CHAPTER 11

# Developing Misinformation Immunity in a Post-Truth World: Human Computer Interaction for Data Literacy

*Elena Musi, Kay L. O'Halloran, Elinor Carmi, and Simeon Yates*

## 11.1 Introduction

One of the major challenges of the current information ecosystem is the rapid spread of fake news through digital media. The phenomenon of fake news is complex and includes at least three types of media distortions. First, disinformation, non-factual information which is spread with the intention of disseminating harmful false information. Second, misinformation, information which is misleading but not created with the intention

E. Musi (✉) • K. L. O'Halloran • S. Yates
Department of Communication and Media, University of Liverpool, Liverpool, UK
e-mail: elena.musi@liverpool.ac.uk; kay.ohalloran@liverpool.ac.uk; scssy@liverpool.ac.uk

E. Carmi
Department of Sociology, City University, London, UK
e-mail: Elinor.Carmi@city.ac.uk

B. Schirrmacher, N. Mousavi (eds.), *Truth Claims Across Media*,
Palgrave Studies in Intermediality,
https://doi.org/10.1007/978-3-031-42064-1_11

of producing harm. Third, malinformation, information that is based in reality, but is used to inflict harm on a person, organisation or country (Carmi et al., 2020).

Even though misinformation is often disseminated unintentionally, it has a wide societal impact. For example, in a sample of 225 cases of fake news collected between January and March 2020, 59% of fake news does not contain fabricated or imposter content, but rather reconfigured misinformation (Brennen et al., 2020). Such information proliferates through social media, the main source of news for infodemically vulnerable citizens. Since it is unintentional, misinformation is not always blocked through internal fact-checking before a news article is published. More generally, the identification of misinformation is also far from being successfully addressed by human third parties fact-checkers, let alone automated techniques aimed at verifying the accuracy of information.

This is largely due to the lack of an agreed upon *truth barometer*, which in turn hinders the creation of datasets to train automatic systems. Continuous updates about Covid-19 from the scientific community, as well as governments and health institutions, often results in media outlets unintentionally disseminating misleading content. What makes these types of news *fake* is not the mere truthfulness of the information conveyed, but the fallacious way the arguments are presented (Musi & Reed, 2022). The news making process is, in fact, a rhetorical and argumentative exercise since it is aimed at gaining the acceptance of a certain interpretation of a news event. Misleading or misrepresentative arguments, if perceived as coherent and trustworthy, might thus crucially affect the processes through which information turns into shared public knowledge. In such an environment, to counter the fake news phenomenon, it is necessary to go beyond the identification of non-factual information towards a reason-checking exercise "evaluating whether the completer argumentative reasoning [underpinning news] is acceptable, relevant and sufficient" (Visser et al., 2020, 38).

As part of our UKRI funded project *Being Alone Together: Developing Fake News Immunity*,[1] we proposed to counter misinformation by providing the means for (a) citizens to act as their own fact-checkers and (b) communication gatekeepers (e.g. journalists and news editors) to avoid creating and spreading misleading news. We did so by combining Fallacy Theory (Carmi et al., 2021) with Human Computer Interaction (HCI).

[1] https://fakenewsimmunity.liverpool.ac.uk/

Drawing from the multi-level annotation of a dataset of 1500 Covid-19 related news web-crawled from 5 English fact-checkers (*Snopes*, *The Ferret*, *Politifact*, *Healthfeedback.org*, and *Fullfact*), we propose a systematic procedure to identify fallacious arguments across different digital media sources and type of claims (e.g. predictions, interpretations). Relying on the analysis of significant correlations (positive $p$ values) among types of fallacies, sources, and claims, we show trends in the way misinformation is constructed and communicated (Musi et al., 2022).

Leveraging the outcomes of our data analysis, we built two chatbots, the *Fake News Immunity Chatbot*[2] and the *Vaccinating News Chatbot*,[3] respectively targeting citizens and communication gatekeepers. Through these chatbots, users learn how to fact-check through fallacies and create fallacy-free news content through interactions with the *fathers of critical thinking* (i.e. Aristotle, Gorgias, and Socrates) and members of the research team. While adhering to default design principles of gamification environments (e.g. progressive game levels), the two chatbots give voice through their dialogical templates to philosophical modes of inquiry (e.g. Socratic maieutic) with the goal of increasing users' learning process in a conversational environment. In this paper, we present four aspects of our work. First, we introduce the notion of data literacy as new form of media literacy. Second, we introduce the heuristics we developed in order to teach how to fact-check misinformation through fallacies (Sect. 11.2.2). Third, we describe the design of a human-computer interaction environment (the *Fake News Immunity Chatbot*) as a pedagogical tool to assist citizens in learning how to reason-check misinformation (Sects. 11.2.3, 11.3.1, 11.3.2). Fourth, we report on the beta testing of the chatbot through survey-based focus groups aimed at eliciting advantages and pitfalls of the devised human-computer interaction tool. The results shed light on multimodal factors which affect users' trust in AI agents, suggesting that HCI can be effectively used to augment rather than replace human skills through a tailored design thinking.[4]

[2] http://fni.arg.tech/

[3] http://fni.arg.tech/?chatbot_type=vaccine

[4] Although the whole paper has been the result of a continuous process of interaction among the authors, Elena Musi is the main responsible of Sects. 11.1, 11.2.2 and 11.3; Elinor Carmi of Sect. 11.2.1 and 11.2.2. Simeon Yates and Kay O'Halloran have contributed to the questionnaire design/analysis and elaboration of recommendations.

## 11.2 Media Literacy in the Post-truth World

### *11.2.1 From Media Literacy to Data Literacy*

People's engagement with and understanding of media devices and digital systems have been intertwined with their levels of trust towards institutions and other people. These two main relations have influenced people's media literacies and their data literacies, particularly during the Covid-19 pandemic. We developed our data literacies framework, which we call Data Citizenship (Carmi et al., 2020), across three main dimensions: (1) *Data doing*—Citizens' everyday engagements with data (for example, using data in an ethical way); (2) *Data thinking*—Citizens' critical understanding of data (for example, verifying information and sources online); and (3) *Data participating*—Citizens' proactive engagement with data and their networks of literacy (for example, helping others with their data literacy through games/chatbots). Trust is a common thread that relates to all these three dimensions and consequently how people navigate the datafied ecosystem.

