



City Research Online

City, University of London Institutional Repository

Citation: Musi, E., Carmi, E., O'Halloran, K. & Yates, S. (2023). Developing misinformation immunity: how to reason-check fallacious news in a human computer interaction environment. *Social Media + Society*, 9(1), 20563051221150407. doi: 10.1177/20563051221150407

This is the published version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/29724/>

Link to published version: <https://doi.org/10.1177/20563051221150407>

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.


City Research Online:

<http://openaccess.city.ac.uk/>

publications@city.ac.uk

Developing Misinformation Immunity: How to Reason-Check Fallacious News in a Human–Computer Interaction Environment

Elena Musi¹ , Elinor Carmi², Chris Reed³, Simeon Yates¹, and Kay O'Halloran¹

Social Media + Society
January–March 2023: 1–18
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20563051221150407
journals.sagepub.com/home/sms


Abstract

To counter the fake news phenomenon, the scholarly community has attempted to debunk and prebunk disinformation. However, misinformation still constitutes a major challenge due to the variety of misleading techniques and their continuous updates which call for the exercise of critical thinking to build resilience. In this study we present two open access chatbots, the *Fake News Immunity Chatbot* and the *Vaccinating News Chatbot*, which combine Fallacy Theory and Human–Computer Interaction to inoculate citizens and communication gatekeepers against misinformation. These chatbots differ from existing tools both in function and form. First, they target misinformation and enhance the identification of fallacious arguments; and second, they are multiagent and leverage discourse theories of persuasion in their conversational design. After having described both their backend and their frontend design, we report on the evaluation of the user interface and impact on users' critical thinking skills through a questionnaire, a crowdsourced survey, and a pilot qualitative experiment. The results shed light on the best practices to design user-friendly active inoculation tools and reveal that the two chatbots are perceived as increasing critical thinking skills in the current misinformation ecosystem.

Keywords

misinformation, fallacies, chatbots, critical thinking, reason-checking, human–computer interaction

Introduction

One of the major challenges of the current information ecosystem is the rapid spread of misinformation through digital media. Differently from disinformation, misinformation can be misleading despite the intention of its authors/spreaders (Carmi et al., 2020). However, this does not make it less dangerous due to its wide societal impact. For example, according to the RISJ 2020 fact sheet about the types, sources, and claims about COVID-19 information, 59% of fake news contains neither fabricated nor imposter content, but rather reconfigured misinformation (Brennen et al., 2020). Similarly, Allen et al. (2020), through the analysis of a multimode dataset of news consumption in the United States, show that blatantly false fake news constitutes 0.15% of Americans' daily news diet, while misinformation driven by agenda setting in mainstream media is largely understudied and misrepresented, together with news avoidance. As underlined by the newly published *Reuters Digital News*

Report (Newman et al., 2022), the proportion of news avoiders has sharply increased across countries. This includes a significant portion of young people and people with lower educational attainment who blame news media for being hard to follow or understand, especially in cases when the information is de-contextualized or confusing language is used.

Misleading information may be communicated by authoritative sources, such as reputable news media outlets or institutional websites (Kyriakidou et al., 2020; Musi et al., 2022). Thus, citizens' skills in assigning trust values

¹University of Liverpool, UK

²City, University of London, UK

³University of Dundee, UK

Corresponding Author:

Elena Musi, Department of Communication and Media, University of Liverpool, 19 Abercromby Square, Liverpool L69 7ZG, UK.
Email: elena.musi@liverpool.ac.uk



to mainstream media sources, hyperpartisan ones, and fake news websites, despite appearing quite well developed in experimental environments (Pennycook & Rand, 2019), are not well enough developed to identify misinformation across different media platforms and in everyday interactions. The situation is exacerbated in today's networked society where information is repurposed from one application to another or centralized among other sources through information aggregators. Furthermore, in a context of epistemological uncertainty, such as the pandemic or the Ukrainian War, besides *authority* and *objectivity*, three other criteria are recommended for information evaluation (Metzger, 2007)—*accuracy*, *currency* (whether the information is up-to-date) and *coverage* (comprehensiveness of the information provided). This requires additional effort when attempting to verify information.

Overall, this makes it difficult for people to identify misinformation and distinguish trustworthy from misleading news within a social media message and leaves citizens vulnerable. As a result, misperceptions have caused significant downstream consequences across multiple domains. For example, in health they have prevented the timely adoption of measures and treatments to counter the epidemic (Freed et al., 2010; Starbird et al., 2020). In relation to climate change, they cause/sustain climate damaging behaviors (McCright & Dunlap, 2011); and in political decisions, where they have helped shape justifications for wars such as the invasion of Iraq (Kull et al., 2003).

Unfortunately, the identification of misinformation is far from being successfully addressed by human fact-checkers, let alone automatic ones where the lack of a common “truth barometer” hinders the creation of datasets to train automatic systems. Thus, debunking through fact-checking, involving the post hoc correction of misleading content circulating through digital media, is far from an efficient means to counter the fake news phenomenon. This situation has brought to the fore the importance of prebunking, which involves preemptively raising citizens' awareness of mis/disinformation techniques to make them resilient toward fake news. So far, inoculation—the exposure of people to weakened doses of techniques used to spread fake news to generate mental antibodies—has proved to be the most effective way of prebunking (Lewandowsky & Van Der Linden, 2021). More specifically, active inoculation (Saleh et al., 2021) through engagement in a digital game has turned out to be particularly effective for cognitive reasons (Pfau et al., 2005).

In light of this, we have designed two chatbots, the *Fake News Immunity chatbot* (<http://fni.arg.tech/>) and the *Vaccinating News chatbot* (fni.arg.tech/?chatbot_type=vaccine), to interactively teach citizens and communication gatekeepers, respectively, how to become their own fact-checkers and how to avoid creating and spreading misinformation through the identification of fallacious arguments (see section “The Fake News Immunity and the Vaccinating News Chatbots”). Newsmaking, especially in situations of epistemological uncertainty such as

the pandemic, involves the argumentative process of gaining the acceptance of a certain interpretation of a news event. An assessment of the quality of arguments which constitute a news claim is, thus, a key factor to exercise critical thinking for the identification of fake news. More specifically, the recognition of arguments which seem valid, but are not—fallacies—results in the gray area of misinformation, where the information conveyed might be factual, but presented in a misleading way through strategies such as cherry picking, false analogies, and hasty generalizations (Musi & Reed, 2022). The two chatbots differ from existing tools in their scope over misinformation rather than disinformation, in the data-driven selection of the scenarios as well as their multiagent infrastructure and front-end features. After having introduced related work and described the design of the two chatbots, we report on their evaluation, focusing both on user experience and efficacy on advancing reason-checking among their users. We do so through the combination of a quantitative online questionnaire, a crowdsourced survey, and two workshops in a qualitative environmental setting. As a result, a preliminary framework to evaluate critical thinking skills tailored to the information ecosystem is provided.¹

Related Work

From Debunking to Prebunking

To counter the spread of fake news, two main types of interventions have been put into place: debunking, the retroactive correction of false beliefs, and prebunking, the preemptive exposure to misinformation and disinformation techniques or sources before they strike. Recent studies have cast doubts on the efficacy of debunking for a plethora of reasons. Park et al. (2021) have, for instance, revealed through a set of randomized surveys (overall sample: 1,145 young adults), that the positive effects of fact-checking are reduced by perception and belief biases. They registered a widespread reluctance to change views when fact-checking reveals that claims initially perceived as negative are true (self-correction bias) and a perception bias in interpreting messages flagged with the rating “Lack of Evidence” as closer to be false than claims marked as “Mixed Evidence” due to our cognitive uncertainty-aversion. As for long-lasting effects, Carey et al. (2022), through a large-scale survey in Canada, the United States, and the United Kingdom show that exposure to fact-checks reduced beliefs in false claims with no spillover effects, but the improvements in accuracy judgments already dissipates after a few weeks.

Regardless of the complexity of our cognitive systems, it is intuitive to think that corrections are more effective when the arguments supporting them are made transparent. For example, a fact-checker rating such as “mixed” or “divided evidence” is not very informative unless readers are explained why the evidence provided is not enough, or potentially misleading. In other words, foregrounding and explaining the

roots of the distortions is more effective than merely flagging them, since it allows prebunking across contexts (Van der Linden et al., 2020). The effects of prebunking promise to be less ephemeral since they enhance critical thinking skills which are neither space- nor time-bounded (Tay et al., 2021). The underpinnings of effective prebunking rest on “inoculation theory” which states that “if people are forewarned that they might be misinformed and are exposed to weakened examples of the ways in which they might be misled, they will become more immune to misinformation” (Lewandowsky & Van Der Linden, 2021, p. 348). Three main ways to prebunk have so far been used (<https://tinyurl.com/2fd9yh7c>): fact-based, logic-based, and source based. It is clear that the skills allowing to disentangle factual information from fakery, identify misleading rhetorical techniques as well as recognize trustworthy sources are core to the exercise of critical thinking, advocated by policymakers as a key pillar for media literacy (see *UNESCO Media and Information Literacy: Policy and Strategy Guidelines*). Comparing various types of intervention, extant research has shown that instructing against misleading rhetorical strategies is a highly efficient way to boost people’s resilience against fake news (Cook et al., 2017) and that content features such as degree of novelty and emotional reactions of recipients are key factors in making fake news spread fast. Such an effort requires first of all a theoretical framework to surface flawed rhetorical techniques and arguments as well as a heuristic to make them publicly recognizable.

Furthermore, in an information society where the medium is more and more the message, the venues that are used to inoculate against misinformation also play a crucial role.

Critical Thinking and Argumentation for Media Literacy

The importance of critical thinking skills for media literacy has been widely acknowledged. Koltay (2011) lists, for example, “Having a critical approach to quality and accuracy of content” among the five stages to build media literacy” (p. 213). Going one step further, during the opening speech of the EC (European Commission) Media and Learning Conference (10 March 2016), Roberto Viola points out that “a key pillar in all possible definitions of media literacy is the development of critical thinking by the user or citizen.” *Critical thinking* has been an object of discussion across disciplines ranging from Philosophy to Psychology to Informal Logic and has become a buzzword in pedagogical settings since the late 20th century (Goodnight, 2009), leading to a proliferation of definitions (e.g., Ennis, 1989; Hatcher and Spencer, 2005; Paul, 1981). In the “Delphi report,” Peter Facione attempted to reach consensus, gathering together 46 scholars who came up with the following definition: “We understand critical thinking to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential,

conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based” (Facione, 1990, p. 2). However, an operationalizable definition of what critical thinking means when applied to data literacy is still missing, together with a methodology to assess critical skills by news consumers.

