Reliability Criteria for News Websites

HENDRIK HEUER, Harvard University, USA and University of Bremen, Germany
ELENA L. GLASSMAN, Harvard University, USA

Misinformation poses a threat to democracy and to people's health. Reliability criteria for news websites can help people identify misinformation. But despite their importance, there has been no empirically substantiated list of criteria for distinguishing reliable from unreliable news websites. We identify reliability criteria, describe how they are applied in practice, and compare them to prior work. Based on our analysis, we distinguish between manipulable and less manipulable criteria and compare politically diverse laypeople as end users and journalists as expert users. We discuss 11 widely recognized criteria, including the following 6 criteria that are difficult to manipulate: content, political alignment, authors, professional standards, what sources are used, and a website's reputation. Finally, we describe how technology may be able to support people in applying these criteria in practice to assess the reliability of websites.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Misinformation, Disinformation, Fake News, Criteria, Media Literacy, Checklist, Online News, Fact-checking, Credibility, Information Quality

1 INTRODUCTION

Many people find it challenging to distinguish reliable from unreliable information. Misinformation played an important role in the COVID-19 pandemic [86], the 2021 United States Capitol attack [46], and the Russian invasion of Ukraine [72]. In their large-scale study, Arechar et al. showed that misinformation is a global problem, and that the psychological factors that underlie misinformation are similar across the globe [5]. In the U.S., nine out of ten adults believe that false information is a challenging problem [42], and a Pew Research Center study showed that most people in the U.S. have difficulty distinguishing real news from opinions [41].

Reliability criteria can empower readers to be more discerning when navigating a landscape that may include misinformation. The goal of this paper was to yield robust news reliability criteria independent of a person's experience with how news is produced. This paper is the first to empirically ground such criteria in a study that combines elements of contextual inquiry [61], *think aloud* study [35], and semi-structured interviews [34]. Our study focuses on reliability criteria for end users of genuine, live news websites with many articles. We report all criteria used by participants while evaluating different news websites in two countries, and discuss why certain criteria are more difficult than others to manipulate.

It is important to investigate news websites in a realistic and live setting, since four out of five people (82%) worldwide use online sources to read news [44]. Almost three out of five (57%) rely on social media, where links to online websites are frequently shared. For both online websites in general and social media in particular, there is little guidance on distinguishing reliable from unreliable news websites. While prior research showed that

Authors' addresses: Hendrik Heuer, hheuer@uni-bremen.de, Harvard University, Cambridge, MA, USA and University of Bremen, Bremen, Germany; Elena L. Glassman, glassman@seas.harvard.edu, Harvard University, Cambridge, MA, USA.

ACM 1073-0516/2023/12-ART
https://doi.org/10.1145/3635147

source labels are effective [27, 33], labels for entire websites are rarely provided. The reliability criteria for news websites presented in this paper can help users recognize websites that, while appearing to provide news, may not adhere to journalistic standards and may provide biased or false information for political or monetary gain.

In the context of misinformation, Wardle distinguishes between seven types of mis- and disinformation [77]: 1) misleading content, 2) false connection, 3) false context, 4) manipulated content, 5) fabricated content that is 100% false, 6) imposter content impersonating genuine sources, and 7) satire. While applicable to all seven types, the reliability criteria for news websites presented in this paper are especially relevant for online websites that provide misleading content, false connection, false context, and manipulated content.

A person's partisanship and political stance have been shown to influence their assessment of news reliability. In the context of the crowdsourced fact-checking program on X, formerly Twitter, for example, Allen et al. found that users tend to selectively question or dispute content posted by people with whom they disagree politically [2]. Work by Pereira et al. indicates that political identity increases the likelihood of a person believing positive news about their ingroup and negative news about their outgroup [55]. Pennycook et al., on the other hand, found that people's assessment of the plausibility of headlines is independent of whether a story is consistent or inconsistent with their political ideology [53]. Overall, prior work indicates that partisanship plays a complex role in the context of misinformation, motivating us to account for partisanship in our study design. A unique aspect of our work is covering the political spectrum in two countries. We provide a comprehensive overview of the criteria used in practice by professional journalists as expert users and laypeople end users across political affiliations. The term *end user*, in this context, refers to people without specialized knowledge about how to assess whether a news website is a reliable source to inform themselves and others.

While initial steps towards reliability criteria have been made [7, 9, 25, 45, 88], and while research indicates that these criteria are helpful in practice [5], which criteria users can apply and which are useful in practice remain open questions. Based on observing politically diverse laypeople as end users and journalists as expert users, we provide news reliability criteria that focus not on individual articles, but on entire news websites.

With far fewer websites than articles, providing reliability criteria for entire websites saves time, as users do not have to evaluate individual news articles and claims. The cost of starting a news website is also higher than the cost of publishing and promoting individual articles. These criteria also mean it is not necessary to engage with individual claims in a story to explain why a website is unreliable, and they can support users before an article containing misinformation is even created and be integrated more easily into social media platforms.

To observe which reliability criteria are actually applied in practice, we studied professional journalists as experts and politically diverse laypeople as end users. With knowledge of how news is produced, professional journalists are experts in assessing the reliability of news websites and in fact-checking claims. End users have less experience with how news is produced. By looking at both experts and end users, we identify criteria 1) that are used by experts, 2) that are within end users' abilities, and 3) that debunk common end users' choices that are less helpful for identifying reliable information. As explained, prior work has shown that the assessment of news reliability can be influenced by a person's partisanship and political stance [2, 53, 55]. To do justice to this complexity, we recruited elected representatives from a state parliament in Germany as end users. The political alignment of elected representatives can be clearly identified and, as a group, they are representative of the many different common political stances within the society. Our sample represents the voting body in the Free Hanseatic City of Bremen, one of 16 federal states in Germany.

We used eleven criteria to determine the reliability of online information: content, political alignment, writing style, authors, self-description, professional standards, advertisements, ownership, sources, reputation, and website design. These criteria are intended to directly help news consumers assess the reliability of news websites they encounter on social media and through search engines. The criteria may also be helpful for those who operate social media platforms like TikTok, Facebook, and Telegram, who could use the criteria to streamline content moderation processes, train content moderators, and show our criteria to users. In addition, the criteria

may support those who witness the sharing of misinformation, e.g., by helping them spot the misinformation and justify why it qualifies as such. Witnesses could include journalists who fact-check a particular website or people who encounter friends or family members sharing questionable links, e.g., in a family group chat. For these situations, we hope our reliability criteria for news websites' practice are a quick and empirically sound way to decide and explain why a given website may be unreliable.

We distinguish between criteria that can be easily manipulated by misinformation providers and other malicious actors and criteria that are more difficult to manipulate. In Section 5.1, we also compare the criteria to prior work based on different methodologies. We conclude by explaining how the criteria we consider difficult to manipulate can be best implemented in practice.

We address the described gaps in the literature by making the following contributions:

- We examine how end users and experts assess the reliability of news websites through an empirical investigation that combines elements of contextual inquiry, think aloud study, and semi-structured interviews (RQ1).
- We present criteria based on our empirical work that could be used by future news consumers to help them assess the reliability of news websites (RQ2).
- We compare the similarities and differences of end users and experts with respect to identifying misinformation (RQ3).

2 BACKGROUND

With the advent and expansion of social media, the distribution of misinformation has become faster and cheaper. Consequently, misinformation has received a great deal of scholarly attention in recent years [3, 37, 38, 50, 58, 66, 73]. In this paper we adapt Wardle et al.'s operationalization of misinformation as false information, false connections, and misleading content [78]. Like Starbird et al., we understand misinformation as collaborative work within online crowds [70]. We explore criteria that can expose misinformation, and discuss the use of technology to support this exposure. Prior research has examined what makes people prone to believe in misinformation [4, 54, 62] and how the spread of misinformation can be studied on social media [21, 47, 65]. Notably, Pennycook et al. found that people's proneness to believe or share misinformation is often not because of motivated reasoning but because people do not pay attention [51, 53]. Prior research also indicates that the role of individual rationality in the context of misinformation may be overstated, and that sharing of misinformation may be influenced by shared group-level narratives [68] and the social media context [51].

It is important here to consider different notions of "sharing" [36]. In the social media context, many activities described as "sharing" are political and value-laden. Facebook, for instance, encourages peer-to-peer sharing to "amass" data about user activities [36]. Considering the findings by Pennycook et al. [51, 53], Facebook's incentivization and users' tendency to "share" information without paying attention could, therefore, increase the misinformation problem.

Considering the challenges posed by misinformation, simple ways of supporting users in determining the reliability of a news website are needed. Shahid et al.'s recent study on misinformation in videos in India showed, for instance, that users are unwilling and lack the skills to make the assessment [63]. In a similar study on WhatsApp in India, Varanasi et al. found that simple and easy-to-process labels are preferable to more complex fact-checks [63].

As an important step towards simple and easy-to-process solutions, we focus on the utility of reliability criteria for entire news websites. Qualitative work by Jahanbakhsh et al. on why people trust individual news stories showed that this trust can be attributed to individual users' knowledge and experience [30]. Their identified criteria include having firsthand knowledge, whether information is consistent with users' experience, whether other sources confirm claims, whether evidence is provided in the article, and whether the story is from a

trusted source. This prior work informed us to explore criteria beyond individual claims, since it can be hard for individuals to verify if they lack experience with news production.

2.1 The Pursuit of Solutions Against Misinformation

Researchers have explored a number of ways to support people in identifying misinformation. The most important prerequisite for this is that people can be persuaded to consistently and reliably change their beliefs — a possibility shown in prior work [1, 60, 85]. Attempts to address the problem of misinformation include checklists that help people consider the accuracy of the content they want to share [30, 51] and reliability ratings for individual stories [33].