When it comes to trust in institutions, this relates to people's critical thinking about how reliable they perceive a specific institution to be (e.g. a news outlet or health institute) and therefore whether they should read or believe their messages. Fletcher et al. (2020), for example, show how during the Covid-19 pandemic people's attitudes in the United Kingdom towards the trustworthiness of news outlets have decreased from 57% to 46%, and their trust levels in the government have declined from 67% to 48%. In addition, as Cushion et al. (2021) have shown, UK citizens have broadly managed to identify "fake news", but they felt confused by the statistics and the neglect of important facts, such as how the pandemic was being handled by the UK government. According to Cushion et al. (2021), this was mainly due to editorial decisions where sufficient details were not conveyed to citizens, with the result that people felt misinformed. Cushion et al. points out both how people's trust in the main sources of information has deteriorated during the pandemic but also, importantly, that the arguments presented to them by mainstream media were confusing and were causing a proliferation of misinformation.

Such a situation cannot be simply solved by relying on fact-checking organisations as arbiters of trustworthy information. Not only is there a great abundance of non-fact checked information but the epistemology of fact-checking is far from standardised and affected by selection effects and other types of biases (Uscinski & Butler, 2013).

Drawing from our research (Yates et al., 2021), it seems that the answer lies in a collective effort: We found that people mainly rely on their personal *networks of literacy* to verify information and learn new data literacy skills. We see this as a modern digital version of the 2-step-flow model of influence, originally conceived by Katz (1957) taking place in citizens networks of literacy. In relation to the *Fake News Immunity Chatbot*, this means that people can develop their *data thinking* skills using the game. This teaches them how to critically use fallacy theory as they search for and identify reliable/trusted sources and how to locate reliable articles on social media. It also encourages them to practice *data participating* by either playing with others in their networks of literacy or through propagating these key ideas through these networks. This can help different people in their networks with lower data literacies.

### *11.2.2 Fallacies as Misperceptions of Truthfulness*

The pandemic has created an epistemological situation in which what counts as true is continuously updated. A vaccine trial could, for instance, by default, offer reliable but constrained truths about potential side-effects, while reporters rely on second-hand evidence due to lockdown restrictions. In such a scenario, the distinction between mediation, the "material prerequisites for representation in media", and representation, "the semiotic operation, that is, the creation of meaning in the mind" (Elleström, 2017, 663) is almost removed since the representation necessarily happens in a mediated environment. That is, our sense-making processes have to rely on truth claims which depend on the reasons provided by a media product supporting their trustworthiness (Elleström, 2014, 2017, 2021).

In the era of citizens' journalism, the news making process is inherently transmedial, and it assumes the shape of a polylogue (Musi & Aakhus, 2018) where multiple users (often anonymous) negotiate different opinions across different venues (from social media to fora). In the absence of a gatekeeping process, the discourse through which a news claim is shaped becomes the main guarantor for its truth. Thus, its persuasiveness through rhetorical strategies impacts on our perception of truthfulness. In light of this, fallacies, i.e. arguments that seem valid but are not (Hamblin, 1970), are likely to support claims which, even if not false, might be misleading and trigger misperceptions of truthfulness.

We call these misinformation news "semi-fake" since "created/shared by the authors with the intention of circulating fabricated information and hard to be flagged by the public through common ground knowledge" (Musi & Reed, 2022, 17). *Fallacies, thus, constitute useful means to identify fake news*. This is especially true when misinformation—information which is misleading but not necessarily non-factual or created with the intention of causing harm—rather than disinformation—i.e. information which is blatantly false—is at stake.

Two caveats have, however, to be considered. Firstly, verifiable news can also be supported by fallacious arguments, and secondly, there is so far no agreed upon taxonomy for fallacies. As to the former, our goal is to make people aware of fallacious arguments as means for scrutiny rather than truth verdicts. Our guidelines for the analysis of fallacies are, in fact, based on critical questions which cast doubt on various aspects of the news. As to the latter, we have adopted an empirical approach in selecting a decalogue of fallacies relevant for the current misinformation ecosystem. We analysed a preliminary set of 40 fact-checking commentaries and their source articles randomly picked from the fact-checker *Healthfeedback.org*. We developed guidelines for 10 fallacy types found in the data, which include a definition, identification questions, and an intuitive example, as presented below.

- **EVADING THE BURDEN OF PROOF**
  Definition: A position is advanced without any arguments supporting it as if it was self-evident.
  Critical Questions:
    1. Does the position express an unassailable fact?
    2. Are there any arguments in support of the statement apart from personal guarantee?

  Example: A politician tweeting that a vaccine for Covid-19 was found without providing proof.

- **STRAWMAN**
  Definition: An intentional misrepresentation of the other side's opinion is attempted.

Critical Questions:

1. Has an opponent's position been misrepresented?
2. Is that misrepresentation the basis for an attack or dismissal of the opponent's claim or argument?

Example: A politician arguing that he does not have to follow the advice of the World Health Organization (WHO) since it did not give positive results in the past, even though that piece of advice was good at that time and in that context.

- **FALSE AUTHORITY**

  Definition: An appeal to authority is made where the source lacks credibility in the discussed matter or (s)he is attributed a statement which has been tweaked.

  Critical Questions:

  1. Is the proposed person or source a genuine/impartial authority?
  2. Did the authority make the attributed claim?
  3. Are the authority and claim made relevant to the subject matter?

  Example: When a politician says he knows that the climate crisis does not exist because he did research on it.

- **RED HERRING**

  Definition: The argument may be formally valid, but its conclusion is irrelevant to the issue at stake.

  Critical Questions:

  1. Has the issue been shifted in the course of an argument to another issue or different aspect of the same issue and not shifted back?
  2. Is the shift irrelevant to addressing the initial issue?

  Example: When a politician is asked to assess the seriousness of the Covid-19 pandemic and replies that corruption is a worse problem.

- **CHERRY-PICKING**

  Definition: The act of choosing among competing evidence that which supports a given position, ignoring or dismissing findings which do not support it.

Critical Questions:

1. Is the evidence reported the only available?
2. Is there any other data available which would bring to a different conclusion?

Example: When a politician announces that schools should be open because one research project indicates that children are less affected by a virus, whilst different research suggests otherwise.

- **FALSE ANALOGY**

  Definition: Since two entities or situations are similar in one or more aspect they mist be similar in other aspects as well.

  Critical Questions:

  1. Are the two situations alike for real?
  2. Are the similarities relevant to derive the conclusion?
  3. Are there any dissimilarities relevant for the conclusion?

  Example: When someone compares Covid-19 with regular flu.

- **HASTY GENERALIZATION**

  Definition: A generalization is drawn from a numerically insufficient sample or a sample that is not representative of the population or a sample which is not applicable to the situation if all the variables/circumstances are considered.

