This suggests participants had some recollection of headline valence. A significant effect of Credibility was also found [$F(1,127) = 11.36$, $p = .001$, $\eta_p^2 = .08$], with participants more confident that headlines were positive when they were non-disputed ($M = 0.34$) than when they were disputed ($M = -2.55$). There was no Valence × Credibility interaction [$F(1,127) = 1.76$, $p = .187$, $\eta_p^2 = .01$].

Recall of headline credibility

Finally, we sought to assess participants' recall of the credibility of the headline associated with each target face. We subjected participants' confidence ratings to ANOVA with Valence (positive, negative) and Credibility (disputed, non-disputed) as within-subjects factors. The analysis revealed a significant main effect of Credibility [$F(1,127) = 7.57$, $p = .007$, $\eta_p^2 = .06$], with participants being more confident that targets were presented with a disputed headline when the headline was disputed ($M = -5.13$) than when the headline was non-disputed ($M = -7.47$). We observed no main effect of Valence [$F(1,127) = 3.50$, $p = .064$, $\eta_p^2 = .03$] and no Valence × Credibility interaction [$F(1,127) = 0.62$, $p = .432$, $\eta_p^2 = .005$].

General discussion

Adults spontaneously attribute a wide range of traits to others based on their facial appearance [1, 2]. Previous findings indicate that this person evaluation is affected by information provided about the targets' previous actions and behaviours [12–15]. The present study sought to determine whether Twitter posts that pair facial images with favourable and unfavourable biographical information afford comparable face-trait learning. We also sought to establish whether any effects of the biographical information provided were attenuated by the presence of the "disputed tag" used by Twitter to mitigate the impact of fake news. To this end, participants were shown fictious tweets that paired strangers' faces with news headlines suggestive of positive or negative behaviours. Half of the tweets were tagged as 'disputed' and half were untagged.

In our first experiment, we found that, in the absence of a disputed tag, participants tended to judge targets associated with positive headlines to be more trustworthy than those associated with negative headlines. Interestingly, when a disputed tag was present, participants judged targets associated with positive and negative headlines as similarly trustworthy. These results appear to suggest that disputed tags are effective in mitigating the potentially harmful effects of fake news on person perception. In our second experiment, however, we found that the 'protective' effects of the disputed tag disappeared when a 10-minute delay was introduced between the study phase and test phase. Under these conditions, we found that headline valence modulated subsequent trust ratings irrespective of whether they were accompanied by a disputed tag.

Broadly speaking, our results–in particular, those of the Experiment 2 –accord with previous findings described by Baum and colleagues which suggest that trait-relevant information influences person evaluation irrespective of source credibility [14, 47]. Although we observed an effect of the disputed tag in Experiment 1, the effect of the credibility manipulation appears to be so short-lived as to have little impact on face-trait learning seen outside the lab. If the presence of a disputed tag has little or no effect after 10 minutes, it is unlikely to have an effect after 10 hours or 10 days. Nevertheless, this pattern of results raises the question: why do the effects of headline-valence survive a 10-mins delay, while the effects of the headline-credibility do not?

Our findings, and those from previous studies [12, 13, 15, 47], suggest that faces readily acquire positive and negative valence when paired with information about the individuals' positive and negative behaviours. This kind of learning may be relatively fast and easy, even to the point of being hard to inhibit. One possibility is that the natural relational structure between "agent" and "action" makes the elements easy to associate (e.g., they can be readily visualised). Another factor may be that the actions used here (e.g., "Local woman saves boy from drowning in river") and elsewhere in the literature (e.g., "Stole money and jewellery from the relatives he was living with"[13]) are emotive and salient. Finally, this kind of learning may occur even where participants encode only the gist or valence of the past behaviour.

While face-action or face-valence learning may be fast and easy, face-credibility or information-credibility learning may be harder. The relational structure between the to-be-learned elements is perhaps less intuitive (e.g., harder to visualise). The disputed tags used by Twitter are also relatively subtle and may fail to capture participants' attention. Indeed, in both experiments, a non-trivial number of participants had to be replaced for having failed to notice the presence of the disputed tags. As a result, we speculate that the learning responsible for the credibility effect in Experiment 1 was vulnerable to subsequent interference and soon started to decay [44].

The present findings have some disturbing implications for the effectiveness of propaganda circulated via social media. Our results suggest that fake news stories that depict an individual alongside negative biographical information (e.g., fictitious descriptions of their previous misdemeanours, crimes, and anti-social behaviour) may well have detrimental effects on the way that person is evaluated by users who encounter this content. The results from our second study suggest that the presence of a disputed tag may do little to reduce the impact on users' perceptions of those depicted. Perceptions of trustworthiness are thought to exert a strong influence on our decision making [49], they may affect how we vote [4], and what kinds of punishments we endorse [11]. Given the reach of social media platforms like Twitter, the dangers posed by this kind of propaganda are obvious.

