From fallacies to semifake news

Musi, Elena ; Reed, Chris

Published in:
Discourse & Society

DOI:
10.1177/09579265221076609

Publication date:
2022

Licence:
CC BY

Document Version
Publisher's PDF, also known as Version of record

Research article

Discourse & Society
1–22

DOI: 10.1177/09579265221076609
journals.sagepub.com/home/das

From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media

Elena Musi
University of Liverpool, UK

Chris Reed
University of Dundee, UK

Abstract

This study tackles the fake news phenomenon during the pandemic from a critical thinking perspective. It addresses the lack of systematic criteria by which to fact-check the grey area of misinformation. As a preliminary step, drawing from fallacy theory, we define what type of fake news convey misinformation. Through a data data driven approach, we then identify 10 fallacious strategies which flag misinformation and we provide a deterministic analysis method by which to recognize them. An annotation study of over 220 news articles about COVID-19 fact-checked by Snopes shows that (i) the strategies work as indicators of misinformation (ii) they are related to digital media affordances (iii) and they can be used as the backbone of more informative fact-checkers' ratings. The results of this study are meant to help citizens to become their own fact-checkers through critical thinking and digital activism.

Keywords

misinformation, fallacies, semi fake news, fact checking, multi-level annotation

Introduction

On 15th February 2020 the WHO Director General Tedros Adhanom Ghebreyesus stated at the Munich Security conference that during COVID-19, 'We're not just fighting an epidemic; we're fighting an infodemic'. Not surprisingly, there has been an outburst in

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

public discussions about the social impact played by the virality of dis- and mis-information. A recent study (Islam et al., 2020) has shown that during the first 3 months of the pandemic, nearly 6000 people around the world were hospitalised due to coronavirus dis- and mis-information. To counter the consequences of the infodemic, it is not sufficient to debunk myths or conspiracy theories (disinformation): as underlined by the last *RISJ* factsheet (Brennen et al., 2020), 59% of fake news contains neither fabricated nor imposter content, but rather reconfigured misinformation (misleading content, false context, manipulated content). Fact-checkers are, thus, not called to check facts only, but have to navigate a highly complex information ecosystem. The lack of systematic and agreed criteria to identify and classify types of misinformation further complicates a crisis scenario where what counts as true information is constantly updated: the list of symptoms characterising the virus, the likelihood of a timely vaccine and even government measures are in constant evolution. As a result, fact-checkers have adopted different sets of labels to flag and make sense of misinformation. Besides being potentially confusing for the readers and frequently not informative as to the roots of misinformation, the presence of diverse and only partially overlapping truth barometers hinders the scalability of misinformation debunking. Automatic fact-checking systems, required to face the fast spread of misinformation on digital media, cannot rely on coherent data provided by fact-checking initiatives to train their systems for the identification of misleading news. Conjoined efforts in the Natural Language Processing Community are addressing the task of claims verification (Barrón-Cedeño et al., 2020), leveraging large datasets of scientific articles and news certified as true by authoritative sources. However, there are currently no systems able to accurately identify and scaffold misinformation (Thorne and Vlachos, 2018).

This study addresses these issues providing a systematic procedure for the analysis and the classification of types of misinformation. It does so by leveraging *Fallacy Theory*, informal logic theory that has its roots in the ancient classic tradition. The main rationale is that fallacies, arguments which seems valid but are not, work as indicators of misinformation, news that seem informative, but are not. After having reviewed the state of the art concept of *fake news* and their problematic classification, we introduce the notion of semi-fake news as news that does not necessarily contain fabricated content, but fallacious arguments; we explain why fallacies work as valuable tools for the misinformation ecosystem and how fallacies types can be mapped onto types of misinformation. To make fallacies operationalisable for fact-checking we propose a data-driven taxonomy of fallacies together with a systematic heuristic key to allow for their identification. We verify and show the relevance of our fallacies taxonomy through an annotation study on a corpus of 220 news fact checked news *Snopes* about misinformation related to COVID-19, encompassing the analysis of types of sources next to fallacy types. The results of the annotation show a significant correlation between fallacies and misinformation, confirming their role as indicators. We then describe the misinformation ecosystem emerging from our analysis looking at the distribution of fallacious news, their correlation with types of sources as well as ratings provided by the fact-checkers.[1]

We provide qualitative analysis on the identified patterns showing how fact-checkers' truth barometers can be informed by fallacies identification. The full dataset containing the gold annotation will be made publicly available to the academic community.

Fake news

An open-ended definition

It is widely recognised that the term *fake news* is an umbrella term rather than a consistently defined notion (Levi, 2018; Tandoc et al., 2018). The common ground among these definitions is that 'fake news' looks like real news without being so. Tandoc et al. (2018) point out that these definitions can be positioned on a continuum along the two dimensions of facticity and intention. Verstrate et al. (2017) also adopt a binary set of features to define a typology of fake news, putting next to intention to deceive, the presence of a financial motivation to create fake news.

Acknowledging the complexity of mis- and dis-information which happen to be shaped by national information environments (Humprecht, 2019), Wardle (2017) strays away from a holistic definition of *fake news*, proposing a clustered approach to characterise the information eco-system encompassing seven different types. What distinguishes the latter taxonomy from the others is that both the types of content created and their (un)intentional dissemination play a crucial role. As remarked by Gelfert (2018), however, in media studies the phenomenon of *fake news* is framed as an epistemological issue, where different criteria to evaluate news trustworthiness rely on the analysis of the news source or the news content, without accounting for the recipients' ability to critically questions the news, especially when not felt dubious according to their judgement (Tandoc et al., 2018). Regardless of the gap between reported facts and reality, according to Gelfert (2018: 108), news is fake news if it manages to deceive at least a part of the audience: 'Fake news is the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design [. . .] either because it aims to instil falsehoods in its target audience [. . .] or because the way it is deliberately operated is objectively likely to mislead its target audience'. As explained by Schwarzenegger (2020), personal epistemologies based on selective criticality, pragmatic trust and competence-confidence make it difficult to predict what target audience might be reached by fake news.

In line with Gelfert (2018) we believe that fake news constitutes a co-constructed concept where the audience's reception is as important as the authors' intention: a fake news recognised as such by readers stops to fulfil its ultimate goal of deception. But this awareness does not bring us to the end of the conundrum since the context of digital media makes it difficult to identify what counts as *deliberation in the first place*. It is, for example, true that in social media such as Facebook (mis)information is tightly correlated with filter bubbles, but users' *online conviviality* plays a crucial role too in spreading fake news (Seargeant and Tagg, 2019), impacting at the same time interpersonal relationships when shared news turn out to be fake (Duffy et al., 2020).

Newsrooms feature a variety of actors involved in the deliberation process of news construction: an editorial move to change the original title of a news to make it more 'clickable' and increase visibility over a crucial matter might, for example, result in unintentional misleading information. This is especially the case when a crisis situation, such as the COVID-19 pandemic, brings us to a post-truth scenario (Collins, 2018): the radical uncertainty underlying preventive measures and future developments goes hand in

hand with scientific advancement rendering objectivity a chimaera. Taking the BBC editorial guidelines as a benchmark (https://tinyurl.com/y6awgk3y) it is evident that the COVID-19 pandemic ticks all the boxes to be deemed not only a 'controversial subject', but even a 'major matter'. However, advocated measures to deal with such a complexity such as giving 'due weight to various opinions or distinguish opinions from fact' are challenging: the results of a scientific study over the efficacy of a drug/vaccine are inherently provisional and might be discarded by larger trials. Thus, in such an unprecedented situation, to understand what counts as mis- or dis-information and goes against the public interest, we start from the analysis of those news that have been flagged as 'fake' by fact-checkers and have so far caused harm to citizens. Fact-checked news constitute a relevant sample of analysis since they have been selected by fact-checkers on the basis of their *newsworthiness*, measured by looking at analytics of their popularity on social media.