Lazer et al. identify two kinds of solutions against misinformation: 1) changes focused on the individual, e.g., empowering them to evaluate the misinformation, and 2) structural changes that prevent individuals from encountering misinformation [37]. With this paper, we hope to contribute to both. Reliability criteria can empower individuals to assess whether a website is reliable, and online platforms and media companies can use the criteria to add warnings to content and to remove certain websites in a transparent and explainable way.

While a large body of research on potential solutions examines the feasibility of using data mining and machine learning to automatically classify articles as misinformation [12, 56, 58, 64, 66, 76], researchers in critical data studies highlight limitations in the quality and availability of training data [6, 28], noting that since ML-based systems rely on correlations, their applicability is limited. For example, while data-driven approaches may help identify potential existing and debunked misinformation stories, their helpfulness for emerging stories is limited.

One of the most promising non-technical solutions against misinformation that has emerged in recent years is the so-called lateral reading technique, pioneered by Wineburg and McGrew [84]. The technique is informed by the finding that professional fact-checkers make better decisions and need less time because they step out of a website and use additional information from other sources to make an informed assessment. For Wineburg and McGrew, this lateral reading means reading less and learning more. Unlike those who stay on a specific page to evaluate a website's reliability, those who read laterally scan a website and then use other websites to judge the original site's credibility. The effectiveness of lateral reading was shown in a large-scale, district-wide field study within high school government classes, with results showing that lateral reading can significantly improve high school students' ability to judge the credibility of digital content [83]. Our reliability criteria support such lateral reading by helping end users decide what aspects to consider when using other websites to judge the reliability of a news website.

2.2 Reliability Criteria

In the early 2000s, Fogg et al. examined people's perception of the credibility of early websites [22]. Elements that were seen to increase the perceived credibility included "real-world feel," "ease of use," and "trustworthiness," among others. Perceived "amateurism" and "commercial implications," on the other hand, were seen as factors that hurt the credibility of websites. Fogg et al. also provide an early example of studying the credibility of live websites with a focus on laypeople [23]. Twenty years after these investigations, information environments have changed dramatically, especially news websites. These differences motivated us to focus on reliability criteria of current news websites.

Based on a literature review of keywords like "reliability criteria" and "credibility criteria," as well as on recommendations by reviewers, we compiled a list of criteria from prior work aimed at helping users distinguish reliable from unreliable news websites. Many lists of criteria include no explanation of how the criteria were determined or of their empirical basis [9, 26, 45, 71]. Two such lists are particularly noteworthy because they are widely used: the Facebook Tips [25] and the News Guard criteria [45] (Table 1). Guess et al. empirically showed the effectiveness of the Facebook Tips in the United States and India [25]. In the United States, the increase in

Table 1. Reliability criteria from prior work on assessing websites. We include the Facebook Tips [25] and the NewsGuard Criteria [45]. Both sets of criteria are widely used, and both lack empirical grounding.

Facebook Tips [25]	**NewsGuard Criteria [45]**
1. Be skeptical of headlines	I. Does not repeatedly publish false content
2. Look closely at the URL	II. Gathers and presents information responsibly
3. Investigate the source	III. Regularly corrects or clarifies errors
4. Watch for unusual formatting	IV. Handles the difference between news and opinion responsibly
5. Consider the photos	V. Avoids deceptive headlines
6. Inspect the dates	VI. Website discloses ownership and financing
7. Check the evidence	VII. Clearly labeling advertising
8. Look at other reports	VIII. Reveals who's in charge
9. Is the story a joke?	IX. The site provides names of content creators along
10. Some stories are intentionally false	with either contact or biographical information

discernment remained measurable after several weeks. These tips were also included in a large-scale investigation by Arechar et al., who surveyed more than 33,000+ people in 16 countries across six continents, finding that combining the tips with subtle prompts to think about accuracy improved the veracity of news people were willing to share across the globe [5].

Other reliability criteria without clear explanations (i.e., the authors wrote that they selected certain criteria without also specifying their selection process) are those used by Bradshaw et al. [9] and the Trust Indicators by The Trust Project [71]. The criteria by Bradshaw et al. [9] include professionalism, i.e., whether best practices of professional journalism are followed; writing style; credibility, including whether false information or conspiracy theories are shared and corrections are provided; bias; and whether others' design and content strategies are imitated ("counterfeit"). The Trust Indicators by The Trust Project [71] include following journalistic best practices, having journalistic expertise, labeling types of content and reducing bias, referencing sources, using transparent reporting methods, locally sourcing stories, including diverse voices, and inviting and listening to feedback.

We identified two examples in which criteria are based on discussions with experts. Zhang et al. [88] used a subset of 16 credibility indicators chosen from over 100 proposed by representatives from journalism and fact-checking groups, research labs, social and annotation platforms, web standards, and others. They distinguish between content and context indicators. *Content indicators* include title representativeness, "clickbait" titles, quotes from outside experts, citations of organizations and studies, calibration of confidence, logical fallacies, and inference (i.e., how correlation and causation are discussed). *Context indicators* include originality, fact-checks, representative citations, the reputation of citations, the number of ads, the number of social calls, "spammy" ads, and the placement of ads. The 16 chosen indicators require no subject matter domain knowledge, off-line investigation, or additional data gathering. Since they focused on articles, Zhang et al. also ignored indicators related to publishers, authors, and multimedia content.

The second example of criteria based on experts' discussions are the Credibility Signals proposed by the Credible Web Community Group [74]. Here, again, criteria were discussed by experts. The following signals were selected after a review in 2020: the date a website was first archived, and whether the site has a corrections policy, has won any awards, has won a Pulitzer Prize, or has won the Ramnath Goenka Excellence in Journalism Award.

Considering the plethora of reliability criteria, it is important to understand whether they are empirically grounded in the practices of end users and experts. We identified only one example of reliability criteria based on a survey that studied this [7]. Bhuiyan et al. focused on individual news stories in the context of climate science and crowdsourcing news credibility assessment [7], comparing the news credibility assessments of students

and crowd workers to those of three scientists and three professional journalists. Based on the responses of the six participants, they identified eight credibility criteria: accuracy, impartiality, completeness of coverage, originality and insight, credible evidence/grounding, publication reputation, professional practices and standards, and website aesthetics. Since these criteria hinge on the domain knowledge of individuals, we expand on them with a focus on news websites for the reasons explained in our introduction. Our study focuses on the whole website, a more extensive breadth of topics, and a much larger sample of consulted experts.

3 METHODS

To determine reliability criteria for news websites, and to ground these criteria in how journalism is done in practice and how end users think about the journalism they consume, we combine elements of contextual inquiry [61], think aloud study [35], and semi-structured interviews [34]. It was necessary to combine these different elements, as contextual inquiry or participant observation alone would not have allowed us to understand the criteria participants use in practice, since these considerations are not observable. Conducting a survey was not feasible, since we wanted to capture participants' reactions to live websites with multiple articles and since our goal was to capture all criteria used in practice. Finally, structured interviews without the live websites and the think aloud method could have limited the insights we obtained, since participants would have had to rely on their memories about past interactions with unreliable websites.

We studied two groups of people: professional journalists as experts, and politically diverse laypeople as end users. Recruiting from these two groups also allowed us to determine participants' political affiliation, which is important considering the complex role that partisanship plays in the assessment of news [2, 20, 32, 53, 55]. For end users, we recruited a representative sample of elected politicians in a Western democracy. This sampling ensured participant diversity and a full range of political views. In theory, the elected politicians in a representative democracy are representative of the voting body. In practice, people with higher education degrees are highly overrepresented in parliaments worldwide [19]. We will reflect on this in the Limitations Section 5.6.

The target audience for our news reliability criteria is the end user. While end users may lack the expertise necessary to follow all of the practices experts use to assess a news site's credibility, experts' practices are worth capturing in that 1) they represent a potential gold standard, 2) some of their practices may be teachable to end users, and 3) some of their practices may be accessible to end users with the assistance of future automated or socio-technical systems. Current end users' practices in evaluating news sites' credibility are valuable for 1) capturing the range of current practices among those not professionally engaged in news production and 2) comparing these current practices to those of experts to better understand the gap that these reliability guidelines are intended to remediate.

Since prior work has shown the importance of cultural and political differences, we identified appropriate field sites in two different countries — the United States and Germany. We selected experts from the United States, a country with high concern about misinformation and high media polarization [29, 43], which these experts had directly experienced. And we selected a politically diverse set of end users that were representative of a federal state within Germany because that is where we had the best access to elected officials from across a wide political spectrum. While this prohibits us from drawing *strong* conclusions about cultural differences or differences in how end users and experts think about evaluating news sites, it allowed us to collect data across two different information and cultural landscapes, which we hope leads to a more generalized and therefore more useful set of criteria.

The research questions that drove our study design are:

- **RQ1:** What reliability ratings do end users and experts provide for news websites?
- **RQ2:** What criteria do end users and experts apply to assess whether news websites are reliable?
- **RQ3:** What differences between end users and experts can be observed?

Table 2. **Study Procedure:** Each participant reviewed three purported news websites following the structure in the table.

Instructions & Questions for the Tasks: Rate the Reliability of This Purported News Website.
1. First, participants were asked **whether they had personally used** the website. 2. Then they were asked to **freely explore the website** in their own browser at their own pace. We encouraged them to **describe the reasoning** behind their actions and decisions (**"think aloud"** [35]). Participants were allowed to read as many articles as they wanted, to visit all sections of the website, and to open other websites like search engines, social media websites, and encyclopedias. 3. While exploring the website, participants were regularly asked whether they **felt comfortable providing a rating**. If they answered affirmatively, they were asked about their rating. If not, they were allowed to continue browsing. Reviewing an individual website took around five minutes. 4. After participants **provided their rating** on a 5-point Likert scale, they were asked to **elaborate on their rationale for the rating**. 5. Finally, participants were asked whether **any information** that would have helped them make their decisions was **missing** during the study.