We should also be mindful of the dangers posed by more subtle propaganda campaigns. As discussed above, it is well-established that participants quickly learn that individuals are trustworthy or untrustworthy when their faces are paired with details about their supposed prosocial and anti-social behaviours. Importantly, however, this kind of face-trait learning is known to generalise to other individuals who resemble those encountered during the study phase [16, 17, 19, 50]. In other words, if we learn that Bob is untrustworthy, we may spontaneously dislike Bob's cousin Fred, with whom he shares a passing resemblance. Together, these findings suggest a sinister possibility; that it may be possible to alter the public perception of a rival by circulating fake news stories on social media that present people who resemble the rival in unflattering terms. Conceivably, this kind of propaganda may pass unnoticed unless one suspected someone was being targeted in this way and was purposely searching for it.

It is also important that future research consider if and how the impact of disputed tags and other warning labels can be enhanced. Recent research suggests that placing warning labels immediately after, rather than alongside, a false headline may have a more powerful and longer lasting effect on its believability. According to the 'concurrent storage hypothesis', valence information and credibility are initially retained but message credibility fades more quickly from memory over time. Presenting the credibility label after the misinformation may help to increase its salience and so retention [44]. Other work has investigated the content of the warnings themselves, showing that the nature and appearance of the warning labels may influence their effectiveness [40] For example, Kirchner & Reuter [41] showed that adding an explanation for why a post had been labelled as disinformation increased its influence on perceived accuracy [41]. It would be interesting to measure whether manipulating the content of warning labels to enhance their salience increases their protective effects on person perception. If warning labels cannot be made more effective, there may be a case for blocking or obscuring the facial images included in disputed stories.

It is important to acknowledge that our highly controlled experimental design differs in important respects from more naturalistic viewing conditions. For example, in our design participants viewed each face separately for a pre-specified time. It would be interesting to investigate face-trait learning in conditions more closely resembling those of twitter where participants can scroll through multiple stories at their leisure. In our design, we also blurred out information regarding the people who sent the tweets. In future research, it would be

interesting to measure how the legitimacy of the source interacts with misinformation tags [51]. It would also be interesting to investigate how engagement with the disputed tag is influenced by the group membership and pre-existing biases of the participants. Finally, previous work has demonstrated that ambiguous behaviours are more likely to be interpreted as negative when the individual depicted has a face that is perceived as untrustworthy [52]. With this in mind, it may be interesting to investigate how the valence of the tweet, the presence of the disputed tag, and the perceived trustworthiness of an individual depicted may interact. It may be that participants would give more credence to the disputed tag when the information was negative and the individual depicted appeared untrustworthy than when the information was negative but the individual depicted appeared trustworthy. Addressing these questions will help us understand how best to combat misinformation online.

Across two experiments, we found that fictitious tweets that paired facial images with details of the person's positive or negative actions affected the extent to which readers subsequently judged the faces depicted to be trustworthy. When the rating phase followed immediately after the study phase, the presence of disputed tag attenuated the effect of the behavioural information. However, when the rating phase was conducted after a 10-mins delay, the presence of disputed tag had no effect. Our findings suggest that disputed tags may have relatively little impact on face-trait learning that occurs via social media. As such, fake news stories may have considerable potential to shape users' person evaluation.

Author Contributions

Conceptualization: Adam Eggleston, Richard Cook, Harriet Over.

Data curation: Adam Eggleston.

Formal analysis: Adam Eggleston.

Funding acquisition: Richard Cook, Harriet Over.

Investigation: Adam Eggleston.

Methodology: Adam Eggleston, Harriet Over.

Project administration: Adam Eggleston, Harriet Over.

Resources: Adam Eggleston.

Supervision: Richard Cook, Harriet Over.

Writing – original draft: Adam Eggleston, Harriet Over.

Writing – review & editing: Adam Eggleston, Richard Cook, Harriet Over.