Fact-checking fake news

One of the challenges in using a notion of truth as the yardstick against which to measure fakery is that it assumes that language can somehow directly reflect truth, and that objectively measurable truth detection is a commonplace. This, indeed, is the assumption that underlies the very institution of the fact-checker: the idea that some stretch of text that can be identified as a fact can then be checked off against truth in the world. With such a conceptually clean picture being presupposed by the activity of fact-checkers, and even by their very name, it is no wonder that the computational sciences have latched on to it with such vigour. Large scale competitions or *shared tasks* such as FEVER (Thorne et al., 2018) set out to pit research labs against one another in developing AI algorithms (typically based in statistical models of language use) that can either identify which passages of text constitute checkable facts, or, more typically, can identify whether or not a given passage of text constitutes something that is true or not. This conception, however, belies the difficulty of identifying fake news, and belies what it is that fact checkers do.

In their computational account, Hanselowski et al. (2019) compare the major fact checking organisations and report that the *only* characteristic that all fact checkers have in common is textual commentary rather than a systematic methodology to identify factuality. The reason is that fact checkers understand full well that objective truth is hard to come by, and, furthermore, that by far the most common category of fakery from Wardle's list is *Misleading Content*. We can see this even in fact checker datasets. *Snopes*, for example, collected into a machine-processable dataset in Hanselowski et al. (2019), comprises 2943 examples marked as false, 659 marked as true and 2890 marked with ratings pointing to not true and not false content. The acknowledgement of fake news as a *continuum* is reflected in the fact-checker's truth *barometer* labels which contain many more values than the mere *false* and *true*, at different degrees of granularity: *Healthfeedback.org*, for example, features *Inappropriate sources* next to the vague *Misleading*, while *Glenn Kessler's fact checker* puts together the *Upside down Pinocchio* ('a statement that represents a clear but unacknowledged flip-flop from a previously held position') and the *Four Pinocchio* ('whoppers'). This lack of consistency in ratings shows that, despite general agreement that some news are more fake than others, a systematic methodology to name and scaffold misinformation is currently missing.

Semi-fake news

To counter misinformation, it is necessary to provide an operationalisable definition of what *fake news* (that we call *semi fake news*) fits this grey misinformation. As remarked in the previous section, the digital era has radically changed the way news is constructed, shaped and distributed: the term *newsworthy* does not refer exclusively to what is considered warranting mention by news media agencies, but it depends on what achieves public visibility on social media and other unofficial information channels. Thus, to define *semi-fake news* we adopt a polyphonic perspective encompassing the authors of the news with their intentionality as well as the audience awareness next to the facticity of the conveyed content (Ofcom weekly research findings: https://tinyurl.com/ybqmmjbh; Brennen et al., 2020).

From the authors' perspective, we define *semi-fake* news in contrast to prototypical fake news ('fabricated information'). Semi-fake news is not constructed by the authors with the deliberate intention of dis-informing the audience. This does not entail that rhetorical strategies cannot be put into place by the journalist in order to attract readers, but it means that the new information conveyed by the media piece is conceived as a genuine representation of the situation. Thus, semi-fake news articles do not contain fabricated information intended to represent states of affairs known by the authors to be in conflict with reality. However, they may contain propositions presented as assertions backed up by sources partially valid or anyways not sufficient to draw conclusions presented as factual: a news title 'Link between blood type and Covid risk', based on the result of a single scientific study might turn out to be true in light of further scientific results, but it is definitely less informative than its mitigated counterpart 'Possible link between blood type and Covid risk'.

The lack of misleading intention and fabricated news make *semi-fake news* good candidates to appear on trustworthy news sources. As a result, *semi-fake news* is that which cannot be disguised by the audience ascertaining the trustworthiness of the information source, avoiding clickbaits or consulting myths deburst archives, as suggested by the majority of public guides fighting misinformation. Even if conveying information less far from truth than patently false information, they are more pernicious at a large scale: a 6-week diary study of news audiences (Kyriakidou et al., 2020) has found out that the public can easily spot disinformation such as conspiracy theories, while they are less skilled in navigating misinformation not immediately suspect, such as UK death rate compared to other countries.

AI-powered fact checking tools are of little help for the public to identify misinformation. The Coronacheck enterprise (https://coronacheck.eurecom.fr/en) created a user-friendly interface to verify statistical claims about the spread and the effect of COVID-19 leveraging fact-checked data from a dataset of official sources. However, the database does not allow checking claims of other nature, such as predictions or policy statements which make up the bulk of news editorials. Other Natural Language Processing methodologies for the automatic recognition of fake news rely on the detection of deception linguistic cues (Conroy et al., 2015) at various levels. Deep language structures analysed through Probabilistic Context Free Grammars have, for example, been leveraged to predict instances of deception (Feng et al., 2012). At the semantic level, negative sentiment (negative emotion words) has been shown to be associated with deceitful messages (e.g. Horne and Adali, 2017; Kwon et al., 2013). At the discourse level, the prominent use of certain rhetorical relations rather than others can be indicative of deception. According to

a preliminary study by Rubin and Lukoianova (2015) for instance, evidence and antithesis are significantly more frequent in truthful rather than deceptive stories. Even though all these features are no doubt of great use to identify fake news, they are of little help in identifying semi-fake news where there is no intention to circulate counterfeit information (Fallis and Mathiesen, 2019). The authors of semi-fake news, instead, do not generally lack good intentions, but commit genuine flaws in the processes of interpreting the available evidence and/or presenting it in the most informative way. In other words, the issue of semi-fakery hinges upon the argumentative nature of news (Zampa, 2017) which express claims whose informativity relies on the quantity and the quality of the arguments supporting them. Evaluating those arguments can, thus, provide us with analytic means to identify *semi fake news*. This is not an easy task since the factors that might make arguments in news fallacious include, at least, the quantity of available information (e.g. presence of sufficient evidence in support of a claim), the unfolding of such information (e.g. malformed inferences linking one statement to the other) and the way the information is expressed (e.g. ambiguous and vague terms). For this sake, we propose to leverage *Fallacy theory*, the study of fallacious arguments, to provide a systematic set of credibility criteria by which to flag semi-fake news.

From (semi) fake news to fallacies

Fake news is news that has the appearance of proper news but is not. Fallacies are generally defined as arguments that seem to be valid but are not (Hamblin, 1970: 12). Thus, fake news pieces are likely hosts for fallacious arguments. The ability to identify fallacious arguments would constitute an asset for the journalist to avoid spreading potentially misleading information and for citizens to interpret them *cum grano salis*. The advent of digital and social media has, in fact, brought to an end the era of the *ipse dixit*, where few news media agencies were responsible for informing the public in a top down manner, in favour of what Sunstein (2018) calls a 'divided democracy'. In such a scenario of news democratisation, social media platforms work as a contemporary *agora* that is subdued by Rhetorics at different levels: what news (fake or true) becomes viral is, for example, ultimately, a matter of rhetoric. Even the way we select our news feed can be critically interpreted through rhetorical lenses. What Negroponte et al. (1997) have long ago coined as the 'Daily Me', our personalised design of the information package we choose to receive, is an instantiation of the cherry picking fallacy: we decide to reinforce the world-view that we like exposing ourselves only to certain topics, certain sources etc. In designing our news architecture we are prompted by the algorithms which propose news liked by our friends, according to the ancient principle of *homophilia*, the tendency to appreciate what our peers appreciate. As underlined by Aristotle, Rhetoric is useful since it allows understanding reasons behind incorrect decision making processes (Rhet. 1355a22-25). In line with this view, if citizens make wrong decisions, it is the ecosystem of (mis)information to be blamed and strengthening their judgement criteria promises to make them better judges.