To tackle this issue, we propose to view newsmaking as a process aimed at gaining the acceptance of a certain interpretation of a news event. In this perspective, newsmaking is, thus, a form of argumentation, intended as “a discourse aimed at convincing a reasonable critic of the acceptability of a standpoint by giving reasons that justify the standpoint” (Van Eemeren & Grootendorst, 2003, p. 1). A critical assessment of the presence and quality of arguments (fallacious or not) provided in support of a news claim, thus facilitates the identification of misleading news. Learning how to identify fallacies in news through the *Fake News Immunity Chatbot* offers an opportunity to inoculate against fake news. In line with Mercier and Sperber’s (2017) argumentative theory of human reason, we believe that dialogical contexts where participants argumentatively exchange information are more likely to facilitate the acquisition of reasoning skills than monological ones.

However, a measurable evaluation of the impact played by learning how to recognize fallacious arguments on identifying misinformation is challenging: a citizen might, for example, learn to be suspicious about causal relations since they might be simple correlations, but still believe in a news stating a causal explanation in the absence of contrary evidence. Furthermore, fallacious arguments are more or less hard to identify, depending on the news context and personal knowledge about that context. Thus, large-scale surveys measuring changes in truth values assigned to a set of news, frequently used to evaluate digital tools (see section “Active Inoculation Through Digital Tools”), do not allow to verify acquired critical thinking skills which relate to the epistemic process rather than the product (e.g., truth judgments). In light of this, we propose an impact evaluation framework which combines quantitative and qualitative components (see section “Impact Evaluation”).

Active Inoculation Through Digital Tools

Active inoculation, differently from passive inoculation prompt engagement: differently from passive inoculation, where both counter-arguments and refutations are provided by the recipient, in active inoculation it is the participants who must produce both pro- and counter-arguments themselves. To enhance the fight against fake news, a suite of digital tools in a gamification environment has been developed. Online quiz-based games such as *Fakey* (<https://fakey.osome.iu.edu/v/>) and *NewsWise* (<https://tinyurl.com/5bmau7jk>) have the goal of teaching users how to recognize misleading sources or headlines through trial and error, while *Real or Photoshop quiz* by Adobe

focuses on the identification of fake images. *BBCireporter* (<https://tinyurl.com/2jnnctrd>) and *NewsFeed Defenders* (<https://tinyurl.com/ms4nar5>) instead, put users in the shoes of communication gatekeepers, simulating their decision-making processes. A more sophisticated generation of digital games is showcased by tools such as *Bad News* (<https://www.getbadnews.com/#intro>), *Go Viral!* (<https://www.goviralgame.com/books/go-viral/>) and *The Harmony Square* (<https://harmonysquare.game/en>), which cast the player as the “fake news spreader” who learns by doing successful misleading strategies used by disinformers. Impact evaluations have shown that these tools are highly advantageous. The findings of Basol et al.’s (2021) study to evaluate the efficacy of the *Go Viral!* game on 1,777 players reveal an increase in skepticism toward both real and fake news right after playing, but an enhanced immunity toward disinformation only after 1 week.

The *Fake News Immunity* and the *Vaccinating News chatbots* differ from state-of-the-art active inoculation tools since they target misinformation rather than disinformation. As such, their primary goal is not teaching users how to disentangle true from fake information—a task not always feasible to perform in crisis situations where information is provisional (e.g., side effects of a vaccine)—but rather to learn what questions to ask to critically consume and create news. We propose to do so by (a) applying the notion of fallacy to the identification of various types of misinformation through a novel heuristic based on critical questions (Musi & Reed, 2022) and (b) leveraging a large-scale data analysis to select those scenarios that turned out to be more prominent and, hence, potentially more dangerous.

As remarked by Almalki and Azeez (2020), a plethora of health chatbots have been developed during the pandemic to “disseminate health information and knowledge; self-triage and personal risk assessment; monitoring exposure and notifications; tracking COVID-19 symptoms and health aspects; combating misinformation and fake news” (p. 244). The latter group of chatbots counter misinformation mostly offering accurate, tailored, and easy-to-access correct information (Altay et al., 2021; Herriman et al., 2020; Siedlikowski et al., 2021), rather than teaching citizens how to recognize misinformation in messages spread across digital media.

Although not addressing the fake news phenomenon, digital tools have been built by the scholarly community to enhance critical thinking through argumentation. More specifically, various computer software packages (e.g., *Araucaria*, Reed & Rowe, 2004; *iLogos*, Harrell, 2008; *Rationale*, Martin Davies, 2009; ter Berg et al., 2013) have been created to support argument mapping through visualizations. The educational efficacy of the argument mapping software *Rationale* has been tested across domains ranging from English as a Foreign Language Context (Eftekhar et al., 2016) to business education (Kunsch et al., 2014). Going one step further, *LiteMap* (De Liddo & Strube, 2021), is a

collaborative tool that besides argument mapping enables visual summarization to help sensemaking of online public debates. With the aim of preventing the formation of misconceptions about genetically modified organisms (GMOs), Altay et al. (2022) have developed a chatbot to provide participants with good arguments rebutting the most common counter arguments against GMOs. Finally, the ArgTech research center (<https://arg-tech.org/>), has showcased how argument technology can be applied to the media sphere, teaching how to improve debate skills (Test Your Argument, <https://www.bbc.co.uk/taster/pilots/moral-maze>) and appraise argumentative structures in news reports (Evidence Toolkit, <https://bbc.in/2FFNQen>) with the goal of instilling those critical literacy skills needed to reduce polarization and strengthen communication persuasive skills.

Drawing on these preliminary results, our chatbot positions argument technology at the forefront in the fight against misinformation.

From Fact-Checking to “Content-Checking” to “Reason-Checking”

Increasingly, the challenge for the general public and specialists alike is shifting away from mere checking of “facts.” In the first place, nuance, subtlety, and open-texture make the veracity of statements that are in principle verifiable much fuzzier than a simple true–false distinction. Claims depend upon context, definitions, deixis, and more, all of which may or may not be explicit, and may lead to significantly different judgments of reliability. As a result, many fact-checking organizations do not check facts as much as provide interpretation, contextualization, and exegesis. Increasingly, they focus not on the ways in which truth is attached to a claim, but on the ways in which truth is maintained or eroded along the passage to a claim from its evidence. The awareness that exposure to facts is not a solution to disinformation spread has induced scholarly communities building digital tools for enhanced content curation: the platforms *SadView* (<https://imi-sad.pages.switch.ch/sadview/>) and *Newteller* (<https://newteller.lsir.ch/>) developed by the *Media Observatory Initiative* (EPFL), respectively, enable journalists and citizens to monitor the propagation of controversies across social media leveraging social network analysis and offer context for news articles combining content, social, and source indicators. Acknowledging that citizens tend to passively adhere to news feeds suggested by social media algorithms, Horne et al. (2019) develop a Trust Nudging Model through a recommendation system that nudges people to make better news consumption choices. In view of the role played by emotion in news interpretation, Sethi et al. (2019) propose a recommender system explaining interface where users’ emotional profiles are factored in the interaction with pedagogical agents who compare and contrast various stances of an issue.

This shift toward a focus on a more relational notion of fact-checking and content curation goes hand in hand with a rise in the role played by critical thinking (Johnson & Blair, 2006) in countering misinformation. For models of critical thinking uniformly reject absolutist notions of truth in favor of contextualized, relativistic conceptions of goodness both epistemologically and inferentially. Thus, for example, acceptability (to an audience), relevance (between pieces of information), and sufficiency (of evidence for claims) substitute for deductive validity, and as a result naturally usher in an approach that focuses upon relations between pieces of information, and between information and context.

Techniques of critical thinking have long been explored in computational environments to provide scaffolding for better quality reasoning in domains such as law (see, e.g., Gordon et al., 2007) and politics. Recently, however, they have been placed front-and-center in a wide-scale deployment of software for the general public, to support an educational program in media literacy with the BBC in the United Kingdom (Visser et al., 2020). The focus in that work is not upon fact-checking, but rather upon *reason-checking*—using theories of critical thinking to scaffold the investigation and interrogation not of claims, simpliciter, but of the connections between claims and their evidential context. It is such a shift of focus that underpins our attempts here to develop tools for enhancing fake news immunity.

Let us consider an example that is part of the knowledge base we created for the *Vaccinating News Chatbot*. One of the four main learning outcomes of the chatbot is that of selecting non-fallacious sources for drafting an editorial. Zooming into the topic of “politicizing the vaccine,” the user is asked to write about Amazon’s offer to help with the U.S. government COVID-19 vaccination program. One of the first steps in a journalist’s activity is that of picking a set of sources to draw upon. To simulate such a procedure a pool of four sources mixed as to origin (social media vs. official news source) is provided:

1. Source 1: <https://archive.is/MQGVE>
2. Source 2: <https://www.whitehouse.gov/briefing-room/press-briefings/2021/01/21/press-briefing-by-press-secretary-jen-psaki-january-21-2021/>
3. Source 3: <https://www.foxnews.com/us/why-did-amazon-wait-until-bidens-inauguration-to-offer-help-with-vaccine-distribution>
4. Source 4: <https://twitter.com/amazonnews/status/1351991663191871491>

Regardless of the digital venue, both the second source (official transcript of a White House press briefing by Press Secretary Jen Psaki) and the fourth source (tweet linking to Amazon’s letter to Biden declaring their intention to assist in the vaccination efforts constitute) offer accurate information to give respectable voice to the governmental response to Amazon’s move and Amazon’s perspective on the matter.