3.1 Procedure

In both the written and the oral briefings before the study, we explained that we had two goals: to understand how people decide whether a website is reliable and, as a consequence, to inform the design of software that can better support users in making this determination. The sessions were scheduled for 30 minutes, but some took considerably longer. At the start of each session, we collected demographic information. All participants were asked their age, gender, and highest completed educational degree. Politicians, as representatives of end users, were also asked to identify their party affiliation, while journalists, as experts, were asked to identify the news organizations for which they currently work and worked for in the past.

During the sessions, each participant evaluated the reliability of three purported news websites. Table 2 describes the study procedure, the task, the instructions we gave, and the questions we asked. Participants freely explored each website in their own browsers, at their own pace, and while thinking aloud. To make the study as realistic as possible, they interacted with live websites. Since these websites were continuously adding content, the participants viewed different articles.

This task, the *rating task*, helped participants reflect on the reasons that make them consider a website reliable or not. For each purported news website, we first asked participants whether they had personally used the site. The answer options included "Never," "Rarely," "Sometimes," "Often," and "Always." We then asked "How do you think the reliability of the source should be rated?" The participants provided their assessment on a 5-point scale, with the options "Very unreliable," "Unreliable," "Partially reliable, partially unreliable," "Reliable," and "Very reliable." Participants were instructed to take as much time as they needed. Akin to iterative interrogative techniques like the five whys, we repeatedly asked participants to explain the reasons for their rating. Our goal was to determine the root reasons for the perceived reliability of a source [80]. The study also included four short questions for a follow-up publication unrelated to reliability criteria.

We transcribed the audio recordings of the think aloud study and the semi-structured interviews, and analyzed the material using qualitative content analysis [39], a method equivalent to thematic analysis [10, 11]. We followed axial coding principles [14], with the first author reading the texts multiple times and, moving back and forth through the material, doing an open coding of the material. The German material was coded in German, and the English material in English. These codes were then discussed in weekly meetings with the second author, which helped us refine a set of well-defined codes. After clustering, splitting, and merging different codes into coherent groups, we identified categories/themes and subcategories, and improved them in weekly sessions

until we reached a unanimous agreement. The responsible authorities granted IRB-equivalent approval for the portion of the study conducted in Germany. The U.S.-based portion of the study was reviewed and accepted by the hosting institution's IRB. We obtained informed consent from all participants.

3.2 Selection of Purported News Sources

Since our participants were located in two different countries, we customized the news sites for each country. The politically diverse end users, all German speakers, were presented with three German news sources, while the experts were presented with three news sources in English from the U.S. All of the sources had been rated by reputable external news analysts as unreliable.

For the end user sessions in Germany, we selected two sources from the 20 German domains most frequently visited by Facebook users [40]. The dataset is based on 40,000 URLs that Facebook verified via third-party fact-checks. Since the Facebook dataset only contained extreme cases of content reported by users, we also included one questionable source that is not extreme on the spectrum from very reliable to very unreliable, selecting a German website that has won a prize for alternative media but that was regularly criticized for publishing conspiracy theories about the COVID-19 virus and pandemic.

For the sessions with experts, we sampled three English sources classified as unreliable in the meta-ranking by Gruppi et al. [24], which combines reliability ratings by Allsides, BuzzFeed, Media Bias / Fact Check, Open Sources, Pew Research Center, PolitiFact, and Wikipedia. Media Bias / Fact Check also provides a more fine-grained website classification. We randomly selected two sources from those that Media Bias / Fact Check rated as "Conspiracy Pseudoscience" and one source from those rated "Questionable Source."

We used the described method to ensure the purported news sources were unreliable. The first author confirmed the content and tone of the different websites to be comparable to those of unreliable websites, based on his five-year experience studying misinformation in German and English. Throughout the paper, we make cultural differences — like the role of the German AfD — transparent when they might influence the understanding of news reliability criteria.

3.3 Participants

Politicians as End Users. We recruited a politically diverse group of 23 elected politicians from one of the sixteen state parliaments of the Federal Republic of Germany. To do so, we contacted the president of the parliament through personal connections. The president authorized and endorsed our study and sent an invitation to parliament members.

For context, we briefly explain the political alignment of Germany's different parties, using Wikipedia's classification of these parties as a reference for their political alignment [79]. According to Wikipedia, the Christian Democratic Union (CDU) is seen as center-right, the Free Democratic Party (FDP) as center to center-right, and the Social Democratic Party (SPD) and Alliance '90/The Greens (GN) as center-left. The party The Left (LT) is classified as left-wing to far-left, and the so-called "Alternative" for Germany (AfD) as far-right. At the time of our study, the SPD, The Greens, and The Left formed the state government. The AfD is the first far-right party to win seats in the German Federal Parliament since the Nazis. It is essential for readers without exposure to German politics to know that, at the time of our study, all other parties had resolutions not to cooperate with the AfD.

We recruited 23 of the 84 elected representatives (27.4%) of the Free Hanseatic City of Bremen. The distribution of party affiliation of the end users is almost proportional to that of the last election. We recruited seven end users from the CDU, six from the SPD, four from GN, one from the FDP, two from the AFD, and one from another far-right party called "Citizens in Rage" (German: "Bürger in Wut"). To avoid identification of the latter, we merged his data with that of the AfD members. Both parties are far-right parties with similar policies.

Fifteen (65.2%) of the end users were male, and eight (34.8%) were female — similar to the percentage of females in the state parliament of the Free Hanseatic City of Bremen (36.9%). The mean age of the end users was 49.10 years old (SD=11.37); the oldest was 65, and the youngest was 30. The sample of end users was diverse in terms of the highest education completed. All had completed the German equivalent of high school, four had received the German certificate of general qualification for university entrance but did not continue to higher education, two finished vocational/professional qualifications, two had a Bachelor's degree, 11 had a Master's degree, one had a PhD, and one had completed the state examination to be a lawyer.

Journalists as Experts. We used two distinct newspaper sampling strategies to recruit 20 journalists working for U.S.-based, English-language newspapers. The first was based on readership, looking at the ten most widely circulated newspapers in the U.S. in 2019 [82]; we recruited seven experts from these newspapers. We refer to them as JP — i.e., JP1 is the first expert we recruited based on the popularity of their employer's newspaper. The second strategy was based on political alignment, allowing us to include a politically diverse collection of news organizations. Using the meta-ranking of Gruppi et al. [24], we identified reliable newspapers classified as left, center, or right. We determined political alignment using Media Bias / Fact Check, an American fact-checking website that rates news sources. While these ratings have been criticized for not meeting the highest standards of rigor and objectivity, Chołoniewski et al. [13] point out that, despite their imperfections, they have been judged as "accurate enough to be used as ground-truth for, e.g., media bias classifiers [48, 50, 51], fake news studies [52–54], and automatic fact-checking systems [56–58]." We sampled seven reliable newspapers identified as politically left or center-left, seven politically in the center or "least biased," and eight politically right or center-right.

For context, we briefly describe the political alignment of the major parties in the United States. Since the 1850s, the U.S. presidential election has been won by either the Democratic Party or the Republican Party. The Pew Research Center recognizes four Democrat groups — Outsider Left (closest to center), Democratic Mainstays, Establishment Liberals, and Progressive Left (furthest from center), and four Republican-oriented groups — Ambivalent Right (closest to center), Populist Right, Committed Conservatives, and Faith and Flag Conservatives (furthest from center) [16]. Due to a lack of available resources on how to map different news sources to these fine-grained criteria, we relied on the criteria by Media Bias / Fact Check.

For each newspaper selected through our sampling scheme, we emailed the experts listed as authors of the most recent news articles posted on the newspaper's official Twitter[1] account. We ignored reporting on sports, celebrities, dining, and cooking. In cases where an email was not listed but a Twitter account was open for direct messages, we contacted them via Twitter from the first author's personal account. After recruiting 11 experts who identified as male, we stopped recruiting experts with stereotypical male names in an attempt to recruit a more gender-balanced sample.

In the final sample, eleven of 20 experts identified as male and nine as female. The mean age was 35.5 (SD=10.27); the oldest expert was 61, and the youngest 21. Seven experts were recruited from among the Top 10 newspapers sampled by readership, and the remaining 13 from newspapers sampled by political stance. We recruited six experts from the left (we refer to them as JL), three from the center (JC), and four from the right (JR). All sessions with experts were conducted in English. The experts in our sample were highly educated; all but one were working towards or had completed university degrees after finishing high school. One was currently working towards a Bachelor's while working as a journalist. Seven stated that a Bachelor's was their highest degree; of these, four degrees were in journalism. Twelve experts had a Master's degree, four in journalism.

The experts worked for a number of news outlets, including newspapers with international coverage and circulation like the New York Times, national papers like Newsday and USA Today, and regional papers like the Star Tribune (Minnesota), The Chicago Tribune, and Capitol News Illinois. Also represented were online-only

[1]Twitter has been rebranded to X. In the Methods and Results sections, we refer to Twitter because this was the name of the service at the time of the investigation. In the Discussion and Conclusion, we refer to X.

outlets like Salon, Slate, Vox, The Observer, and CQ RollCall. "Right center" (rated by Media Bias / Fact Check as center-right) outlets included the Washington Examiner and HotAir. Two or more participating experts represented the following outlets in our sample: The Chicago Tribune, the Boston Globe, Foreign Policy, Media Matters, Newsday, Politico, and Vox.