From fallacies to semi-fake news

In 1847, Augustus De Morgan begins the chapter 'On Fallacies' in his *Formal Logic*, with the statement, 'There *is* no such thing as a classification of the ways in which men

may arrive at an error: it is much to be doubted whether there ever *can be*' (de Morgan, 1847: 237; emphasis in the original). This attitude that the study and attempt to taxonomise fallacies was inherently doomed was an attitude that prevailed for more than a century. It was not until the Australian philosopher Hamblin (1970) addressed the issue in *Fallacies* that the field started upon its road to academic rehabilitation. Compendious approaches to fallacies vary widely, but perhaps a common core can be identified that have been dubbed by Woods (2004) the 'gang of eighteen': equivocation, amphiboly, composition, division, petition, complex question, post hoc ergo propter hoc, ignoratio elenchi, ad verecundiam, ad populum, ad baculum, ad misericordiam, ad hominem, faulty analogy, slippery slope, affirming the consequent and denying the antecedent. The final two are representative of the class of *formal fallacies*, and reflect deep cognitive biases that have occupied psychologists for more than half a century (Wason, 1968). It is a subset of the remaining gang members that are of interest to us here: the *informal fallacies.* What is intriguing about informal fallacies is the way in which their invalidity is fuzzy. Whilst from a classical, logical perspective, a piece of reasoning may be incorrect or faulty, that same reasoning might (with appropriate context) be a perfectly sensible way of proceeding. Philosophers such as Walton (1996) have explored in detail the ways in which instances of what might classically be regarded as fallacies can be used justifiably, rationally and effectively – with more recent linguistic exploration uncovering the extraordinarily wide extent to which these patterns are used (Visser et al., 2020a). The fuzziness of the delineation of fallacious reasoning is precisely what underpins the notion of semi-fake news and what makes recognising such news so demanding. We found our approach on insights from one of the most extensive and incisive writers on fallacies. Bentham (2015), writing over 200 years ago, offers a scaffold that can help us. Talking in the context of fallacies in parliament, he distinguishes the social-discursive roles involved in fallacy (Bentham, 2015: 38), *viz*., the fabricator, utterer and acceptor. His explication rests on an analogy with counterfeit currency: the fabricator creates and first employs the deceit and may trigger a chain of subsequent acceptance and re-use, with acceptors receiving the fakery either cognisantly or not, and then re-using the fake similarly. The parallel with fake news in social media is striking. Fakery is first fabricated, then uttered (or tweeted) and then re-uttered (or re-tweeted) to acceptors (or followers). Bentham (2015) also issues an engagingly modern call in his demand for annotation of fallacies (in his case, in Hansard, the parliamentary record): 'in each instance in which the use of any such [fallacy] is discoverable, let him at the bottom of the page. . .give intimation of it' (p. 74). Bentham (2015), of course, was imagining a manual process of such intimation, yet his eloquent call to arms (which here is referred to as, 'the faculty which Detection has of stripping Deception of her power' (p. 79)) is one to which, two centuries later, we can respond with a combination of manual and automatic means.

The first challenge for any empirical study of informal fallacies is a reliable taxonomy that can serve as the foundation for analysis and annotation. Whilst Aristotle, for example, in his *Sophistical Refutations*, distinguishes fallacies dependent on language (*in dictione*) from those not dependent on language (*extra dictionem*), Whately (1875) cleaves logical fallacies from semi-logical ones which require world knowledge to be interpreted. Pragmatic frameworks classify fallacies as infringements of the rules of an ideal critical discussion (Van Eemeren and Grootendorst, 2004). Regardless of the taxonomic approach, the key issue at stake though is the *Fallacy Fork* (Boudry et al., 2015):

cut-and-dry compendia of fallacies are unlikely to be reflected in real life discourse. As a result, we have adopted a bottom up approach, seeded by a contemporary taxonomic account.

The Snopes case

Dataset and annotation procedure

As a case study to investigate interrelations between fallacies and fake news we have considered the fact-checker *Snopes* (https://www.snopes.com/about-snopes/). We have chosen to focus on this fact-checker to showcase the relation between fallacies and misinformation during the pandemic for two main reasons. First, *Snopes* unlike other fact checkers, does not focus on political issues (e.g. governmental responses), but attempts to cover a variety of topics depending on readers' demand emerging from digital media. Second, it leverages a spectrum of ratings which acknowledges the presence and the importance of nuances in the information ecosystem (https://tinyurl.com/y96ahae6). Next to ratings such as *true, false, mostly true, scam*, it also includes labels such as *unproven* or *misattribution* that point to the reasons behind (un)trustworthiness of the news. When it comes to COVID-19, *Snopes* has assembled a coronavirus collection currently spread across 18 topics. The number of news considered for each topic reflects readers' interest determined on the basis of reader email submissions and posts on the fact-checker's social media accounts as well as general popularity on Google and other social media venues.

As a dataset, we have collected through web scraping all the fact-checked news appearing in the collection till the end of June 2020 distributed across all different topics at that time for an overall number of 220 articles (Table 1):

Table 1. Composition of the *Snopes* dataset.

Collection topics	Number of articles
Business	11
Conspiracy	19
Entertainment	8
Gates-Foundation	11
History	5
International response	14
Memes	24
Origins	18
Pandemic	2
Predictions	7
Prevention	21
Protests	10
Trump	41
USA response	38
TOT	220

Even though the numbers are too low to allow for any significant correlation between attested fallacies/types of misinformation and subject matters, the range of topics reduces the risk of contextual bias. For each article we have collected (i) claims (ii) text of the source article (iii) weblink to the source (iv) factchecker's comment (v) rating associated by the fact-checker (vi) collection type. The data have been archived in CSV files. It has to be noted that this dataset does not allow us to check for privileged associations of fallacies with cases of misinformation rather than disinformation or information. We leave such endeavour for future work, while shedding light on the relevance of certain types of fallacies to flag misinformation news.

Annotation guidelines and output

To investigate in a systematic manner correlations between fallacies and types of misinformation we have conducted an annotation study on the entire dataset. The annotation has been carried out by two students in Communication and Media without any previous knowledge in Argumentation Theory or Informal Logic. We, in fact, wanted to verify whether the task necessarily entails an institutional training and whether our guidelines and intersubjective. They have been introduced to the main notions of fallacy theory during a training session of 1.5 hours. They have been given the same set of 220 CSV files and asked to (i) read both the origin (comment from the fact-checker) and the source text (the text of the fact-checked source) (ii) identify the type of source according to a predefined set of categories (iii) identify the portion of text from the fact-checked source that triggers the critique according to the fact-checker (target critique) (iv) annotate (if any) the fallacy at stake in the original text according to the guidelines.

The categories that we provided for the classification of sources are the following: *Broadcast news* available only through a digital channel; *Broadcast news* multi-channel; *Government* and official political sources; *Social media; Blog posts* (personal or official); *Scientific Papers*. The rationale in adopting such categories is centred around the presence/absence and type of gatekeeping process at stake.

The guidelines for the annotation of fallacies contain a taxonomy of 10 fallacies. We have developed this typology using a bottom up approach: our starting point is Tindale (2007), which gathers the most common fallacies discussed in the informal logic tradition. Using this as a basis, we have commissioned expert analysis of 40 fact-checking commentaries and their source articles randomly picked from the dataset in order to identify which fallacies have been called out by the fact checker. We then summarise the most common fallacies identified.