  Critical Questions:

  1. Is the considered sample quantitatively large enough?
  2. Is the considered sample representative of a population or it has been selected in a biased way?
  3. Is the considered sample relevant due to the circumstances of a present situation or does it constitute an exception?

  Example: Arguing that all people from a specific race are more likely to refuse to wear face-masks because of one incident.

- **POST HOC**

  Definition: It is assumed that because B happens after A, it happens because of A. In other words a causal relation is attributed where, instead, a simple correlation is at stake.

Critical Questions:

1. Is there a correlation supporting the causal claim? That is, are there a number of cases on which the claim is grounded?
2. Can the move from the correlation to the alleged causal link be explained by coincidence?

Example: Claiming that 5G is causing Covid-19.

- **FALSE CAUSE**
  Definition: X is identified as the cause of Y when another factor Z causes both X and Y or X is considered the cause of Y when actually it is the opposite.
  Critical Questions:
  1. Is the causal claim itself credible? That is, are the cause and effect correctly identified and has an underlying common cause of both clearly been ruled out?

  Example: When someone claims that ibuprofen makes Covid-19 worse.

- **AMBIGUITY/VAGUENESS**
  Definition: A word/a concept or a sentence structure which are ambiguous are shifted in meaning in the process of arguing or are left vague being potentially subject to skewed interpretations.
  Critical Questions:
  1. Have key terms, concepts, or phrases retained their initial meanings throughout the argument?
  2. Does a word, concept, or phrase have no clear meaning in the context in which it arises?
  3. Does that vagueness prevent us from being able to judge whether an argument has occurred or what it might be?

  Example: A council stating that there is a fair number of available swabs without specifying what "fair" means.

Using these ten fallacy classifications, we undertook an annotation exercise involving two minimally trained students and an expert annotator to solve cases of disagreement to analyse a dataset of news web scraped by five English factcheckers (*Snopes*, *The Ferret*, *FullFact*, *Politifact*, and *Healthfeedback.org*) from January 2020 to end of June 2020 and from January 2021 to end of March 2021. The resulting dataset consisted of 1500 claims.

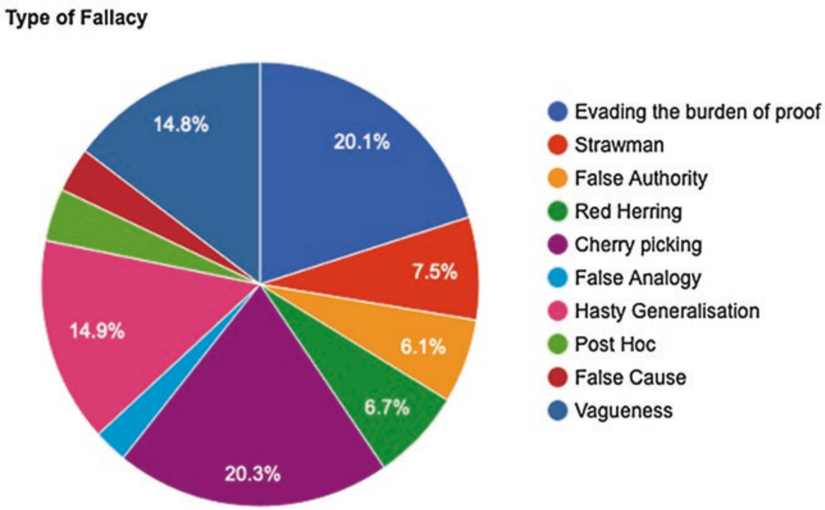


**Fig. 11.1** Distribution of fallacy types in the dataset

The results show that the 10 fallacies are associated with all cases of misinformation according to the distributions displayed in Fig. 11.1. We developed series of recommendations on how to identify such fallacies, which are publicly available in a simplified format.[5]

The annotation experiment has shown that non-experts are able to successfully identify the majority fallacies when aided with a set of heuristic guidelines. However, the task is highly complex and cognitively taxing, as shown by the cases of disagreement in the annotation exercise. For this reason, an active learning environment grounded in educational technology promises to foster engagement while retaining a focus on the task at hand, as discussed below.

### *11.2.3 Human Computer Interaction as an Educational Tool for Data Literacy*

According to the report *Testing and Refining Criteria to Assess Media Literacy Levels in Europe* (Shapiro & Celot, 2011) commissioned by the

[5] https://fakenewsimmunity.liverpool.ac.uk/wp-content/uploads/2021/03/Fake-News-Immunity-Liverpool-Uni-project.pdf

European Union, one of the factors which hinders critically reflective skills is screen time since it affects the way we access information (in a fragmented rather than holistic vein). Such a negative correlation is particularly problematic in times such as the pandemic where educational settings are constrained to virtual environments. The situation is further complicated by the proliferation of fake news which is frequently hard to pinpoint, especially as it proliferates through social media. In such a scenario digital fluency is hard to achieve. Teaching people how to identify reliable sources of information is, in fact, not enough in the current (mis)information ecosystem since official news venues can convey misleading information regardless of their intentions.

Relying on fact-checkers' comments as material to develop their critical media literacy (Kellner & Share, 2007) also has its limitations. Besides struggling to keep up with the abundance of information spread online, fact checkers make use of ratings that do not inform the readers about the basis for the misinformation (e.g. *Mostly False/Half True*). Rather, these online factor checkers simply apply a veracity value with little pedagogical relevance. For example, the following news was analysed by the factchecker *Snopes*: "In January 2021, the World Health Organization warned that pregnant women should avoid the COVID-19 vaccine". The rating assigned by the fact-checker is *Mixture*, which points to the presence of both elements of truth and of falsity without, however, instructing on how to identify misleading aspects. For example, is it true that pregnant women should avoid the vaccine or is it not accurate to state that the WHO expressed a view upon the issue?

Crisis situation such as the pandemic bring to the fore two challenges identified by Kline (2016) in relation to the critical media literacy framework. Firstly, while the pedagogical focus is on teaching how the content of media messages responds to political and economic agendas, the problem of media at the level of form is underestimated. Secondly, when it comes to misinformation, media affordances play a crucial role. Word limits imposed by a social media platform can, for instance, cause cherry-picking behaviours of information. Rendering a headline more clickable might bring to misrepresentations of various kinds, while the "sharing without caring" widespread attitude might make misleading content viral.