4 RESULTS

Through a combination of contextual inquiry, think aloud study, and semi-structured interviews, we identified the reliability ratings our participants provided for different news websites (RQ1). Based on our observations of and their reflections on their process, we compiled a set of criteria participants used for arriving at those ratings (RQ2). We also examined the similarities and differences in how end users and experts approach this critical task (RQ3).

We refer to the three German websites as DE1, DE2, and DE3, and the three U.S.-based websites as EN1, EN2, and EN3.

4.1 Reliability Ratings Participants Gave to Purported News Websites (RQ1)

Personal Usage. In general, the end users and experts in our sample were not personally using the three purported news sources we asked them to assess. Most of the German end users had never used the websites, with 22% of end users at least rarely using DE1, 17% using DE2, and 13% using DE3. One end user (AFD2) stated that he frequently visited DE1, while another (SPD1) frequently used DE2. Even fewer experts reported that they personally used any of the sources we asked them to review. Seventeen of the 20 had never used EN1, while the remaining 3 (JP7, JL2, JR2) stated that they used it rarely. None of the experts had personally used EN2 or EN3.

Reliability Ratings. We found that end users and experts rarely rated the purported news sites as reliable. Only one end user (3%) considered DE1 to be reliable (AFD1), while more than half (52%) rated it as unreliable (35%) or very unreliable (17%). Three end users (13%) rated DE2 as reliable (CDU4, GN3, SPD1), but almost six of ten (57%) found it unreliable (22%) or very unreliable (35%). DE3 was never rated as reliable, and almost four out of five end users (78%) perceived DE3 as unreliable (26%) or very unreliable (52%). Table 5 in the Appendix provides all individual ratings, along with the party of the end user rating the websites.

The ratings of the experts for the English sources were even more extreme along the spectrum from reliable to unreliable. Only one expert (JP5) rated EN1 as reliable (5%), and only one (JC1) rated EN2 as reliable. Nobody perceived EN3 as reliable. Fourteen of the 20 experts (70%) rated EN1 to be either unreliable (25%) or very unreliable (45%), 17 of the 20 (85%) considered EN2 either unreliable (30%) or very unreliable (55%), and 16 of the 20 (80%) perceived EN3 as unreliable (20%) or very unreliable (60%). A detailed overview of the individual ratings of each expert and information on how the expert was recruited can be found in the Appendix in Table 6.

Overall, we found that the ratings in the U.S. corresponded to the meta-ranking of reliability by Gruppi et al. [24], while the ratings in Germany were consistent with the crowd-sourced ratings validated by fact-checkers [40]. The two populations were not the same, however, in their recognition of the sources' unreliability: experts more frequently labeled sources as very unreliable or unreliable, while end users chose the "partially reliable, partially unreliable" option more frequently.

4.2 Criteria To Determine the Credibility of News Websites (RQ2)

The qualitative coding of the 43 sessions yielded 11 criteria used by participants to determine the reliability of news websites. Figure 1 shows how many end users and experts used each criterion to evaluate the reliability of a purported news website; the black lines show the means of both groups. In order of aggregated use across both end users and experts, the most widely applied criterion was the website's content, followed by its political alignment and its writing style. Both end users and experts also took into account the authors of the hosted articles

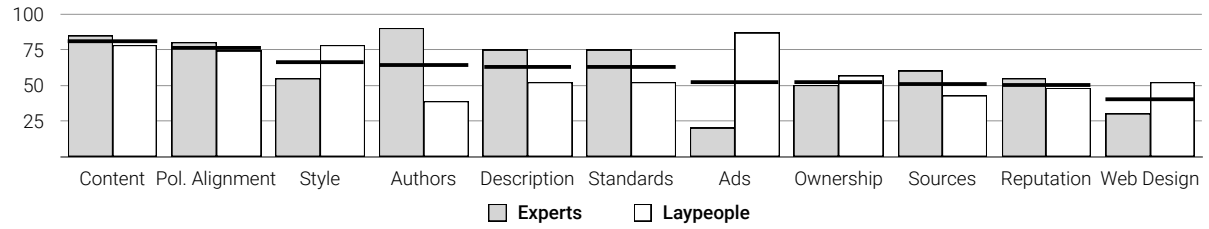

Fig. 1. Percentage of the 23 politically diverse end users from Germany and the 20 experts from the U.S. who used the eleven reliability criteria when assessing the reliability of a news source. The black line indicates the mean of both groups.

and the website's self-description. Other important factors included whether a website followed professional standards, the site's advertisements, and its owners. Participants also considered the sources used by the website, the site's reputation, and its design.

We next describe each criterion in detail, distinguishing between end users and experts to highlight differences in how they approached each criterion.

4.3 Content

End Users. Eighteen of the 23 end users (78%) took the articles' content and topics into account when rating a news source's reliability. Important aspects in this regard included whether more than one side of an issue was represented (FDP1, GN2), and how independent (GN1, LT2), neutral (CDU1), and objective (GN1) the reporting was. End users considered how critically and thoughtfully something was discussed (FDP1), how controversial topics were presented (GN1), and, more generally, the quality of the content (SPD5). End users also found it problematic if opinions were disguised as news (GN2) and if websites read like a blog (SPD4).

The content criterion connects to whether the information is compatible with people's knowledge and whether they see it as realistic (AFD1, AFD2, AFD3, CDU3, CDU6, LT2, SPD1, SPD2, SPD4, SPD5). CDU6, for instance, considered whether she had heard about an incident from different people: "When everyone reports the same thing, then it must be real." SPD5 described this as a plausibility check. SPD1 used topics he is knowledgeable about to judge source reliability. Topics mentioned by end users included COVID-19 (LT1, SPD3) and mandatory vaccinations (FDP1), anti-Israeli or antisemitic attitudes (GN2, SPD4, SPD4), and right-wing conspiracy theories (GN1, LT1). FDP1 found it unusual that one of the websites did not cover topics like politics, sports, and society.

GN2 distinguished between misinformation that is completely made up, e.g., an invented incident involving foreigners and crimes, and articles that reinforce stereotypes and provide certain opinions in one-sided accounts. SPD3 included whether facts are twisted as an essential criterion. The end users looked at whether an article was from a news website or a blog (CDU3, GN2, GN4, LT1, SPD3, SPD4), and whether it was an op-ed or a source spreading propaganda (LT1). The amount of content was also an important criterion. SPD4 argued that if a source would produce that much content, "then one would have heard from this source already".

End users also commented specifically on whether websites were perceived as spreading conspiracies or promoting conspiratorial thinking, including right-wing conspiracy ideologies or propaganda (GN1, GN4), especially antisemitic and Anti-Israeli positions (SPD4).

Experts. Seventeen of the 20 experts (85%) recognized the content a news website publishes and the topics covered as important criteria. JP5, for instance, commented on the selection of stories, expecting that important current news stories would be covered. JL5 considered whether a website covered a story in the same way as a reliable source and whether the facts stated in the article fit what she knew from other sources.

When assessing the reliability of news sources, the experts took the type of website into account. They distinguished between sources that provide breaking news (JP5, JC1, JR1) and websites that do editorial, commentary, or opinion writing (JP1, JP5, JC2, JC3, JR1, JR4) or provide analyses (JR4). Experts also differentiated between traditional or mainstream media sites with editorial processes (JC1, JC2, JR3), curated websites that collect articles from other sources (JL1, JL3), and blogs (JP1, JC3). JC2 and JP2 expected a website to clearly indicate that it is providing an opinion. The experts also took into account whether the content is actually relevant for the stated target group, e.g., military veterans (JR1). JR1 believed that news sources must be neutral. JL1 believed that reliable websites give news free of personal opinions. While he believed that there are stories that have four or five sides, there are many others with only one side: the earth is round, man did land on the moon, and the Holocaust happened. JL1 reviewed all articles on one website and found that "these are all lies — I mean, these aren't even good lies." For JP2 and JP3, opinions need to be labeled as such. JP2 felt that opinion and news "bled into each [other]" in one of the sources. Experts criticized websites they perceived as passing opinions as news (JP2, JP3, JL2, JL4, JL6).

JL6 commented that not everything written by one of the websites was inaccurate but that he was still skeptical about the sites. JC1 wanted to understand whether a news source had an accurate grasp of the subjects they reported on. To quickly check a website, she focused on statistics she could easily cross-check, e.g., via government websites. Considering the breadth of content on one of the sites, JP2 wondered whether a source had reporters in all countries covered or, if not, where the information came from. JL6 found that ideologically driven sites like the one he was reviewing might write something factually correct, but without context. He argued that such sites do a lot of "framing by omission," which made him skeptical.

The experts took several other content-related factors into account, including whether a well-established news outlet would cover a story (JP2), and a website's categories and how they were determined (JP7, JL2). JL2 perceived news sites with categories like "Disloyal Military," "Progressive Fascists," and "Racist Mayors" as unreliable. JL6 doubted the reliability of the reporters' fact-finding and whether it was based on original research, and JP7 found the information highlighted on the sites to be false or delivered with a "spin" (JP7).

Several experts (JP3, JP5, JP6, JL2, JL6, JR2, JR3) commented on the conspiracy-related and antisemitic content of one of the websites, noting conspiracies like the so-called "great reset" (JP7, JR2) and the "deep state" (JP5) as well as misinformation about vaccines (JP6). JR2, writing for a center-right news source, described these topics as connections to "the crazier sides of the conservative ecosystem." Experts noticed a general conspiratorial feeling (JP5) or pointed out specific topics like Freemansory (JR2, JR3) and Satanism (JL6). JP1 and JP2 recognized right-wing tropes. Specific content made experts suspicious, e.g., alleged plans for gun permits based on political views (JR3), allegations that U.S. President Biden pays to promote atheism worldwide (JR1), the alleged involvement of NATO in Sri Lanka's current financial crisis (JC1), and Ukraine "becoming the new Israel" (JL1). JP2 felt that a story on Russia and China announcing a new global reserve currency preyed on fears held by "evangelical Christian circles" and people on the far-right end of the American political spectrum.