The resulting annotation schema includes 10 types of fallacies organised into four classes related to the distribution of arguments and supported claims: evading the burden of proof (EBP); the (un)intentional diversion of the attention from the issue at hand: strawman (ST), false authority (FAUT), red herring (RH) and cherry picking (CP); the argument schemes at play: false analogy (FA), hasty generalisation (HG), post hoc (PH) and false cause (FC); and the language used: vagueness (VAG). As a verification step, we have analysed the definitions of the different verdicts and labels employed by main fact-checkers in English (snopes.com; healthfeedback.org; politifact.com; fullfact.org; theferret.scot;) to see whether critiques might point to fallacious moves different from the ones identified.

Our set covers the fact-checkers ratings completely: even if not exhaustively representing the universe of fallacies, our sub-selection covers the most frequent fallacious moves executed in online news. To offer a degree of systematicity, the fallacies have been arranged starting from those having to do with the quantity of information provided (structural fallacies), followed by those related to aspects external to the issue discussed (fallacies from diversion); logical fallacies come into place after the other two classes are excluded. This order echoes the one provided by the pragma-dialectic rules for a critical discussion (Van Eemeren and Snoeck Henkemans, 2016), a series of principles which are meant to guarantee the reasonableness of an ideal discussion: the violations of rule 8 (Argument Scheme Rule) follow the violations of rule 2 (Burden- of- Proof Rule), rule 3 (Standpoint Rule) and rule 4 (Relevance Rule). It is, in fact, not worth looking at the reasoning at play if the information conveyed in the arguments is irrelevant for the conclusion. It has to be remarked that the PragmaDialectic definition of fallacies differs for the standard one of Informal Logic: fallacies are speech acts which hinder the resolution of a dispute due to the violation of the rules of a critical discussion (Van Eemeren and Grootendorst, 1987). Even though the validity of an argument goes beyond those rules, the normativity of the framework offers a rationale to prioritise certain critical questions over others.

The broad fallacy class encompassing vagueness, equivocation and ambiguity occupies the final position in the list, as a catch-all category for when the other options are excluded.

In our guidelines, each fallacy is first defined and then associated with an example and accompanied by one or more identification questions, which have turned out to be useful means to evaluate arguments (Song et al., 2014). In this way, the annotator can go through the identification questions in a binary way stopping when one of the identification questions applies. The pattern here is to exploit the notion of a dichotomous key, developed in taxonomic biology (Dallwitz, 1980) and applied successfully in the notoriously demanding task of identification of argumentation schemes (Visser et al., 2020b). The annotators were also given the option to use the label ‘none’ when no identification would apply, but they still thought that some type of misinformation would be in place. Let’s consider a random instance of news claim from our dataset accompanied by the fact-checker comment:

> Example (https://tinyurl.com/y6syu2un)
>
> Claim: The COVID-19 coronavirus disease is ‘spreading quickly from gas pumps’.
>
> Source (social media): ‘FYI just spoke with a friend who got called into an emergency meeting at his hospital. He said the virus is spreading quickly from gas pumps’
>
> Fact-checker comment:
>
> ‘What’s True
>
> Gas pump handles are a potential source of surface contact transmission of the COVID-19 coronavirus.
>
> What’s False
>
> Gas pumps are only one of many commonly handled objects that could transmit the COVID-19 coronavirus, and we have found no substantiated reports of anyone having been infected in that fashion yet.

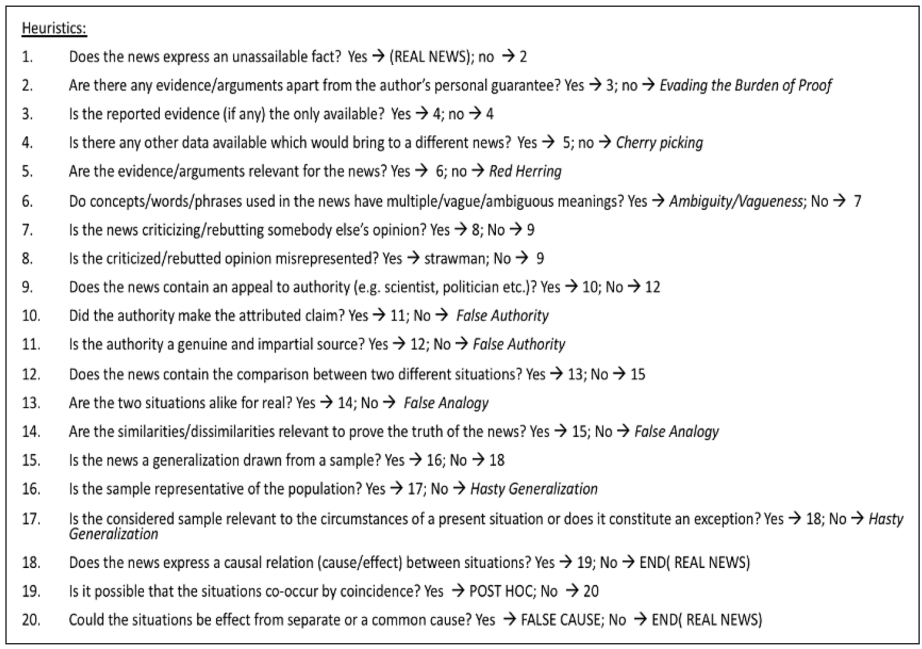

Figure 1. Heuristic identification key for fallacies.

> [. . .]
>
> "At this time, we are not aware of any studies that support the claim that the virus can be transmitted via contact with a gas pump," wrote API [National Institutes of Health and Princeton University]. However, the level of risk associated with contracting the virus from a gas pump is no different than the risk associated with touching other common surfaces like grocery store carts or door handles'.
>
> [. . .]

Annotators, after having read both the news claims and the comment, are asked to decide whether and which fallacy is at stake browsing the heuristics in Figure 1:

> In the considered example, annotators would stop their inquiry at points 10 and 11, identifying the presence of a False Authority fallacy.

In order to evaluate the reliability of the annotations we have first calculated the inter-annotator agreement (IAA) using Cohen's (1960) coefficient. For the type of source, the *kappa* value amounts to 0.988 which corresponds to perfect agreement according to Landis and Koch (1977). The annotation both of the type of source and of fallacies has been carried out on top of the entire dataset (220 news). The annotation of fallacy types has been carried out on top of 73/220 news: the excluded news are those labelled by the fact checkers as false (94 cases), true (36 cases), correct attribution (10 cases), scam (3 cases) and satire (4 cases).

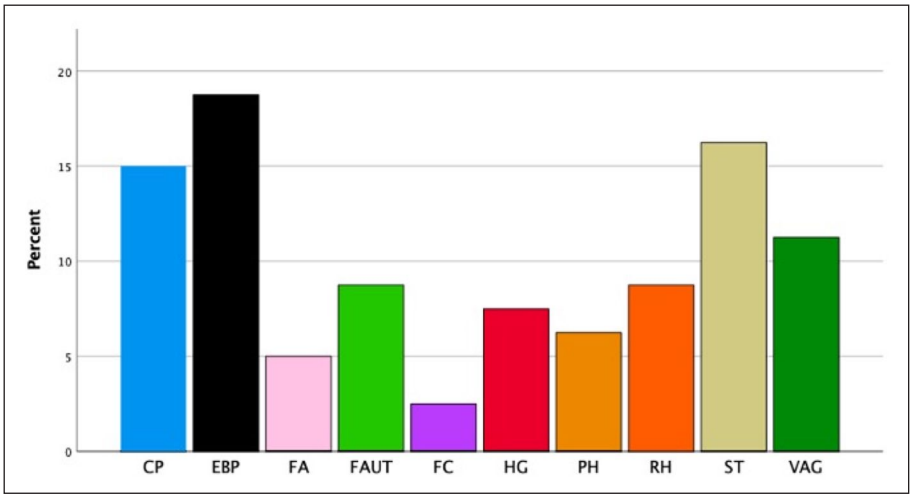

Figure 2. Distribution of fallacies across the misinformation ecosystem.