In relation to fallacies, two main interconnected educational challenges imposed by the virtual medium need to be considered to guarantee an effective learning environment: (1) learning how to recognize fallacies, as any other critical thinking skill, implies a thinking slow process (Kahneman,

2011) which clashes with the *scrolling* behaviours of today's news readers; and (2) the digital medium tends to reduce the attention span if not involving interaction. To cope with these issues, we developed the *Fake News Immunity Chatbot* which is an online publicly accessible game. This game aligns with the pedagogical hallmark of boosting active learning methodologies. That is, through the game, users learn how to fact-check news using fallacies by talking to a set of characters which embody ancient philosophers. Gamification has, in fact, been proven to foster students' commitment and motivation, which may lead to improvement of critical skills (Huang & Soman, 2013). For example, the *GoViralGame*[6] was recently launched by Cambridge researchers to introduce players to four tactics (e.g. using charged language) used to spread fake news online. The concept behind the game is that exposing people to a mild dose of the ways used to disseminate false information will help to generate *inoculation*. Inoculation Theory (Compton, 2013; McGuire & Papageorgis, 1961) is based on the assumption that "just as vaccines generate antibodies to resist future viruses, inoculation messages equip people with counter arguments that potentially convey resistance to future misinformation, even if the misinformation is congruent with pre-existing attitudes" (Cook et al., 2017, 4).

In our case, the *Fake News Immunity Chatbot* leverage fallacies as a *vaccine* to *inoculate* against fake news. As opposed to *GoViralGame*, it is focused on misinformation rather than disinformation and on identifying triggers, which can make a news misleading even when the author did not mean to spread false content. To assess the efficacy of gamification to teach critical thinking, we tested the *Fake News Immunity Chatbot* with two cohorts of postgraduate students (36 participants overall). After having played with the chatbot, students are asked to complete a questionnaire to evaluate the chatbot at different levels. The choice of having students self-reflect upon various aspects of the chatbot instead of merely tracking their interaction times and behaviours is driven by two main reasons: (a) It increases the perception of their role and responsibility as beta testers in a research-led teaching environment, and (b) it serves the learning outcome of making students mull over the role that human computer interaction might have in facilitating rather than replacing human decision making and reasoning (Vinuesa et al., 2020). The gamification exercise is discussed below.

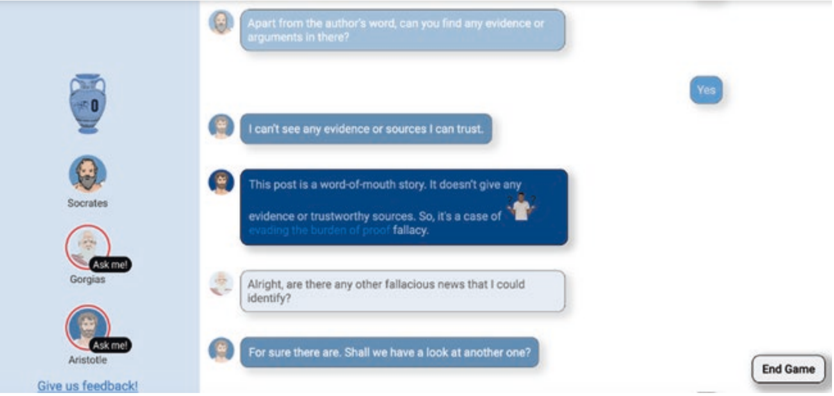
[6] https://www.goviralgame.com/en

## 11.3 The *Fake News Immunity Chatbot*

### *11.3.1 Chatbot Design*

The *Fake News Immunity Chatbot* has been created with the overall goal of reverse-engineering the manipulation of information. It is designed to use Fallacy Theory to teach citizens how to act as fact-checkers by training them in critical thinking. As users, citizens are in essence signing up to be students of a fact-checking initiative. After having been fronted with a summary of the chatbot rationale, users are introduced to the other participants, the three avatars of the Ancient philosophers Aristotle, Gorgias, and Socrates and the avatars of the members of the research team, among whom they are asked to select an interlocutor. The conversation begins *in media res*, with the selected avatar asking the user to assess the reliability of a news item, then cross-checking it with the fact-checker's answer. After this first prompt, if the user decides to be willing to learn how to fact-check through fallacies, they are asked to read a news item and answer questions by the philosophers, while challenged in their decision-making process. For example, Fig. 11.2 shows an example of how the user is guided through the fact-checking process by identifying fallacies.

As the chatbot is educational, its dynamics beyond the conversational outline have been designed to match those gamification principles that have turned out to be most pedagogically effective (Stott & Neustaedter, 2013):



Apart from the author's word, can you find any evidence or arguments in there?

Yes

I can't see any evidence or sources I can trust.

This post is a word-of-mouth story. It doesn't give any evidence or trustworthy sources. So, it's a case of evading the burden of proof fallacy.

Alright, are there any other fallacious news that I could identify?

For sure there are. Shall we have a look at another one?

**Fig. 11.2** Example of guided fact-checking through fallacies identification

- *Freedom to Fail*: Users can fail any of the identification questions without having to start all over again, provided that they read the explanation provided by the philosophers and amend their choices. Accordingly, each step of the decision-making process works as a formative assessment during which they assess their digital literacy while interacting with *experts* in the field.
- *Rapid Feedback*: As underlined by Kapp (2012), feedback is a critical element in learning that is especially effective when targeted. We have thus, ensured that users receive continuous and fast paced feedback since their answers are immediately commented by the philosophers in the format of an argumentative discussion. Furthermore, the students can ask for help to Aristotle, Socrates, or Gorgias before making a choice at any stage of the game.
- *Progression*: Users are invited to follow a progression path across three incremental levels (i.e. credulous, skeptic, and agnostic) to keep track of their learning process in an organized manner. For every eight correct answers they receive a point as a reward for becoming an expert in recognising fallacious news.
- *Storytelling*: Since the misinformation ecosystem during the pandemic can be overwhelming, we decided to build a narrative centred around ancient Greece as *the cradle* of Critical Thinking. The main participants are among the fathers of informal Logic—Aristotle, Socrates, and Gorgias—represented through avatars and dialogical patterns which mirror their historical portraits and philosophies. In addition, users can choose their own avatar among those representing the members of the research team. The points earned throughout the game are represented through gadflies in honour of Socrates, described by Plato as a gadfly who stings people with his questions to keep them on track in the pursuit of virtue. Such a setting is also meant to induce users realizing the evergreen role played by ancient philosophy and critical thinking to solve contemporary issues.