4.4 Political Alignment

End Users. The political alignment of a news source was taken into account by 17 of the 23 politically diverse end users (74%), who found it especially relevant if a source was perceived to be affiliated with the political right (AFD3, GN1, GN2, GN4, SPD1, SPD3, SPD4, SPD6) — in particular, the far-right AfD (CDU1, CDU4, FDP1, GN1, GN4, LT1, LT2, SPD6). End users also commented on a news site's alignment with the political left (GN1, LT4, SPD3, SPD1, SPD6) and with political parties in general (CDU4, GN4). Other topics related to political alignment included the relationship with and portrayal of Russia (AFD1, AFD3, CDU5, CDU7, FDP1, SPD3); SPD3, for example, criticized one of the sources for blindly supporting Russian politics, and CDU7 found ads for any party problematic. More generally, end users highlighted the importance of a website's neutrality (CDU4, CDU7),

looked at who was sharing a source — with CDU1 noting that a certain source was shared by people critical of immigration — and considered the political alignment of those cited or interviewed within the articles (SPD3).

Experts. Seventeen of the 20 experts (80%) used the political alignment of a source to determine its reliability. JR1 argued that news websites should be about disseminating information free of bias, slant, and political mission. JL1 argued that a good journalist is a disinterested third-party observer, elaborating "The journalists that are far-right: not journalists! The journalists that are far-left: not journalists. The journalists who always try to be in the middle: may not be journalists either." JP5 negatively perceived one of the sources as having an agenda and publishing opinionated content. The perceived political alignment included specific groups like the political right-wing or far-right (JP1, JP2, JP3, JP4, JP7, JL1, JL2, JL3, JL5, JR2), conservatives (JP2, JP5, JL3, JL4, JR1, JR3), Islamophobes (JP1, JP2, JP3, JP7, JL1, JC3, JR1), antisemites (JP3, JL2, JL5, JL6), and pro-Russians (JP3, JP6, JL5). Experts considered whether sources positioned themselves against progressives (JP4, JP5, JP6, JL1, JL4, JL5, JR1, JR2) or the left (JP1, JR2, JP7, JC3), and referred to fringe content or extremism in general (JP3) or to other groups (JP2, JL1, JL5). JR1 saw political alignment as problematic, saying newspapers should not have a political position. Experts also viewed patriotism and statements that a source is "100% pro-USA" as problematic (JP4, JP6). JL2, however, stated that just because a source has a different ideological perspective does not mean it is not useful or is automatically unreliable. JR3 distinguished between sources that want to give people reliable news and sources that support certain political causes, but how she made this distinction was not captured. Experts also discussed how the ideological lens of a website shapes its motivations (JL2, JL3), with JL2 finding that one of the sources was not even pretending to present non-biased news but was instead responding to the news of the day with a very specific ideological lens.

Experts described a number of ways in which they determined the alignment of websites and authors. The Southern Poverty Law Center, for example, an organization that monitors extremism, informed the perception of a website's political affiliation (JP1, JP3, JL5). Experts also recognized political alignment through tropes they associated with political parties or partisans (JP1, JP4). Experts recognized a right-wing political bent (JP1, JP2), e.g., in stories about crime in big cities with progressive mayors, considered a familiar tactic by right-wing media (JP1).

4.5 Writing Style

End Users. Eighteen of the 23 end users (78%) used the writing style of a source as a criterion to judge reliability, describing wording as "lurid" (AFD3, FDP1, SPD2, SPD3, SPD4, SPD6), "populist" (CDU4, CDU5, LT1), "sensationalist" (CDU5, CDU6, CDU7, GN3, SPD4), and "propaganda" (GN4, LT1). They also perceived the websites as "polemic" (CDU2), "manipulative" (GN3), "heretic" (CDU5), "clickbait" (SPD2), and "right-wing agitation" (GN4). AFD3 and GN4 argued that a headline's writing style can sometimes be sufficient evidence of "fake news" or political bias, citing examples such as "Media go to war against Russia" (AFD3, CDU2) and "Compulsory masks are child abuse" (GN2, LT1). For GN4, the keywords used by the political right would lead him to judge a source as "very unreliable," while SPD5 also took into account the terminology's precision and subtlety.

GN3 commented on specific techniques like exaggeration, talking down, and trivialization. AFD3 criticized allusions like "Russia's Red Lines and Who's Really Escalating." AFD3 and CDU2 commented on the practice of posing questions, with CDU2 citing the question "Is this what modern synagogues look like?" as an example that made her cautious.

For SPD4, certain words reminded him of other far-right conspiracy websites. GN3 said the term "lust for war" alone stopped her from reading. Similar examples included "vaccination enthusiast" (GN1, GN4), "lab rats" (GN1), "war-mongering" (SPD4), and "attitude journalism" (GN4). LT2 highlighted the term "genocide of Russians" as biased. CDU7 criticized terms like "resistance to [the vaccination] terror," which she considered factually wrong. Several end users compared the wording of the websites to a commonly criticized German tabloid paper (GN3,

SPD4, CDU6, CDU7). LT1 found the wording of one source so populist that it did not allow an objective political debate.

Experts. Eleven of the 20 experts (55%) used writing style as a criterion. They took tone into account (JP2), evaluated the language's objectivity and neutrality (JP1, JR1), and criticized the writing for not being clear (JP6, JC2), e.g., when a source did not indicate whether authors spoke to a source or were citing a secondary source (JC2). JP6 criticized a source when language dehumanized and villanized people, and JC1 looked at whether the writing went straight to the point or was "rambling" (JC1).

Experts described the language of the reviewed articles as "distorted" (JR4), "inflammatory" (JP1), "biased" (JC3), "provocative" (JC2), "hyperbolic" (JR2, JR4), "overwrought" (JR4), and "clickbait" (JC2). JR4 saw the hyperbole as enforcing a certain point of view. Experts also negatively commented on headlines they perceived to have an ideological slant (JL2, JR3). For JL6, the editorializing of the headlines was the reason he considered one source unreliable. JP2 described the tone of one website as right-leaning. Experts criticized the use of adjectives, e.g., in "delusional Biden" (JC1), and of superlatives, e.g., the allegations that U.S. President Biden "destroyed the greatest military in the world" (JP1, JP2, JL3) or that somebody made the "worst pro-abortion argument ever" (JP5).

Experts criticized statements that were not objective (JP4, JP6, JR2), and criticized the source if a headline did not tell the story (JR3). They also commented on punctuation (JC1, JR3) — e.g., the use of question marks (JR3) and exclamation marks (JL6, JC1) — and on capitalization (JR1) and ellipses (JC1), which were seen as diverging from standard journalitic practices (JC1). JC1 argued that exclamation marks should be avoided in journalism because they "imply an emotion." JP2 interpreted the headline "We did it, Joe" as a snarky statement that held U.S. President Biden responsible for the recession. The wording of the headline led her to believe that the source was biased.

The experts also highlighted vocabulary (JP4, JR2, JR3), connecting unreliability to certain terms like "progressive fascists" (JP1, JP6, JR2), "disloyal military" (JP1, JR2), "tyranny of woke intentions" (JR1), "great reset" (JR2), and "Zio-globalist handlers" (JP1), and to neologisms like "Fauci-baric" (JR1). Experts perceived terms such as "manifesto" as "loaded" and "politicized" (JP4), and were also triggered by terms such as "Freemasonry" (JR2), "fascist" (JP6), and "racist" (JP6). Experts noted tropes of political parties and political partisans (JP4) or conspiracies (GN1). JR1 criticized how one source referred to paid advertisements as "the Information," attempting to cover up that the article was actually a paid advertisement.

4.6 Authors

End Users. Only 9 of 23 end users (39%) took the authors of articles on a news website into account. These end users criticized websites on which authorship was not clear (CDU6) or information about the author was hard to find (SPD4). CDU3 connected the question of who is writing to their motivations. LT1 looked at whether multiple journalists were responsible for the content (LT1), and noted whether a journalist's background and photo were provided (LT1). End users also took into account whether an author worked for other, more widely known publications (LT1), and considered whether they were familiar with certain authors (LT1), who the guest authors were (SPD6), and whether freelancers were employed (CDU3). In addition, they viewed authors favorably if readers could provide input and leads via email (SPD6).

Experts. The author of a news story was the criterion used by most experts. Eighteen of the 20 experts (90%) used it to judge a source's reliability. The byline was seen as crucial (JP1, JP2, JP4, JP6, JP7, JL1, JC2, JC3), with experts using information about the authors to understand where they had worked (JP1, JL3, JC2), their "beat" (JC2), and whether they were professional writers (JC2). For JL1, it was important to "know why [the authors] are saying what they're saying." Experts also thought it important to know the site's editors and owners (JR4) and how the authors were funded (JL4). They examined whether an author worked for a "known organization," whether the work was funded by a known organization (JC1), and whether an author was affiliated with a university. They

perceived an article as reliable if more than one author was involved (JP2, JC3) and criticized sources centered around one personality (JL5, JR1). Experts were put off by the profile of one of the authors (JP6, JC3), because it showed a "random picture of a baby smoking a cigarette."

Experts criticized a source if no author information (e.g., a biography) was available (JP6, JC2, JC3) and if the author's qualifications were not presented at the top or bottom of a page (JP1). JP2 found it important to be able to verify that somebody is a real person with an education and background in journalism. JL1 appreciated knowing where articles were curated from. Articles written anonymously or under a pseudonym were seen as problematic (JP1, JP2, JP4, JP6, JP7, JL4, JL6, JC2, JC3) because it was difficult to judge the credibility of an individual author and their previous record (JP1).