However, selecting bits and pieces of these sources could result in “cherry picking” behavior and foreground facets, which suggest a defeasible interpretation of state affairs not far from the one expressed by Source 1. The tweet, fact-checked by *Snopes*, does not contain non-factual-information *per se*, but it puts forward a misleading interpretation of facts: the fact that Amazon announced their help *after* Biden’s inauguration does not mean that they did it *because of* Biden’s inauguration, so that Biden takes credit for it instead of Trump. A similar interpretation, even if not asserted but simply alluded to by the question in the title (“Why did Amazon wait until Biden’s inauguration to offer help with vaccine distribution?”), is suggested by the Fox News article. Such an instance of *post hoc* fallacy, establishing a causal connection when a simple correlation is at stake, is not a matter of facts but calls for a critical evaluation of the inferential links linking available evidence to the standpoints put forward.

The Fake News Immunity and the Vaccinating News Chatbots

Design

The design of the system is founded upon three tenets: first, that identifying misinformation rests critically upon critiquing the passage from premises to conclusions, from evidence to claims; that is, upon processes of reason-checking; second, that a powerful mechanism for reason-checking is to actively engage in dialogue, in multi-perspective exchange that puts inferential steps under a dialectical microscope; and third, that the process of dialogue can be conceptually and practically disentangled from the informational substrate over which it acts. These three tenets are explored first empirically through an analysis of data-informed cases, and second, through the design of the computational infrastructure by which such dialogue can be mediated and executed.

Data-Informed Cases. The selected cases of misinformation addressed by the tools are news that actually circulated across digital media. This choice is reminiscent of the fact that authentic problems are a crucial factor when teaching critical thinking (Abrami et al., 2015). More specifically, the chosen cases come from a dataset of 1,500 news web-crawled from five English fact-checkers (*Snopes*, *The Ferret*, *PolitiFact*, *Healthfeedback.org*, *Fullfact*) in two time spans: from January 2020 till June 2020 (1,135 news items) and from September 2020 to December 2020 (365 news) to include news about the vaccine. This dataset has been systematically analyzed through a multilevel manual annotation encompassing (a) type of semantic claim expressed in the headline, (b) type of source (e.g., social media) for the entire dataset, and (c) type of fallacies. The statistical analysis of the results (Musi et al., 2022) shows that while social media are privileged sources for disinformation, misinformation is spread

across the board and that a set of 10 fallacies emerged from the data analysis allows to explain the misleading roots of the attested misinformation cases (see <https://tinyurl.com/2p86ptxs> for an explanation of the fallacy types). While some fallacies (e.g., evading the burden of proof) are significantly more frequent than others (e.g., false analogy), different types of fallacies do not pattern significantly with different types of sources. However, the interlevel analysis suggests a significant correlation between type of fallacies and type of claims where interpretations pattern with false cause; evaluation emotional with false analogy and predictions with evading the burden of proof. Drawing from this analysis, we have selected from the dataset misinformation news items with the most significant configurations of features (e.g., prediction claim—evading the burden of proof fallacy—social media source), assuming that they would resemble actual news read by citizens. More specifically, we have chosen 20 news for the *Fake News Immunity chatbot* and 16 for the *Vaccinating News Chatbot*. In the design of the Fake News Immunity (FNI) chatbot, to diminish bias in the news topic due to fact-checkers' editorial choices, we have picked the same number of news (4) from each fact-checkers for the *Fake News Immunity chatbot*. For the *Vaccinating News Chatbot*, we have first identified 4 popular topics related to the vaccine according to the World Health Organization (WHO; adverse reactions to vaccine; vaccine, immunity and transmission; vaccine manufacturers; politicizing the vaccine) and we have then selected 4 news from each topic evenly distributed across fact-checkers.

Infrastructure. The FNI chatbot and the Vaccinating News platforms can be conceived as computational executions of dialogues, in a gamification format. The infrastructural architecture is represented in Figure 1.

The structure of the dialogue game, written in the tailored dialogue game programming language DGDL (Wells & Reed, 2012), is detached from the knowledge over which the game is to be played. This makes updating and revising the underlying data a straightforward task that is independent of the structure of the interaction. To design the frontend of the *Fake News Chatbot* and the *Vaccinating News Chatbot* we used gamification principles that have been proved to be advantageous to enhance critical thinking (Stott & Neustaedter, 2013) which include (a) freedom to fail, (b) rapid feedback, (c) sense of progression, and (d) storytelling. Starting from the latter, we have chosen as a setting ancient Greece through the aid of multimodal features in the graphic design and the choice of Aristotle, Gorgias, and Socrates, fathers of critical thinking, as avatars. To allow for a sense of progression we have created a reward system where players receive a “gadfly” in their jars whenever they accomplish eight correct answers. Furthermore, the *Fake News Chatbot* contains three levels of increasing complexity (credulous, skeptic, and agnostic), while the *Vaccinating News Chatbot*

allows for user to progressively select different tasks (write fallacy-free headlines; select fallacy-free sources; write fallacy-free articles; and write fallacy-free news on social media). No penalties are involved in the scoring system while each conversational turn by the user is followed by a prompt reaction from one of the avatars, to whom the user is allowed to ask for help at any stage of the decision-making process. Besides the three philosophers, a fourth avatar is a member of the research team that is selected by the player.

The locution types expressed by the avatars are typified in accordance to their philosophical personalities: Socrates asks maieutic questions (e.g., “Does the news express an unsailable fact?”) aimed at eliciting doubts and new concepts previously latent in the users' minds when reading a news; Aristotle explains notions and concepts through assertions (e.g., “An argument is relevant if it provides information that makes the claim more likely to be true”); Gorgias challenges users' answers as well as common ground opinions through rhetorical questions and witty comments (e.g., “no other opinions are mentioned, how can the post criticize someone else's opinion?”). This stylistic choice is motivated by three main factors: first, interacting with the philosophers' users inductively learn their dialectical techniques acquiring historical knowledge; second, research shows that building software agents as dialogical *personas* increases users' engagement (Tsai et al., 2021); third we wanted to test (see feedback questionnaire section “Conclusion”) what character and, hence, dialectical style, is preferred by users.

The structural rules underlying users/avatars interactions respond to the two chatbots learning outcomes (learn how to reason-check through fallacies; learn how to write fallacy-free news). Both the *Fake News Chatbot* and the *Vaccinating News Chatbot* start with a request to the user to assess the reliability of a news article explaining their rationale, paired with access to the fact-checker's verdict. After this self-assessment moment, in the *Fake News Chatbot*, the user is fronted with instances of news and guided by Socrates through heuristics meant to teach users how to identify potential fallacies. The heuristics is, in fact, composed of a set of *critical questions*, which are conceived in Argumentation Theory (Walton et al., 2008) as those questions that scrutinize the soundness of the reasoning expressed by the arguments (e.g., “Is the reported evidence [if any] the only available?” to verify whether cherry picking is at stake). The user is asked to take dyadic choices (yes/no) as an answer and (s)he is explained in detail the reasons underlying the right choice when the incorrect answer is picked (Figure 2).

In the *Vaccinating News Chatbot* the user, who is meant to simulate the decision-making processes of a journalist/communication gatekeeper, has a more agentive role: she has to select an option out a series available (e.g., select one headline out of five) and justify the choice, while being challenged/prompted by the avatars (Figure 3).

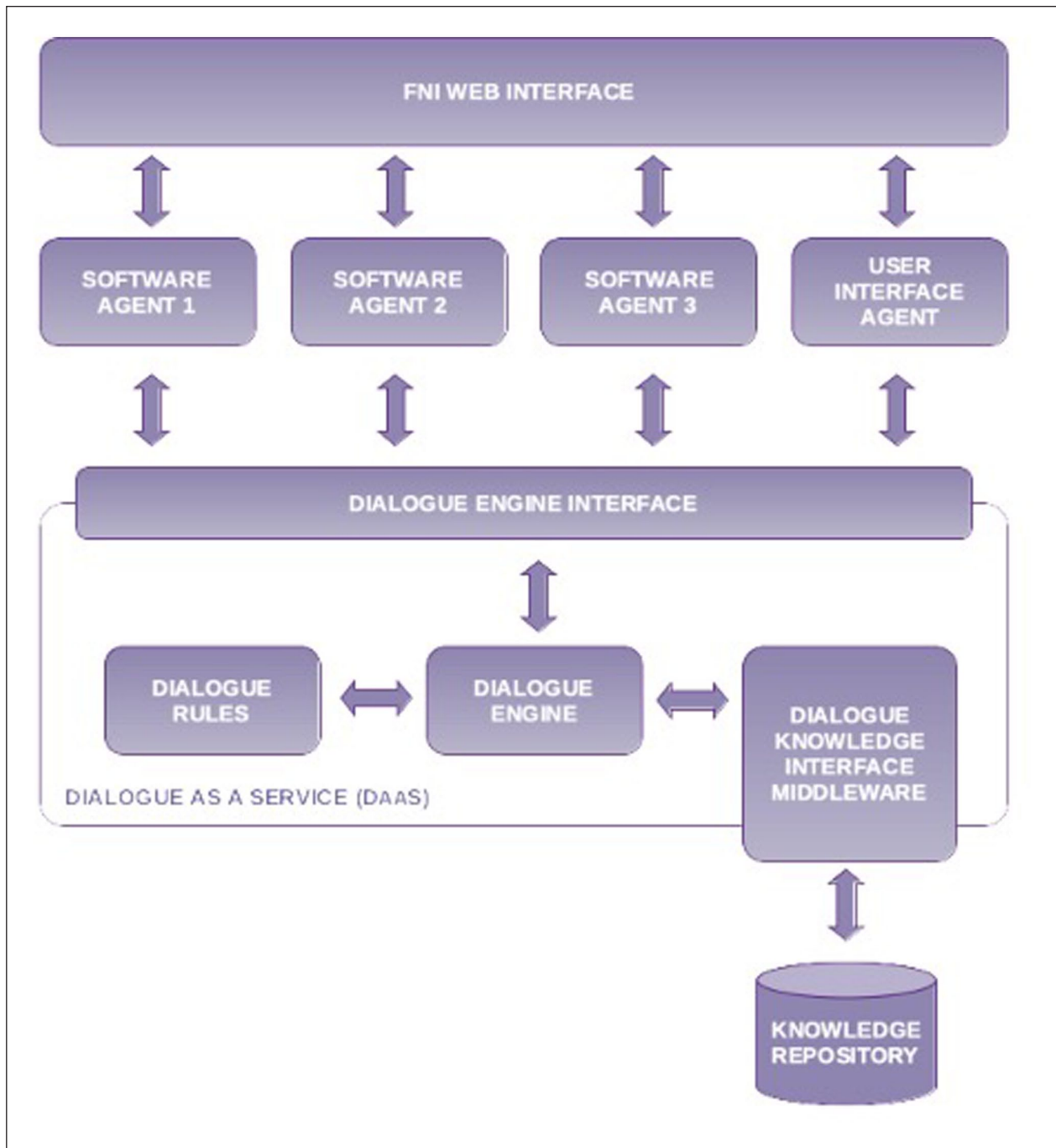


Figure 1. System architecture.

