Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment

SAUMYA PAREEK, The University of Melbourne, Australia
NIELS VAN BERKEL, Aalborg University, Denmark
EDUARDO VELLOSO, The University of Melbourne, Australia
JORGE GONCALVES, The University of Melbourne, Australia

As misinformation increasingly proliferates on social media platforms, it has become crucial to explore how to best convey automated news credibility assessments to end-users, and foster trust in fact-checking AIs. In this paper, we investigate how model-agnostic, natural language explanations influence trust and reliance on a fact-checking AI. We construct explanations from four Conceptualisation Validations (CVs) – namely *consensual*, *expert*, *internal (logical)*, and *empirical* – which are foundational units of evidence that humans utilise to validate and accept new information. Our results show that providing explanations significantly enhances trust in AI, even in a fact-checking context where influencing pre-existing beliefs is often challenging, with different CVs causing varying degrees of reliance. We find *consensual* explanations to be the least influential, with *expert*, *internal*, and *empirical* explanations exerting twice as much influence. However, we also find that users could not discern whether the AI directed them towards the truth, highlighting the dual nature of explanations to both guide and potentially mislead. Further, we uncover the presence of automation bias and aversion during collaborative fact-checking, indicating how users' previously established trust in AI can moderate their reliance on AI judgements. We also observe the manifestation of a 'boomerang'/backfire effect often seen in traditional corrections to misinformation, with individuals who perceive AI as biased or untrustworthy doubling down and reinforcing their existing (in)correct beliefs when challenged by the AI. We conclude by presenting nuanced insights into the dynamics of user behaviour during AI-based fact-checking, offering important lessons for social media platforms.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

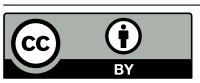
Additional Key Words and Phrases: misinformation, trust, reliance, credibility assessment, artificial intelligence, conceptualisation validations, human-AI interaction

ACM Reference Format:
Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 383 (November 2024), 31 pages. https://doi.org/10.1145/3686922

1 Introduction

Misinformation—false or misleading information with a credible appearance—has been increasingly infiltrating online information consumption [53]. Importantly, misinformation on social media travels "farther, faster, deeper, and more broadly" than the truth [112], making these platforms avenues where it proliferates the most. In response, these platforms have tested several interventions to identify and treat misinformation. These range from centralised, in-house moderation undertaken

Authors' Contact Information: Saumya Pareek, saumya.pareek@student.unimelb.edu.au, The University of Melbourne, Australia; Niels van Berkel, nielsvanberkel@cs.aau.dk, Aalborg University, Denmark; Eduardo Velloso, eduardo.velloso@sydney.edu.au, The University of Melbourne, Australia; Jorge Goncalves, jorge.goncalves@unimelb.edu.au, The University of Melbourne, Australia.



ACM 2573-0142/2024/11-ART383
https://doi.org/10.1145/3686922

by the platforms themselves [95] to employing third-party fact-checkers to assess information and assign credibility labels based on its propensity for harm [45, 93]. However, these approaches have shortcomings. Platform-led interventions make platforms the ultimate arbiters of truth. This practice is criticised by scholars and users alike, as it can impinge on an individual's autonomy in choosing the content they consume, with some users finding this approach "punitive and patronising" [95]. Furthermore, manual fact-checking fails to keep pace with the rapid creation of misinformation [25], offering limited scalability and allowing non-credible content to freely circulate on social media.

Given the difficulty of scaling interventions, recent research has increasingly explored automated approaches to assess content credibility [62, 94] and signal this to end-users [92, 100, 122]. However, mixed evidence exists regarding the effectiveness of AI-based credibility indicators, irrespective of detection accuracy. While some research demonstrates that they indeed enhance users' ability to distinguish between factual and fake news [41, 100], others find no discernible impact [122]. The effectiveness of such approaches hinges upon how well they regulate users' trust in AI decisions, influencing users' reliance on its recommendations. This phenomenon is so pivotal that several regulations recommend the 'right to explanation' – for example, disclosing to users "*the existence of automated decision making, including [...] meaningful information about the logic involved [...]*" [28].

To promote this understanding and foster trust, researchers have proposed various explanation methods adjoining automated credibility decisions. However, much of the existing research either investigates AI-based credibility indicators in an overly simplified setting by presenting explicit decisions without explanations [63, 122] or delves into highly technical, model-centric explanations like saliency maps and task-decision pairs [57, 72, 85]. While the latter is an improvement over opaque AI aids, it only exacerbates the pre-existing dissonance between human understanding and machine explanations. These approaches do not assist users in forming mental models of the AI's decision-making process, a crucial element when deciding whether and how to incorporate its advice [4]. To empower individuals to trust AI-based credibility indicators, it is thus imperative to design explanations that possess a strong undertone of human reasoning and convey a model's decision in terms of how humans construct and revise theories.

Jaccard and Jacoby [46] identify four foundational approaches to validating information, termed **conceptualisation validations (CVs)**. These CVs offer a systematic framework for how humans evaluate and accept information, shaping the formulation and assimilation of beliefs. *Consensual* validation equates a concept's worth to the level of acceptance (or consensus) it garners from the masses. *Expert* validation suggests acceptance when experts with relevant knowledge endorse information. *Internal* validation requires information to survive logical assessment and be without logical inconsistencies. Finally, *empirical* validation accepts information supported by rigorous and systematic empirical evidence. In this work, we are interested in how their use by AI during collaborative credibility assessment can shape users' decision-making and reliance on the AI.

Our investigation also explores the impact of two headline characteristics. The first, **scientificness**, distinguishes between the headline's message being scientific or non-scientific in nature. The second, **political congruence**, encompasses the alignment of the headline with participants' partisan beliefs and biases, distinguishing between congruent, incongruent, and altogether non-political headlines. By considering these factors, we aim to understand how AI-based credibility indicators can augment the perceived accuracy of (in)congruent headlines with differing scientificness.

We hypothesise these characteristics may require different types of evidence during fact-checking. For instance, the credibility of *scientific* news, such as medical research, may be verified through empirical data [70, 97]. In contrast, *political* news may be verified by examining the biases of those endorsing it [51, 113]. Importantly, belief in *political* news (but not *scientific* news) is often subject to motivated reasoning, where information congruent with one's partisan attitudes is deemed more credible [107].

Addressing the aforementioned research gaps, we seek to answer the following research questions:

RQ1: How do the *presence* and *type* of CV-based explanations influence reliance on an AI-based credibility indicator?
RQ2: How does the effectiveness of CV-based explanations vary based on the characteristics (scientificness and political congruence) of the headline being fact-checked?
RQ3: How do personal and contextual factors, such as an individual's confidence in their judgement and trust in AI, influence the impact of CV-based explanations?

To answer these questions, we conducted a survey-based study with 320 participants. We showed participants both factual and fake news headlines, each accompanied by an AI-based credibility indicator and an explanation whose presence and type varied between treatments. For each headline, we measured participants' credibility judgements and their confidence in those judgements twice – once before and once after displaying the indicator. Our experimental design manipulated the *AI judgement* (i.e. agreeing or disagreeing with the user's assessment), the *scientificness* of the headline (i.e. scientific or non-scientific), the *political congruence* of the headline (i.e. congruent, incongruent, non-political) and the *explanation Conceptualisation Validation (CV)* (i.e. Control (no explanation), Consensual, Expert, Internal, or Empirical).

We found that both the *presence* and the *type* of explanation influences trust in the AI. Providing explanations with AI assessments resulted in higher levels of trust compared to situations with no explanations. However, CVs differed in their ability to induce trust in the AI, with Consensual explanations being the least effective piece of information supplied. In contrast, Expert, Internal, and Empirical explanations were almost twice as effective, despite lacking external sources to corroborate their claims. Overall, explanations were highly effective irrespective of the AI's correctness – participants could not detect when they were being guided towards the truth. This suggests the potential of such indicators to (in)correctly guide people when the AI's judgement is (in)accurate. Further, we find no evidence that a headline's scientificness and political congruence influence switching behaviour, suggesting that individuals aligned their judgement with the AI for both attitude-affirming and challenging headlines. Finally, we observed both automation bias and aversion manifest in this experiment. Participants with higher trust in AI relied more on its judgements and perceived it as superior. However, some were reluctant to trust the AI irrespective of its accuracy, embracing their initial (in)correct beliefs when presented with a dissenting AI, mirroring the 'boomerang effect' observed in traditional corrections to misinformation.

This study has four major contributions. First, we highlight how providing model-agnostic, natural language, CV-based explanations enhances trust in an AI, even in the challenging context of fact-checking, where prior beliefs are often difficult to influence. Second, we demonstrate that the framing of explanations matters—identical AI judgements explained with different conceptualisation validations lead to different levels of reliance. We underscore how user behaviours such as conformity and epistemic dependence shape these trust dynamics. Third, we show that belief in both attitude-congruent and incongruent headlines can be influenced by a fact-checking AI and its explanations, suggesting its utility in addressing political misconceptions. Lastly, we underscore the dual nature of AI-based indicators, demonstrating their capacity to guide users both towards the truth and away from it. We identify the presence of automation bias and aversion during collaborative fact-checking, quantifying how users' pre-established trust in AI causes different reliance on the same credibility judgements. We also observe a 'boomerang' effect—one often seen in traditional corrections to misinformation—wherein individuals perceiving AI as biased embrace their (in)correct beliefs with increased conviction when fact-checked by the AI. We conclude by discussing these implications.

2 Related Work

To design effective AI aids that facilitate credibility evaluation, understanding factors regulating the influence of automated advice on human decision-making is essential. In the following sections, we

summarise the research undertaken on this topic. First, we highlight current platform-based misinformation interventions and explore reasons for their limited effectiveness. Next, we examine factors influencing human-AI decision-making, and explore the role of explanations in promoting understanding of automated decisions. Lastly, we discuss a Conceptualisation Validation (CV) framework, which outlines the different types of information humans seek to validate and accept new concepts.

2.1 Misinformation: The Story so Far

Misinformation has become rampant in today's online spaces, especially on social media platforms where user beliefs and partisan narratives shape content creation and dissemination [58]. This steady influx of unverified content, coupled with its propensity for widespread harm [1, 37, 91] has attracted interest from platforms and researchers alike. Despite being breeding grounds for misinformation, social media platforms also carry the potential to combat the relentless spread of misinformation within their communities. Credibility signals on disputed social media posts can notify entire networks of users viewing unverified content, besides alerting the original authors. This phenomenon, termed 'observational correction' [115], furthers the reach of corrective efforts by communicating credibility to those who may be unwilling or unable to verify content themselves. Moreover, alerting individuals to the veracity of posts directly at reading time is more optimal than offering it afterwards, since corrective messages tend to propagate slower than misinformation [112] and may not reach all exposed to it. Therefore, corrective efforts on social media have the advantage of clinging onto problematic content, ensuring greater reach.

2.1.1 Existing interventions to combat misinformation. The current landscape of efforts against online misinformation varies along two dimensions—the *central authority on truth* (who determines what content classifies as misinformation) and the *treatment of non-truths* (how detected misinformation is dealt with). Social media platforms mostly follow a centralised approach, utilising on-site moderation techniques such as policy-based fact-checking [69, 95, 110], and employing a combination of AI and third-party human fact-checkers [68] to identify and review potential misinformation.

While seemingly promising, platforms moderating content and deciding what millions of users can consume makes them the supreme arbiters of truth—an outcome that users and researchers alike have heavily criticised [95]. Harsher platform responses towards misinformation, such as downranking or altogether removing disputed content [74], can conflict with freedom of speech, with some viewing it as disguised censorship. Further, centralised decisions to remove content may go against end-users' needs, as some users may nevertheless wish to view unverified content to assess it independently, and stay informed of what their online peers share, irrespective of its veracity [49, 90].

In contrast, more lenient interventions that display credibility labels or warning flags aim to empower end-users to assess information themselves, without compromising user autonomy. However, empirical evidence for their effectiveness is mixed. While some research reports that labels enhance users' ability to discern truth from fiction and reduce their willingness to share misinformation [67, 122], others find labels to have only a minuscule effect [18, 78]. Notably, users are often reluctant to trust platform-supplied interventions. For instance, Saltz et al. [96] report that participants perceived platform-assigned fact-checking overlays as "punitive and patronising", finding it paternalistic and judgemental when the platform warned them about content credibility.