### *11.3.2 Design of the Gamification Experience*

Drawing from Huang and Soman (2013), we designed the gamification experience accounting for: (1) target audience and context, (2) learning objectives, (3) structure of the experience, and (4) gamification elements. As to the audience, the students that participated in the study were enrolled in the postgraduate modules “Artificial Intelligence and

Communication" (MSc in Data Science and Artificial Intelligence) and "Discourse, Rhetoric and Society" (MSc in Strategic Communication) at the University of Liverpool in the United Kingdom. Both student cohorts were introduced during the modules to key concepts related to the (mis) information ecosystem, such as the distinction between misinformation, disinformation, and malinformation (Carmi et al., 2020), the blurred notion of fake news (Tandoc et al., 2018) and the fact-checking process. Both modules took the form of workshops and followed a blended learning approach during which students were presented with notions followed by active learning activities and discussions in small groups. The modules were hosted on Zoom due to the pandemic, which resulted in various challenges, including Zoom fatigue. Students were also provided with an overview of the scope, the goals, and the methodologies adopted within the research project. However, they were not presented with a thorough explanation of fallacy theory and its relevance for misinformation before playing with the *Fake News Immunity Chatbot*. They were, instead, instructed about the role played by their feedback as beta testers in improving the chatbot in view of its launch to foster their sense of self-ownership and motivation.

The learning objectives of the sessions hosting the gamification experience were to: (1) learn how to fact-check news through fallacies and (2) reflect upon the role played by human computer interaction in learning skills. From the analysts' point of view, we wanted to understand (1) whether the heuristic implemented in the chatbot is effective in teaching how to recognize fallacies, and (2) what are best strategies to guarantee human-computer trust in the context of the misinformation ecosystem. The latter aspect is crucial in building effective pedagogical digital interventions for data literacy, but it has so far been under-investigated. The majority of studies have tackled the need for human-computer trust scales which cut across domains (Gulati et al., 2019).

However, the information ecosystem is peculiar since social bots are generally discussed as fake news spreaders. They are, therefore, potentially associated with dis-information by the larger public, rather than as gatekeepers of truth. Furthermore, while persuasive technology is usually associated with human-likeliness, this tendency might not apply to a context where radical uncertainty makes peers somewhat unreliable as experts.

The gamification experience was structured in the same way for both cohorts. That is, students had 15 minutes to freely play with the chatbot, after which they were asked to fill in a questionnaire on *Qualtrics*[7] to evaluate the chatbot for an estimated time of 10 minutes. Even though students were then encouraged to discuss their experience with their peers, the training activity was mainly self-led to avoid face-threatening situations (e.g. a student might be faster in correctly completing a stage). The students could select which game level (i.e. credulous, skeptic, and agnostic) to start with to allow them to build their own strategies in training as factcheckers (Simões et al., 2013).

### *11.3.3 Questionnaire Design*

The questionnaire has been designed to assess four different facets of the chatbot in line with evaluation criteria set up by Jain et al. (2018). The aspects considered, the associated questions and their underlying rationale are the following:

- **Conversational Intelligence**
  Q1: How was the rhythm of the conversation flow? Too slow/slow/just right/fast/too fast.
  Q2: How did you find the tone? Too formal/formal/ just right/informal.
  Q3: The conversation had contributions from several participants. Sometimes the AI participants talked amongst themselves. Did you find this: interesting/confusing/informative/boring.
  Q4: Did you feel that you managed to actively participate in the conversation? Active/just right/not active/sometimes.
  Q5: Pick one or more of these sentences if you agree:
    - These philosophers talk the same way to everyone…They did not even remember my name.
    - Finally a chatbot with no “bro language”.
    - Everyone there is so serious… Cheer up guys!
    - Having questions to ask yourself while reading really helped me out!
    - Sometimes I did not feel ready to choose yes or no…. The world is not black and white!

[7] https://www.qualtrics.com/

*Rationale:* The questions are aimed at understanding what conversational features prompt students' engagement in a virtual setting. Even though the ultimate goal of chatbot developers is that of achieving human-like conversation, the natural language patterns with the strongest educational value in a gamification environment have been understudied. For instance, the presence of options that allow to delay choices (e.g. "maybe later", "I do not know") seem to increase users' engagement in commercial chatbots (Valério et al., 2020), but might have a different outcome in an educational setting.

- **Chatbot Personality**
  Q6: Who is your favourite participant? Aristotle/Gorgias/Socrates.
  Q7: Why do you like them? Pick three adjectives that apply: humorous, knowledgeable, nosy, smart, expert, reliable, friendly, helpful, open minded, provocative, organized, unpredictable.
  Q8: Did you ask help more frequently from your favourite character? Yes/no/sometimes.
  Q9: What do you think are the three most important qualities in a teacher? (Open question).

*Rationale*: The questions are aimed at eliciting what personality traits are perceived by the students as positive in a virtual pedagogical interaction, thus facilitating learning. In designing the questions, we used findings from persuasive technologies studies. For example, as explained by Fogg (2002): a) We are more likely to be persuaded by computing technology that we perceive similar to us (principle of similarity), b) we tend to be persuaded by computing technology that offers us praise of some sort (principle of praise), and c) we tend to feel the need to reciprocate when computing technology has provided some benefits to us (principle of reciprocity).

- **Chatbot Interface**
  Q10: How does the interface make you feel? Relaxed/bored/overwhelmed/amused/Other.
  Q11: What would you change? Font/colours/example/avatars/Other.
  Q12: If you would change an avatar, which one and why? (Open Question).
  Q13: Which participant looks more trustworthy? And why? (Open Question).

*Rationale*: The questions are aimed at getting information as to the role played by multimodal input in creating a favourable learning setting. In particular, Q13 it is formulated on the basis of the hypothesis that computing technology that shows the role of authority is generally perceived as more trustworthy and, thus, persuasive.

- **Functionality**
  Q14: What fallacy did you discover?
  Q15: Are you able to describe it? Please write in 1–2 sentences.
  Q16: Do you think you might be able to recognize this fallacy in news you might read in the future? Yes, maybe, no, if no, why not?

*Rationale:* The questions are meant to check whether the learning outcome of being able to identify fallacies across news has been achieved. Specifically, Q16 explores whether the participants think they have been, at least partially, inoculated.

### *11.3.4 Beta Testing Results*

The results of the questionnaire show that some facets of the chatbot design have been deemed as more controversial than others. To start with, the rhythm of the conversation has variously been perceived as "slow" (40.54%), "just right" (29.73%) or "fast" (27.03%), and "too slow" by a minority (2.70%). No one perceived the rhythm to be "too fast". The turn-taking has been designed to mimic human behaviour, adding a 250 milliseconds delay per word in coming up with a new conversational move. As to the tone, that has been meant to be colloquial but lexically rigorous, it has been assessed by the majority as "just right" (64.86%). More varied has been the assessment of the presence of a multi-agent conversation, considered "interesting" (38.30%), "informative" (27.66%), but also "confusing" (29.79%) and rarely "boring" (4.26%). Overall, the participation in the conversation has been felt by the majority of the participants as sometimes "active" (37.84%) or "just right" (32.43%), but less frequently as "not active" (18.92%) or "active" (10.81%). As to Q5, 20% of respondents remarked that the philosophers did not remember their name, 16% recognized the usefulness of having a set of questions to help them in the fact-checking process, and 50% declared that the binary choice was sometimes difficult to make.