Impact Evaluation

Since their launch in November 2020, the *Fake News Immunity* and the *Vaccinating News Chatbots* have registered 1,700 users across 10 countries (United States, 490; United Kingdom, 375; Italy, 122; Germany, 78; Netherlands, 56; Switzerland, 53; Canada, 39), with an average engagement time of 3 min and 14 s.

User Experience Design. To evaluate and improve the user experience design, we have embedded a voluntary, anonymous questionnaire composed of 18 multiple-choice and open-ended questions accessible at: https://liverpoolcomms-media.fra1.qualtrics.com/jfe/form/SV_cumhqbTOi8rJTud. Overall, 8% of the users provided feedback, amounting to 211 completed questionnaires. The questions have been designed to account for the four chatbot evaluation criteria

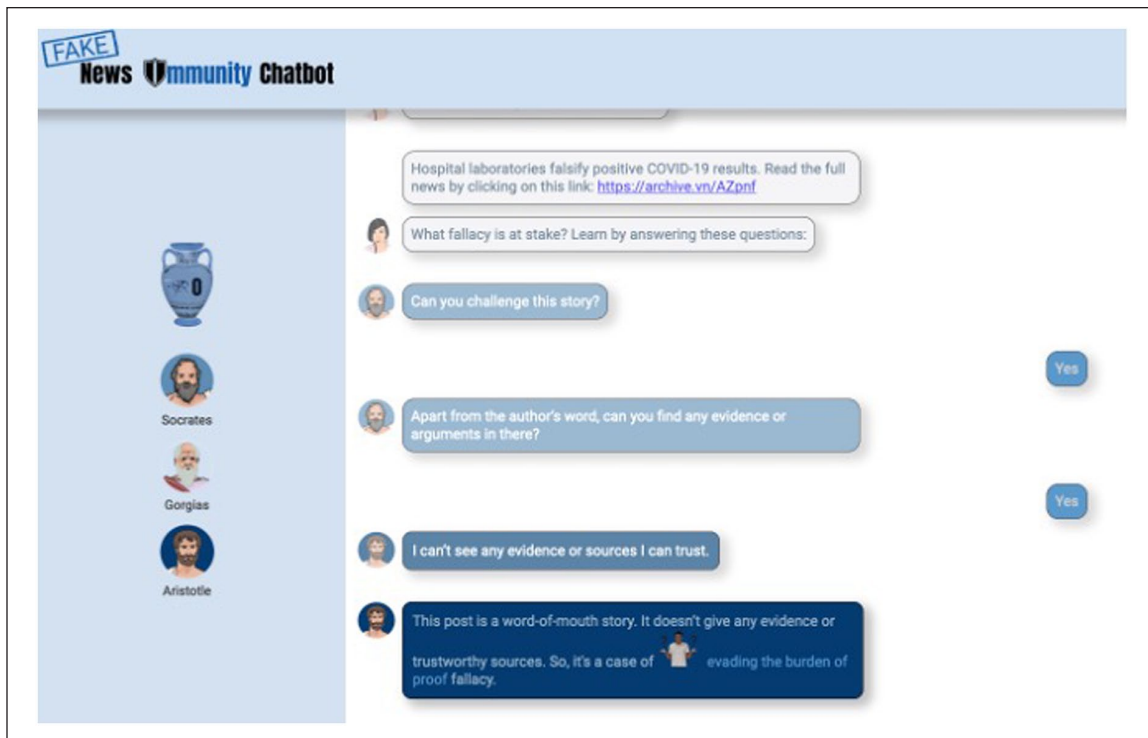


Figure 2. Example of reason-checking interaction through the Fake-News Immunity platform.



Figure 3. Example of reason-checking interaction through the Vaccinating News Chatbot.

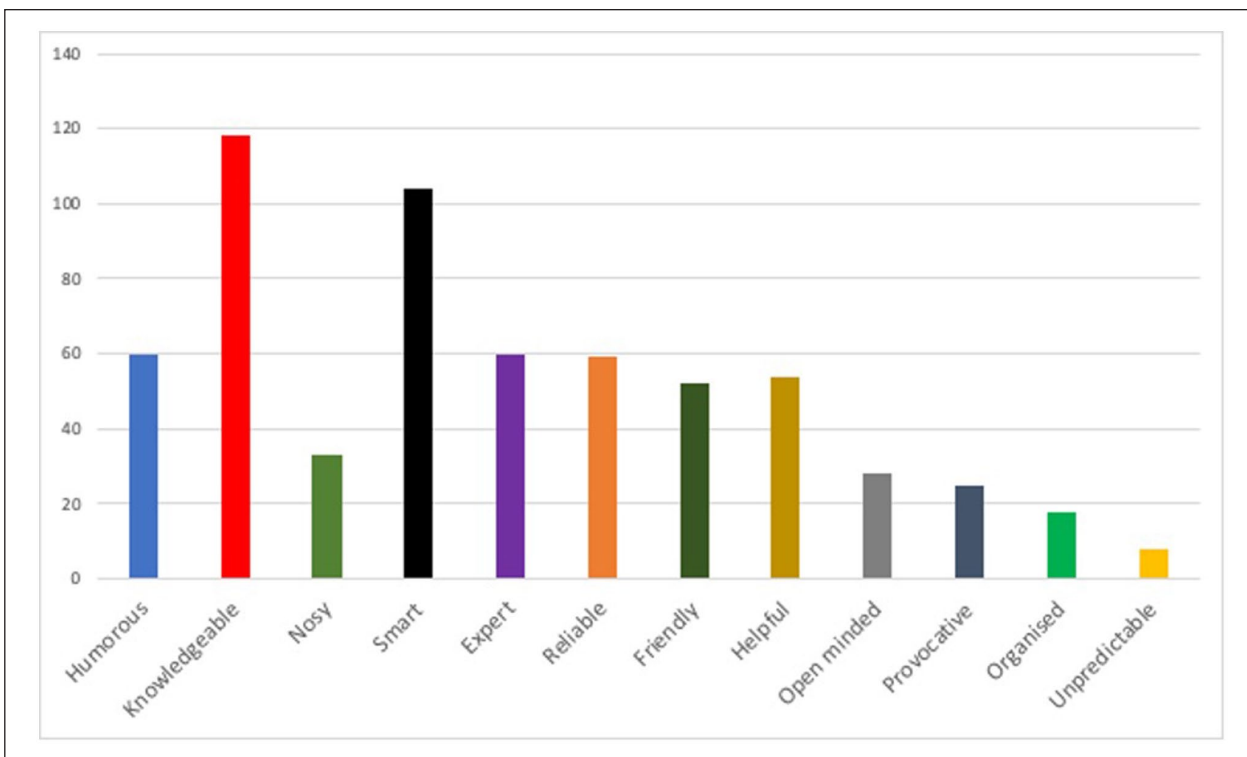


Figure 4. Personality traits of avatars preferred by the users.

identified by Jain et al. (2018): Chatbot Conversational intelligence (Q2–6), Chatbot Personality (Q7–10), Chatbot Interface (Q11–14) and Chatbot Functionality (Q15–17), with the addition of an onboarding question (“How many news did you fact-check?”), and a final open-ended question (“Overall, if you could change something in the chatbot what would you do differently?”).

From the first question, it emerges that half of the users were able to fact-check either three or four news items, having spent more than the average engagement time on the chatbot (we estimate 2 min per news item). As for conversational intelligence, we received overall positive feedback with some hints for reflection: around half of the respondents considered both conversational rhythm and tone “just right,” while the majority of the other were scattered, respectively, across “slow” (17%) / “fast” (25%), and “formal” (25%), “informal” (14%), showing that the way conversational flow is perceived is highly subjective. Our fourth question was meant to assess the design of the multi-software agents’ interaction: to simulate a peers’ discussion, we did not limit software agents’ conversations to interactions with the users (1—>many), but we added conversational turns between the software agents. Users found the fact that “Sometime the AI participants talked amongst themselves” to be interesting (42%), informative (28%), and a minority confusing (26%) or boring (4%), suggesting that multiparty conversations shall be further explored in multiagent chatbots. Overall, the perceived active participation by the users

could still be improved since 41% of users felt that their participation was “just right” and 26% felt “active,” but 15% rated their participation as “sometimes active” and 18% as “inactive.” To increase perceived agency, we are planning to allow for more unconstrained questions on the part of the user. As shown by responses to Q6, almost half of the respondents (85) agreed with the statement, “Sometimes I did not feel ready to choose yes or no . . . The world is not black and white!,” highlighting the difficulty of making straight diadic choices.

Zooming into chatbot’s personality, the most favored software agent (Q7) is Aristotle (42%) followed by Socrates (36%) and Gorgias (22%). This line of preference matches with the choice of adjectives picked by the users to motivate their choices (Figure 4).