This grave distrust in platform-assigned credibility indicators stems from users' awareness of the conflict of interest faced by these platforms. Social media giants, where misinformation proliferates, are profit-driven and often algorithmically promote content that generates higher user engagement. Thus, combating misinformation—a category of content that attracts more attention than factual content [101]—appears to occur at the expense of revenue. This can cause platforms' fact-checking efforts to appear contradictory to their ethos [96]. Employing third-party human fact-checkers and

moderators, which are perceived as more neutral, may reduce this institutional distrust. However, manual fact-checking initiatives cannot scale to keep pace with the rapid generation of misinformation [25]. Thus, there is an urgent need to identify and design automated aids that can act as more neutral authorities, scale efficiently, and alert users to the questionable veracity of posts.

2.2 AI-based Credibility Assessment

A rich variety of empirical research has examined how Artificial Intelligence (AI) can collaborate with humans during decision-making [9, 15]. AI agents have successfully improved human decision-making across domains, including agricultural productivity [79], recidivism prediction [60], and medical diagnoses [44, 88]. Correspondingly, estimating the veracity of new information can be perceived as akin to the aforementioned decision-making tasks, where users leverage prior knowledge, heuristics, and contextual information to formulate judgements, potentially benefiting from AI assistance. While most research on automated fact-checking has focused on detecting and classifying misinformation [7, 66], there is a growing need to study how best to present automated credibility outcomes to end users, with fact-checks being increasingly automated. Thus, researchers have started investigating the utility of AI-based credibility indicators in signalling misinformation and influencing users' discernment abilities, an approach offering greater scalability and efficiency than manual fact-checking.

Current research on AI-based credibility indicators indicates a spectrum of effectiveness [63, 77, 100, 122]. Notably, people's accuracy in spotting fake and factual news increases when presented with a machine learning model's warning. However, increased discernment does not necessarily translate to increased trust in the model [14, 77, 100]. Following this finding, researchers have emphasised the significance of utilising transparent, human-understandable models. Similarly, Seo et al. [100] observed a stark disconnect between intervention accuracy and trust—participants placed more trust in a less accurate but more familiar fact-checking indicator (presented as the statement "Disputed by Snopes.com and PolitiFact.com[1]"), compared to a more accurate machine learning indicator [100]. In a similar vein, Yaqub et al. [122] compared AI-based credibility indicators with traditional, more commonplace indicators such as those involving fact-checkers and news media, and found that the AI-based indicators were the least effective. Of note is the common characteristic of AI-based indicators in the previous research: they were presented without any explanations. Because explanations and decision interpretability are prerequisites to trust in any intelligent system's decision [40], it is unsurprising that automated credibility decisions without explanations minimally impacted end-users.

2.2.1 The pivotal role of explanations. Recent literature on human-AI collaborative decision-making has identified several factors shaping user trust and reliance on automated systems. A multitude of contextual and personal determinants, such as risk perception [35, 36], AI literacy [17], confidence [124], and level of anthropomorphism of the automated aid [54], have been found to influence trust. Perhaps most important is the widespread finding that model explanations play a significant role in building trust [61, 124]. Explanations can enhance understanding of automated decisions, with increased model transparency heavily influencing users' perceptions of automated systems [27].

Explanations of AI decisions range from being feature-based, such as saliency maps which highlight input features utilised by the AI to make predictions [42, 57], to being example-based, where task-outcome pairs are presented to convey the rationale the AI might be following [85]. Recently, researchers have also proposed data-centric explanations, which describe the data the AI was trained on, including the collection and labelling process, pre-processing details, sample diversity, and recommended use-cases [3]. While these approaches can be more transparent than AI decisions without

[1]Snopes and PolitiFact are popular fact-checking websites.

explanation, they are either overly technical and model-centric, or are data-centric and require significant cognitive effort to parse [3]. End-users interfacing with AI aids should not be assumed to be technical experts and thus may not appropriately understand these complex, technical explanations [103].

The exploration of explanations that are not excessively technical remains limited in the context of automated fact-checking. Horne et al. [42] examined an explanation incorporating surface-level features of news content, such as word usage and tone, and found it to enhance users' ability to assess news veracity. However, other styles of non-technical explanations that attempt to reason with humans exhibit limited effectiveness. Rader et al. [89] examined 'How', 'Why', 'What', and 'Objective' (unbiased) explanations on users' perception of algorithmic transparency. These explanations improved users' awareness of algorithm involvement but not their understanding of it. Further, Epstein et al. [24] tested a headline-agnostic, static explanation, which communicated that the credibility decision was made through automated analysis of human labellers' input. While this explanation reduced the sharing of disputed posts, it failed to foster users' trust in the AI's credibility judgements.

Taken together, these works emphasise the need for designing explanations that not only promote understanding of automated credibility decisions, but also foster trust in these decisions. Especially in the context of credibility assessment, where AI-based indicators are expected to influence beliefs and attitudes, users need to understand the rationale behind the AI aid's advice. They should be able to trust it to incorporate its judgement during decision-making since users are more likely to accept the recommendations of automated aids when they have a sound and complete understanding of the reasoning behind them [26, 55].

2.3 Appropriate Reliance and Biases in AI-assisted Decision-Making

Despite their promise, explanations may not consistently guide users in determining when to accept or reject AI recommendations. This phenomenon where users are able to discern the accuracy of AI advice and rely on it only when warranted is termed appropriate reliance [98]. In Human-AI collaborative scenarios, the *accuracy* of reliance on AI significantly impacts decision-making outcomes [104, 108]. Since overreliance can lead to poorer human-AI team performance [5, 12, 124], it becomes crucial to understand factors that influence the degree of human reliance on AI. However, several biases come into play when quantifying reliance on automated aids. The perceived trustworthiness of AI can be influenced by users' dispositional trust (or lack thereof) in automated systems. Individuals may excessively trust automated advice and incorporate it into their decision-making without critical evaluation. This tendency to perceive automated systems as more knowledgeable than oneself and exhibiting unwarranted trust is referred to as *automation bias* [33, 75]. Similarly, individuals lacking trust in fellow humans may prefer automated content moderation over that by humans [73]. On the other hand, undue scepticism or distrust towards automated systems, known as *algorithm aversion* [52, 87], may lead users to solely rely on their own judgements, disregarding automated advice. However, while explanations aim to enhance user understanding, they may paradoxically lead to overreliance on AI decisions, creating challenges in fostering appropriate reliance on AI systems [19, 59]. Such undue effects, termed 'explanation pitfalls' [23], have been found to exist, albeit in a different human-AI decision-making context than this study examines. Therefore, investigating how automation bias and aversion may manifest in an AI-assisted credibility assessment scenario and how they regulate the effectiveness of explanations is crucial.

2.4 Conceptualisation Validations

Jaccard and Jacoby [46] propose four fundamental approaches individuals use for validating information, termed *conceptualisation validations* (CVs). These CVs form an essential component of theory construction and, together, formulate a structured system on how humans evaluate and accept information, ultimately influencing the formation and integration of beliefs. The four CVs are:

(1) **Consensual validation:** This validation type assesses the worth of a concept by the level of acceptance or consensus it receives from the masses. Consensual validation recognises the importance of social influence in shaping beliefs, attitudes, and behaviours.
(2) **Expert validation:** This validation type relies on the endorsement of experts with relevant knowledge and experience to confirm the validity of a concept. Expert validation recognises the critical role of authority figures and domain experts and the trust individuals place in them.
(3) **Internal validation:** This validation type suggests a concept can be accepted if it withstands logical scrutiny and is free of logical inconsistencies. Internal validation recognises the significance of logical coherence and consistency in shaping beliefs and attitudes.
(4) **Empirical validation:** This validation type asserts that concepts are valid if they are supported by rigorous and systematic empirical evidence. Empirical validation recognises the importance of evidence-based arguments and scientific rigour in regulating beliefs and attitudes.

The existing literature underscores the importance of explanations in the realm of automated fact-checking, yet there is a notable dearth of explanations that have been successful at explaining and fostering trust in automated credibility assessments. Consequently, our identified research gaps centre on understanding how best to present complex automated credibility assessments to end-users, in a manner that resonates with their reasoning and enables them to trust the AI's decisions. Given that CVs encompass various building blocks of human understanding and are associated with different types of evidence, they present a unique opportunity to investigate how explanations rooted in different CVs can regulate users' reliance on AI-based credibility indicators. Therefore, in this study, we examine whether and to what extent humans incorporate an AI's fact-checking advice into their decision-making, when accompanied by explanations derived from different CVs. Additionally, we explore how automation bias and aversion may manifest in the context of AI-assisted credibility assessment, potentially mediating the effectiveness of these CV-based explanations. We aim to contribute insights into the design of effective and comprehensible explanations for AI-based fact-checking systems.

3 Method

Through the lens of a news assessment scenario, we seek to examine how explanations rooted in different ***conceptualisation validations*** **(CVs)** [46] can impact understanding of an AI aid's decision and subsequently influence people's judgement of the news. To achieve this objective, we deployed an online survey-based experiment. The following sections describe our experimental setup, the explanations, participant recruitment, and experimental procedure.

3.1 Experimental Setup

To assess the effectiveness of explanations, participants evaluated the credibility of news headlines in a two-stage decision-making process. In the first stage, participants viewed a news headline (see Figure 1 - Step 1/4) and provided an initial binary credibility assessment along with their confidence in this assessment (see Figure 1 - Step 2/4). In the second stage, the same headline was re-displayed, with the addition of an AI-based credibility indicator (see Figure 1 - Step 3/4). This indicator comprised a binary decision and an explanation whose presence and formulation varied between treatments as per the four CVs. The CONTROL condition presented an explicit judgement without any explanation. Participants then had the option to revise their initial judgement by making a final credibility judgement, incorporating or rejecting the AI's advice (see Figure 1 - Step 4/4).

This approach is akin to a pretest-posttest experimental design, commonly employed in similar experiments measuring the effectiveness of credibility indicators on belief fluctuations in news articles [32, 39, 80, 109]. Moreover, it aligns with the research finding that individuals tend to form independent decisions before considering an automated aid's recommendations [34]. This setup

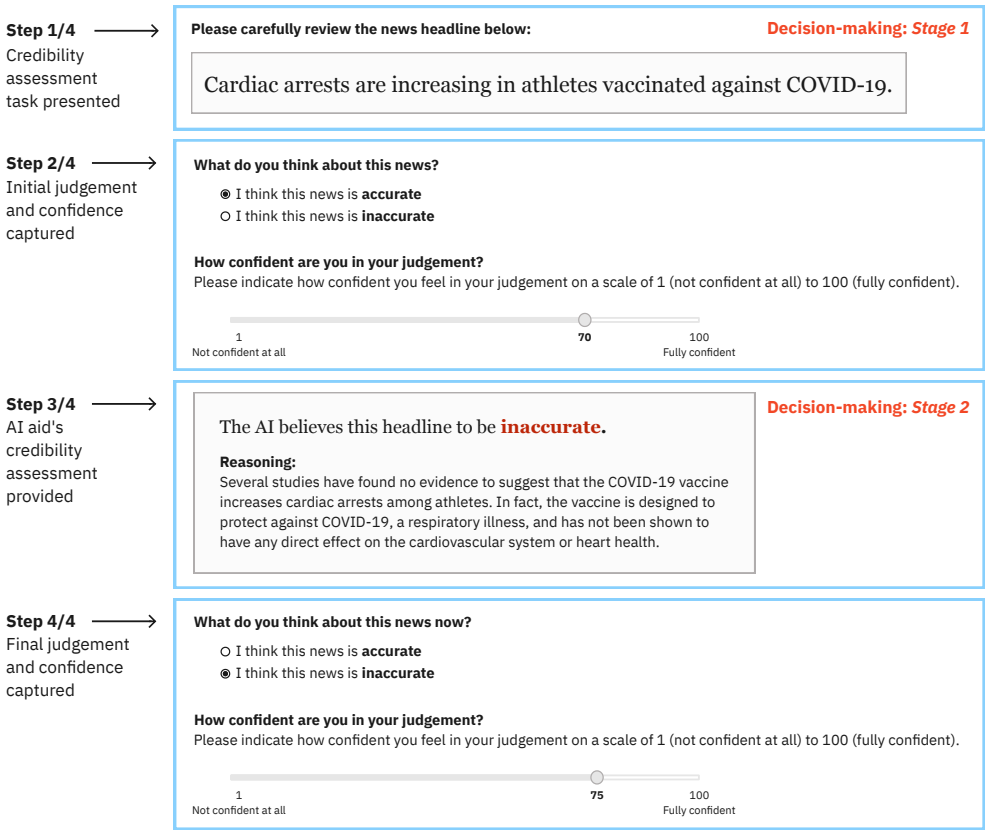


Fig. 1. An example of a task sequence where the AI aid disagrees with the user, progressively presenting each step.

also mimics the 'update' experimental condition in Green and Chen [35], which similarly examines reliance through a two-stage decision-making process.