For fallacy types, we have achieved a 0.41 *kappa*, which corresponds to *moderate* agreement. Such a result outperforms those obtained in tasks of similar complexity such as the annotation of argument schemes (Musi et al., 2016). All cases of disagreement have been checked by an expert, achieving an accurate gold annotation. Our dataset is the first annotated as to fallacies in the misinformation environment.

Results and discussion

Distribution of fallacious news

To test our hypothesis that fallacies work as indicators of misinformation, we have first looked at the distribution of annotated fallacies across ratings pointing to misinformation (*misattribution, mostly true, unproven, mixture, miscaptioned, labelled satire* and *mostly false*):

As shown in Figure 2, according to the gold annotation, for each of the news articles tagged with a misinformation rating, a fallacy from our taxonomy has been identified, suggesting that our taxonomy exhaustively accounts for cases of misinformation in our sample. Overall, while the structural fallacy *evading the burden of proof* is the most frequent (15%); fallacies from diversion (*strawman*, 14%; *false authority*, 4%; *cherry picking*, 11%; *red herring*, 8%) and language (*vagueness*, 8%) also occupy a prominent position, leaving logical fallacies to a lower frequency on average (*post hoc*, 2%; *false cause*, 4%; *hasty generalisation*, 7%; *false analogy*, 4%).

To check whether certain news venues tend to be associated with specific fallacious news types, we have looked at the types of sources that host them, keeping cases where no fallacies is at stake out of the picture (Figure 3):

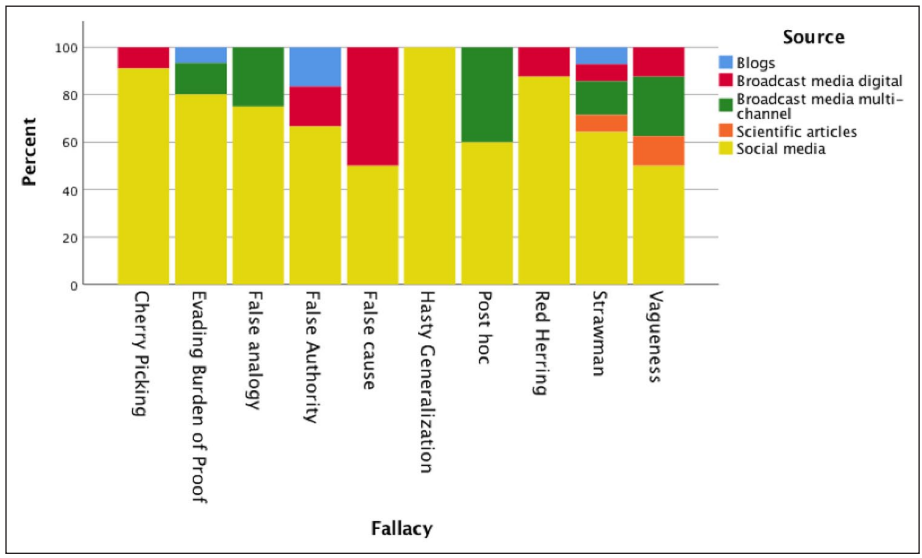

Figure 3. Distribution of fallacy per type of source.

The size of our dataset is too small to surmise any significant correlation between the two variables and call for verification over a larger dataset. However, major encountered trends can be explained through the design of different types of media.

The fallacies of *hasty generalisation, cherry picking, red herring* and *evading the burden of proof* are contained in at least 80% of cases in social media. This tendency well resonates with the presence of constraints in shaping and distributing news imposed by the gatekeeping process of official news media venues, regardless their (inter)national scope. Most codes of practices have in common as a core principle both the accuracy and the reliability of information. The *Global Charter of Ethics for Journalists*, for example, lists a set of rules that would be flouted when one of the above-mentioned fallacies is at stake: rule 2 stresses the importance of accounting for the difference between facts, which do not call for supplementary evidence, and opinions that have to be backed up by arguments ('He/she will make sure to clearly distinguish factual information from commentary and criticism'); rule 3 calls out the importance of picking relevant information and avoiding cherry-picking behaviours ('The journalist shall report only in accordance with facts of which he/ she knows the origin. The journalist shall not suppress essential information [. . .]'). Rule 5 reiterates these points underlying that 'The notion of urgency or immediacy in the dissemination of information shall not take precedence over the verification of facts, sources and/or the offer of a reply', while social media typically respond to users' urgency of spreading information of more or less personal and emotional relevance.

It is thus not surprising that news such as 'COVID-19 contact-tracing apps like Healthy Together and ABTraceTogether are tracking you and also the people in your phone contacts and Facebook friends lists' have been spread through Facebook posts rather than news editorials (https://tinyurl.com/y3x8zy8z0). The suppression of the

crucial information that contact numbers, Facebook lists or users' locations are collected only under voluntary provision would have been deemed against the journalist's ethics in an official news outlet. Similarly, official news media outlets have specific rules on how to select images which do not breach editorial standards. As a result, a video that portrays anti-surveillance protesters tearing down a 'smart' lamppost in Hong Kong (https://youtu.be/u1Ji7wonUhE) could have not been used to support the claim that 5G played a role in the pandemic, as had been suggested on the Instagram page of the actor Woody Harrelson (https://archive.vn/MzYv9), leading to a red herring fallacy. Instead, the video was correctly captioned by *The Guardian* on 2019 as evidence for protesters in Hong Kong in August 2019 tearing down 'smart' lamp posts used for surveillance (https://tinyurl.com/56ahuv5a).

It still might, of course, be the case that journalists happen to un-intentionally foreground certain types of evidence or draw illegitimate generalisations; however, the presence of standards to be ideally followed is likely to work as a nudging force in avoid such fallacious moves.

When it comes to flawed ways of reasoning in shaping a piece of news given a set of factual evidence, no guidance is formally supplied to or by news media agencies. As a result, professional journalists and social media users have the same epistemological starting points when trying to make sense of scientific studies about the symptoms of the virus or the readiness of a vaccine as well as when reporting about situations with uncertain causes. This might explain why broadcast media constitute almost 50% of the sources containing fallacies related to causal reasoning (*false cause* fallacy and *post hoc* fallacy). The headline 'Teen Who Died of COVID-19 Was Denied Treatment Because He Didn't Have Health Insurance' published by Gizmodo (https://tinyurl.com/yw6fsjjw) was based on the initial claim advanced by the mayor of Lancaster in a video posted on YouTube and, thus, publicly available to journalists as well as to the larger public. The story quickly became viral since the teenager was the first to be considered killed by the complications of COVID-19 in the country.

However, as clarified by *Snopes* (https://www.snopes.com/fact-check/teen-insurance-coronavirus/?collection-id=244110), the mayor then retracted his claim explaining that there was a misunderstanding probably due to language barriers, but the family had health insurance and the teenager received emergency treatment. In such a confused scenario, it is difficult to blame Gizmodo journalists for having committed a 'false cause' fallacy: given that urgent care clinics in USA can legally deny treatment to uninsured people and that 27 millions of Americans do not own insurance, the causal relation could have easily been true. Although news outlets, to help the public making sense of reality, cannot avoid adopting the abductive way of reasoning of finding the best possible explanations for state of affairs in view of the available evidence, extra caution should be devoted to avoiding fearmongering when reporting situations that are likely to be felt close and emblematic by the audience. A possible strategy, even if not optimal, is that of modulating the degree of certainty through the use of modal verbs, especially when correlations might be, but are not necessarily, causations. An example is offered by the Mirror which originally published an article entitled 'Netflix's Tiger King star Joe Exotic hospitalised after contracting coronavirus in prison' (https://tinyurl.com/kvwu83kw), but then had then to update its headline specifying that Exotic 'could' have contracted the virus but was still not officially diagnosed with it.