Experts consulted other sources to understand who the authors were (JP1, JL1, JL3, JC1), with one using the "About" page (JP4), while another consulted the Wikipedia article about an author (JP1), also noting whether the Wikipedia article was "comprehensive" and "well-cited" (JP1). Experts also considered how institutions they trusted — like the Southern Poverty Law Center — described the authors (JP1), and whether the content posted by an author was a repost from another website, e.g., the author's personal blog (JP1). Other criteria included how frequently something was reposted (JP1) and whether the authors had professional-looking email addresses from the website for which they were writing (JC3).

4.7 Self-Description

End Users. Twelve of 23 end users (52%) evaluated a source's self-description in determining reliability, with SPD6 noting that it is worth reading self-descriptions to understand what a website describes as its mission. Some found it problematic when sources described themselves as "critical" (CDU6, FDP1, SPD2), "dissenting" (CDU6), or "independent media" (CDU1, GN1), and perceived slogans like "Truth and Tradition" as dubious and untrustworthy (FDP1, GN4, LT1, SPD2). For FDP1 and GN4, explicit statements that a source was telling the truth made them skeptical. The slogan "The critical website" was interpreted as implying "The others are not as critical as we are. We claim to be more critical." (CDU2), which was seen as problematic (SPD2). GN1 and GN2 commented on the slogan "For those who still have their own thoughts;" for GN1, it insinuated that the source's authors and readership were the only ones who still thought freely, a self-description that made him very skeptical. GN4 criticized sources that discredited others to make themselves appear better, and end users also commented on the name of a source itself (AFD2, CDU6).

Experts. Fifteen of 20 experts (75%) used a website's self-description to determine reliability, seeing it as a way to understand editorial processes (JP6, JP7, JL4, JR1, JR3), the editorial line or mission (JP4, JL4, JL6, JC2, JR1, JR3), and the website's ownership and independence (JL3, JL4, JL6, JC3). Experts used the description to understand the goals and motives behind the site (JL3, JR3), "who these people are" (JP4, JP7), what the website stood for (JR3), and whether it had a political alignment (JL6). JR1 felt that a mission statement should be about delivering news, JP4 expected a commitment to journalistic objectivity, and JR1 argued that the mission statement should be free of bias, slant, and a political mission. JP4 and JC2 noted which corporate websites and charity groups were mentioned. Experts also commented on the opinions and perspectives they found in the self-description (JC3, JR2), using them to understand a source's target group (JL3, JR3). Statements deemed by experts as not objective frequently related to certain political groups, like the political left (JC3, JR2) and Islam (JC3, JR1, JR2), as well as to claims that a source was "100% pro-USA" (JP4, JP6) or defending "free society" (JR1).

Unlike end users, experts rarely commented on the name of a source. Some commented on slogans (JL3, JL5, JR1), like those saying a source was serving the clandestine community or veterans (JL3), or on a banner that said that "inside every progressive is a totalitarian screaming to get out" (JP5, JL4, JL5, JR1, JR2).

Surprisingly, with the exception of JL3, JL6, and JC3, experts rarely commented on the reliability of the self-disclosed information. JL3 stated that "there's no way for me to know that for sure," JC3 commented "I guess I

don't know if it's real," and JL6 believed that even when the "About" page includes self-flattering things about a source, it can provide "some insight as to how they perceive themselves."

4.8 Professional Standards for Journalists

End Users. Eleven of the 23 end users (48%) used whether a source followed professional standards as a criterion in judging reliability, considering how "independent" (GN1, LT1), "objective" (GN1), or "critical" (FDP1) the website's perspective was. They also examined whether multiple sides of a story were discussed (FDP1, GN2). SPD5, for instance, highlighted the importance of a "journalistic quality of the preparation and presentation of content." SPD4 argued that one website he reviewed felt like a blog, and found this problematic because he believed that certain journalistic and self-regulatory aspects do not apply to blogs. In relation to this criterion, end users took into account whether they found something to be a "journalistic contribution" (GN1) and whether a website was perceived as hiding or selling opinions in the reporting (GN2, SPD5). For LT1, this meant "how close" the writers of a website were to the subjects. SPD6 also considered whether a source was a member of journalistic associations which subscribe to certain ethical professional standards.

Experts. Thirteen of the 20 experts (75%) explicitly commented on professional standards, the neutrality of a news source, and the clear distinction between opinions and facts (JP1, JP4, JP7, JL1, JL2, JC2, JR1). They criticized the reflection of a journalist's personal opinions or political ideology in the writing (JP2, JP4, JL1, JC3). Experts also commented on the bias of one source they reviewed (JL4, JC3). JL1 believed that a good reporter assesses information individually and does not take sides, and a reliable source should provide news without a viewpoint. JP7 believed that journalism should show a set of facts. JP4 named striving for journalistic objectivity as the "normal standard." For JP7, it was a red flag if only one side of the political spectrum was represented and if the content was portrayed through a political lens; he perceived one of the sources as adding a spin to the content. JC2 argued that a website has to indicate if it provides opinions only for a particular segment of the political spectrum. He saw claims of patriotism or pro-USA sentiment as indications of not being critical enough of the U.S. government (JP4).

JL1 highlighted the importance of factual, vetted information free of opinions and preconceived notions, and stressed that opinions should be clearly labeled. It was important to experts like JL1 and JR1 that knowledgeable copy editors check the writing. JL1 argued that, when gathering information, it is important to always attack it "fresh." As somebody working in the White House, he perceived his job as giving "every president shit" and questioning power. In his personal experience, he found that "because I went after Trump, I was [perceived as] far-left — and now, because I go after Biden, I'm [perceived as] far-right."

One news source wrote that editors and writers were not employees or contractors of the site and were not paid; this was seen as a red flag by a number of professional experts (JP7, JL4, JC2, JR1, JR3), who linked it to a lack of editorial control and vetting (JP7, JC2, JR1). Experts found it problematic that there were no editors checking content and claims in the articles for veracity, and no copyediting on the writing itself (JC2, JR3, JP7, JL4).

4.9 Advertisements

End Users. The criterion that end users used most frequently to assess a source's reliability was advertising, with 20 of the 23 end users (87%) referring to ads when explaining their reasoning. They commented frequently on the number (CDU3, CDU4, FDP1, SPD4, SPD6), content (CDU6, FDP1, GN3, SPD4), presentation (FDP1, GN3, SPD3), and size (CDU4, SPD4) of ads, and on apparent disconnects between a site's content and its ads (FDP1, GN3). Ads were considered "dubious" (CDU1, CDU6), e.g., if they featured diet pills (CDU6) or "tricks" to repair eyesight (FDP1). CDU3 perceived a large number of ads as indicative of financial dependence and political bias. End users

also commented on how ads were integrated into a website's design (CDU4, FDP1). FDP1 criticized how ads were not labeled as such, and layouts that made it hard to distinguish between content and ads.

End users considered whose ads were displayed (CDU1, FDP1, GN1, GN4, LT2). One of the websites evaluated included ads for a book by politicians from the far-right AfD party; this was criticized by end users from across the political spectrum (CDU1, CDU4, CDU7, FDP1, GN1, GN2, GN4, LT2, SPD3). CDU7 believed that such ads limited the source's neutrality, saying she would also criticize an ad if the book was from her party. For GN1, ads for the AfD gave him an "absolutely clear" impression of the source because he has a clear opinion of the AfD as a party. LT1, too, argued that since the AfD is not a "respectable" party, a website advertising a book related to the AfD is not "respectable." For CDU1 and LT2, the AfD ad meant that the website was "obviously far-right." For them, a populist ad was reason alone to leave a site and an indication that a source was related to the political right. The same was true for far-right publishers (GN2, LT1).

Experts. Only 4 of the 20 experts (20%) referred to advertisements when making their assessments. The experts required websites to clearly distinguish between ads and content (JP7, JR1), criticizing sites if an ad was the first thing shown (JL1) and if the ad looked like an article (JC3). The appearance of one ad was seen as an indication of a "typical unreliable website" (JL1). Another important criterion for experts was whether major ad networks cooperated with a website (JC3, JR1). JC3 argued that "if it gets to the point where the ad networks are like — we don't want to work with you — that kind of raises a red flag." JR1 noticed that one of the sources described itself as paid advertisers who publish favorable information, and saw this as unreliable.

4.10 Ownership

End Users. A website's ownership was used as a criterion by 13 of the 23 end users (57%), who asked questions such as "Who is behind it or who owns it?" (AFD1, CDU5, GN2, SPD4, SPD6), "What interests are behind it?" (CDU5), and "Who finances it?" (CDU5). SPD6 looked at what type of company ran the website, whether the company's goal was to make a profit, and whether the website was financed through ads or subscriptions or was fee-financed under public law. SPD2 considered whether a publishing company was behind a site and whether she saw this company as trustworthy. End users like CDU6 stated that they regularly look at the *Impressum* of a website. In Germany and other German-speaking countries, the *Impressum* is a legally mandated statement of a document's ownership and authorship [81]. This information was often challenging to find (CDU5, CDU6, LT1, SPD4), a fact criticized by CDU5. End users also took into account whether they were familiar with the website's publisher (SPD3).

Experts. Ten of the 20 experts (50%) took ownership into account when assessing a website's reliability, finding it helpful to understand whether a website was funded by a known organization (JC1) and whether that organization had an agenda (JL4). They wanted to understand the owner, their motivations, and their prior work (JL3, JR1). In one source, the owner was clearly named, and this was noted by several experts (JP1, JP3, JL3, JL4, JR1, JR3). JP3 had professional interactions with this owner, and did not hold him in high regard. Experts researched the website owners on Wikipedia, Google, and Twitter (JP1, JP3, JR3, JP4, JL4), taking into account what was reported by institutions or sources they trusted (JP1) and who the owner associated with on Twitter (JR3).