3.1.1 Credibility assessment tasks. The headline characteristic of *scientificness* – i.e. scientific or non-scientific – and *political congruence* – i.e. congruent (supporting partisan beliefs), incongruent (contradicting partisan beliefs), or non-political (non-partisan news) – present six possible combinations. The purpose of classifying headlines as having or missing threads of political congruence, and being scientific or non-scientific, was to comprehensively investigate whether one explanation type induces more reliance than others based on headline characteristics **(RQ2)**. We collected headlines from Politifact and ImproveTheNews[2]. The final 16 headlines were selected based on three inclusion criteria:

(1) Headlines had to be verifiably and objectively true or false so that the AI's explanations could be woven around the ground truth;
(2) Headlines had to be relevant to the U.S. social and political climate at the time, such that they were not too 'stale' and their evaluation not too obvious, so participants may experience the need to rely on the AI;
(3) Headlines had to contain a singular claim, allowing a binary credibility judgement and avoiding situations where a headline is partially accurate.

[2]ImproveTheNews is a popular aggregator of trustworthy news.

Table 1. Headlines for each *Scientificness* × *Political Congruence* pairing, as perceived by Republican participants.

Scientificness	Political Congruence	Example headline
Scientific	Congruent	"Athletes vaccinated against COVID-19 are experiencing a higher rate of cardiac arrests."
	Incongruent	"Coronavirus can be transported by both cigarette and e-cigarette smoke."
	Non-Political	"Erythritol, an artificial sweetener, increases risk of heart attacks and strokes."
Non-Scientific	Congruent	"Mass shootings during Biden's presidency surpass Trump's entire term."
	Incongruent	"Firearms are now the leading cause of death among children ages 1-19."
	Non-Political	"Disney announces retirement of Mickey Mouse as official mascot in 2024."

Each headline pertained to one of the six *Scientificness* × *Political Congruence* pairings, with half being factual and the other half fake (see Table 1). The veracity of headlines was determined by the credibility ratings given by Politifact and ImproveTheNews. We further solidified our verification by cross-referencing with the multiple independent and credible fact-checking sources listed by these organisations in their fact-checking reports for our chosen headlines.

We adopt the characterisation of ***scientific*** misinformation by Southwell et al. [105], defining it as news contradicting the best available scientific evidence and established scientific principles. For scientific headlines, we selected medical and health-related news, following other works examining scientific misinformation [21]. We selected non-scientific headlines which represent a diverse range of non-scientific topics encountered in everyday lives, including entertainment, agriculture, and social media. Further, the congruence of a political headline is a factor of both the headline (its slant) and the participant viewing it (their political orientation). Thus, to operationalise ***political congruence***, we included an equal number of Republican-congruent and Democrat-congruent headlines, as well as entirely non-political headlines. To ensure participants perceive congruence as intended, we recruited only those identifying as Republicans or Democrats. This ensured that participants were exposed to both attitude-affirming and attitude-contradicting political claims. By considering the congruence between an individual's political orientation and the headline's slant, we aim to enhance the likelihood of our findings being applicable to a broader range of political stances beyond just Republicans and Democrats. We ascertained this political slant based on prior knowledge of the deep disagreements between Republicans and Democrats on polarised issues in the U.S., such as gun ownership. The authors validated this by cross-checking trustworthy online sources of data on political polarisation, such as the Pew Research Center [82–84]. The complete set of headlines and our classification is included in the supplementary material.

3.1.2 AI-based Credibility Indicator and Explanations. After capturing participants' initial independent assessment, we presented a credibility indicator evaluating the headline in question. Although these indicators and explanations were created by us, we explicitly informed participants at the survey's outset that they would be interfacing with a fact-checking AI aid. To further enhance the perceived authenticity of AI involvement, and following experiments involving an actual fact-checking AI [63], we phrased the indicator to attribute its judgements to an AI.

Since reliance on the AI can only be effectuated when there is a disagreement between the AI and the user's initial assessment [39], we designed our indicator to randomly disagree with participants' initial judgements half the time throughout the experiment, irrespective of the ground truth. Agreement and disagreement with the user were operationalised indirectly by presenting the AI's judgement of the headline to mirror or oppose the user's. Although the resulting eight headlines for which the AI agreed with the participant do not contribute to our understanding of reliance, this balancing was necessary for two reasons. Firstly, an automated credibility assessment system that contradicted participants'

Table 2. Eight explanations for a headline, highlighting the difference between CVs and the AI's judgement as either accurate (✅) and inaccurate (❌). Each participant only saw one of these explanations for a headline.

	Headline: "Wisconsin emerges as the top producer of cranberries."
Consensual	✅: Of the individuals taking this survey with you, 71% have rated it as accurate while 29% have rated it as inaccurate. ❌: Of the individuals taking this survey with you, 71% have rated it as inaccurate, while 29% have rated it as accurate.
Expert	✅: According to an expert in the agricultural industry, Wisconsin enjoys its dual status as both a dairy and cranberry powerhouse, ranking at the top. ❌: According to an expert in the agricultural industry, cranberries are Wisconsin's number one fruit crop, but it is not in fact the nation's leading producer.
Logical	✅: Cranberries flourish in Wisconsin, given its unique geography and climate. Cranberries thrive in wet, acidic soil, and Wisconsin's northern regions are dotted with thousands of shallow, marshy bogs that provide the perfect growing conditions. ❌: While Wisconsin may have the ideal conditions for growing cranberries, that does not necessarily translate to being the nation's top producer. Other factors, such as farm size, infrastructure, and access to markets, also play a role in reducing the state's yield.
Empirical	✅: Wisconsin produced a record-breaking 5.38 million barrels of cranberries in 2020. This represents over 60% of the total US cranberry production for the year, solidifying Wisconsin's position as the nation's top cranberry producer. ❌: Wisconsin produced a record-breaking 2 million barrels of cranberries in 2020, representing around 35% of the total US cranberry production for the year. Wisconsin is the leading producer of dairy in the country, but it ranks only fourth in cranberry production.

every decision would appear inherently untrustworthy and may interfere with any causal relationships we infer between explanations and observed user behaviour. Secondly, such a setup may also shatter the believability of AI involvement, making the experiment appear synthetic to participants.

In the CONTROL condition, to establish a baseline, the AI provided an overt binary credibility decision (i.e. *accurate* or *inaccurate*) without any explanation. In the four treatment conditions, the AI additionally displayed an explanation rationalising how it arrived at its decision. For each headline, we designed four explanations, each rooted in a conceptualisation validation (CV) – CONSENSUAL, EXPERT, LOGICAL, and EMPIRICAL [46]. We leveraged ChatGPT (GPT-3.5) to design explanations that better resemble model outputs and offer less variability between headlines. Being a language model, we instructed it to generate eight explanations for each headline, one set of four corroborating the headline and another set of four refuting it, with each set framed according to the four CVs. For each explanation type, we thoroughly outlined the characteristics of its underlying CV in our prompts to ChatGPT.

The authors vetted the generated explanations over several passes, ensuring explanations constructed from each CV have no unintended variability between headlines. For instance, for the EXPERT CV-based explanations, we ensured that the 16 explanations that support our 16 headlines and the 16 explanations that contradict them were identical in structure and framing, only varying the specific details of the headline incorporated into the explanation, and whether or not the experts supported the claim. This process helped us clearly operationalise the CVs, ensuring that any observed differences in participant behaviour between explanation types are strictly owing to the nature of the explanation (see Table 2 for an example of generated explanations). The complete set of headlines and explanations is included in the supplementary material.

3.2 Experimental Procedure

Figure 2 provides an overview of our experimental design. We utilised a 2 (**Headline Scientificness:** Scientific and Non-Scientific) × 3 (**Political Congruence:** Congruent, Incongruent, and

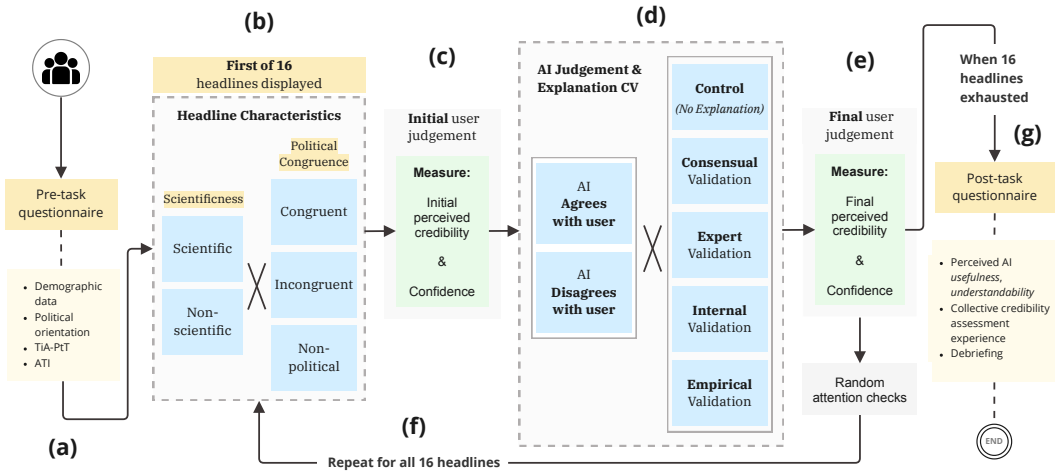


Fig. 2. The full experiment flow. Headline characteristics (6) and explanation type (5) vary between treatments. **(a)**: Pre-task questionnaire. **(b)**: First of 16 headline displayed **(c)**: Measurement of initial judgement and confidence. **(d)**: Same headline presented again with the AI aid's binary judgement and an explanation type. **(e)**: Final measurement of participant's judgement and confidence. **(f)**: Procedure repeated for all 16 headlines. **(g)**: Open-ended questions and debriefing related to participants' exposure to misinformation.

Non-Political) × 4 (**Explanation CV:** Consensual, Expert, Internal, Empirical) within-subjects factorial design, as represented in Figure 2 (b) and (d). Additionally, participants also saw a CONTROL condition where no explanation was displayed.

3.2.1 Participants. We published our study on Prolific and screened participants prior to recruitment, restricting their political orientation to the two dominant in current U.S. electoral politics. Thus, participants were required to be born and located in the U.S., and identify either as a Republican or Democrat. By considering the congruence between participants and the headline slant, rather than absolute political orientations, we can gain insights that may extend beyond the specific political orientations of the participants we recruit. We presented the survey to participants with an approval rate ≥97% on Prolific and ensured no participant took part in our study more than once. Overall, we recruited valid data from 320 participants, equally divided between Republicans and Democrats.

3.2.2 Procedure. Participants were randomly assigned a counterbalanced *headline characteristics* and *explanation CV* sequence, with the *AI Judgement* being in favour of participants' assessment for half the headlines. We informed participants that they would work with a credibility assessment AI aid. The survey began with a pre-task questionnaire, which collected participants' demographic details and political orientation, and administered the TiA-PtT (Trust in Automation – Propensity to Trust subscale [56], measures dispositional trust in automation) and ATI (Affinity for Technology Interaction [31]) questionnaires (Figure 2 (a)). Next, we displayed the first randomised headline (Figure 2 (b)) and collected participants' initial, unassisted binary credibility judgement (Figure 2 (c)). We also measured their confidence in their reported judgement, i.e., how certain they were about their credibility evaluation, on a sliding scale of 1 to 100, with a higher score indicating higher confidence. To nullify the potential bias generated by the starting position of the anchor on a slider [99], our sliders started unmarked, with an anchor appearing only after users clicked on the slider's range. In studies examining changes in beliefs or opinions, including research on misinformation [47, 118] and conformity [120, 121], initial confidence commonly serves as a proxy for participants' prior

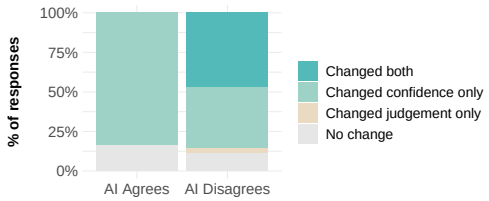


Fig. 3. Distribution of user behaviour during cases of agreement and disagreement with the AI.

knowledge or uncertainty about the task, and influences their inclination to rely on the AI [64, 81, 117]. Collecting initial confidence allowed us to effectively account for participants' prior knowledge of the headlines during our analyses. Following this, the AI aid either agreed or disagreed with the users' initial judgement, for example, disagreeing with the participant by judging the headline to be accurate if the participant had judged it to be inaccurate. This system behaviour aimed to understand how participants' reliance on the AI's judgement varied between explanations when there was an initial disagreement. Thus, in conditions other than CONTROL, the AI judgement was also accompanied by an explanation, whose CV varied between treatments (Figure 2 (d)).