The fallacy of ambiguity and vagueness is, in our dataset, the one associated with the most varied set of sources. As to the former (ambiguity) our fact-checked cases show that frequently social media have tightened the tone of recommendations provided by governments and institutions, circulating panic. This mostly happens when the original source provides vague indications about the scope of applicability of the proposed measures. This is the case, for example, of a graphic containing recommendations from the U.S. Centers for Disease Control and Prevention (CDC) for reopening schools amid the COVID-19 coronavirus pandemic entitled 'here we go' (https://tinyurl.com/y4oh2uqh) that went viral on Facebook. While the graphic would suggest that the proposed set of rules is stringent, the CDS guidance stated that schools should determine whether and how it is feasible to implement such recommendations (https://tinyurl.com/y7k7ofnw). In our dataset, vagueness, when identified in broadcast media, often resides in the titles of the news, which allows for twisted interpretations. An example is offered by the *Raw Story* report entitled 'Here's how the Kushner family is cashing in on the coronavirus' (https://tinyurl.com/y3t5ba5p), where the phrasal verb 'cashing in on' could be interpreted both as merely getting financial revenue from a situation or taking advantage of a situation in an unfair way. As underlined by *Snopes*, while it is true that the Kushner brothers are co-founders of the health insurance start-up Oscar, which released an online tool to locate COVID-19 testing centres in some areas, there is no evidence that the startup is linked to any public damage. Therefore, the use of polysemous terms in news titles shall be avoided since potentially misleading for the majority of readers who are used to getting their daily news feed scrolling through news titles.

Towards a fallacy truth barometer

In order to investigate how fallacies entangle the information ecosystem we have looked at the correlations between types of rating and fallacies types (Figure 4):

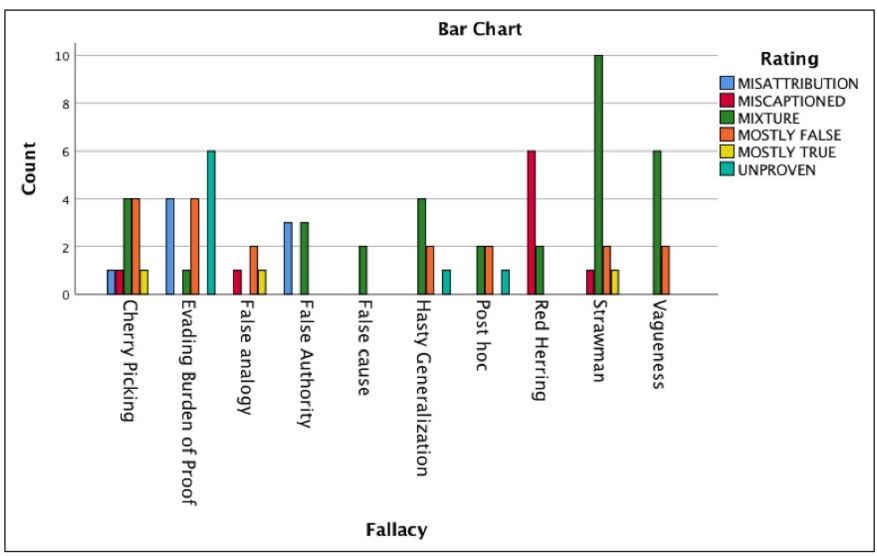

Figure 4. Fallacy types and *Snopes* ratings.

The categories *mostly false* and *mixture* are compatible with a wide array of fallacies from all four classes: the two ratings point to a degree of veridicality rather than the trigger of potentially misleading content. The rating *mostly true*, defined as a counterpart to *mostly false*, is attested in correspondence of three types of fallacies only, but it could, in principle, be compatible with almost any kind. The other three ratings correspond, instead, to a more constrained set of fallacies. When the rating *miscaptioned* is marked, what the fact checkers are pointing at is mostly the presence of a *red herring* fallacy: multi-modal content (images, videos) is generally used in news to provide arguments for a claim easy to process; the presence of a miscaption that falsely describes the origin, the context or the meaning of an image invalidates its relevance as a source of evidence. A common example is a photograph of empty food shelves except for vegan products used to back up claims such as 'Vegan Foods are Left Unsold During the COVID-19 Pandemic' (https://tinyurl.com/yy7h4wo3): the pictures were not taken during the pandemic, being available online years before. While representing vegan food products, such pictures are irrelevant to conclude anything related to food consumption habits during the pandemic.

The rating *misattribution* includes fallacies from false authority where an authority is attributed a claim that has been either tweaked or never uttered. A typical example is the letter, spread on social media, attributed to Johns Hopkins University containing an 'excellent summary' of advice on how to avoid catching COVID-19 (https://tinyurl.com/y6cz6mbd), while the content did not originate within the university.

Finally, the rating *Unproven* co-occurs with the fallacy of *evading the burden of proof* since no or poor evidence for the claim is offered apart from an individual's guarantee. An example is Dr. Vladimir Zelenko's claim about the treatment of COVID-19 coronavirus patients with a cocktail of hydroxychloroquine, azithromycin and zinc, which rests solely on his words (https://tinyurl.com/y2k86xg4).

The identification of fallacies could be used as a backbone to build more informative fact-checking ratings. Let's compare two examples both tagged by *Snopes* as *Mixture* (Table 2):

Table 2. Examples rated as *Mixture* by *Snopes*..

Claims	What is true	What is false	Source
1. Dr. Anthony Fauci said there was 'nothing to worry about' in late February 2020 in regards to COVID-19 and it was 'safe' to do things like go to the movies and the gym. (https://tinyurl.com/yy7b5j85)	During a Feb. 29, 2020, interview, Dr. Fauci said that at that time and under the circumstances pertaining to that date, Americans didn't need to change their behaviour patterns.	However, Fauci did not say there was 'nothing to worry about', and although he stated that Americans did not yet need to change their behaviours, he noted that what was then classified as the COVID-19 outbreak could require that to change	Social media
2. Amid a nationwide COVID-19 lockdown, Italians reported seeing wildlife such as swans and dolphins 'returning' to newly tranquil waterways, ports and canals. (https://tinyurl.com/y4dlxqg8)	Dolphins and swans were indeed spotted in some of Italy's waterways after the nationwide lockdown was imposed.	Dolphins and swans swimming in Italy's waterways were not necessarily new phenomena related to reduced human activity during the COVID-19 lockdown.	Social media

As correctly flagged by *Snopes* both claims contain elements of truth and of falsity. However, the origin and the nature of the false information differs wildly. Claim 1 reports a quote wrongly attributed to Dr Fauci: the sentence 'there is nothing to worry about' is a misleading rephrase of his statement uttered on February 29th during the NBC morning talk show *Today* that, at that moment in time, 'the risk is still low, but this could change'. By neglecting the provisional tone of Dr Fauci's assertion explicitly bound to the circumstances, social media posts from Trump's supporters have misrepresented Dr Fauci's position. The final goal of the critique is that of blaming the *National Institute of Allergy and Infectious Disease*s rather than Trump's administration for mishandling the response to the COVID-19 pandemic. This claim constitutes a clear instance of *strawman fallacy* since an opponent's point of view (Dr Fauci for Trump supporters) is distorted to make it easily attackable.