While all the avatars have been construed as to portray reliability, Aristotle, *qua* father of fallacy theory, has been presented as the most knowledgeable and, together with Socrates, smart. Although the most humorous, Gorgias’ unpredictable and provocative personality traits have turned out not to be the most appreciated. This might be due to the disclosed educational nature of the chatbot, which positions the avatars in a pedagogical setting as teachers rather than peers. Another possible explanation lies in the reputation cognitive heuristics according to which “people are likely to believe a source whose name they recognize as more credible compared to an unfamiliar source” (Metzger & Flanagin, 2013, p. 214). Both Aristotle and Socrates are names of

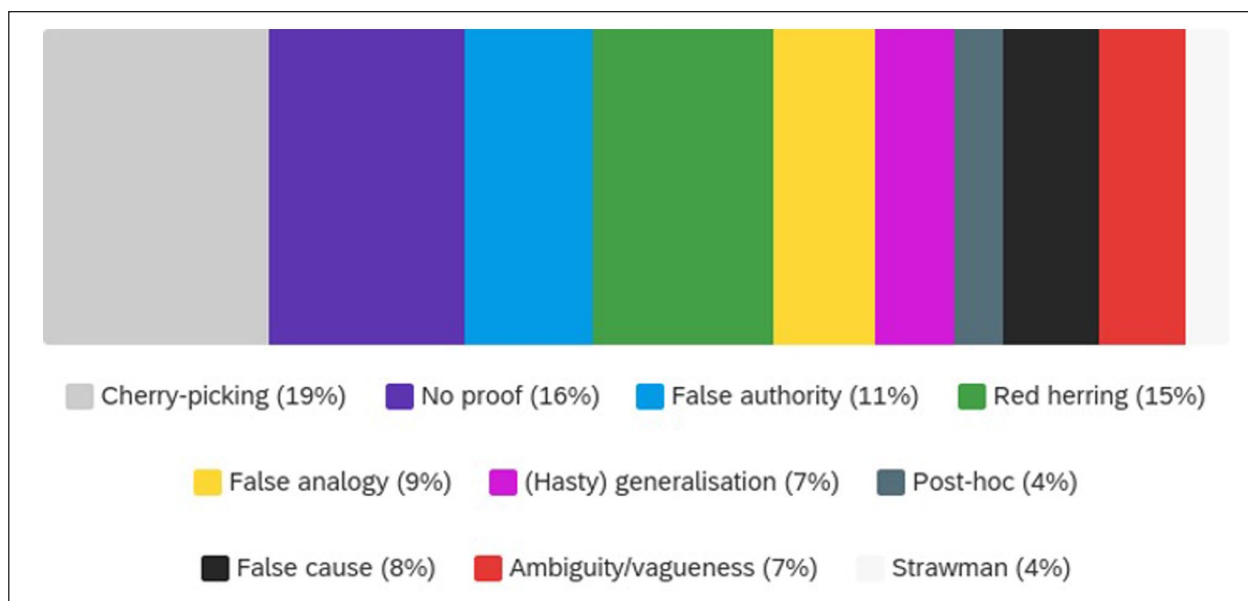


Figure 5. Fallacies discovered by users while playing.

philosophers which the majority of users would recognize, while the same does not apply to Gorgias.

The analysis of the open answers to Q8 (“What do you think are the three most important qualities in a teacher?”) reveals, in fact, that the semantic domain of knowledgeability (tokens “knowledge,” “someone who knows”; “knowledgeable”) is the most frequent (50 mentions). Furthermore, the avatar considered most trustworthy by the participants who answered question Q11 (“Which participant looks more trustworthy? And why?”) is again Aristotle since “knowledgeable” and “intelligent.” These results seem to challenge a key component of the social media trust framework, namely, that social media users tend first of all to trust their peers rather than institutions, equating familiarity with credibility (Shareef et al., 2020), what Yates & Carmi (2022), call “networks of literacy.” However, this might not be the case in crisis scenarios or educational settings in which familiarity, a behavioral component of trust, does not reduce uncertainty, while competence does. A supplementary survey is needed in future work to shed light on the features which enhance at once likeability and trust.

Interestingly, however, 40% of users claimed that they did ask for help more frequently from their favorite character, while 44% said they did not and 16% did only sometimes. This self-reported info matches with the trends tracked over the two platforms which register 553 “help” clicks on Gorgias’ avatar and 460 on Aristotle’s one. It thus appears that recognizing high pedagogical *ethos* to an avatar does not translate in a propensity to ask for direct help, may be since perceived as face threatening with respect to an authority in the field.

As for the interface, the majority of users declared that it made them feel “relaxed” compared with “bored” (17%),

“overwhelmed” (18%), or “amused” (15%). This was our intention to prompt users to adopt a thinking-slow process, which is generally hampered by the overwhelming and fast proliferation of information. A recurrent aspect that users would have changed is that the opportunity of getting help from the avatars during the decision-making process was not apparent. To compensate for this issue, we have added to the question marks next to the avatars’ portraits, a flashing light to capture users’ attention. As for the question pointing to which avatar looks more trustworthy, the top choice has been Aristotle followed by Socrates and then Gorgias, with similar arguments to the ones supporting the choice for a favorite character.

Finally, turning to the chatbot functionality, the most frequently encountered fallacies have been cherry picking, evading the burden of proof and strawman, as displayed in Figure 5.

The descriptions of the discovered fallacies provided by some of the respondents were all accurate, suggesting that they learnt their meaning. When asked whether they would be able to recognize the fallacies in the future, 50% respondents answered they would maybe be able to, while 44% were more resolute (“yes”); the minority who was doubtful explicated as a reason the lack of required focus due to the fast-paced flow which features our digital lives.²

To better understand factors that might prompt users’ interactions with avatars, we plan to make Q8 (“Why do you like them? Pick 3 adjectives that apply”) an open-ended question to directly crowdsource properties which trigger avatars’ likeability.

Crowdsourced Survey. Besides evaluating the user interface, we also wanted to verify whether the chatbots actually exercised users’ critical thinking for media literacy. Since critical

Table 1. Sensemaking Scale for Critical Thinking Self-Assessment.

Critical thinking factors (De Liddo et al., 2021)	Definition (De Liddo et al., 2021)	Survey prompts
Reflection	Capability to think back and in depth	I found that the chatbot made me reflect more deeply upon the news I read
Insights	Capability to get unexpected ideas or make unexpected inferences	I found that the chatbot provided me with unexpected insights on the issues discussed in the news
Focus	Capability to see different angles and aspects in the debates	I found that the chatbot made me focus on different aspects of the news that I would have otherwise neglected
Argumentation	Capability to reconstruct the arguments that the speakers make	I found that the chatbots helped me reconstruct the arguments that the author made
Explanation	Capability to identify and explain issues	I found that the chatbots helped me decide whether a news is trustworthy
Assess facts and evidence	Capability to assess presented facts and evidence	I found that the chatbot provided me with new ways to evaluate the interplay of facts and evidence in the news
Distinguishing	Capability to make a difference between the speakers' claims and the options proposed	The chatbot helped me distinguishing different types of misleading information
Change Assumptions	Capability to change one's own mind	Using the chatbot I changed some initial assumptions I had before-head

thinking goes beyond the capacity of assessing news reliability and relies first of all on awareness about the need for analytic parameters, we decided to assess users' self-reported perceptions of changes in critical thinking skills. To recruit respondents with diverse demographic features, we have set up a crowdsourced survey using the *Amazon Mechanical Turk* platform, aiming for a sample of 150 participants. The *Amazon Mechanical Turk* platform has been fruitfully employed to gather respondents for a range of tasks across domains ranging from social sciences to computer science (Strickland & Stoops, 2019).

The task consisted of playing with the chatbots for 15 min and then filling in a questionnaire accessible on *Qualtrics* consisting of 10 questions. Each participant was provided with an incentive of 5 GBP to complete the task. Due to the remote and anonymous nature of the experiment, the first two multiple-choice questions were used to ascertain that the users played with the chatbots before completing the survey, asking about the levels of the *Fake News Immunity chatbot* and the way fact-checking is taught (through fallacies). After having discarded users who did not meet this requirement, we have obtained 142 answers. The design of the other questions was based on identifying parameters that are symptomatic of critical thinking in the context of news consumption. To this aim, existing tests for the evaluation of critical thinking so far proposed in the educational literature (e.g., California Critical Thinking Skills Test, Cornell Critical Thinking Test) were not suitable since they address general cognitive skills (e.g., deduction/induction) which are tangential but do not have scope over media literacy.

Assuming that critical thinking in the media ecosystem implies a process of sensemaking the information accessed through the news (Grasso & Convertino, 2012), we have