References

1. Oosterhof NN, Todorov A. The functional basis of face evaluation. Proceedings of the National Academy of Sciences. 2008; 105(32):11087–92. https://doi.org/10.1073/pnas.0805664105 PMID: 18685089
2. Todorov A, Olivola CY, Dotsch R, Mende-Siedlecki P. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. Annu Rev Psychol. 2015; 66:519–45. https://doi.org/10.1146/annurev-psych-113011-143831 PMID: 25196277
3. Sutherland CAM, Oldmeadow JA, Santos IM, Towler J, Burt DM, Young AW. Social inferences from faces: Ambient images generate a three-dimensional model. Cognition. 2013; 127(1):105–18. https://doi.org/10.1016/j.cognition.2012.12.001 PMID: 23376296
4. Ballew CC, Todorov A. Predicting political elections from rapid and unreflective face judgments. Proc Natl Acad Sci U S A. 2007; 104(46):17948–53. https://doi.org/10.1073/pnas.0705435104 PMID: 17959769

5. Willis J, Todorov A. First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. Psychol Sci. 2006; 17(7):592–8. https://doi.org/10.1111/j.1467-9280.2006.01750.x PMID: 16866745
6. Todorov A, Pakrashi M, Oosterhof NN. Evaluating faces on trustworthiness after minimal time exposure. Soc Cogn. 2009; 27(6):813–33.
7. Dilger A, Muller J, Muller M. Is Trustworthiness Written on the Face? SSRN Electronic Journal. 2017;
8. Efferson C, Vogt S. Viewing men's faces does not lead to accurate predictions of trustworthiness. Sci Rep. 2013; 3.
9. Lavan N, Mileva M, Burton AM, Young AW, McGettigan C. Trait evaluations of faces and voices: Comparing within- and between-person variability. J Exp Psychol Gen. 2021; https://doi.org/10.1037/xge0001019 PMID: 33734774
10. Todorov A, Porter JM. Misleading First Impressions. Psychol Sci. 2014; 25(7):1404–17.
11. Wilson JP, Rule NO. Facial Trustworthiness Predicts Extreme Criminal-Sentencing Outcomes. Psychol Sci. 2015; 26(8):1325–31. https://doi.org/10.1177/0956797615590992 PMID: 26162847
12. Bliss-Moreau E, Barrett LF, Wright CI. Individual Differences in Learning the Affective Value of Others Under Minimal Conditions. Emotion. 2008; 8(4):479–93. https://doi.org/10.1037/1528-3542.8.4.479 PMID: 18729580
13. Falvello V, Vinson M, Ferrari C, Todorov A. The robustness of learning about the trustworthiness of other people. Soc Cogn. 2015; 33(5):368–86.
14. Baum J, Rahman RA. Emotional news affects social judgments independent of perceived media credibility. Soc Cogn Affect Neurosci. 2021; 16(3):280–91. https://doi.org/10.1093/scan/nsaa164 PMID: 33274748
15. Rydell RJ, Gawronski B. I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. Cogn Emot. 2009; 23(6):1118–52.
16. FeldmanHall O, Dunsmoor JE, Tompary A, Hunter LE, Todorov A, Phelps EA. Stimulus generalization as a mechanism for learning to trust. 2018; 115(7):E1690–7.
17. Verosky SC, Todorov A. When physical similarity matters: Mechanisms underlying affective learning generalization to the evaluation of novel faces. J Exp Soc Psychol. 2013; 49(4):661–9.
18. Eggleston A, Geangu E, Tipper SP, Cook R, Over H. Young children learn first impressions of faces through social referencing. Scientific Reports 2021 11:1. 2021; 11(1):1–8.
19. Gawronski B, Quinn KA. Guilty by mere similarity: Assimilative effects of facial resemblance on automatic evaluation. J Exp Soc Psychol. 2013; 49(1):120–5.
20. Gawronski B, Bodenhausen G v. Implicit and explicit evaluation: A brief review of the associative-propositional evaluation model. Soc Personal Psychol Compass. 2014; 8(8).
21. Shen X, Mann TC, Ferguson MJ. Beware a dishonest face?: Updating face-based implicit impressions using diagnostic behavioral information. J Exp Soc Psychol. 2020;86.
22. Shen X, Ferguson MJ. How resistant are implicit impressions of facial trustworthiness? When new evidence leads to durable updating. J Exp Soc Psychol. 2021; 97:104219.
23. Gregg AP, Seibt B, Banaji MR. Easier done than undone: Asymmetry in the malleability of implicit preferences. J Pers Soc Psychol. 2006; 90(1):1–20. https://doi.org/10.1037/0022-3514.90.1.1 PMID: 16448307
24. Petty RE, Briñol P, Tormala ZL, Blair W, Jarvis G. Implicit ambivalence from attitude change: An exploration of the PAST model. J Pers Soc Psychol. 2006; 90(1). https://doi.org/10.1037/0022-3514.90.1.21 PMID: 16448308
25. Rydell RJ, McConnell AR. Understanding implicit and explicit attitude change: A systems of reasoning analysis. J Pers Soc Psychol. 2006; 91(6).
26. Boukes M. Social network sites and acquiring current affairs knowledge: The impact of Twitter and Facebook usage on learning about the news. Journal of Information Technology & Politics. 2019; 16 (1):36–51.
27. Flintham M, Karner C, Creswick H, Gupta N, Moran S, Bachour K. Falling for Fake News: Investigating the Consumption of News via Social Media. 2018;
28. Shahi GK, Dirkson A, Majchrzak TA. An Exploratory Study of COVID-19 Misinformation on Twitter [Internet]. Vol. 22, arXiv. arXiv; 2020. p. 100104.
29. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science (1979). 2018; 359 (6380):1146–51. https://doi.org/10.1126/science.aap9559 PMID: 29590045
30. Peng Y. Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision. Article in Journal of Communication. 2018;

31. Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav. 2021;1–12.

32. Kanozia R, Arya R. "Fake news", religion, and COVID-19 vaccine hesitancy in India, Pakistan, and Bangladesh. Vol. 48, Media Asia. 2021. p. 313–21.

33. Allcott H, Gentzkow M. Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives. 2017; 31(2):211–36.

34. Mutahi P, Kimari B. Fake News and the 2017 Kenyan Elections. Communicatio. 2020; 46(4):31–49.

35. Ncube L. Digital Media, Fake News and Pro-Movement for Democratic Change (MDC) Alliance Cyber-Propaganda during the 2018 Zimbabwe Election. https://doi.org/101080/2374367020191670225. 2019;40(4):44–61.

36. Allcott H, Gentzkow M, Yu C. Trends in the diffusion of misinformation on social media. Research and Politics. 2019; 6(2).

37. Costa D. "The TrustedNews Chrome extension can save you from reading fake news." [Internet]. Know-Techie. 2018. p. Noember 27.

38. Roth Y, Pickles N. Updating our approach to misleading information [Internet]. Twitter Blog. 2020.

39. Mena P. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. Policy Internet. 2020; 12(2):165–83.

40. Clayton K, Blair S, Busam J, Forstner S, Behavior JG-P, 2020 undefined. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Springer. 2020; 42(4):1073–95.

41. Kirchner J, Reuter C. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. In: Proceedings of the ACM on Human-Computer Interaction. Association for Computing Machinery; 2020. p. 1–27.

42. Lee J. The effect of web add-on correction and narrative correction on belief in misinformation depending on motivations for using social media. Behaviour and Information Technology. 2020;

43. Chan M pui S, Jones CR, Hall Jamieson K, Albarracín D. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. Psychol Sci. 2017; 28(11):1531–46. https://doi.org/10.1177/0956797617714579 PMID: 28895452

44. Brashier NM, Pennycook G, Berinsky AJ, Rand DG. Timing matters when correcting fake news. Proc Natl Acad Sci U S A. 2021;118(5). https://doi.org/10.1073/pnas.2020043118 PMID: 33495336

45. Pennycook G, Cannon TD, Rand DG. Prior exposure increases perceived accuracy of fake news. J Exp Psychol Gen. 2018; 147(12):1865–80. https://doi.org/10.1037/xge0000465 PMID: 30247057

46. Smelter TJ, Calvillo DP. Pictures and repeated exposure increase perceived accuracy of news headlines. Appl Cogn Psychol. 2020; 34(5):1061–71.

47. Baum J, Rabovsky M, Rose SB, Rahman RA. Clear Judgments Based on Unclear Evidence: Person Evaluation Is Strongly Influenced by Untrustworthy Gossip. Emotion. 2020; 20(2):248. https://doi.org/10.1037/emo0000545 PMID: 30589302

48. Baum J, Rahman R. Negative news dominates fast and slow brain responses and social judgments even after source credibility evaluation. Neuroimage. 2021; 244:118572. https://doi.org/10.1016/j.neuroimage.2021.118572 PMID: 34508894

49. Olivola CY, Funk F, Todorov A. Social attributions from faces bias human choices. Vol. 18, Trends in Cognitive Sciences. Elsevier Ltd; 2014. p. 566–70.

50. Lee R, Flavell JC, Tipper SP, Cook R, Over H. Spontaneous first impressions emerge from brief training. Sci Rep. 2021; 11(1):1–13.

51. Ferguson MJ, Mann TC, Cone J, Shen X. When and How Implicit First Impressions Can Be Updated. Curr Dir Psychol Sci. 2019; 28(4).

52. Thierry SM, Mondloch CJ. First impressions of child faces: Facial trustworthiness influences adults' interpretations of children's behavior in ambiguous situations. J Exp Child Psychol. 2021; 208:105153. https://doi.org/10.1016/j.jecp.2021.105153 PMID: 33905972