Upon seeing the AI aid's advice, participants were asked to provide their final credibility assessment and confidence (Figure 2 (e)), incorporating or neglecting the AI's judgement. This setup was repeated for all 16 headlines for each participant (Figure 2 (f)). Lastly, using a post-task questionnaire, we probed participants through open-ended questions to obtain insights about their collaborative credibility assessment experience and their trust in the AI (Figure 2 (g)). We were also interested in understanding why they may have changed their judgement to align with the AI or perhaps resisted its influence. Moreover, since participants were exposed to some misinformation, they were debriefed at the end of the study. The debrief message was designed to thoroughly and clearly communicate the necessary information to participants, and was revised based on feedback from a pilot study. The debrief also included links to Politifact reports comprehensively debunking any misinformation shown during the study. Lastly, to ensure comprehension and detect inattentive participants, we randomly presented attention check questions throughout the survey, which followed best academic practices and solely measured attentiveness [43] rather than memory or knowledge. Those who failed both checks (N = 11) were removed from our final dataset, and we recruited additional participants until we reached a valid dataset with 320 participants.

The Ethics Committee of our university approved the study. Participants took a median time of approximately 14 minutes to complete the survey and received around US$4 for participation.

4 Results

We recruited 320 participants (156 men, 159 women, 3 non-binary, and 2 preferred not to specify) for this study. Each participant evaluated 16 unique headlines, resulting in 5120 initial and an equal number of final credibility responses. The AI-based indicator mirrored and opposed the participants' judgements an equal number of times. Thus, there were 2560 instances where the AI's judgement directly opposed the participants'. We note that our intention was not to compare user behaviour between human-AI agreement and disagreement scenarios but to examine the influence of the presence and type of CV-based explanations on the adoption of AI advice in cases of disagreement.

4.1 Robustness and Manipulation Check

All but three participants changed their initial judgement or confidence at least once after seeing the AI disagree with them. We observed 1300 changed judgements in total, with an average of 4.06

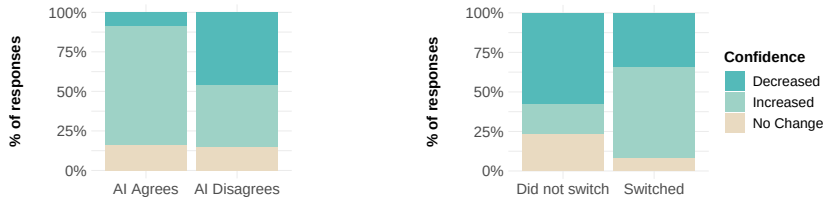


Fig. 4. Distribution of changes in user confidence. *Left:* During cases of agreement and disagreement with the AI; *Right:* Solely displaying cases of disagreement with the AI, for when users changed their judgement versus when they did not.

changes (SD = 2.39) per participant, out of the eight instances of disagreement between the participant and the AI. The distribution of participant behaviour after viewing the AI's judgement is depicted in Figure 3. During *AI agreement*, participants changed their credibility judgement in 0% of cases, made no changes to both judgement and confidence in 15.9% of cases, while revised their confidence in 84.1% of cases. Conversely, during *AI disagreement*, participants changed both their confidence and judgement in 46.9% of cases, while revising only their confidence in 38% of cases.

Figure 4 (*left*) illustrates this change of confidence in detail. When the AI agreed with participants, they increased their confidence in 75% of instances (M = 20.2, SD = 16.4) after observing the AI reinforce their judgement. Conversely, when the AI disagreed, we observed two distinct behaviours — participants decreased their confidence in around 46% of cases (M = -17.8, SD = 16), whereas they also increased their confidence in 39% of cases (M = 18.2, SD = 16.1). This divergence in confidence is explained when we analyse it along with switching behaviour during *AI Disagreement* (Figure 4 (*right*)). When the AI disagreed with participants, but they chose to maintain their judgement, their confidence decreased in 58% of instances (M = -16.8, SD = 15.5). However, when they did switch to align with the AI's judgement, they increased their confidence in almost 58% of instances (M = 21.3, SD = 16.3). These results firmly establish that participants were more likely to switch their judgement and reduce their confidence when the AI's judgement did not match theirs and increase their confidence when the AI did align with them, not randomly but under the influence of our experimental conditions, confirming the validity of our results.

To verify whether the congruence between participants' political orientation and a headline's slant was perceived by participants as we intended, we examined participants' initial, independent belief in the headlines. We hypothesised that participants would perceive politically congruent headlines to be accurate more frequently than politically incongruent ones, while neutral headlines would be evaluated similarly across participants. A chi-square test of independence reveals a significant relationship between headline slant and participant political orientation (χ^2 = 63.58, df = 2, p < 0.001). As expected, Republican participants initially perceived a much higher percentage (65.8%) of Republican-slanted headlines as accurate compared to Democrat-slanted headlines (39.5%). Likewise, Democrat participants found a much higher percentage (63.9%) of Democrat-slanted headlines accurate compared to Republican-slanted headlines (44.2%). Both groups perceived neutral headlines similarly, with approximately 42.1% of Republicans and 39.8% of Democrats assessing them as accurate. Together, these results affirm the effectiveness of our manipulation of political congruence, providing evidence that headlines categorised as favouring or opposing specific political views were indeed perceived as such by participants identifying with that political group.

Finally, as the AI randomly agreed and disagreed with participants, we determined if our participants detected this and consequently experienced a decline in trust in the AI as the experiment progressed. Instead, we found a slight positive correlation between the likelihood for participants to

switch and the trial number (β = 0.003, SE = 0.001, p = 0.025). Participants exhibited a minute increase in reliance on the AI as the experiment progressed. These findings increase the likelihood of any observed switching behaviour being attributed to our experimental manipulations, rather than extraneous factors, allowing us to make causal inferences about the influence of our manipulated variables.

4.2 Model Construction

We measured participants' reliance on the AI aid using **Switch Fraction**, a metric widely used in the literature [64, 123] and suitable for a two-stage decision-making process like ours [111]. In our study, the switch fraction captures the instances where a participant changed their credibility judgement to align with the AI aid's prediction following an initial disagreement between the two. Since cases without initial disagreement do not contribute to this measure, we subset our data during analysis to reflect only those cases with an initial disagreement between the participant and the AI aid (50% of cases). Moreover, our research objective was to understand how different CV-based explanations influence trust and reliance. Thus, we deliberately chose not to report any performance or accuracy measures. This was also necessary because the AI aid's accuracy is not fixed and instead is a function of the accuracy of the users themselves — it agrees with the users' judgement half the time and disagrees with the other half, irrespective of the ground truth. We investigated the impact of the following eight predictor variables on participants' switching behaviour:

- **Scientificness:** Whether or not the headline message was scientific (possible values: *scientific* or *non-scientific*).
- **Political Congruence:** Political alignment between a headline's slant and participants' political orientation (possible values: *congruent* (where the headline aligns with the participant's political orientation), *incongruent* (where the headline opposes the participant's political orientation), and *non-political*).
- **Conceptualisation Validation (CV):** The presence and type of the AI's explanation CV (possible values: CONTROL (no explanation), CONSENSUAL, EXPERT, INTERNAL, and *empirical*).
- **Confidence$_{initial}$:** Participants' initial confidence in their judgement before seeing the AI's assessment, allowing us to account for their prior knowledge of the headline (ranging from 0–100, higher values denoting higher confidence).
- **AI Accuracy:** Whether or not the AI's credibility judgement was accurate according to the ground truth. This variable is determined by comparing the AI's judgement with the veracity of the headline (fake or factual), establishing whether the AI made correct or incorrect recommendations (possible values: *accurate* or *inaccurate*).
- **Trust in Automation (TiA):** A set of validated questionnaire scales to measure subjective trust [56], from which we adopted the Propensity to Trust (TiA-PtT) subscale to account for any effects of dispositional trust on user reliance, following similar studies [39, 59]. The responses were collected on a Likert scale ranging from 1: *Strongly Disagree* to 5: *Strongly Agree*, subsequently aggregated into a single measure.
- **Affinity for Technology Interaction (ATI):** A validated instrument to consider any possible effects of participants' affinity for technology [31] on the reliance they exhibit, following similar experiments [39]. The responses were collected on a Likert scale ranging from 1: *Completely Disagree* to 6: *Completely Agree*, subsequently aggregated into a single measure.

We utilised the statistical R package `lme4` [6] to construct a generalised linear mixed-effects model (GLMM) of the relationship between the aforementioned predictor variables and switching behaviour, using a `logit` link function. This enabled us to determine the impact of a group of predictor variables on an outcome variable (switched or not) with a non-normal distribution. We specified participant

Table 3. Effect of predictors on participants' switching behaviour. Statistically significant main effects and interactions (p <0.05) are in bold. The sign of the estimate (+/-) denotes the direction of the relationship between the predictor and switching behaviour.

Variable	Estimate	Odds Ratio	95% CI	p-value
Baseline: Condition = Control				
Condition = Treatment	0.826	2.29	[1.48, 3.52]	< **0.001**
Baselines: ***Scientificness = Non-scientific, Political Congruence = Congruent, CV = Control, AI Accuracy = Inaccurate***				
Headline characteristics				
Political Congruence = Incongruent	-0.128	0.87	[0.44, 1.70]	0.707
Political Congruence = Neutral	0.017	1.00	[0.55, 1.82]	0.955
Scientificness = Scientific	-0.222	0.81	[0.51, 1.31]	0.358
Conceptualisation validations (CVs)				
CV = Consensual	0.410	1.50	[0.61, 3.69]	0.370
CV = Expert	1.158	3.18	[1.31, 7.72]	**0.010**
CV = Internal	1.581	4.87	[1.99, 11.93]	< **0.001**
CV = Empirical	1.591	4.93	[2.04, 11.93]	< **0.001**
Participant characteristics				
TiA-PtT	0.021	1.35	[1.09, 1.67]	**0.005**
ATI	-0.006	0.90	[0.74, 1.10]	0.304
Confidence$_{initial}$	-0.044	0.33	[0.29, 0.38]	< **0.001**
AI characteristics				
AI Accuracy	0.167	1.18	[0.95, 1.47]	0.125
Interaction effects				
Political Congruence = Incongruent : CV = Consensual	0.095	1.13	[0.42, 2.99]	0.848
Political Congruence = Incongruent : CV = Expert	0.477	1.64	[0.63, 4.25]	0.325
Political Congruence = Incongruent : CV = Internal	-0.243	0.79	[0.30, 2.09]	0.621
Political Congruence = Incongruent : CV = Empirical	-0.015	1.00	[0.38, 2.60]	0.974
Political Congruence = Neutral : CV = Consensual	0.256	1.32	[0.58, 3.03]	0.544
Political Congruence = Neutral : CV = Expert	0.209	1.26	[0.55, 2.83]	0.616
Political Congruence = Neutral : CV = Internal	-0.480	0.63	[0.27, 1.43]	0.253
Political Congruence = Neutral : CV = Empirical	0.250	1.30	[0.57, 2.94]	0.549
Scientificness = Scientific : CV = Consensual	-0.374	0.67	[0.35, 1.30]	0.266
Scientificness = Scientific : CV = Expert	0.230	1.24	[0.64, 2.38]	0.489
Scientificness = Scientific : CV = Internal	0.434	1.52	[0.79, 2.94]	0.195
Scientificness = Scientific : CV = Empirical	0.168	1.16	[0.63, 2.25]	0.617

IDs as a random effect in our statistical model. This accounted for individual differences as well as any correlation amongst repeated measurements from the same participant.

We conducted a likelihood ratio test between our final model and the null model to ascertain the goodness of fit [8]. Our model provides a significantly better fit to the data compared to the null model ($\chi^2 = 423.6$, $p < 0.001$) and accounts for 44.9% of the variance in switching behaviour ($R^2 = 0.449$).

We observed significant main effects of conceptualisation validations (CVs) **(RQ1)** (Figure 5) and participants' initial confidence **(RQ3)** (Figure 6a, 6b) on switching behaviour. Participants' TiA-PtT **(RQ3)** (Cronbach's $\alpha = 0.75$) also had a main influence on switching behaviour (Figure 6c). No effects were observed for participants' ATI **(RQ3)** (Cronbach's $\alpha = 0.89$), or the characteristics of headlines **(RQ2)** in this study. In the following sections, we describe these findings in detail.

4.3 The Effect of Conceptualisation Validations (CVs) on Trust

As presented in Table 3, we found a statistically significant difference in switching behaviour between the Control and all treatment conditions combined ($\beta = 0.826$, $SE = 0.220$, $p < 0.001$) **(RQ1)**. Participants were more likely to switch their credibility judgement to align with the AI in the treatment conditions (containing an explanation) compared to the Control (no explanation), with an odds ratio of 2.29 (95% CI between 1.48 and 3.52). These results demonstrate that CV-based explanations increased reliance on the AI's judgement.