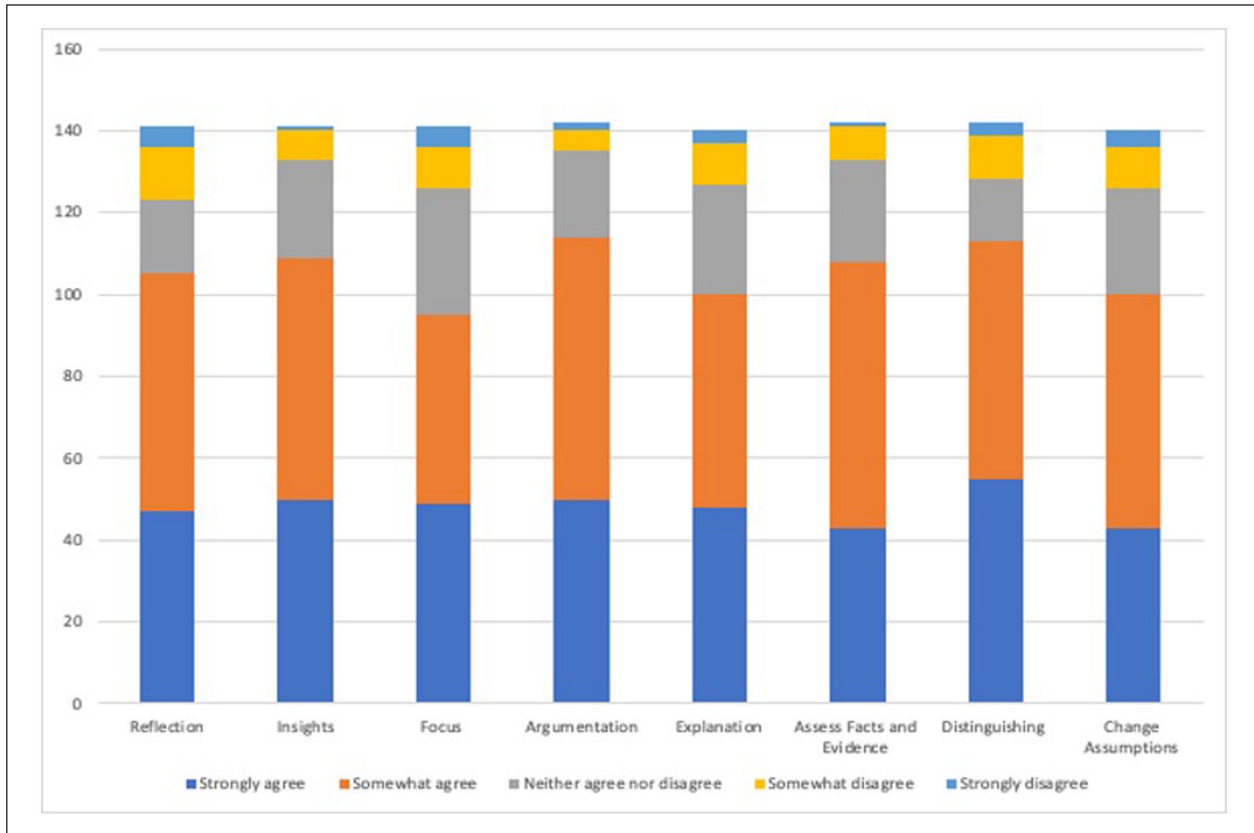
taken as a starting point the sensemaking scale developed by De Liddo et al. (2021). Their scale has been used to evaluate the efficacy of *Democratic Replay*, a platform meant to enhance televised election debates with interactive visualizations of speakers' arguments, dialogical performance, and public reactions. The scale encompasses nine factors based on Alsufiani and Attfield (2018) theory: reflection, insight, focus, argumentation, explanation, assess facts and evidence, assess assumptions, and change assumptions. We developed eight prompts, adjusted to the context of news consumption, one per each of the factors with the exception of "assess assumption." Differently from the context of political debates, news reading does not foreground the assessment of personal ideas, but rather a change in opinion deriving from consumed information. The factors, their definition, and the matched survey prompts are displayed in Table 1.

For each prompt, users had to express their agreement on a 5-point Likert-type scale. The breakdown of the answers per factor is reported in Table 2 and visualized in Figure 6.

The comparison of the results across the eight factors shows similar trends in users' responses, with a mean among values per each factor, which oscillates between 0.88 and 1.08. A third of users agreed that the use of the chatbots increased their skills across the eight factors. The ratio between *strongly agree* (max value: 39% for *Distinguishing*; min value: 30% for *Assess Facts and Evidence*) and *somewhat agree* (max value: 46% for *Assess Facts and Evidence*; min value: 33% for *Insights*) is in favor of a less convinced stance (*somewhat agree*) across the board. The highest effects are found to correspond with the factors "Distinguishing" and "Argumentation." This result is not surprising since the identification of fallacies, the main target of the chatbot, itself involves identifying the different types

Table 2. Users' Perceived Enhanced Critical Skills per Prompt.

	Reflection (%)	Insights (%)	Focus (%)	Argumentation (%)	Explanation (%)	Assess facts and evidence (%)	Distinguishing (%)	Change Assumptions (%)
Strongly agree	35.46	35.46	34.75	35.21	34.29	30.28	38.73	30.71
Somewhat agree	41.84	41.84	32.62	45.07	37.14	45.77	40.85	40.71
Neither agree nor disagree	17.02	17.02	21.99	14.79	19.29	17.61	10.56	18.57
Somewhat disagree	4.96	4.96	7.09	3.52	7.14	5.63	7.75	7.14
Strongly disagree	0.71	0.71	3.55	1.41	2.14	0.70	2.11	2.86

**Figure 6.** Users' responses to the critical thinking for media literacy scale.

of misinformation and calls for a preliminary identification of the main standpoints and arguments making up the news. We are aware that self-reported information might not directly translate into behaviors and, in our case, into increased capabilities of identifying misleading information. However, crowdsourcing platforms do not offer a suitable environment to assess citizens' reasons behind their truth assessments, which require open-ended feedback, proficient English, and a population varied as to demographic features. In a preliminary experiment encompassing pre- and post-intervention surveys with the same population of workers, we have encountered issues such as fraudulent, nonsensical, and partial responding, which are common for complex tasks

that require a willingness to engage (Chmielewski & Kucker, 2020), but they do not allow for a valid assessment. As remarked by Garcia-Molina et al. (2016), macro-tasks pose more challenges than micro-tasks in a crowdsourcing environment where workers' starting points in terms of focus and knowledge are not transparent as well as their motivations to participate. In our case, a high number of workers, for instance, encountered difficulties in carrying out the two surveys in the right order and ended up finalizing one only.

To test a framework to measure the impact of the chatbots on the users' news interpretation processes, we have conducted a qualitative pilot experiment encompassing pre- and post-intervention feedbacks, rather than redesign the crowdsourcing

experiment. We, in fact, realized that it is hard to prompt users' engagement in such a transactional environment that does not guarantee an unbiased environment.

Pilot Qualitative Experiment. To investigate the impact of the chatbot on enhanced critical thinking skills, we recruited 20 participants with the help of the Pook FieldWork recruitment agency. The participants were balanced as to gender, half below and half over 45 years old, and with mixed socio-economic features (ABC1 & C2DE grades). The study took place in two workshop sessions on Zoom (40 min each), featuring 10 participants per session. The design of the first session was as follows: (a) pre-interaction phase during which participants have been asked through a Qualtrics questionnaire about whether they would believe five different news claims on a scale from 0 (*not at all*) to 100 (*completely*), and then explain their answer ("Please explain why you feel the claim is believable, unbelievable or what further information you would need to decide?"). It has to be noted that we did not ask them to rate a discrete truth value (e.g., "Pick one of the following option: True, Somewhat True, Mixed, Somewhat False, False"), since such a fact-checking task does not mirror the news consumption process where citizens are asked to decide whether to believe or not in the news they read with limited time capacity and knowledge about related facts. The chosen news claims were mixed as to topic, source, reliability, and presence of multimodal features (the full questionnaire is available at: https://liverpoolcommsmedia.fra1.qualtrics.com/jfe/form/SV_8uYG2fGcdalLl8a); (b) an interaction phase of 15 min during which they played with the FNI chatbot; (c) a post-interaction phase during which they completed the again the questionnaire in (a).

The second session shared the same design with the exception of phase (b), during which instead of playing with the chatbot, participants were asked to read a booklet of media literacy recommendations explaining the decalogue of fallacies and how to recognize them (the booklet is freely accessible at: <https://fakenewsimmunity.liverpool.ac.uk/wp-content/uploads/2021/03/Fake-News-Immunity-Liverpool-Uni-project.pdf>). We, in fact, wanted to test whether (a) learning fallacies affected news interpretation patterns and (b) a human-computer interaction environment has more or less influence compared with a static intervention.

Due to last-minute issues, one participant from the first group did not manage to join the session, while two participants from the second group did not fully complete the tasks (csv files showing the full results of the experiments are available on github folder X).

Based on the credibility scores, it is apparent that participants assign on average less credibility to the news when post-intervention (Figure 7).

The mean values of participants' scores are, in fact, lower after having played with the *Fake News Immunity Chatbot* or having read the booklet, even though the drop in trustworthiness is not statistically significant. The increased skepticism

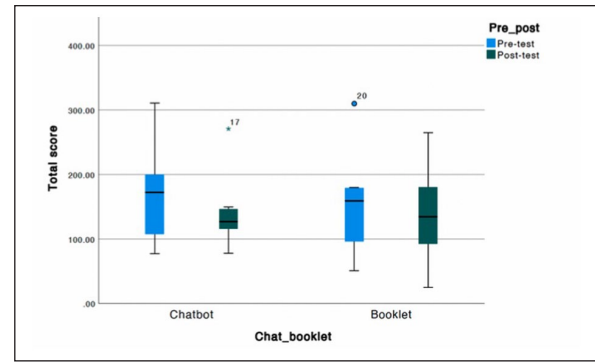


Figure 7. Clustered Boxplot of total scores pre- and post-interventions.

Table 3. Mean Values of Credibility Scores Pre- and Post-Interventions.

News claims	Chatbot pre	Chatbot post	Booklet pre	Booklet post
Q1	15.10	6.89	15.00	14.13
Q2	22.50	29.78	24.63	14.13
Q3	71.60	75.22	53.38	43.13
Q4	36.20	23.33	18.25	22.88
Q5	20.30	8.00	42.50	44.00

applies to the majority of the news claims, as displayed in Table 3.

A significant different behavior among the two groups is attested in response to Q5:

This is an Instagram post that became viral across social media: It claims that "Worldwide shortages of oil, gas, paper, milk, grain and other raw materials are not because of Ukraine."

Participants from the first group were already more skeptical pre-intervention compared with those of the second group and then radicalized their views instead of taming them after the intervention. Looking at the explanations it seems that after playing with the chatbot participants tended to classify the Instagram post as "opinion" rather than a fact, pointing to its defeasible nature (Figure 8).

The qualitative analysis of the open-ended questions shows some consistent changes in the post-interventions explanations underpinning participants' credibility rates along these lines:

- In both groups pre-intervention explanations pointed mostly to the trustworthiness of the sources (e.g., "unknown source, personal account unable to verify"; "Anything you see on social media can be true or false") without taking into account the actual content of the news claims; in post-intervention explanations, instead, more focus is paid on the information which is



Figure 8. Example of fallacious news from social media (Instagram).

conveyed both in terms of number of arguments (e.g., “Nothing to back up their view. Need more evidence that this is just due to Ukraine.”), their facticity (e.g., “Again its just a claim so not FACT [. . .]”) and their formal aspects (“it’s a bit of a strong allegation”).

- In both groups, post-intervention explanations contain element of skepticism, absent in the pre-intervention ones, leading to a suspension of judgment (e.g., “no way of knowing if true or false”; “just not sure”) and awareness that further information is required to assess the reliability of the claim (e.g., “Not sure, would need to read the article,” “[. . .] I would need to research myself as I am unaware of the number of diseases which humans have had over the years”).
- Post-interventions explanations by participants of the first group point to fallacious arguments which were not identified before the intervention and which are not clearly enucleated by group two (pre and post). It is, for example, the case for the participants who answered to Q5 as follows.