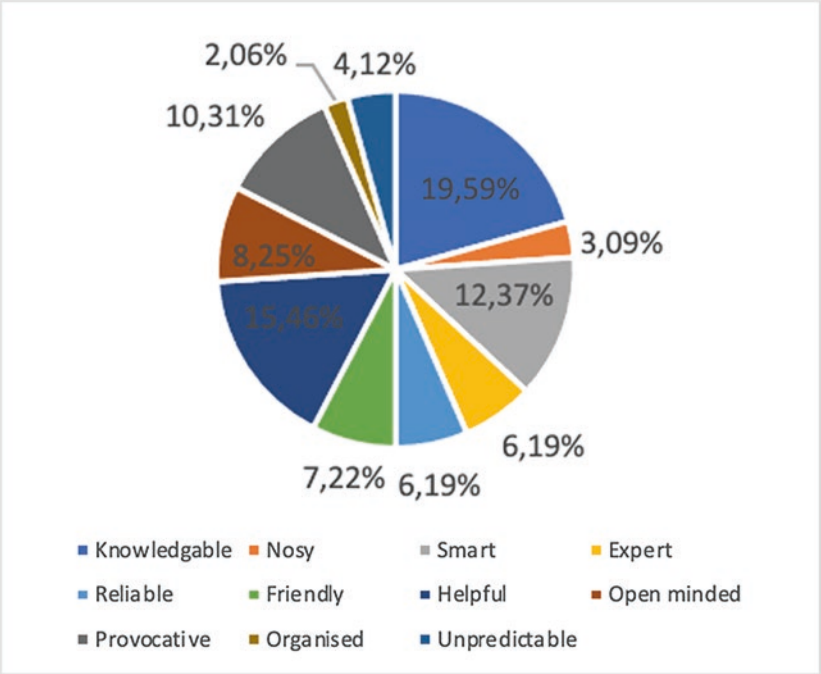


**Fig. 11.3** Frequency of adjectives describing reasons for liking a philosopher avatar

Turning to the chatbot personality, there is no clear preference for one philosopher over another. The adjectives providing reasons for the positive sentiment are variously distributed as displayed in Fig. 11.3, with the most common adjective being “knowledgeable” (19.59%.).

The majority of the respondents answered “no” to Q8 (64.86%), suggesting that liking a character does not increase the likelihood of reaching out to them to seek help, most likely to avoid face-threatening feelings. From the open question Q9, it emerges that five qualities (frequency > 5) are commonly perceived as characterizing a good teacher beyond the fallacy scenario: namely, knowledgeable, friendly, helpful, clarity, and passion, as displayed in Fig. 11.4. Therefore it seems that folk values associated to quality in an ideal pedagogical setting underpin positive attitudes towards different participants in the praxis of a digital game.

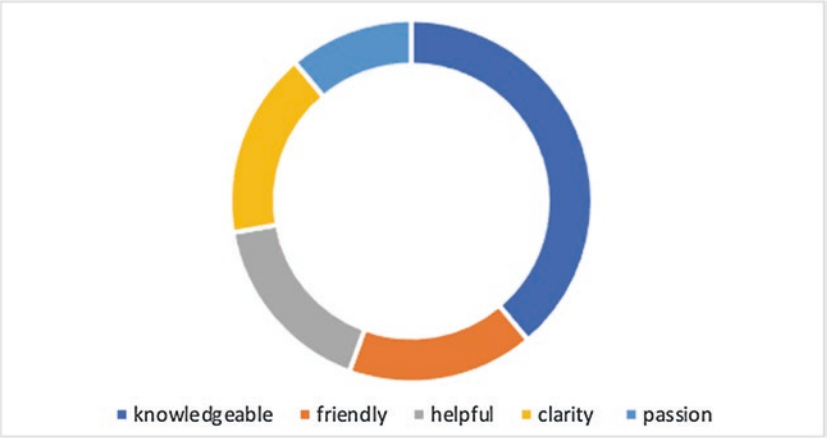


**Fig. 11.4** Most frequent qualities defining a good teacher

Moving to questions related to the interface, there is variation in terms of associated feelings (Fig. 11.5), with no single feeling emerging as significant. In fact, the most common response was "other", closely followed by "relaxed", "overwhelmed", "bored", and "amused."

Browsing through the explanations provided for "other", a feeling of confusion regarding actions to take is the one most frequently voiced. Similarly, the advocated changes in the interface cut across different facets, including stylistic features such as font, colours, examples, language, avatars, and others (Fig. 11.6).

Focusing specifically on the avatars, the large majority of respondents would have not changed any of them (Fig. 11.6). As to trustworthiness (Q11), only 21 students provided an answer. One third of the responses pointed to a lack of preference, while among the chosen avatars, the ones most frequently selected are Aristotle, the Principal Investigator of the project, and Socrates. Unfortunately, few participants provided a justification for their choice. When reasons were given, these related to *familiarity* for the Principal Investigator and *helpfulness* for Socrates (e.g. "Socrates, even if he didn't provide straight answers, he pointed me in the direction that I had to look in").

In terms of functionality, students encountered a wide range of fallacies, with cherry- picking (25.30%), hasty generalization (13.25%), and