We observe that Expert, Internal, and Empirical CV-based explanations were more effective than both Consensual explanations and having no explanation (Control) **(RQ1)**. In the Control

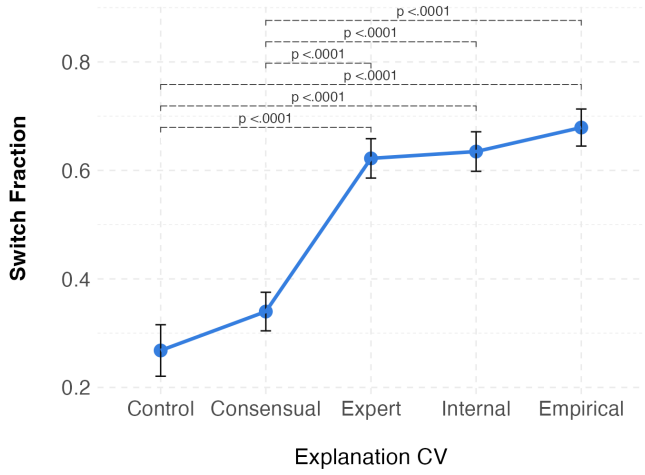


Fig. 5. The effect of the presence and type of explanation on switch fraction. Error bars denote standard error (SE).

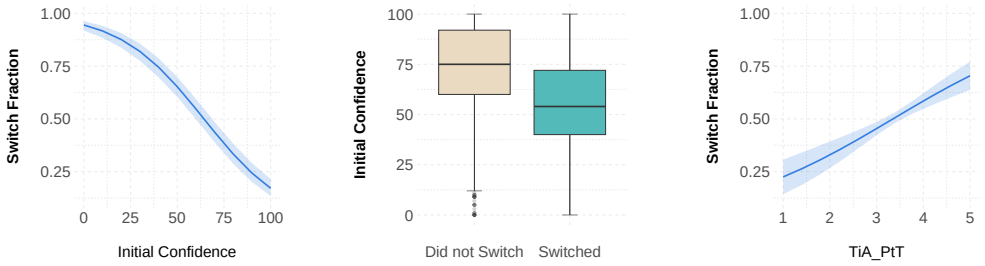
condition, participants switched in 26.8% of instances, seeing an explicit AI judgement with no explanation. However, compared to the Control, greater switching was observed in instances involving CV-based explanations – participants switched 34% of times for Consensual (β = 0.410, SE = 0.458, p = 0.370), 62.2% for Expert (β = 1.158, SE = 0.458, p = 0.010), 63.5% for Internal (β = 1.581, SE = 0.457, p < 0.001), and 68% for Empirical explanations (β = 1.591, SE = 0.450, p < 0.001). In other words, Consensual explanations were the least effective form of explanation supplied to participants, while the other three were almost double as effective. The difference between the relative effectiveness of CVs and the Control is depicted in Figure 5.

Further, we performed a post-hoc analysis by obtaining pairwise contrasts between the different explanation types, and found statistically significant differences when comparing Control and Consensual with the other variants, further highlighting the greater relative effectiveness of Expert, Internal, and Empirical CVs. All explanations except Consensual performed better than Control: Control vs Consensual (p = 0.775), Control vs Expert (β = -1.503, SE = 0.295, p < 0.001), Control vs Internal (β = -1.557, SE = 0.297, p < 0.001), and Control vs Empirical (β = -1.754, SE = 0.297, p < 0.001). Further, Consensual was the least influential CV in causing participants to switch their judgement: Consensual vs Expert (β = -1.163, SE = 0.179, p < 0.001), Consensual vs Internal (β = -1.217, SE = 0.181, p < 0.001), and Consensual vs Empirical (β = -1.413, SE = 0.181, p < 0.001). However, pairwise contrasts revealed no significant differences between Expert and Internal (p = 0.998), Expert and Empirical (p = 0.601), and Internal and Empirical (p = 0.799). Figure 5 illustrates these comparisons.

4.4 Initial Confidence, Trust in Automation, Headline Characteristics, and AI Accuracy

Our results indicate a significant main effect of initial confidence on switching behaviour (β = -0.044, SE = 0.002, p < 0.001) **(RQ3)**. Participants with higher confidence in their own judgement prior to seeing the AI aid were less likely to be impacted by the explanations to switch their judgement to align with the AI's (see Figure 6a). While the initial confidence values of participants ranged between 0-100 during both cases of switching and no switching, we observed a difference in the median values. Participants who did not switch demonstrated a median initial confidence of 75, whereas those who switched displayed a lower median value of 54 (see Figure 6b).

We also observed a main effect of participants' propensity to trust automation (TiA-PtT) on switching behaviour (β = 0.021, SE = 0.007, p = 0.005) **(RQ3)**. Participants with a higher trust in automation were more likely to switch their responses to align with the AI (see Figure 6c). Those



(a) (b) (c)

Fig. 6. (a) The effect of participants' initial confidence on switching behaviour. The shaded area denotes 95% confidence intervals (CI); (b) Distribution of initial confidence when participants did not switch versus when they switched; (c) The effect of participants' Trust in Automation (Propensity to Trust subscale) on switch fraction. The shaded area denotes standard error (SE).

who scored higher on the TiA-PtT scale were 35% more likely to switch, with an odds ratio of 1.35 (95% CI between 1.09 and 1.67).

We manipulated the headline characteristics across two axes: *scientificness* and *political congruence*. We did not find a significant main effect of the headline being scientific, nor of political congruence, on switch fraction in this study **(RQ2)**. We also do not find a significant interaction effect between CVs and headline characteristics.

Lastly, we did not find a significant main effect of AI accuracy on switching behaviour, suggesting that whether the AI's credibility judgement aligned with the ground truth did not impact users' decisions to switch their judgement ($p = 0.125$).

4.5 Qualitative Analysis

At the survey's conclusion, participants answered open-ended questions about factors pertaining to the AI-based credibility indicator that may have influenced their decisions to revise or maintain their judgements. We sought insights into how CV-based explanations impacted participants' trust in the AI and their perceptions of it. We systematically coded the responses following a deductive thematic analysis approach [10]. We aimed to understand how different explanation CVs and other factors influenced reliance on the AI. To conduct the deductive thematic analysis, we developed a coding framework based on pre-established themes derived from relevant literature and our research objectives. We first gained a holistic understanding of our data and then identified and labelled segments of participants' responses corresponding to our pre-determined themes. The coding process involved systematically assigning these segments to their respective themes, ensuring consistency in our analysis. In the following sections, we present the five main themes that originated from our analysis.

4.5.1 ***The Impact of Explanations.*** Existing literature highlights the pivotal role of explanations in fostering trust in AI decisions. Our qualitative results reemphasise this impact, as when the AI did not provide explanations, participants were reluctant to trust it; *"I would like the AI to provide the information that allowed it to make a final decision. Without that it's harder to trust it."* (P301). Conversely, explanation compelled participants to rely on the AI; *"[The explanations] helped me switch as they were always very explicit and clear."* (P67), and update their mental models; *"It helped having information cited to you instead of getting an 'inaccurate' or 'accurate' rating. If my reasoning was wrong, it helped to have the information to correct it."* (P72). Explanations also gave rise to anthropomorphic descriptions, such as perceived transparency, honesty, and a lack of deceptive intentions; *"It was nice*

to see how [the AI] came up with answers. It made me feel like it was transparent and honest. […] it was not trying to fool me or hide anything." (P62).

Further, participants highlighted the credibility assessment AI as a tool for promoting deliberation, as it made them pause and critically reflect upon headline accuracy, *"giving it more thought"* (P136). The explanations also reminded participants about the potential fallibility of their judgements; *"It reminded me of the possibility that I could be wrong."* (P83). Lastly, the explanations brought to participants' attention novel information or reasoning approaches; *"[The explanations] presented something I had not thought of."* (P78).

4.5.2 ***How Initial Confidence affects Reliance.*** Mirroring the findings of research on individuals' confidence and their likelihood of conforming to automated advice, we found that low initial confidence led to more switching. Participants mentioned their lack of headline knowledge as a precursor to low confidence and as an indicator of greater reliance on the AI; *"Where I didn't know too much about the subject, I may have deferred to the AI a little more."* (P1). In contrast, when participants were confident in their assessment, they were less likely to accept the AI's advice; *"[I did not switch] because I was very confident in my initial answer."* (P221).

4.5.3 ***Explanation Conceptualisation Validations (CVs).*** We identified **Consensual** explanations to be the least effective form of explanation, a finding also salient in our quantitative results; *"In some cases the AI's reasoning was based on how many people prior answered that way instead of giving details on the headline. In these cases I stuck with my initial judgement."* (P60). Participants being unaware of the knowledge of other respondents was the primary reason for reduced trust in Consensual explanations; *"[…] it did not increase my confidence because I do not know the knowledge of the participants on any of the subjects. It's possible they have no accurate information and are guessing."* (P162).

Conversely, some participants felt positive about Consensual explanations and were more likely to rely on the AI if it presented an opposing opinion of the majority; *"[I switched] particularly when it gave percentages of other users who agreed or disagreed with it. I felt if a large percentage voted one way then that was most likely the correct answer."* (P113).

A significant majority of our participants highlighted the greater effectiveness of the other three CVs – namely Expert, Internal, and Empirical – compared to Consensual; *"If the AI mentioned statistics that showed how its reasoning was influenced, or if a specific study was mentioned or if a doctor or other professional was cited, it was more likely to change my judgement. If it just mentioned what other people chose in the study, it didn't really change my mind."* (P116).

For **Expert** explanations, statements from field experts played a major role in persuading participants to switch; *"I tended to agree with the AI when the information was cited as coming from experts in a relevant field."* (P219). Further, Expert explanations also instilled greater trust in the AI for scientific headlines; *"I was more likely to trust the AI when it provided explanations by researchers or health professionals."* (P312), and; *"The inclusion of statements from health professionals [made me switch]."* (P242).

For **Internal** explanations, participants communicated the importance of being able to gauge the logical (in)consistency of claims, both for credibility assessment and in enhancing their trust in the AI; *"[…] logical arguments were my main motivating factors when switching answers."* (P155). Such explanations provided an alternative rationale that participants had not considered; *"Usually when the AI […] suggested a more plausible rationale than I had been using – I considered switching my answer."* (P214).

Lastly, most participants found the statistical evidence included in **Empirical** explanations to be influential and informative; *"I [switched] more when specific statistics, facts, theories and studies were cited."* (P26). Reflecting our quantitative findings, Empirical explanations were more convincing than others; *"When the AI had cold hard facts—that's when I listened. Otherwise it's just guessing."* (P199). Factual information also increased trust in the AI; *"Evidence increased my level of trust because there*

was a foundation for the AI influence, not just a statement of fact." (P16), and made the AI appear more trustworthy; *"When the AI provided specific data points I thought it was more trustworthy."* (P162).

4.5.4 ***The Impact of AI Judgement on Final Confidence.*** While some participants maintained their judgements, they experienced reduced confidence after the AI disagreed with them. This suggests that when the AI could not sway judgements, it did induce some uncertainty; *"I don't think I switched much, I just lowered the percentage of my confidence."* (P98). Further, some participants only changed their confidence levels, not the overall judgement; *"None of them made me completely change my opinion, but it did increase or decrease my confidence level."* (P202). Interestingly, some participants experienced decreased confidence in their new judgement after switching; *"If the AI thought the answer was different than mine I would switch sides, but be less confident about my answers."* (P180). These qualitative results align with our quantitative findings. Conversely, the AI mirroring participants' judgement served as a confident booster; *"If I already believed it to be true or not and the AI confirmed my choice, then my confidence increased."* (P54), and *"[...] the AI agreeing made me more confident in my opinion."* (P72).

4.5.5 ***Bias Towards AI and Aversion to Being Fact-checked.*** Automation **bias** can manifest when users excessively rely on the AI's judgements without critically evaluating them. Some participants unquestionably accepted the AI's advice due to the perception that AI systems are highly accurate and superior; *"I always assumed the AI would give me correct answers based off of factual information, so I changed my mind because of that."* (P64).

Conversely, some participants exhibited an **aversion** towards automation, wherein they were hesitant to fully trust the AI irrespective of the accuracy of its explanations; *"AI is great but I still don't have complete trust in it."* (P158), and *"[The explanations] didn't really affect how much I trust AI. I don't really trust AI, and I would rather just rely on my own judgement [...]"* (P116). A few participants blamed their aversion on the AI gathering information from the internet where inaccurate information exists; *"The AI is picking up information off the internet and there is a lot of misinformation out there."* (P16).