Example 1

Pre-intervention explanation (rating 70): “The war in Ukraine has obviously had an impact on materials being transported.”

Post-intervention explanation (rating 80): “It is misleading and doesn’t give the whole picture of the situation.”

The participant already noticed pre-intervention that Ukraine necessarily had an effect on supply issues for certain products; however, it is only after the intervention that she or he has been able to point to the cherry picking behavior of the post which provides as evidence the country’s geographical dimensions, neglecting the complexity of the political picture.

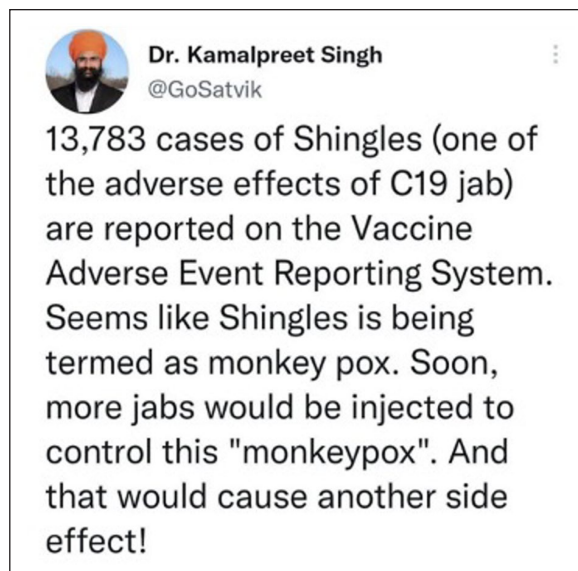


Figure 9. Example of fallacious news from social media (Twitter).

The role played by the chatbot experience has also to do with the number of criteria considered when making a reliability judgment. Let us consider, for instance, the explanations provided by a participant in response to Q1.

Example 2

This is a post on Instagram which shows a screenshot of a tweet:

The tweet claims: “13,783 cases of Shingles (one of the adverse effects of C19 jab) are reported on the Vaccine Adverse Event Reporting System. Seems like Shingles is being termed as monkey pox.”

Pre-intervention explanation (rating 10): “I wouldn’t believe that the two are related” (Figure 9).

Post-intervention explanation (rating 2): “Despite being specific in terms of numbers—no source, potential exaggerated and not verified poster.”

Although suspicion about the presence of causal relations also before intervention, post intervention the explanation is not presented as a personal belief, but supported by arguments that pinpoint various aspects of the message. It has to be noted that new means for epistemic vigilance acquired interacting with the chatbot do not necessarily correspond to changes in the assigned reliability values.

Example 3

This is a tweet that became popular on Twitter:

It claims that “Biden gave Americans the cheapest gas prices on Earth.”

Pre-intervention explanation (rating 30):

I don't really believe it as i feel i would have heard more about this online and on the news if this was true, however i know there has been a big issue with gas prices so it may be that the US is cheaper but not that cheap

Post-intervention explanation (rating 31):

I would need further information on the "rest of the world." How can you believe a claim that doesn't provide the data on the "cheapest prices on the Earth." Only prices shown are a few countries. The claim that "Biden gave" doesn't have any evidence either. Just sounds like somebody who supports Biden giving a biased opinion with very selective data to back it up.

Although the participant does not modify reliability score pre- and post-intervention, his or her cognitive heuristics are significantly updated: before the intervention hearsay and popularity on the news are perceived as truth benchmarks (with a clear risk for bandwagon effects); after the intervention, the participant is, instead, able to identify the partiality of the data reported in the chart (hasty generalization fallacy and cherry picking) as well as the lack of enough evidence (evading the burden of proof), other than partisanship, to attribute to Biden the responsibility for gas prices in the United States.

Conclusion

The advent of digitization has crucially changed the way we access, consume, and share news. The online information ecosystem has created new participatory models of news creation and consumption, but it has also widened the array of existing media distortions. More specifically, it has fueled misinformation, information that is misleading without necessarily the intention of being so. The gray area of misinformation is hard to debunk both by human and automatic fact-checkers due to the variety of distortions in place, which proliferate across digital media and that cannot be reduced to a binary problem of true versus false information. Prebunking efforts have been proved to be more effective, but have so far mostly targeted disinformation. As advocated by scholars and policymakers, what is needed to counter misinformation is critical thinking skills. The act of critical thinking news, that we call *reason-checking*, implies an assessment of the quality of the arguments that support a news claim, especially in situations such as the pandemic, where limited factual information is available. Our theoretical starting point is that the presence of flawed arguments—fallacies—works as an indicator of misleading information.

Drawing from Fallacy Theory and extant research and tools for active inoculation, we present two open access chatbots, the *Fake News Immunity Chatbot* (<http://fni.arg.tech/>) and the *Vaccinating News Chatbot* (fni.arg.tech/?chatbot_type=vaccine), to, respectively, teach citizens

and communication gatekeepers how to avoid believing, creating, and spreading misinformation. These tools differ with respect to state-of-the-art digital tools for active inoculation both in terms of design and learning outcomes. Their educational goal is not that of enhancing users' ability to directly disentangle truth from fakery, which might not be possible in uncertain epistemological scenarios, but to exercise users' critical thinking skills in questioning news' reliability. The design responds to this goal both from a backend and a frontend perspective. The knowledge base that underpins the scenarios portrayed in the chatbot is based on the multilevel analysis of 1,500 fact-checked news to surface fallacious arguments, which feature misinformation and their distribution across sources and types of news claims.

This reason-checking preliminary activity is aimed at prioritizing those types of fallacious arguments that are more frequent in the actual misinformation ecosystem, adopting a bottom-up approach. The underlying infrastructure keeps the knowledge base separated from the dialogue process to allow for updates in the informational substrate keeping the conversational dynamics. The process of reason-checking is, in fact, taught through a dialogical exchange with multiple users who engage in a group discussion; the underlying idea is that of simulating the current *media agora*, where multiple parties are engaged in the process of news construction. While the frontend follows state-of-the-art gamification principles, it also proposes a new heuristic for the identification of fallacies leveraging *critical questions* and philosophically inspired dialectic profiles for different software agents.

To evaluate the impact of the two chatbots, which have reached 1,700 users over 100 countries, we have made available a UX experience questionnaire to be voluntarily completed by users and we have conducted a crowdsourcing experiment. The questionnaire, so far filled in by 211 users, has revealed an (a) overall positive attitude toward conversational intelligence and interface; (b) trends in users' preferences (and reasons therein) for agents' personality types which, however, do not correlate with increased outreach. Aristotle is, in fact, perceived as the character preferred by the majority since knowledgeable; while such preference correlates with trustworthiness judgments, it doesn't make Aristotle the character to which users most frequently ask help for, suggesting that *perceived* authority might inhibit communication. Finally (c) users deem to have learnt fallacies and be likely to remember them, being able to describe their meaning in an accurate way. The crowdsourced survey was designed to assess self-reported perceptions of changes in critical thinking skills. We developed the first sensemaking scale for critical thinking self-assessment as news consumers and/or producers, drawing from factors identified in the context of public collective deliberation. The survey results show that users perceived an increase in each of the eight identified factors (Reflection, Insights, Focus, Argumentation, Assess Facts and Evidence, Distinguishing, Change Assumptions).

The pilot qualitative experiment to assess pre- and post-intervention changes in assessing news reliability has revealed an overall increased skepticism accompanied with an increased ability to identify fallacious arguments (especially after having used the chatbot), which promise to enhance epistemic vigilance against misinformation. These three-tiered pipelines to assess enhanced critical thinking skills for media literacy through pedagogical chatbots calls for further experiments to confirm the attested results. In particular, further impact evaluations through qualitative experiments are needed to assess whether users' perceptions translate into behaviors in the long term, while more research as to the viability of scaled-up evaluations is required. An option could be that of embedding an evaluation stage in the chatbot design. Despite its limitations, the *Fake News immunity Chatbot* and the *Vaccinating News Chatbot* are deemed to open doors for a new generation of digital tools to advance critical thinking for media literacy.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the UK Research and Innovation Economic and Social Research Council (Grant No. ES/V003909/1).

ORCID iD

Elena Musi  <https://orcid.org/0000-0003-2431-455X>

Notes

1. Although the whole article has been a result of continuous process of interaction among the authors, E.M. is the main contributor responsible for the design of the theoretical frameworks, the impact evaluations, and their analysis (sections 1, 2.3-4, 3.1.2, 3.2.1, 3.2.2, 4); E.C. contributed to the theoretical framework (sections 2.1; 2.2); C.R. contributed to the infrastructure of the chatbot (3.1.3); S.Y. contributed to the statistical analysis of the results of the qualitative experiment (3.2.3); and K.O. contributed as a mentor in the analytic pipeline.
2. Monthly updated reports of the questionnaire answers will be available on the authors' Github.