One important criterion was how a source was financed (JR1). A website claiming to be independent was perceived positively because "they are not owned by any billionaire" (JL3). However, not taking money from outside media sources, the government, and sponsors was also seen as unreliable by one of the experts, JR3, who trusted traditional media partners, and also mentioned the board of an organization. JP4 said he would try to access IRS forms to see who the funders and officers were, and how the money was spent.

4.11 Sources

End Users. Ten of the 23 end users (43%) noted the sources that were cited and linked and the type of supporting evidence provided. End users believed that reliable sources should transparently document and cite evidence for reports, studies, and comments (FDP1, LT2). Examples of reliable sources mentioned by end users included public broadcasting, newspapers of record, and public authorities. SPD2 highlighted the importance of scientific findings, e.g., regarding the COVID-19 pandemic. Other criteria included whether more than one source was cited and the critical engagement with these sources. Some end users also appreciated the citation of sources they considered reliable (LT1, SPD2).

Experts. Twelve of the 20 experts (60%) referred to the sources used by a website when evaluating its reliability. They highlighted the importance of original reporting (JP5, JP7, JL4) — in particular, first-hand accounts (JP3, JP5, JL1), interviews, and direct quotes (JP1, JP3, JL1, JL3), especially with named sources (JL1, JC1). JL1, for instance, argued that the only way to determine if a news story is real is to witness the event or to talk to somebody who witnessed it. Experts also mentioned direct consultations with domain experts (JP1, JP2, JL1, JC3) and the use of wire services (JP1, JP5, JL1, JR3) and other newspapers (JP3, JP7, JL1, JR3), especially those which are reliable or centrist (JP3). They considered where a source obtained its material (JR1) and how the information was vetted (JL1). When vetting sources, the experts considered who the sources were (JL4) and what connections the sources had, e.g., whether they were affiliated with a government or a certain "oligarch" (JP5). The experts referred to authorities like the government or the U.S. Federal Reserve System as reliable sources (JP1, JP5, JR3). Regarding first-hand accounts, JP5 highlighted the importance of news organizations having their own photographers. JL1 argued that when anonymous sources are used, the information needs to be confirmed independently. JL5, who worked as a professional fact-checker, said that she always got experts on the phone to verify the information in a story and trace the provenance of claims. Considering the breadth of one source's reporting, JP2 wanted to know whether the website actually had reporters in all the countries they were covering or whether they pulled their information from other outlets. JP5 warned that few organizations can afford global reporting with people on the ground and experts on all topics.

4.12 Reputation

End Users. Eleven of the 23 end users (48%) took the reputation of a source into account. They also commented on whether they already knew a source (CDU3, SPD6), and relied on their experience doing research for their work as members of parliament (SPD2, SPD4). CDU3 said he is "extremely skeptical" about sources he does not know. End users also referred to the reputation of the sources and took their personal experience with a source into account. SPD2 argued that the more abstruse a story is and the less familiar she is with a source, the less credible she considers the source. SPD6 mentioned witnessing "an extensive discussion about [one of the sources] in the social left," using this to judge the sources. GN1 knew one of the sources because they wrote a "stupid" story about him that resulted in a "shitstorm" because other far-right websites picked up the story.

Many end users (CDU5, GN2, GN4, SPD3) used search engines like Google to research news sources, while GN2 and GN4 used Wikipedia to inform themselves about a website's reputation. More generally, GN1 used other websites and references that he trusts and perceives as neutral and objective, GN4 said he would search Twitter to understand what people he knows and respects write about a source, and SPD2 argued that she would ask many people, including colleagues, experts, journalists, and the State Media Authority, before making her assessment. AFD2 and AFD3 said they would use their personal networks.

Experts. Eleven of the 20 experts (55%) considered the reputation of a website and its authors. JC2 argued that "somebody or something has to tell you that it is a credible website." JL3 considered the reputation of the sources used by a website. For JR1, the fact that ad networks refused to work with one of the sources was "setting off

alarms." For the economic context, JC2 argued that if the writer had worked as a financial analyst for a large bank, she would have trusted him.

Experts also considered what they had heard about a source (JP4, JL1, JL4). If they knew the author, they took into account their opinion of that author and what they knew about the author's reporting and political bent (JP4, JL4, JL6, JR2). They also looked at whether an author was well-established and knowledgeable (JP1). One author was described as unreliable and politically motivated by JR2, and as "intellectually dishonest" by JP4. Despite sharing a political alignment, JR2 said he "would never consider" the author a reliable source. JP3 knew one website author, having previously interviewed him. Based on this professional interaction, he felt "very uncomfortable" about using him in a story he wrote in the past. JR2 knew that information from one source had been overstated or falsely stated in the past, and found this problematic even though the website matched his understanding of the current economic market at the time of the study.

Affiliation with reputable sources was seen by experts as an indicator of reliability (JP1, JC2). They also considered what others, especially peers, thought about a source (JL4, JL5, JR1, JR3). JL3 referred to the kinds of people who share an article as an indicator of reliability, and JR1 noted the perceived reliability of the peers and the sources in which they published. Also considered was how many followers an author had (JR3), whether experts and experts in their personal social network followed, quoted, or shared content by the author (JR3), whether a site was banned by Facebook (JP3), and whether unreliable websites were presented in the related search tab on Google Search (JL5).

The experts mentioned several ways of learning more about the reputation of a website (JP1, JL1, JL4, JL5), e.g., using the name of the website and the term "bias" as keywords on Google Search (JP1) and consulting Wikipedia (JP1, JP3, JL5). They also took into account what was published about a source by reliable or centrist mainstream sources like the New York Times (JP1, JP3, JL1, JL5) and by institutions like the Southern Poverty Law Center (JP3, JL5).

4.13 Design of Website

End Users. Twelve of the 23 end users (52%) used a website's design and structure as criteria for evaluating its reliability, noting whether a site was cluttered and well-structured (CDU4, CDU6, SPD4) and if the site and design "looked reputable" (CDU3, CDU6). SPD4 explained that, while fringe websites used to look like "bad PowerPoint presentations," today many look more professional. End users noted a site looking "wild" or "turbulent" (CDU6, FDP1) or resembling a blog (LT1, SPD4). Other design-related criteria included whether a source was "colorful" (CDU6, FDP1), the fonts used (SPD4), consistency (FDP1), the number of buttons (FDP1), and the overall impression given by a site (FDP1). CDU3 said that he "immediately saw" whether the first two sources were reliable. FDP1 and SPD2 considered the similarity of the design to that of the largest tabloid paper in Germany, and SPD3 noted how similar the design was to that of other unreliable sources.

Experts. Far fewer professional experts commented on a website's design, with only 6 of the 20 (30%) using design as a criterion. Those who did noted layout (JP5, JP7, JC1, JC3), formatting (JR1), fonts (JC1), and relative size of images (JC3) as indicators. JC1 also considered similarity of the design to that of reliable websites like the New York Times. JC1 noted whether a website looked "messy," "clean," or "serious" (JC1), and JP3 criticized one website for looking like a personal blog.

4.14 Differences Between Experts and Non-Experts (RQ3)

An exploration of the differences between experts and non-experts is important because it 1) helps us identify criteria used by people with experience in the production of news, 2) enables us to identify criteria within end users' abilities that are well-founded, and 3) allows us to debunk common choices of end users.

Figure 1 indicates the proportion of end users and experts who used a criterion to determine whether a purported news website was reliable. Overall, we found strong similarities between the two groups, even though they reviewed different news sources in different languages and came from different countries and political backgrounds. We identified four criteria with little difference between end users and experts: political alignment (6% difference), ownership (7%), content (7%), and reputation (7%). Note that content and political alignment are the two most frequently used criteria; we see this as corroboration of the ecological validity of the criteria, as these criteria are frequently invoked by experienced news producers and are within end users' abilities.

Other criteria were considered helpful by one group but not the other. The top five criteria for end users were advertisements, content, writing style, political alignment, and ownership, while the top five for experts were authors, content, political alignment, self-description, and professional standards. For two criteria the difference between the two groups was greater than 50%. The criterion with the most considerable difference was advertising (67%). While almost 20 of the 23 end users (87%) took advertising into account, only 4 of the 20 experts (20%) did so. For the author criterion, with a 51% difference, only 9 of the 23 end users (39%) referred to the authors, compared to 18 of the 20 experts (90%). We believe that a likely explanation for this difference is that experts as journalists have experience being authors, with a better understanding of authorship and the influence held by the author of a news piece. End users who have never worked as journalists may be less familiar with how news websites are produced and may not fully realize this impact.

5 DISCUSSION

In this paper, we show that the overwhelming majority of end users and experts can identify unreliable websites (RQ1), i.e., the ratings they provide align with what we and Gruppi et al. [24] perceive as unreliable websites. This agreement is noteworthy, since end users and experts had different skill levels and were from different countries with different media landscapes. We also learned that the unreliable websites we used in this study were rarely visited by participants. We identified eleven reliability criteria to assess whether a website is reliable (RQ2) and found differences in the applicability of the criteria between politically diverse end users and experts (RQ3).

We acknowledge concerns voiced by scholars such as danah boyd that media literacy lessons may "backfire" and that even well-executed lessons on critical thinking can lead to "ask[ing] people to doubt what they see," which can make them doubt both unreliable and reliable sources [8]. Informed by these concerns, we focused on reliability criteria beyond simple heuristics that can be easily manipulated. Like boyd, we do not view media literacy as a panacea. Criteria such as the kind of content displayed on a website could be used to teach users what they should pay close attention to when evaluating the reliability of news websites.