Some participants were also reluctant to be fact-checked by the AI because they perceived it to be politically biased. A few participants, all identifying as Republicans, disregarded the AI's advice on the grounds of it being liberal and biased towards the left of the political spectrum; *"I did not switch. [...] I know it is a FACT any AI involved in fact-checking will be liberally biased. Or manipulatively spin context into the next galaxy."* (P118), and *"I generally stayed with my initial judgement. I think the AI was biased to the left."* (P47). We note that we had an equal number of headlines with Republican and Democrat congruent and incongruent themes, as well as non-political themes, seen by all participants. As the AI judged headlines as being inaccurate or accurate in an equally random manner for all headlines and participants, its judgements were not more favourable towards one end of the political spectrum.

5 Discussion

5.1 Explanations Induce Greater Trust in a Fact-checking AI

Existing literature has long emphasised the critical role of explanations in fostering trust in automated decisions [61, 124]. However, fact-checking is notoriously tricky, misconceptions are hard to revert, and people tend to prefer information that aligns with their long-held beliefs to avoid cognitive dissonance [30, 109]. Arguably, automated fact-checking aids must do more persuasive 'work' than support aids in other contexts, such as agricultural productivity and medical diagnoses, because they must be compelling enough to fight against users' political biases and misconceptions, enabling them to overcome motivated reasoning. Therefore, we set out to investigate how to best design explanations to promote reliance and trustworthiness in a fact-checking AI. We designed model-agnostic, natural language explanations woven around the four Conceptualisation Validations (CVs) [46] – Consensual, Expert, Internal, and Empirical.

Our results demonstrate the effectiveness of detailed explanations in building trust in AI, even in a fact-checking context. Participants trusted the AI more after reading its explanations and preferred it over an explicit binary judgement **(RQ1)** because explanations helped augment their reasoning, as shown in our qualitative results. When AI judgements lacked explanations, participants switched in only 26.8% of cases, indicating that the absence of explanations made relying on the AI challenging. Notably, our findings align with and extend existing research, which found AI-based credibility indicators to exhibit limited effectiveness [100, 122], either due to the absence of explanations [14, 77], or the provision of explanations that were challenging to comprehend and could not improve users' understanding of the AI's rationale [3, 24, 89]. In contrast, when our AI presented explanations constructed from the different types of evidence humans utilise to validate new concepts, specifically Expert, Internal, or Empirical CV-based explanations, users could understand the AI's reasoning, and followed its recommendations much more frequently (62-68%). Consensual explanations were perceived as the least compelling, albeit still more compelling than no explanation, with participants following these recommendations in 34% of cases. This underscores the vital role of natural language explanations crafted from CVs in shaping human decision-making during credibility assessment, specifically highlighting the potential of Expert, Internal, and Empirical explanations to enhance the effectiveness of AI-based fact-checking aids in countering misinformation.

5.2 CV-based Explanations Influence Reliance Differently

There are various plausible explanations for the observed influence of different CVs in our study **(RQ1)**. **Consensual** validation emphasises the role of social influence in fostering the acceptance of new information. Social influence literature identifies two motivations for individuals to conform to a majority opinion: *normative* influences (conforming to gain approval and fit in) and *informational* influences (conforming to be more correct) [20]. In our study, when Consensual explanations were provided, participants encountered a dissenting AI that based its decision on the (simulated) judgement of the majority of other survey takers. It is plausible that those who aligned their judgement with the AI might have been subject to informational influence, a notion supported by our qualitative findings. Research on online social conformity reports a conformity rate of around 30-33% [119, 121], which is close to the switching rate we observed for Consensual explanations (34%). We note that we did not incentivise participants' accuracy during our experiment, so this desire to be 'right' may also persist on social media platforms deploying Consensual explanations, and further efforts are required to closely examine its influence.

In addition, our participants did not have an estimate of the knowledge of other survey takers and, as a result, could not gauge the trustworthiness of their judgement. This resonates with our qualitative findings and existing literature, which highlight the influence of source knowledge and credibility in influencing beliefs [22, 76]. The expertise of a source of information is often a heuristic that people rely on to ascertain trustworthiness in the absence of other information [71]. Because participants lacked knowledge about other survey takers' expertise, they more frequently chose to maintain their original judgements instead of trusting the AI's decision. Nevertheless, from our qualitative analysis, we found that the inclusion of explanations—even Consensual ones—motivated individuals to pause and critically re-examine the headline presented. This suggests that such an intervention can promote a more deliberative information-assessment habit on social media platforms, even among individuals who may initially be reluctant to engage in such practices, and should not be dismissed as entirely ineffective. Future work can investigate scenarios where Consensual explanations can offer greater utility, such as in cases where the public perception of a claim might be of greater importance.

We found **Expert** explanations to cause switching in over 62% of instances when they were displayed. Expert validation is rooted in the trust that individuals place in domain experts. Our findings demonstrate that participants relied on judgements accompanied by statements from field

experts, such as health professionals and agriculturalists, resulting in greater trust in the AI. Notably, people often blur the line between their own knowledge and the knowledge they acquire from their community. This reliance on communal knowledge, known as *epistemic dependence*, plays a crucial role in accepting the unknown [38]. Sloman and Rabb [102] report that "knowing that experts understand a phenomenon gives individuals the sense that they understand it better themselves, but only if they believe they have access to the experts' explanation." In our fact-checking scenario, participants likely relied on expert comments to make epistemic judgements about the credibility of headlines. This reliance on communal knowledge also extends to contentious political matters, where people incorporate others' knowledge in the same manner as their own [29] and could explain why we observed no difference in switching behaviour between politically congruent and incongruent headlines **(RQ2)**. Our findings resonate with prior literature and extend it by showing that such an epistemic dependence persists even when an automated aid acts as the messenger of the expert's knowledge.

Further, **Internal**, or logic-based explanations, made participants switch in over 63% of instances because they often provided participants with an alternate line of reasoning which they had not previously considered. In a similar vein, **Empirical** explanations induced switching in 68% of cases when they were presented. Empirical evidence can be perceived as more reliable as it presents concrete data rather than subjective opinions and is less influenced by personal biases.

These findings present two important implications for the design of fact-checking AI aids. On the one hand, they highlight the potential of CV-based explanations to foster trust in a fact-checking AI and present credibility assessments on social media platforms, overcoming the scalability problem of manual interventions. Users appreciated the informational nature of the explanations and could understand how the AI arrived at its decision, which increased their trust in the AI.

On the other hand, we also found that individuals could not detect whether the AI was guiding them towards the truth, leading them to indiscriminately align with the AI irrespective of its accuracy. These results support the findings of Lu et al. [63], who identified a similar indiscriminate reliance on AI-based credibility indicators, albeit in a context involving social influence and without presenting any explanations for the AI's judgements. This underscores the dual nature of AI-based credibility indicators, showcasing their potential to guide individuals towards the truth while also posing the risk of misleading them and contributing to misinformation. However, addressing this challenge requires recognising the two-stage nature of fact-checking: initially classifying information as true or false, and subsequently communicating this decision to end-users. Existing research often focuses on the first stage—detecting misinformation—while our study contributes insights to the second stage. Once the veracity of news has been confidently assessed, our results showcase effective ways to present these verified credibility outcomes to end-users in a manner that resonates with their understanding and reasoning processes. This involves designing tailored explanations that provide expert opinions, logical reasoning, or scientific data, bridging the gap between machine explanation and user comprehension. To further promote responsible use, platforms providing additional information such as AI accuracy metrics or confidence scores [64] can empower users to make informed judgements and mitigate the potential misuse of these indicators. Together, these findings illuminate the intricate dynamics of user interactions with automated fact-checking aids, underscoring the pivotal role of explanations and highlighting the pressing need to assist individuals in effectively assessing the accuracy of AI-based credibility indicators.

5.3 Automation Bias, Overreliance, and Automation Aversion Impact the Effectiveness of AI-based Fact-Checking

5.3.1 Automation Bias. In the context of AI-based credibility assessment, **automation bias** can occur when users overly rely on AI judgements without critically evaluating them. Our study revealed that some participants were subject to automation bias. Further, our results show that a higher

propensity to trust automation increases one's likelihood of switching to the AI judgement **(RQ3)**. Through our qualitative findings, we observe that this behaviour stemmed from users' belief that AI systems are highly accurate and capable, and this perception manifested in various ways.

First, explanations remained influential despite the absence of corroborating sources. It is worth noting that Expert explanations neither named a specific expert nor supplied a source to support the expert's comment. Despite this, participants switched their judgement in over 62% of cases involving an Expert explanation. Similarly, Empirical explanations did not provide a source for the supporting data or any identifying attributes of the organisation conducting the research to lend it credibility. Nevertheless, our participants switched to align with the AI in 68% of cases presenting Empirical explanations. However, when the AI based its decision on the opinion of one's peers, the rate of switching dramatically decreased. This indicates that participants had lower trust in their peers compared to the trust they placed in the AI's explanations derived from online sources, even in cases where the sources were not provided and the explanations were, therefore, not verifiable.

Second, participants' confidence in their judgements increased when the AI agreed with them — regardless of the AI accuracy. This finding contradicts the results of Lu et al. [63], who observed that an agreeing fact-checking AI could not bolster users' confidence. Critically, their participants were subject to social influence, could see the veracity of the judgements of participants preceding them, and were shown no explanations. As such, we hypothesise that the presence of explanations in our study provided additional supporting information, thereby increasing participants' confidence in their assessment. It is thus an essential line of future work to examine how (in)accurate agreements with a fact-checking AI augment the perception of *correct* news, in addition to misinformation, and how contextual factors such as social influence impact this phenomenon.

Third, the perceived superiority of the AI's credibility decisions, coupled with the informational nature of its explanations, superseded the impact of participants' political biases. We found that the political congruence of headlines had a negligible impact on switching behaviour. This suggests that participants were similarly persuaded to align their judgement with the AI, irrespective of whether the headline was attitude-affirming, attitude-challenging, or altogether non-political **(RQ2)**. Our manipulation check further validates this finding by confirming that political congruence and incongruence did manifest in our experiment.

Prior work highlights several possible factors causing this phenomenon. Both source expertise and trustworthiness serve as heuristics for individuals to evaluate source credibility [71], and messages from highly credible sources are more believable than those from sources with lower perceived credibility [2, 116]. Consequently, traditional fact-checking interventions are often limited in their ability to persuade belief change for political news if the source of credibility information is perceived as untrustworthy and unreliable [16, 48, 86]. This strong influence of source perceptions on participants' willingness to follow advice is also highly prevalent in the case of political misinformation—such as that surrounding public health [114]—a category also examined in this work.

Furthermore, studies have shown that the impact of source perceptions extends to cases where an AI serves as the information source, as individuals influenced by automation bias tend to regard algorithmic decisions as superior to those made by humans [33, 73, 75]. We hypothesise that this bias led our participants to perceive the AI as a knowledgeable, expert entity, and its credibility decisions as accurate, as evident in our qualitative results. Importantly, the factual and informational nature of explanations likely enhanced the perceived objectiveness of the AI, reinforcing participants' view of the AI as a reliable source of credibility information. Participants valued the AI's effort to reason with them by providing expert opinions, logical arguments, or empirical evidence, and in a manner that helped update their decision-making accordingly. Thus, the compelling influence of CV-based explanations likely outweighed the influence of participants' partisan attitudes. Overall, participants' trust in the AI's judgement, regardless of political congruence, suggests that well-crafted

and human-centred fact-checking explanations could effectively tackle the challenges associated with partisan attitudes in conventional fact-checking interventions.

5.3.2 Overreliance. Our participants exhibited unwarranted trust and overreliance on the AI because they believed it to be comprehensive, gathering credible information, and making accurate judgements, as suggested by our qualitative results. This raises the question of how including sources in explanations would influence reliance, if users already find themselves compelled to switch even without explicit sources. It is plausible that the addition of external sources to corroborate explanations may promote more appropriate reliance. However, prior work finds that people tend to engage with online news based only on the headline, often disregarding the source entirely [50, 65]. Empirical evidence also suggests the inverse — how an individual perceives the news source can influence their belief in it [106]. In our study, participants relied on unsubstantiated explanations, highlighting how the authoritative influence of a fact-checking AI can cause overreliance, in the absence of any AI accuracy metrics or explanation sources. We call upon future work to systematically examine whether and how to include sources in AI explanations to promote informed decision-making and further steer users towards appropriate reliance.