Claim 2, by contrast, gives voice to a non-legitimate causal inference between the lockdown and a revitalised animal wildlife in Italy. Probably eager to come up with good news, social media users have reframed a simple correlation as a causation, falling into a post hoc fallacy. There is, in fact, no scientific evidence that reduced human activity has caused dolphins and swans to go back to Venice, while it is clear that people had more time to notice their presence during the lockdown. In other words, in Claim 2 what is defeasible is the inference linking two states of affairs ('lockdown with reduced human activity' and 'wildlife around'), while in Claim 2 the defeasibility lies in the attack towards Dr Fauci, since based on inaccurate data (a statement that he literally never uttered). Thus, even if both claims contain propositions involving both truth and falsity, the mechanisms underlying the misinformation are inherently different and can be distinguished to provide readers with an appropriate warning and cultivate their critical skills. In this respect, we believe that the broad classes of fallacies identified (structural fallacies, fallacies from diversion, logical fallacies and language fallacies) can serve as clusters for types of misinformation: the questions proposed in our heuristic identification key can serve as systematic criteria for the identification of the misinformation triggers. Going through this set of questions fact-checkers would have the means to target the roots of misinformation, enabling citizens to strengthen their critical skills.

Conclusion

This study investigates the misinformation ecosystem during the pandemic. it focusses on the grey area of misinformation for three main reasons: misinformation is at least as harmful as disinformation for society, since it impacts a wider pool of people in post-truth scenario such as the pandemic; human fact-checkers currently lack a systematic approach to classify and disguise misinformation; automatic fact-checkers are unable to identity misinformation due to the lack of suitable training data to feed their systems. To tackle these issues, we start from defining which fake news, that we call *semi-fake news*, constitute vehicles of mis- rather than dis-information. Adopting a polyphonic perspective and drawing from current scholarly debate, we define semi-fake news as news not created/shared by the authors with the intention of circulating fabricated information and hard to be flagged by the public through common ground knowledge. What makes these news articles misleading is a variety of factors that underlie the notion of *fallacious*

arguments. Fallacies are arguments that seem valid but are not and, thus, constitute good candidates to convey misleading content.

Besides showing the relevance of the concept of fallacy for investigating misinformation, we propose a typology of 10 fallacious strategies to categorise misinformation about COVID-19, through a preliminary analysis of a set of COVID-19 news fact-checked by five English fact-checkers. This decalogue encompasses three main levels: presence and quantity of evidence supporting the claim (structural fallacies), relevance and (mis)representation of sources (diversion fallacies), (un)sound reasonings in drawing conclusions from available evidence (logical fallacies) and (un)clear narrative (language fallacies). To make these strategies operationalisable outside the academic sphere and by fact-checkers, we offer a heuristic identification key made up of binary questions as guidance for recognising fallacy structures. To further verify and showcase the descriptive power of fallacies for the misinformation ecosystem we conduct an annotation study over a corpus of 220 fact-checked news by *Snopes* related to COVID-19, encompassing the coding of types of sources next to that of fallacies. We accompany the annotation carried out by two students with no previous experience with that of an expert annotator to account for disagreements and provide an accurate dataset. Our dataset is the first containing fallacy annotation in fake news; its release will work as a seed to facilitate development of further research investigating the correlations between types of misinformation and types of fake news.

The results of the annotation show that the ten fallacies strongly correlate with news labelled by *Snopes* with ratings in between true and false. They, moreover, account for all the types of misinformation in such a grey area, thus working as indicators of misinformation. The analysis reveals that the lack of sufficient evidence to back up claims (evading the burden of proof fallacy), their arbitrary selection (cherry picking fallacy) or lack of relevance (red herring fallacy) constitute the most frequent roots of misinformation in our dataset. While these fallacious strategies are mostly hosted by social media due to the lack of editorial processes, issues related to the misrepresentation of sources (e.g. strawman fallacy), language ambiguities and defeasible reasonings (e.g. false cause fallacy) cut across all digital media. We thus, believe that our decalogue of fallacies could be used as a backbone to build a more solid gatekeeping process to counter misinformation. To investigate whether fallacious arguments are more frequently associated with misinformation rather than disinformation or information, we are planning to scale up our annotation study over a balanced corpus of valid news versus blatantly false news (disinformation) and misleading news (misinformation).

Regardless of its distinctive association with misinformation, our decalogue can be leveraged to create a standardised truth barometer based on systematic and non overlapping criteria. Ratings such as ‘mixture’ or ‘half true’ inform about the hybrid nature of the news, but do not focus on the roots of the misinformation. They, thus, leave the public with uncertainty and they fail to inoculate readers in view of similar misleading news which stray away from the fact checking system.

The Covid-19 pandemic has offered a potent environment for the development and spread of fake news. Complex, poorly understood epidemiology combined with inaccessible probabilistic expositions and rapid changes in scientific insight have led to widespread confusion in the general public. On top of that, hair-trigger political decision

making and perceived vested interests have set the charges on an explosion of unfounded theories and conspiracies which have exploited the ubiquity of the pandemic to reach huge audiences. The fine-grained, manual analysis of the anatomy of such fake news presented here provides a starting point for tackling this infodemic. As seen in tools for supporting public understanding and recognition of fakery such as the BBC's Evidence Toolkit (Visser et al., 2020a), even such small-scale, scalpel-like dissection can deliver significant value for large audiences. To have widespread impact, however, what's needed is the kind of rapid reaction to unseen data that can only be delivered through automated means. Automated fallacy detection has heretofore lain strictly beyond the state of the art. The integration of empirically and theoretically driven facets presented here opens up new computational venues for the detection of misinformation.

The integration of empirically and theoretically driven facets presented here opens up new computational venues for the detection of misinformation. The relatively low inter-annotator agreement in the task demonstrates just how demanding it is for humans to reliably identify fallacies, and presents significant challenge not only to automated techniques in general, but also to their application in specific applications. The key at this stage is to identify the lower-hanging fruit of problem use cases in which any automated recognition can deliver value for end users. This kind of 'problem engineering' is exactly the approach taken by the Evidence Toolkit. Though the techniques of automated argument mining (Lawrence and Reed, 2020) are, in general, tackling an enormously demanding linguistic challenge with general-purpose results at very modest levels, their application in the Evidence Toolkit focusses on providing advice for users – advice that is easily over-ridden and leads, whether accurate or not, to the kind of deep critical thinking that represents the goal of the application. And so it must be for automated fallacy identification: though accuracy levels may currently be much more modest than other natural language processing tasks, there are still a wide range of real-world problems that can be tackled with extant algorithms, so long as the precise mise-en-scène of those algorithms into specific applications delivers concrete benefits despite noisy performance.

The next steps are to build upon these foundations and scale up to allow not only the automated recognition of fallacious reasoning in the wild, but from there, the detection of semi-fake news at the point of consumption, empowering the general public to recognise and sift accurate reporting from pernicious, titillating and yet ultimately, deeply destructive fakery. Before such an automatic gatekeeping process becomes available, argumentation technology can still be leveraged for educational purposes. In the ESRC project 'Being Alone together: Developing Fake News Immunity' (https://fakenewsimmunity.liverpool.ac.uk/) we have developed the *Fake News Immunity Chatbot* to interactively teach citizens in a gamified environment how to recognise fallacies in news. Overall, we believe that fallacies can help citizens becoming their own fact-checkers and join the digital activism venture.

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 number ES/V003909/1].

Notes

1. Although the whole paper has been the result of a continuous process of interaction between the two authors, Elena Musi is the main responsible of Sections 1, 2, 4 and 6 while Chris Reed of Sections 3 and 5.

References

Barrón-Cedeño A, Elsayed T, Nakov P, et al. (2020) CheckThat! at CLEF 2020: Enabling the automatic identification and verification of claims in social media. In: Jose JM, Yilmaz E, Magalhães J, et al. (eds) *European Conference on Information Retrieval*. Cham: Springer, pp.499–507.