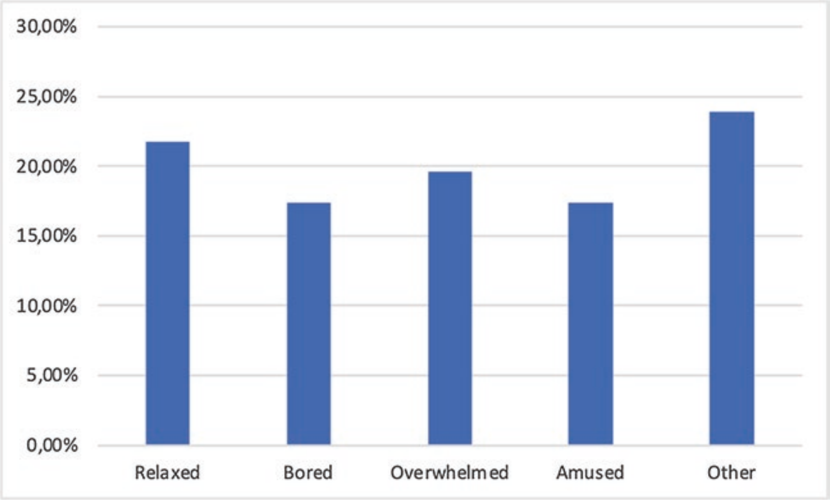


**Fig. 11.5** Feelings triggered by the chatbot interface

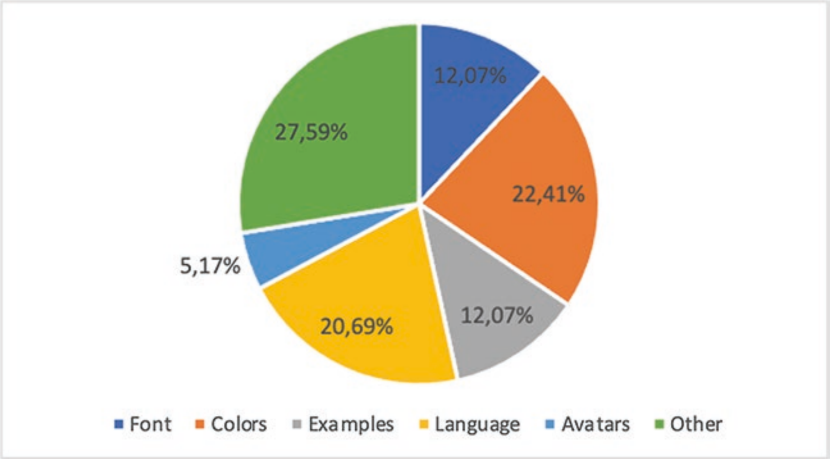


**Fig. 11.6** Aspects that could be improved in the chatbot interface

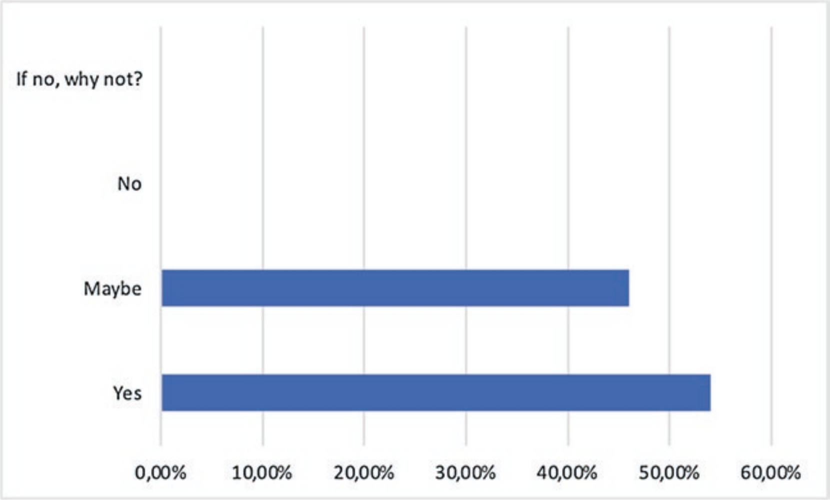


**Fig. 11.7** Perceived ability of recognize fallacies across different context

red herring (10.84%) on top of the frequency scale. Apart from one student, they all declared they were able to describe the fallacy they encountered. Manually checking the answers, they turned out to be all correct. Furthermore, differences in phrasing used by different students to describe the same fallacy show that they did come up with personal elaborations without stemming from dictionary-like definitions (e.g. cherry-picking: “Selectively picking supports to provide a predecided argument”; “The information might be chosen for a specific purpose, but may not tell the whole picture or may misguide you”). Interestingly, the students tend to think they might be able to recognize the fallacy they learnt in different contexts, such as different news (Fig. 11.7). Even though the responses were overall positive, there is uncertainty as to whether they will be able to fully transfer the learnt knowledge.

## 11.4 Conclusion

The results of the beta testing sessions provide insights as to the role played by human-computer interaction to teach critical media literacy. As remarked by social constructivists (see Vygotsky & Cole, 1978), language and dialogue constitute our most powerful semiotic mediators that assist us in the process of developing reasoning patterns and selective skills to make sense of our realities. However, differences between a classroom discussion among humans and an online dialogue game from the students' perspective have so far been under-investigated (Ravenscroft & McAlister, 2005).

The results of the questionnaire provide insights as to the role played by human-computer interaction to teach critical media literacy. First, they show that certain aspects such as the rhythm of the conversational flow and the feelings triggered by digital interfaces or liked avatars are highly subjective and hard to standardise, even though these factors impact on the learning experience. While acknowledging and accounting for differences among individual learners is a core value in education, it is a challenging goal to achieve in a gamification environment where a wide array of settings is predefined. On the other hand, the analysis of the answers revealed some clear trends which are relevant in face-to-face teaching. For example, the attributes more frequently associated to the preferred avatars broadly match those ascribed to a good teacher beyond the digital setting: knowledgeable/smart/expert knowledgeable, friendly, helpful, smart, reliable/organized, and clarity. Not surprisingly, the avatar considered the most trustworthy is Aristotle, introduced as the inventor of Fallacy Theory (thus the most knowledgeable), followed by the avatar of the PI who (having been their lecturer) is most probably perceived as friendly and reliable, and finally Socrates, deemed as helpful.

The main take-away is that competence, among the three main components of trust of competence, benevolence, and integrity (Schoorman et al., 2007) plays a crucial role in the reason-checking context. This is probably due to the scarcity of authoritative sources of information that characterize the misinformation ecosystem. Furthermore, authority and expertise do not necessarily pattern with a perception of "peer hood" in a pedagogical context where asymmetric knowledge is a value. For example while we are likely to trust a friend with tastes similar to ours in choosing what restaurant to reserve, we'd rather trust an expert in matters outside common ground knowledge. The quest for human-like avatars prosecuted in human-computer interaction design does not appear to be a priority for educational contexts. On the other side, the digital infrastructure offers the opportunity to make philosophical ideas and theories accessible, applicable, and usable for contemporary tasks, opening up new venues for active learning.