It is noteworthy that we observed indistinguishable reliance on AI when it was accurate compared to when it was not, emphasising the potential risk of over-reliance. Buçinca et al. [12] propose three cognitive forcing functions to reduce over-reliance on an explainable AI, to be implemented during the decision-making stage to disrupt heuristic reasoning and promote analytical thinking. Our study design implemented two of these three cognitive forcing functions. Namely asking the user to make an independent judgement before seeing AI advice and delaying presenting AI advice until after the decision-making task has been presented. Despite this, we observed over-reliance on the AI, highlighting the pitfalls of explanations [23] and the potential role of automation bias in fostering indiscriminate trust in AI. To address these challenges, future research should explore ways of promoting appropriate reliance and mitigating inappropriate reliance on fact-checking AIs. This includes investigating the effectiveness of Buçinca et al. [12]'s third cognitive forcing function—offering users the agency to choose when to see AI advice.

5.3.3 Automation Aversion. We also found some participants who exhibited **aversion** towards being fact-checked by the AI. They were hesitant to trust the AI and disregarded its judgements. Additionally, when challenged by the AI, some who did not switch their judgement increased their confidence in it, doubling down to defend it. This behaviour parallels the '*boomerang*' effect observed in traditional corrective efforts against misinformation [13]. This phenomenon can be triggered by individuals engaging in psychological rebellion, wherein they perceive fact-checking messages as a threat to their intellectual abilities and core beliefs, actively embracing incorrect beliefs with greater intensity [11, 13]. In a real-world scenario, corrections provided by a fact-checking AI could potentially reinforce the misguided beliefs of individuals who already distrust AI. Investigating the impact of a fact-checking AI on the entrenched beliefs of those who exhibit scepticism towards AI can provide valuable insights into understanding the dynamics of trust and belief formation.

Finally, we did not disclose the accuracy of our AI-based credibility indicator to the users. However, it is crucial to provide such metrics to promote appropriate reliance on an AI system. Interestingly, prior research suggests that users' trust in an AI is influenced more by its observed accuracy during interaction rather than its stated accuracy [123]. Future research endeavours can explore how users calibrate their trust in a fact-checking AI when presented with performance cues, allowing for a deeper understanding of trust dynamics in such a fact-checking context.

5.4 Limitations and Future Work

Our study evaluated the effectiveness and perception of a fact-checking AI using a sample of U.S.-based Prolific workers who identified as either Republicans or Democrats. Rather than analysing the absolute impact of political orientations during our investigation, we adopted a more widely applicable notion of *congruence* between an individual's political orientation and the slant of a headline. This broader perspective increases the likelihood of our findings applying to a broader spectrum of political stances beyond Republicans and Democrats. Nevertheless, our findings may not fully represent the attitudes and behaviours of a more diverse population or individuals from different cultural backgrounds, including non-partisans. Therefore, further research is necessary to assess the generalisability of our findings to participants with different demographics and cultural contexts.

We also acknowledge that a binary credibility judgement cannot capture the full spectrum of accuracy. However, we opted for this approach, recognising that a binary categorisation is commonly employed in misinformation research [47, 63]. Future studies need to explore methods that can capture a broader spectrum of information credibility, investigating headlines that are partially accurate, missing context, or exaggerated.

We also identify the limitation posed by measuring participants' confidence in their judgement using a single-item question, and using it as a proxy for their prior knowledge. Future work can utilise different questionnaires to better measure such constructs. Additionally, future work can consider participant demographics in their analyses, as they may influence how individuals respond to fact-checking, for example, by influencing their understanding of news feed algorithms of platforms such as Facebook [89].

Furthermore, our headlines encompassed scientific, political, both, or unrelated topics. While these categories enabled a broad spectrum of headline themes to be studied, individuals in real-world scenarios can encounter diverse topics, such as sports or entertainment. Future research should investigate CV-based explanations for headlines with greater thematic diversity.

Finally, our study did not investigate the long-term effects of engaging with fact-checking AI or the potential influence of habituation over time. Participants' limited exposure to the AI might have influenced their trust and decision-making tendencies, and the novelty of interacting with a fact-checking AI could have impacted their perceptions. Future research could adopt a longitudinal approach to examine trust dynamics more comprehensively. Additionally, future work can also delve into the temporal aspects of participants' evolving relationship with the AI, investigating how past reliance on the AI influences reliance during subsequent interactions.

6 Conclusion

In this study, we examine how trust in a fact-checking AI can be influenced by model-agnostic, natural language explanations constructed from the different types of evidence that humans utilise to validate new concepts. We find that providing explanations significantly boosts reliance on an AI, even in the context of fact-checking where influencing prior beliefs is often challenging. Our results reveal varying impacts of different CVs, with Consensual explanations having the least impact, and Expert, Internal, and Empirical explanations being almost twice as influential.

These findings underscore the potential of CVs as a means to foster trust in fact-checking AIs on social media platforms. However, we also find individuals being unable to discern whether the AI led them towards the truth, emphasising the dual nature of such influential indicators to both guide and mislead users. Additionally, we find headline characteristics (scientificness and political congruence) do not significantly impact reliance, indicating that belief in both attitude-congruent and incongruent headlines can be influenced by a fact-checking AI offering explanations. Interestingly, we observe both automation bias and aversion manifest during collaborative fact-checking, moderating uptake of

the AI's advice. Individuals with a greater dispositional trust in AI perceived it as superior and relied heavily on its (in)accurate judgements, while some demonstrated reluctance to rely on it irrespective of accuracy, perceiving the AI as incompetent or biased. In these cases, corrections often backfired, compelling individuals to embrace their prior beliefs with increased vigour when challenged by the AI. We offer nuanced insights into the dynamics of user behaviours during collaborative fact-checking with an AI and discuss how explanations augment the perception of such an AI. Overall, our results provide important lessons for social media platforms in designing adequate explanations for AI-based fact-checking approaches to combat misinformation while highlighting their pitfalls.

References

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–236. https://doi.org/10.1257/jep.31.2.211

[2] Michelle Amazeen and Arunima Krishna. 2020. Correcting Vaccine Misinformation: Recognition and Effects of Source Type on Misinformation via Perceived Motivations and Credibility. https://doi.org/10.2139/ssrn.3698102

[3] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445736

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445717

[6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67 (Oct. 2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[7] Sirisha Bojjireddy, Soon Ae Chun, and James Geller. 2021. Machine Learning Approach to Detect Fake News, Misinformation in COVID-19 Pandemic. In *DG.O2021: The 22nd Annual International Conference on Digital Government Research (DG.O'21)*. Association for Computing Machinery, New York, NY, USA, 575–578. https://doi.org/10.1145/3463677.3463762

[8] Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24, 3 (March 2009), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

[9] Claus Bossen and Kathleen H. Pine. 2023. Batman and Robin in Healthcare Knowledge Work: Human-AI Collaboration by Clinical Documentation Integrity Specialists. *ACM Transactions on Computer-Human Interaction* 30, 2 (March 2023), 26:1–26:29. https://doi.org/10.1145/3569892

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[11] Jack W. Brehm. 1966. *A theory of psychological reactance.* Academic Press, Oxford, England.

[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 188:1–188:21. https://doi.org/10.1145/3449287

[13] Sahara Byrne and Philip Solomon Hart. 2009. The Boomerang Effect A Synthesis of Findings and a Preliminary Theoretical Framework. *Annals of the International Communication Association* 33, 1 (Jan. 2009), 3–37. https://doi.org/10.1080/23808985.2009.11679083

[14] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 136:1–136:21. https://doi.org/10.1145/3579612

[15] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 471:1–471:23. https://doi.org/10.1145/3555572

[16] Santanu Chakrabarti, Lucile Stengel, and Sapna Solanki. 2018. Duty, identity, credibility: Fake news and the ordinary citizen in India. *BBC World Service Audiences Research* (2018).

[17] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 148–161. https://doi.org/10.1145/3490099.3511121

[18] Katherine Clayton, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. 2020. 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* 42, 4 (Dec. 2020), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

[19] Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics* 12, 2 (May 2020), 459–478. https://doi.org/10.1007/s12369-019-00596-x

[20] Morton Deutsch and Harold B. Gerard. 1955. A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology* 51 (1955), 629–636. https://doi.org/10.1037/h0046408

[21] Graham N. Dixon, Brooke Weberling McKeever, Avery E. Holton, Christopher Clarke, and Gina Eosco. 2015. The Power of a Picture: Overcoming Scientific Misinformation by Communicating Weight-of-Evidence Information with Visual Exemplars. *Journal of Communication* 65, 4 (2015), 639–659. https://doi.org/10.1111/jcom.12159 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcom.12159.

[22] Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* 1, 1 (Jan. 2022), 13–29. https://doi.org/10.1038/s44159-021-00006-y

[23] Upol Ehsan and Mark O. Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. https://doi.org/10.48550/arXiv.2109.12480

[24] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. Do Explanations Increase the Effectiveness of AI-Crowd Generated Fake News Warnings? *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May 2022), 183–193. https://doi.org/10.1609/icwsm.v16i1.19283

[25] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376232

[26] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For What It's Worth: Humans Overwrite Their Economic Self-interest to Avoid Bargaining With AI Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3491102.3517734

[27] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300724

[28] European Union. 2016. General Data Protection Regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[29] Philip Fernbach and Steven Sloman. 2017. Opinion | Why We Believe Obvious Untruths. *The New York Times* (March 2017). https://www.nytimes.com/2017/03/03/opinion/sunday/why-we-believe-obvious-untruths.html

[30] Leon Festinger. 1962. *A Theory of Cognitive Dissonance.* Stanford University Press.

[31] Thomas Franke, Christiane Attig, and Daniel Wessel. 2018. *A personal resource for technology interaction: Development and validation of the Affinity for Technology Interaction (ATI) scale.*

[32] R. Kelly Garrett, Erik C. Nisbet, and Emily K. Lynch. 2013. Undermining the Corrective Effects of Media-Based Political Fact Checking? The Role of Contextual Cues and Naïve Theory. *Journal of Communication* 63, 4 (Aug. 2013), 617–637. https://doi.org/10.1111/jcom.12038

[33] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2014. Automation bias: empirical results assessing influencing factors. *International Journal of Medical Informatics* 83, 5 (May 2014), 368–375. https://doi.org/10.1016/j.ijmedinf.2014.01.001

[34] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 90–99. https://doi.org/10.1145/3287560.3287563

[35] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 50:1–50:24. https://doi.org/10.1145/3359152

[36] Ben Green and Yiling Chen. 2020. Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts. https://arxiv.org/abs/2012.05370v2

[37] Md Mahfuzul Haque, Mohammad Yousuf, Ahmed Shatil Alam, Pratyasha Saha, Syed Ishtiaque Ahmed, and Naeemul Hassan. 2020. Combating Misinformation in Bangladesh: Roles and Responsibilities as Perceived by Journalists, Fact-checkers, and Users. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 130:1–130:32. https://doi.org/10.1145/3415201

[38] John Hardwig. 1985. Epistemic Dependence. *The Journal of Philosophy* 82, 7 (July 1985), 335–349. https://doi.org/10.2307/2026523
[39] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581025
[40] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User Trust in Intelligent Systems: A Journey Over Time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. Association for Computing Machinery, New York, NY, USA, 164–168. https://doi.org/10.1145/2856767.2856811
[41] Benjamin D. Horne, Dorit Nevo, Sibel Adali, Lydia Manikonda, and Clare Arrington. 2020. Tailoring heuristics and timing AI interventions for supporting news veracity assessments. *Computers in Human Behavior Reports* 2 (Aug. 2020), 100043. https://doi.org/10.1016/j.chbr.2020.100043
[42] Benjamin D. Horne, Dorit Nevo, John O'Donovan, Jin-Hee Cho, and Sibel Adalı. 2019. Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone? *Proceedings of the International AAAI Conference on Web and Social Media* 13 (July 2019), 247–256. https://doi.org/10.1609/icwsm.v13i01.3226
[43] Jason L. Huang, Paul G. Curran, Jessica Keeney, Elizabeth M. Poposki, and Richard P. DeShon. 2012. Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology* 27, 1 (March 2012), 99–114. https://doi.org/10.1007/s10869-011-9231-8
[44] Shigao Huang, Jie Yang, Simon Fong, and Qi Zhao. 2020. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters* 471 (Feb. 2020), 61–71. https://doi.org/10.1016/j.canlet.2019.12.007
[45] Elle Hunt. 2017. 'Disputed by multiple fact-checkers': Facebook rolls out new alert to combat fake news. *The Guardian* (March 2017). https://www.theguardian.com/technology/2017/mar/22/facebook-fact-checking-tool-fake-news
[46] James Jaccard and Jacob Jacoby. 2019. *Theory Construction and Model-Building Skills: A Practical Guide for Social Scientists*. Guilford Publications.
[47] Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–27. https://doi.org/10.1145/3544548.3581219
[48] Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. 2021. Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 18:1–18:42. https://doi.org/10.1145/3449092
[49] Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 524:1–524:40. https://doi.org/10.1145/3555637
[50] Maurice Jakesch, Moran Koren, Anna Evtushenko, and Mor Naaman. 2018. The Role of Source, Headline and Expressive Responding in Political News Evaluation. https://doi.org/10.2139/ssrn.3306403
[51] Thomas J. Johnson, Barbara K. Kaye, Shannon L. Bichard, and W. Joann Wong. 2007. Every Blog Has Its Day: Politically-interested Internet Users' Perceptions of Blog Credibility. *Journal of Computer-Mediated Communication* 13, 1 (Oct. 2007), 100–122. https://doi.org/10.1111/j.1083-6101.2007.00388.x
[52] S Mo Jones-Jang and Yong Jin Park. 2023. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication* 28, 1 (Jan. 2023), zmac029. https://doi.org/10.1093/jcmc/zmac029
[53] N.A. Karlova and K.E. Fisher. 2013. A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research* 18 (Jan. 2013).
[54] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 163:1–163:26. https://doi.org/10.1145/3415234
[55] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. https://doi.org/10.1109/VLHCC.2013.6645235
[56] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation.
[57] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590
[58] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (March 2018), 1094–1096. https://doi.org/10.1126/science.aao2998