Bentham J (2015) *The Book of Fallacies. Collected Works of Jeremy Bent*. Oxford: Clarendon.

Boudry M, Paglieri F and Pigliucci M (2015) The fake, the flimsy, and the fallacious: Demarcating arguments in real life. *Argumentation* 29(4): 431–456.

Brennen JS, Simon F, Howard PN, et al. (2020) Types, sources, and claims of COVID-19 misinformation. *Report, Reuters Institute*, 7.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.

Collins J (2018) "The facts don't work": The EU referendum campaign and the journalistic construction of 'post-truth politics'. *Discourse, Context and Media* 27(3): 15–21.

Conroy NK, Rubin VL and Chen Y (2015) Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52(1): 1–4.

Dallwitz MJ (1980) A general system for coding taxonomic descriptions. *Taxon* 29: 41–46.

De Morgan A (1847) *Formal logic: Or, the calculus of inference, necessary and probable*. Taylor and Walton.

Duffy A, Tandoc E and Ling R (2020) Too good to be true, too good not to share: The social utility of fake news. *Information, Communication and Society* 23(13): 1965–1979.

Fallis D and Mathiesen K (2019) Fake news is counterfeit news. *Inquiry* 1–20.

Feng S, Banerjee R and Choi Y (2012) Syntactic stylometry for deception detection. In: *Proceedings of the 50th annual meeting of the association for computational linguistics*, Jeju, Republic of Korea, 8–14 July 2012, vol. 2: Short Papers, pp.171–175.

Gelfert A (2018) Fake news: A definition. *Informal Logic* 38(1): 84–117.

Hamblin CL (1970) *Fallacies*. London: Methuen and Co Ltd.

Hanselowski A, Stab C, Schulz C, et al. (2019) A richly annotated corpus for different tasks in automated fact-checking. In: *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, Hong Kong, China, 3–4 November 2019, pp.493–503. Hong Kong: Association for Computational Linguistics.

Horne B and Adali S (2017) This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In: *Proceedings of the international AAAI conference on web and social media*, Montreal, Canada, 3 May 2017.

Humprecht E (2019) Where 'fake news' flourishes: A comparison across four Western democracies. *Information, Communication and Society* 22(13): 1973–1988.

Islam MS, Sarkar T, Khan SH, et al. (2020) COVID-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene* 103(4): 1621–1629.

Kwon S, Cha M, Jung K, et al. (2013) Prominent features of rumor propagation in online social media. In: *2013 IEEE 13th international conference on data mining*, Dallas, TX, 7–10 December 2013, pp.1103–1108. New York: IEEE.

Kyriakidou M, Morani M, Soo N, et al. (2020) Government and media misinformation about COVID-19 is confusing the public. *In: LSE COVID-19 blog*. Available at: https://blogs.lse.ac.uk/covid19/2020/05/07/government-and-media-misinformation-about-covid-19-is-confusing-the-public/ (accessed 1 June 2021).

Landis JR and Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 159–174.

Lawrence J and Reed C (2020) Argument mining: A survey. *Computational Linguistics* 45(4): 765–818.

Levi L (2018) Real fake news and fake fake news. *First Amendment Law Review* 16: 232–327.

Musi E, Ghosh D and Muresan S (2016) Towards feasible guidelines for the annotation of argument schemes. In: *Proceedings of the third workshop on argument mining (ArgMining2016)*, Berlin, Germany, 7–12 August 2016, pp.82–93.

Negroponte N, Harrington R, McKay SR, et al. (1997) Being digital. *Computers in Physics* 11(3): 261–262.

Rubin VL and Lukoianova T (2015) Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology* 66(5): 905–917.

Schwarzenegger C (2020) Personal epistemologies of the media: Selective criticality, pragmatic trust, and competence-confidence in navigating media repertoires in the digital age. *New Media & Society* 22(2): 361–377.

Seargeant P and Tagg C (2019) Social media and the future of open debate: A user-oriented approach to Facebook's filter bubble conundrum. *Discourse, Context and Media* 27: 41–48.

Song Y, Heilman M, Klebanov BB, et al. (2014) Applying argumentation schemes for essay scoring. In: *Proceedings of the first workshop on argumentation mining*, Baltimore, MD, 26 June 2014, pp.69–78.

Sunstein CR (2018) *# Republic: Divided Democracy in the Age of Social Media*. Princeton, NJ: Princeton University Press.

Tandoc E, Ling R, Westlund O, et al. (2018) Audiences' acts of authentication in the age of fake news: A conceptual framework. *New Media & Society* 20(8): 2745–2763.

Tandoc EC Jr (2019) The facts of fake news: A research review. *Sociology Compass* 13(9): 1–9.

Thorne J and Vlachos A (2018) Automated fact checking: Task formulations, methods and future directions. In: *Proceedings of the 27th international conference on computational linguistics*, Santa Fe, New Mexico, USA, 20–26 August 2018, pp.3346–3359.

Thorne J, Vlachos A, Christodoulopoulos C, et al. (2018) Fever: A large-scale dataset for fact extraction and verification. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, 1–6 June, pp.809–819. Association for Computational Linguistics.

Tindale CW (2007) *Fallacies and Argument Appraisal*. Cambridge: Cambridge University Press.

Van Eemeren F and Grootendorst R (2004) *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge: Cambridge University Press.

Van Eemeren FH and Grootendorst R (1987) Fallacies in pragma-dialectical perspective. *Argumentation* 1(3): 283–301.

Van Eemeren FH and Snoeck Henkemans AF (2016) *Argumentation: Analysis and Evaluation*. New York, NY and London: Taylor and Francis.

Visser J, Lawrence J and Reed C (2020a) Reason-checking fake news. *Communications of the ACM* 63(11): 38–40.

Visser J, Lawrence J, Reed C, et al. (2020b) Annotating argument schemes. *Argumentation* 35: 1–39.

Walton DN (1996) *Argumentation Schemes for Presumptive Reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.
Wardle C (2017) Fake news. It's complicated. In: *First Draft 16*. Available at: https://firstdraftnews.org/articles/fake-news-complicated/ (accessed 1 June 2021).
Wason PC (1968) Reasoning about a rule. *Quarterly Journal of Experimental Psychology* 20(3): 273–281.
Whately R (1875) *Elements of Logic*, 9th edn. London: Longmans, Green and Company.
Woods J (2004) *The Death of Argument*. Dordrecht: Springer.
Zampa M (2017) *Argumentation in the Newsroom*, vol. 13. Amsterdam: John Benjamins Publishing Company.

Author biographies

Elena Musi is a Senior Lecturer (Associate Professor) in Communication and Media at the University of Liverpool and lead of the Digital inclusion cluster at the Centre for digital Humanities. Before joining the University of Liverpool, Elena worked as the Language Engineer for Alexa in Italian (AMDS team, Cambridge, Mass). Elena's research interweaves Artificial Intelligence, Linguistics and Communication Sciences with the broad aim of tracing back in a critical perspective debates about new technologies and their global impact, with particular focus on (mis)information and human-computer interaction. She has been PI on the UKRI project 'Staying Alone together: Developing Fake News Immunity'.

Chris Reed is Professor of Computer Science and Philosophy at the University of Dundee in Scotland, where he heads the Centre for Argument Technology (www.arg.tech). Chris has been working at the overlap between argumentation theory and artificial intelligence for two decades and specialises in the theory, practice and commercialisation of argument technology. He has won over £8m in funding from government, charity and commercial sources, has over 200 peer-reviewed papers in the area including five books, and has served as a director of several technology companies.