References

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85(2), 275–314.
- Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14), 1–6.
- Almalki, M., & Azeez, F. (2020). Health chatbots for fighting COVID-19: A scoping review. *Acta Informatica Medica*, 28(4), 241–247.
- Alsufiani, K., & Attfield, S. (2018, July). An exploratory study of computer supported sensemaking: Relating representational structure to triage. In *Proceedings of the 32nd international BCS human computer interaction conference* (pp. 1–5). Association for Computing Machinery.
- Altay, S., Hacquin, A. S., Chevallier, C., & Mercier, H. (2021). Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*. Advance online publication. <https://doi.org/10.1037/xap0000400>
- Altay, S., Schwartz, M., Hacquin, A. S., Allard, A., Blancke, S., & Mercier, H. (2022). Scaling up interactive argumentation by providing counterarguments with a chatbot. *Nature Human Behaviour*, 6, 579–592.
- Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & Linden, S. V. D. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, 8(1), Article 1013868.
- Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). *Types, sources, and claims of COVID-19 misinformation* [Doctoral dissertation]. University of Oxford.
- Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B., & Reifler, J. (2022). The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature Human Behaviour*, 6, 236–243.
- Carmi, E., Yates, S. J., Lockley, E., & Pawluczuk, A. (2020). Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, 9(2), 1–22.
- Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4), 464–473.
- Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE*, 12(5), Article e0175799.
- De Liddo, A., Souto, N. P., & Plüss, B. (2021). Let's replay the political debate: Hypervideo technology for visual sensemaking of televised election debates. *International Journal of Human-computer Studies*, 145, Article 102537.
- De Liddo, A., & Strube, R. (2021, June). Understanding failures and potentials of argumentation tools for public deliberation. In *C&T'21: Proceedings of the 10th international conference on communities & technologies-wicked problems in the age of tech* (pp. 75–88). Association for Computing Machinery.
- Eftekhari, M., Sotoudehnama, E., & Marandi, S. S. (2016). Computer-aided argument mapping in an EFL setting: Does technology precede traditional paper and pencil approach in developing critical thinking? *Educational Technology Research and Development*, 64(2), 339–357.
- Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher*, 18(3), 4–10.
- Facione, P. A. (1990). *The California critical thinking skills test: College level. Technical report #1—Experimental validation and content validity*. https://www.researchgate.net/profile/Peter-Facione/publication/268295703_The_California_Critical_Thinking_Skills_Test_College_

- Level_Experimental_Validation_and_Content_VValidity/links/587677f908aebf17d3b9a162/The-California-Critical-Thinking-Skills-Test-Co
- Freed, G. L., Clark, S. J., Butchart, A. T., Singer, D. C., & Davis, M. M. (2010). Parental vaccine safety concerns in 2009. *Pediatrics*, 125(4), 654–659.
- Garcia-Molina, H., Joglekar, M., Marcus, A., Parameswaran, A., & Verroios, V. (2016). Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering*, 28(4), 901–911.
- Goodnight, G. T. (2009). *Critical thinking in a digital age: Argumentation and the projects of new media literacy*. <https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1143&context=ossaarchive>
- Gordon, T. F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10–15), 875–896.
- Grasso, A., & Convertino, G. (2012). Collective intelligence in organizations: Tools and studies. *Computer Supported Cooperative Work*, 21(4), 357–369.
- Harrell, M. (2008). *iLogos* [Computer software]. http://www.phil.cmu.edu/projects/argument_mapping
- Hatcher, D. L., & Spencer, L. A. (2005). *Reasoning and writing: From critical thinking to composition*. American Press.
- Herriman, M., Meer, E., Rosin, R., Lee, V., Washington, V., & Volpp, K. G. (2020). *Asked and answered: Building a chatbot to address COVID-19-related concerns*. NEJM Catalyst Innovations in Care Delivery. <https://catalyst.nejm.org/doi/full/10.1056/CAT.20.0230>
- Horne, B. D., Gruppi, M., & Adali, S. (2019, December). Trustworthy misinformation mitigation with soft information nudging. In *2019 first IEEE international conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)* (pp. 245–254). Institute of Electrical and Electronics Engineers.
- Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018m, June). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference* (pp. 895–906). Association for Computing Machinery.
- Johnson, R. H., & Blair, J. A. (2006). *Logical self-defense*. Idea.
- Koltay, T. (2011). The media and the literacies: Media literacy, information literacy, digital literacy. *Media, Culture & Society*, 33(2), 211–221.
- Kull, S., Ramsay, C., & Lewis, E. (2003). Misperceptions, the media, and the Iraq war. *Political Science Quarterly*, 118(4), 569–598.
- Kunsch, D. W., Schnarr, K., & van Tyle, R. (2014). The use of argument mapping to enhance critical thinking skills in business education. *Journal of Education for Business*, 89(8), 403–410.
- Kyriakidou, M., Morani, M., Soo, N., & Cushion, S. (2020). Government and media misinformation about COVID-19 is confusing the public. *LSE Covid-19 Blog*. <https://blogs.lse.ac.uk/covid19/2020/05/07/government-and-media-misinformation-about-covid-19-is-confusing-the-public/>
- Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 32(2), 348–384.
- Martin Davies, W. (2009). Computer-assisted argument mapping: A rationale approach. *Higher Education*, 58(6), 799–820.
- McCright, A. M., & Dunlap, R. E. (2011). The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly*, 52(2), 155–194.
- Mercier, H., & Sperber, D. (2017). *The enigma of reason: A new theory of human understanding*. Allen Lane.
- Metzger, M. J. (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), 2078–2091.
- Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59, 210–220.
- Musi, E., Aloumpi, M., Carmi, E., Yates, S., & O'Halloran, K. (2022). Developing fake news immunity: Fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies*, 12(3), Article e202217.
- Musi, E., & Reed, C. (2022). From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society*, 33(3), 349–370.
- Newman, N., Fletcher, R., Robertson, C., Eddy, K., & Nielsen, R. (2022). *Reuters Institute digital news report 2022*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf
- Park, S., Park, J. Y., Kang, J. H., & Cha, M. (2021, January 27). The presence of unexpected biases in online fact-checking. *The Harvard Kennedy School Misinformation Review*. <https://misinformationreview.hks.harvard.edu/article/the-presence-of-unexpected-biases-in-online-fact-checking/>
- Paul, R. (1981). Teaching critical thinking in the “strong” sense: A focus on self-deception, world views, and a dialectical mode of analysis. *Informal Logic*, 4(2), Article 2766.
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50.
- Pfau, M., Ivanov, B., Houston, B., Haigh, M., Sims, J., Gilchrist, E., Russell, J., Wigley, S., Eckstein, J., & Richert, N. (2005). Inoculation and mental processing: The instrumental role of associative networks in the process of resistance to counterattitudinal influence. *Communication Monographs*, 72(4), 414–441.
- Reed, C., & Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04), 961–979.
- Saleh, N. F., Roozenbeek, J., Makki, F. A., McClanahan, W. P., & Van der Linden, S. (2021). Active inoculation boosts attitudinal resistance against extremist persuasion techniques: A novel approach towards the prevention of violent extremism. *Behavioural Public Policy*. Advance online publication.
- Sethi, R. J., Rangaraju, R., & Shurts, B. (2019, June). Fact checking misinformation using recommendations from emotional pedagogical agents. In *International conference on intelligent tutoring systems* (pp. 99–104). Springer.
- Shareef, M. A., Kapoor, K. K., Mukerji, B., Dwivedi, R., & Dwivedi, Y. K. (2020). Group behavior in social media: Antecedents of initial trust formation. *Computers in Human Behavior*, 105, Article 106225.
- Siedlikowski, S., Noël, L. P., Moynihan, S. A., & Robin, M. (2021). Chloe for COVID-19: Evolution of an intelligent

- conversational agent to address infodemic management needs during the COVID-19 pandemic. *Journal of Medical Internet Research*, 23(9), Article e27283.
- Starbird, K., Spiro, E. S., & Koltai, K. (2020, June 20). *Misinformation, crisis, and public health: Reviewing the literature V1.0*. Mediawell—Social Science Research Council.
- Stott, A., & Neustaedter, C. (2013). *Analysis of gamification in education*. <http://clab.iat.sfu.ca/pubs/Stott-Gamification.pdf>
- Strickland, J. C., & Stoops, W. W. (2019). The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. *Experimental and Clinical Psychopharmacology*, 27(1), 1–18.
- Tay, L. Q., Hurlstone, M. J., Kurz, T., & Ecker, U. K. (2021). A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology*, 113, 591–607.
- ter Berg, T., van der Brugge, E., & Cullen, S. (2013). Teaching critical thinking with rationale. In 7th International Technology, Education and Development Conference. https://www.reasoninglab.com/wp-content/uploads/2013/10/INTED2013-Ter-Berg_Van-der-Brugge.pdf
- Tsai, W. H. S., Liu, Y., & Chuan, C. H. (2021). How chatbots' social presence communication enhances consumer engagement: The mediating role of parasocial interaction and dialogue. *Journal of Research in Interactive Marketing*, 15, 460–482.
- Van Der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology*, 11, Article 2928.
- Van Eemeren, F. H., & Grootendorst, R. (2003). A pragma-dialectical procedure for a critical discussion. *Argumentation*, 17(4), 365–386.
- Visser, J., Lawrence, J., & Reed, C. (2020). Reason-checking fake news. *Communications of the ACM*, 63(11), 38–40.
- Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.
- Wells, S., & Reed, C. A. (2012). A domain specific language for describing diverse systems of dialogue. *Journal of Applied Logic*, 10(4), 309–329.
- Yates, S. J., & Carmi, E. (2022). Citizens' networks of digital and data literacy. In J. McDougall & K. Fowler-Watt (Eds.), *The*

Palgrave handbook of media misinformation (pp. 191–205). Palgrave Macmillan.

Author Biographies

Elena Musi (PhD, Università della Svizzera italiana) is a Senior Lecturer in Communication and Media at the University of Liverpool where she is Program lead of the MSc in Data Science and Communication. Her research interests rest at the interface between artificial intelligence and communication, including theoretical and applied argumentation, (mis)information, and the development of human–computer interaction technologies to advance critical thinking skills.

Elinor Carmi (PhD, Goldsmiths, University of London) is a Lecturer in Digital Society at the Department of Sociology and Criminology at City, University of London, UK. Her research interests include data politics, data literacies, data feminism, data justice, digital rights, and sound studies. She is working on several funded projects around data literacies and is a parliamentary academic fellow for the Department of Digital, Culture, Media and Sports Committee in the United Kingdom.

Chris Reed (PhD, University College London) is Professor of Computer Science and Philosophy at the University of Dundee where he is Director of the Centre for Argument Technology (www.arg.tech). His research interests include philosophical foundations of argumentation, linguistics and dynamics of debate, and AI applications for representing, navigating, visualizing, and enhancing arguments at scale.

Simeon Yates (PhD BA, BSc, DipNatSci) is a Professor of Digital Culture at the University of Liverpool. His research interests include digital inequalities, digital media, and broad questions about the social impacts of digital systems.

Kay O'Halloran (PhD, Murdoch University) is a Professor of Communication and Media at the University of Liverpool. Her research interests include multimodal analysis and the development of digital tools and techniques for analyzing multimodal datasets at scale.