[59] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
[60] Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The limits of human predictions of recidivism. *Science Advances* 6, 7 (2020), eaaz0652. https://doi.org/10.1126/sciadv.aaz0652
[61] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 408:1–408:45. https://doi.org/10.1145/3479552
[62] Yang Liu and Yi-Fang Wu. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (April 2018). https://doi.org/10.1609/aaai.v32i1.11268
[63] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 461:1–461:27. https://doi.org/10.1145/3555562
[64] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445562
[65] Long Ma, Dion Goh, and Chei Sian Lee. 2014. Understanding News Sharing in Social Media: An Explanation from the Diffusion of Innovations Theory. *Online Information Review* 38 (Jan. 2014), 598–615.
[66] Elio Masciari, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. 2020. Detecting fake news by image analysis. In *Proceedings of the 24th Symposium on International Database Engineering & Applications (IDEAS '20)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3410566.3410599
[67] Paul Mena. 2020. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet* 12, 2 (2020), 165–183. https://doi.org/10.1002/poi3.214
[68] Meta. 2021. How Meta's third-party fact-checking program works. https://www.facebook.com/facebookmedia
[69] Meta. 2023. About fact-checking on Facebook. https://en-gb.facebook.com/business/help/2593586717571940
[70] Miriam J. Metzger. 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 2078–2091. https://doi.org/10.1002/asi.20672
[71] Miriam J. Metzger, Andrew J. Flanagin, and Ryan B. Medders. 2010. Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication* 60, 3 (Sept. 2010), 413–439. https://doi.org/10.1111/j.1460-2466.2010.01488.x
[72] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric Ragan. 2020. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. https://doi.org/10.48550/arXiv.2007.12358
[73] Maria D. Molina and S. Shyam Sundar. 2022. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society* (June 2022), 14614448221103534. https://doi.org/10.1177/14614448221103534
[74] Monika Bickert. 2019. Combatting Vaccine Misinformation. https://about.fb.com/news/2019/03/combatting-vaccine-misinformation/
[75] Kathleen L. Mosier and Linda J. Skitka. 1999. Automation Use and Automation Bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 43, 3 (Sept. 1999), 344–348. https://doi.org/10.1177/154193129904300346
[76] Lena Nadarevic, Rolf Reber, Anne Josephine Helmecke, and Dilara Köse. 2020. 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* 5, 1 (Nov. 2020), 56. https://doi.org/10.1186/s41235-020-00251-4
[77] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 189–199. https://doi.org/10.1145/3242587.3242666
[78] Anne Oeldorf-Hirsch, Mike Schmierbach, Alyssa Appelman, and Michael P. Boyle. 2020. The Ineffectiveness of Fact-Checking Labels on News Memes and Articles. *Mass Communication and Society* 23, 5 (Sept. 2020), 682–704. https://doi.org/10.1080/15205436.2020.1733613
[79] Harikumar Pallathadka, Malik Jawarneh, Domenic Sanchez, Guna Sajja, Sanjeev Gour, and Mohd Naved. 2021. The Impact of Machine Learning on Management, Healthcare, and Agriculture. (July 2021).
[80] Sungkyu Park, Jamie Yejean Park, Hyojin Chin, Jeong-han Kang, and Meeyoung Cha. 2021. An Experimental Study to Understand User Experience and Perception Bias Occurred by Fact-checking Messages. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2769–2780. https://doi.org/10.1145/3442381.3450121
[81] Niccolò Pescetelli, Anna-Katharina Hauperich, and Nick Yeung. 2021. Confidence, advice seeking and changes of mind in decision making. *Cognition* 215 (Oct. 2021), 104810. https://doi.org/10.1016/j.cognition.2021.104810

[82] Pew Research Center. 2014. Political Polarization in the American Public. https://www.pewresearch.org/politics/2014/06/12/section-4-political-compromise-and-divisive-policy-debates/
[83] Pew Research Center. 2019. In a Politically Polarized Era, Sharp Divides in Both Partisan Coalitions. https://www.pewresearch.org/politics/2019/12/17/7-domestic-policy-taxes-environment-health-care/
[84] Pew Research Center. 2020. America is exceptional in the nature of its political divide. https://www.pewresearch.org/short-reads/2020/11/13/america-is-exceptional-in-the-nature-of-its-political-divide/
[85] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A Credibility Lens for Analyzing and Explaining Misinformation. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 155–158. https://doi.org/10.1145/3184558.3186967
[86] Chanthika Pornpitakpan. 2004. The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence. *Journal of Applied Social Psychology* 34, 2 (2004), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x
[87] Andrew Prahl and Lyn Van Swol. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36, 6 (2017), 691–702. https://doi.org/10.1002/for.2464
[88] Julia M. Puaschunder, Josef Mantl, and Bernd Plank. 2020. Medicine of the Future: the Power of Artificial Intelligence (AI) and Big Data in Healthcare. https://doi.org/10.2139/ssrn.3607616
[89] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677
[90] Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/2702123.2702174
[91] David N. Rapp and Nikita A. Salovich. 2018. Can't We Just Disregard Fake News? The Consequences of Exposure to Inaccurate Information. *Policy Insights from the Behavioral and Brain Sciences* 5, 2 (Oct. 2018), 232–239. https://doi.org/10.1177/2372732218785193
[92] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22)*. Association for Computing Machinery, New York, NY, USA, 223–233. https://doi.org/10.1145/3503252.3531311
[93] Yoel Roth and Nick Pickles. 2020. Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information
[94] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3132847.3132877
[95] Emily Saltz, Soubhik Barari, Claire Leibowicz, and Claire Wardle. 2021. Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School Misinformation Review* (Oct. 2021). https://doi.org/10.37016/mr-2020-81
[96] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3411763.3451807
[97] Luis Sanz-Menéndez and Laura Cruz-Castro. 2019. The credibility of scientific communication sources regarding climate change: A population-based survey experiment. *Public Understanding of Science* 28, 5 (July 2019), 534–553. https://doi.org/10.1177/0963662519840946
[98] Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422. https://doi.org/10.1145/3581641.3584066 arXiv:2302.02187 [cs].
[99] Ron Sellers. 2013. How Sliders Bias Survey Data. *MRA's Alert* 53, 3 (2013), 56–57. https://greymatterresearch.com/wp-content/uploads/2019/09/Alert-Sliders-2013.pdf
[100] Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 265–274. https://doi.org/10.1145/3292522.3326012
[101] Megha Sharma, Kapil Yadav, Nitika Yadav, and Keith C. Ferdinand. 2017. Zika virus pandemic—analysis of Facebook as a social media health information platform. *American Journal of Infection Control* 45, 3 (March 2017), 301–302. https://doi.org/10.1016/j.ajic.2016.08.022
[102] Steven A. Sloman and Nathaniel Rabb. 2016. Your Understanding Is My Understanding: Evidence for a Community of Knowledge. *Psychological Science* 27, 11 (Nov. 2016), 1451–1460. https://doi.org/10.1177/0956797616662271

[103] Alison Marie Smith-Renner, Styliani Kleanthous Loizou, Jonathan Dodge, Casey Dugan, Min Kyung Lee, Brian Y Lim, Tsvi Kuflik, Advait Sarkar, Avital Shulner-Tal, and Simone Stumpf. 2021. TExSS: Transparency and Explanations in Smart Systems. In *26th International Conference on Intelligent User Interfaces - Companion (IUI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 24–25. https://doi.org/10.1145/3397482.3450705

[104] Elizabeth Solberg, Magnhild Kaarstad, Maren H. Rø Eitrheim, Rossella Bisio, Kine Reegård, and Marten Bloch. 2022. A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids. *Group & Organization Management* 47, 2 (April 2022), 187–222. https://doi.org/10.1177/10596011221081238 Publisher: SAGE Publications Inc.

[105] Brian G. Southwell, J. Scott Babwah Brennen, Ryan Paquin, Vanessa Boudewyns, and Jing Zeng. 2022. Defining and Measuring Scientific Misinformation. *The ANNALS of the American Academy of Political and Social Science* 700, 1 (March 2022), 98–111. https://doi.org/10.1177/00027162221084709 Publisher: SAGE Publications Inc.

[106] Briony Swire, Adam J. Berinsky, Stephan Lewandowsky, and Ullrich K. H. Ecker. 2017. Processing political misinformation: comprehending the Trump phenomenon. *Royal Society Open Science* 4, 3 (March 2017), 160802. https://doi.org/10.1098/rsos.160802

[107] Charles S. Taber and Milton Lodge. 2006. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science* 50, 3 (2006), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

[108] Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. 2022. An Empirical Investigation of Reliance on AI-Assistance in a Noisy-Image Classification Task. In *HHAI2022: Augmenting Human Intellect*. IOS Press, 225–237. https://doi.org/10.3233/FAIA220201

[109] Emily Thorson. 2016. Belief Echoes: The Persistent Effects of Corrected Misinformation. *Political Communication* 33, 3 (July 2016), 460–480. https://doi.org/10.1080/10584609.2015.1102187

[110] Twitter. 2023. How we address misinformation on Twitter. https://help.twitter.com/en/resources/addressing-misleading-info

[111] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 327:1–327:39. https://doi.org/10.1145/3476068

[112] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (March 2018), 1146–1151. https://doi.org/10.1126/science.aap9559

[113] Emily Vraga, Teresa Myers, John Kotcher, Lindsey Beall, and Ed Maibach. 2018. Scientific risk communication about controversial issues influences public perceptions of scientists' political orientations and credibility. *Royal Society Open Science* 5, 2 (Feb. 2018), 170505. https://doi.org/10.1098/rsos.170505

[114] Emily K. Vraga and Leticia Bode. 2018. I do not believe you: how providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society* 21, 10 (Oct. 2018), 1337–1353. https://doi.org/10.1080/1369118X.2017.1313883

[115] Emily K. Vraga and Leticia Bode. 2020. Correction as a Solution for Health Misinformation on Social Media. *American Journal of Public Health* 110, S3 (Oct. 2020), S278–S280. https://doi.org/10.2105/AJPH.2020.305916

[116] Nathan Walter and Riva Tukachinsky. 2020. A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It? *Communication Research* 47, 2 (March 2020), 155–177. https://doi.org/10.1177/0093650219854600

[117] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 1697–1708. https://doi.org/10.1145/3485447.3512240

[118] Senuri Wijenayake, Danula Hettiachchi, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2021. Effect of Conformity on Perceived Trustworthiness of News in Social Media. *IEEE Internet Computing* 25, 1 (Jan. 2021), 12–19. https://doi.org/10.1109/MIC.2020.3032410 Conference Name: IEEE Internet Computing.

[119] Senuri Wijenayake, Jolan Hu, Vassilis Kostakos, and Jorge Goncalves. 2021. Quantifying the Effects of Age-Related Stereotypes on Online Social Conformity. In *Human-Computer Interaction – INTERACT 2021 (Lecture Notes in Computer Science)*. Springer International Publishing, Cham, 451–475. https://doi.org/10.1007/978-3-030-85610-6_26

[120] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2019. Measuring the Effects of Gender on Online Social Conformity. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 145:1–145:24. https://doi.org/10.1145/3359247

[121] Senuri Wijenayake, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. Quantifying the Effect of Social Presence on Online Social Conformity. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 55:1–55:22. https://doi.org/10.1145/3392863

[122] Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376213

[123] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[124] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

Received July 2023; revised January 2024; accepted March 2024