To improve the chatbot, we plan to better shape avatars' personality to decrease the potential confusion caused by multi-agent interaction and to more clearly define the epistemological contribution provided by each character. Besides the human computer interaction component, the answers to the questionnaire suggest that the *Fake News Immunity Chatbot* constitutes a useful tool to teach critical thinking through fallacies, given that twenty minutes of play enabled students to learn several fallacies. From their descriptions, it emerges that they learnt how to use these fallacies as lenses to interpret the digital media context (e.g. "The no proof one is the easiest one for me to spot - when a politician makes a tweet but has no evidence for the claim"). Furthermore, the identification of fallacy seems to have triggered further critical thoughts about the complexity of the misinformation ecosystem (e.g. "Also, in my opinion it is not always enough to throw in reference there, because even choosing the source (e.g. scientist) over another may still be biased"). Such an awareness shows that "inoculation through fallacy theory" in a gamification environment serves the primary goal of making users more sceptical towards information quality and coherence. The ability of asking the right questions constitutes the kernel of critical thinking and provides an asset to deal with scenarios of radical uncertainty in a post-truth world.FundingThis work was supported by the UK Research and Innovation Economic and Social Research Council [grant number ES/V003909/1].

## References

Brennen, J. S., Simon, F., Howard, P. N., & Nielsen, R. K. (2020). Types, sources, and claims of COVID-19 misinformation. Accessed November 24, 2022, from https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation.

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*, 1–22.

Carmi, E., Musi, E., & Aloumpi, M. (2021). The rule of truth: How fallacies can help stem the Covid-19 Infodemic. Impact of Social Sciences Blog. https://blogs.lse.ac.uk/impactofsocialsciences/2021/01/08/the-rule-of-truth-how-fallacies-can-help-stem-the-covid-19-infodemic/.

Compton, J. (2013). Inoculation theory. In J. P. Dillard & L. Shen (Eds.), *The Sage Handbook of Persuasion: Developments in Theory and Practice* (pp. 220–237). Sage.

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One, 12*, 1–21. https://doi.org/10.1371/journal.pone.0175799

Cushion, S., Morani, M., Kyriakidou, M., & Soo, N. (2021). (Mis)understanding the coronavirus and how it was handled in the UK: An analysis of public knowledge and the information environment. *Journalism Studies, 23*, 1–19. https://doi.org/10.1080/1461670X.2021.1950564

Elleström, L. (2014). *Media transformation: The transfer of media characteristics among media*. Palgrave Macmillan.

Elleström, L. (2017). Transfer of media characteristics among dissimilar media. *Palabra Clave, 20*, 663–685. https://doi.org/10.5294/pacla.2017.20.3.4

Elleström, L. (2021). Media transformation: The transfer of media characteristics between media. In M. Deckert, M. Kocot, & A. Majdzińska-Koczorowicz (Eds.), *Moving between modes: Papers in Intersemiotic translation* (pp. 27–42). Wydawnictwo Uniwersytetu Lodzkiego.

Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2020). Trust in UK government and news media COVID-19 information down, concerns over misinformation from government and politicians up. *Reuters Institute for the Study of Journalism*. Accessed November 24, 2022, from https://reutersinstitute.politics.ox.ac.uk/trust-uk-government-and-news-media-covid-19-information-down-concerns-over-misinformation.

Fogg, B. J. (2002). *Persuasive technology: Using computers to change what we think and do*. Morgan Kaufmann.

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology, 38*, 1004–1015.

Hamblin, C. L. (1970). Fallacies. *Tijdschrift Voor Filosofie, 33*, 183–188.

Huang, W. H. Y., & Soman, D. (2013). Gamification of education. *Report Series: Behavioural Economics in Action, 29*(4), 37.

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pp. 895–906. https://doi.org/10.1145/3196709.3196735.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kapp, K. M. (2012). Games, gamification, and the quest for learner engagement. *T+ D, 66*, 64–68.

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly, 21*(1), 61–78.

Kellner, D., & Share, J. (2007). Critical media literacy: Crucial policy choices for a twenty-first-century democracy. *Policy Futures in Education, 5*, 59–69.

Kline, K. (2016). Jean Baudrillard and the limits of critical media literacy. *Educational Theory, 66*, 641–656.

McGuire, W. J., & Papageorgis, D. (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *The Journal of Abnormal and Social Psychology, 62*, 327.

Musi, E., & Aakhus, M. (2018). Discovering argumentative patterns in energy Polylogues: A macroscope for argument mining. *Argumentation, 32*, 397–430.

Musi, E., & Reed, C. (2022). From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society, 33*, 1–22.

Musi, E., Yates, S., Carmi, E., O'Halloran, K., & Aloumpi, M. (2022). Developing fake news immunity: Fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies, 12*, 1–18. https://doi.org/10.30935/ojcmt/12083

Ravenscroft, A., & McAlister, S. (2005). Dialogue games and e-learning: The Interloc approach. In Looi, C.-K., Jonassen, D., & Ikeda, M. (Eds.), *Towards Sustainable and Scalable Educational Innovations Informed by the Learning Sciences – Sharing Good Practices of Research, Experimentation and Innovation, Proceedings of the 13th International Conference on Computers in Education, ICCE, Volume 133 of Frontiers in Artificial Intelligence and Applications* (pp. 355–362). IOS Press.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management Review, 32*, 344–354.

Shapiro, H., & Celot, P. (2011). Testing and refining criteria to assess media literacy levels in Europe. Final Report. https://eavi.eu/wp-content/uploads/2017/08/study_testing_and_refining_ml_levels_in_europe.pdf.

Simões, J., Redondo, R. D., & Vilas, A. F. (2013). A social gamification framework for a K-6 learning platform. *Computers in Human Behavior, 29*, 345–353.

Stott, A., & Neustaedter, C. (2013). Analysis of gamification in education, technical report 2013–0422-01, Connections Lab, Simon Fraser University, Surrey, BC, Canada, April. http://clab.iat.sfu.ca/pubs/Stott-Gamification.pdf.

Tandoc, E. C., Jr., Lim, Z. W., & Ling, R. (2018). Defining "fake news:" a typology of scholarly definitions. *Digital Journalism, 6*, 137–153.

Uscinski, J. E., & Butler, R. W. (2013). The epistemology of fact checking. *Critical Review, 25*, 162–180.

Valério, F. A. M., Guimarães, T. G., Prates, R. O., & Candello, H. (2020). Comparing users. Perception of different chatbot interaction paradigms: A case study. In *Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3424953.3426501

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Nerini, F. F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications, 11*, 1–10.

Visser, J., Lawrence, J., & Reed, C. (2020). Reason-checking fake news. *Communications of the ACM, 63*, 38–40.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Yates, S. J., Carmi, E., Lockley, E., Wessels, B., & Pawluczuk, A. (2021). *Me and my big data final report: Understanding citizens' data literacies.* University of Liverpool. Accessed November 24, 2022, from https://www.liverpool.ac.uk/humanities-and-social-sciences/research/research-themes/centre-for-digital-humanities/projects/big-data/publications/.