The criteria presented in this paper can support the type of critical thinking and evaluation that Wineburg and McGrew refer to as lateral reading [84]. Our work corroborates and extends their lateral reading technique of scanning a website and consulting other websites [84] whose effectiveness is empirically shown [83]. Criteria such as political alignment and authorship can help users assess a source's political alignment and research an author's background.

Table 3 indicates which criteria are partially covered by prior work. In this discussion, we compare our criteria to those seen in prior work, introduce an analytical distinction between manipulable and less manipulable criteria, and describe how the less manipulable criteria can be augmented through technology.

5.1 Comparison to Prior Work

An important motivation for this work is to provide an empirical basis for reliability criteria. Table 3 shows which of the criteria we observed are also mentioned in prior work. Unlike prior work, which based these criteria on surveys [7] or discussions [74, 88], or which do not explain how the criteria were determined [9, 26, 45, 71], our criteria are based on a user study employing think aloud methods and semi-structured interviews. These criteria

Table 3. The eleven reliability criteria for news websites that we observed in practice based on contextual inquiry, think aloud study, and semi-structured interviews. We compare this to prior work based on surveys, discussions with stakeholders who are not the end users, or an unknown empirical basis.

Reliability Criteria based on:	Survey [7]	Expert Discussion [88]	[74]	Not Explained [9]	[26]	[45]	[71]
01. Content	✓	✓			✓	✓	
02. Political Alignment							
03. Writing Style	✓	✓		✓		✓	
04. Authors						✓	✓
05. Self-Description							✓
06. Professional Standards	✓			✓		✓	✓
07. Advertisements		✓				✓	
08. Ownership						✓	
09. Sources	✓	✓			✓	✓	✓
10. Reputation	✓						
11. Design of Website	✓			✓	✓		✓
Other Criteria	✓	✓	✓	✓	✓	✓	✓

may be closer to what people use in practice. Compared to prior work, we recruited a larger and more diverse sample, including experts and non-experts and controlling for participants' political alignment.

Table 3 shows that some of the criteria we identified in practice are frequently covered in prior work, while others are rarely included. Frequently included criteria are content, writing style, professional standards, sources, and website design. Criteria covered by only one or two other lists include authors, self-description, advertisements, ownership, and reputation. We are the first to describe political alignment, which is surprising since it is one of the most frequently used criteria in our investigation. The largest overlap is between our criteria and those of NewsGuard [45], covering 7 of the 11 criteria we identified. Six of the 11 were also covered by Bhuiyan et al. Our empirical study showed that criteria like writing style and website design, which are frequently included in criteria catalogs, are more frequently used by end users than by experts.

Another noteworthy finding is how few of our criteria are covered by those proposed by the Credible Web Community Group [74]. Their criteria focus on awards, correction policies, and how old a website is. In our study, neither end users nor experts used these criteria during their review of different websites.

Our comparison of end users and experts also showed that some criteria, like authors, self-description, and professional standards, are used frequently, especially by experts, but are rarely found in other compilations of reliability criteria. There is also no other work that covers all criteria.

5.2 Criteria Where Manipulation Is Less Difficult

The reliability criteria presented as results in Section 4 are based on extensive empirical work on how purported news websites are evaluated in practice. Here we reflect on the generalizability of these criteria and discuss how resilient they might be against malicious actors. This is particularly important because adversaries will likely try to use the criteria for malicious intent. Before integrating reliability criteria into online platforms and other socio-technical systems aimed at supporting users, it is, therefore, crucial to critically examine these criteria. For this discussion, we distinguish between manipulable and less manipulable criteria. While we are aware that there may be exceptions for the less manipulable criteria, based on our experience from 43 semi-structured interviews with 23 end users from Germany and 20 experts from the United States, we think that our classification can help

guide efforts to support users. We recommend that the criteria for which manipulation is less difficult should not be taught to laypeople and should not be integrated into socio-technical tools, at least not without additional precautions to overcome manipulation risks.

The theoretical embedding of our classification is signaling theory [69, 87], which is aimed at explaining why certain signals are reliable or not. Following Donath [18], signaling theory distinguishes assessment signals that are "inherently reliable" and conventional signals, where "the link between signal and quality is arbitrary, a matter of social convention." An example of an assessment signal is the relationship between observing somebody lifting a heavy object and physical strength. An example of a conventional signal is self-descriptions in online social networks, where users can easily pretend to be older or younger.

Informed by the theoretical framework of signaling theory, we consider less manipulable criteria as assessment signals and more manipulable criteria as conventional signals. Like Donath [17], we recognize that very few signals are impossible to manipulate, given sufficient motivation. Nevertheless, introducing a distinction between manipulable criteria, which are comparatively easily manipulated, and less manipulable criteria, which are less difficult to manipulate, can help practitioners and others.

Following an approach from cybersecurity [49], we asked ourselves "What is the worst an attacker can do to you?" and brainstormed how somebody operating an unreliable website might apply the criteria to appear more reliable. In this section, we present the criteria for which we identified ways in which the criteria can be manipulated.

The writing style of a website was used by both end users and experts to determine a source's reliability. While research indicates that writing style does affect the success of some misinformation and whether content has a strong chance of going viral [31, 57], wording and writing style — e.g., avoiding certain keywords and changing headlines — do not affect the veracity of the content. Therefore, writing style does not help people confidently assess whether a website is reliable.

Other criteria that can be easily manipulated include a website's self-description and information about its ownership. Following signaling theory, these are conventional signals in which the link between reality and what is presented about self-image or ownership is a mere social convention [18]. While end users and experts used both criteria, these are criteria that can be easily manipulated. In our study, the claims made in a website's self-description and information on alleged ownership were also rarely verified by participants.

The advertisements shown on a website were another criterion used, especially by end users. There could, however, be many unreliable websites that do not feature problematic advertisements, especially if those who run sites do not do so for personal gain, e.g., to push a political agenda. While forging this criterion would come at a cost for unreliable websites that are for-profit, it is too easily manipulated. Website design is also easily manipulated. With the availability of website builders and templates, it is comparatively easy to set up professional-looking websites, even for people without experience with HTML and CSS. This criterion should therefore not be included in guidelines.

As illustrated by Figure 1, the criteria for which manipulation is less difficult were those predominantly used by end users, including writing style, advertisements, and website design. One explanation for this could be users' limited understanding of news production processes. In their lived experience as end users, they may have experienced that reliable news is commonly presented in a professional writing style, and may also assume that advertisements and website design are difficult to manipulate, though this has become very easy.

5.3 Criteria Where Manipulation Is Difficult

Based on our brainstorming of ways to deceive users, we identify criteria for which manipulation is difficult, and invite others to join us in validating this empirically. The first such criterion is the content covered on a website. As our empirical material showed, if a website spreads right-wing conspiracy theories or antisemitic content, this

Table 4. The eleven reliability criteria for news websites observed in practice via contextual inquiry, a think aloud study, and semi-structured interviews. We highlight which criteria we consider challenging to manipulate based on our analysis.

Reliability Criteria	Manipulation is difficult
01. Content	✓
02. Political Alignment	✓
03. Writing Style	
04. Authors	✓
05. Self-Description	
06. Professional Standards	✓
07. Advertisements	
08. Ownership	✓
09. Sources	✓
10. Reputation	✓
11. Design of Website	

is hard evidence that a website is unreliable. Another aspect of the content criterion is whether only one side of an issue is represented. If an unreliable website tries to optimize for this criterion, e.g., by providing a balanced account that considers all sides of a discussion, this would improve the website and decrease its unreliability.

Another less manipulable criterion is a source's political alignment. Akin to content, the featuring of only certain political actors or ideas by a news website is hard evidence that the site is unreliable. A particularly useful and especially hard-to-forge aspect of political alignment is who is sharing links to a website on social media platforms like X or Mastodon, or on any other websites where people post text and links. Since this is beyond the control of a website, it is difficult to manipulate.

Which authors write for a source is another criterion that is difficult to manipulate. With available search engines, it is comparatively easy to research whether authors exist and have received proper training, and whether the content and opinions shown on a website are consistent with the author's prior work.

Another important criterion that is difficult to manipulate is whether a website adheres to professional standards. This can be assessed by seeing if a website clearly distinguishes between opinions and facts and presents different sides of a story. As our investigation showed, this can be done based on a topic that a user is already familiar with.

The sources named by a website can be easily checked, allowing the information provided — especially statistics — to be easily verified. And the sources used directly connect to the criterion of reputation, which is also difficult to manipulate. A reliable third party like Wikipedia or Snopes voting to classify a source as unreliable can be considered hard evidence of that unreliability, as it would be difficult to manipulate such third-party assessments.

Figure 1 indicates that the most frequently used criteria, content and political alignment, are used equally frequently by both end users and experts. There are three criteria that are difficult to manipulate and more frequently used by experts: authors, professional standards, and sources. One possible explanation as to why these three are used more frequently is the lived experienced of experts.

5.4 Implementing the Reliability Criteria for News Websites

Table 4 shows that content, political alignment, authors, professional standards, sources, and reputation are the six criteria that, based on signaling theory and on our theoretical insights on how to manipulate these criteria, we consider to be most helpful in that they are a) difficult to manipulate in practice, b) used by experts, and c) applicable by end users. We therefore think that these criteria should be evaluated further. After successful empirical evaluations, they could be taught widely in media literacy and media education contexts.

The following section describes how HCI can help users assess these criteria. Our goal is to augment users' ability to recognize unreliable online information. The technical proposals presented in this section, however, are only part of a larger socio-technical assemblage, and depend on users who are aware of the complexities of misinformation. For these users, they could become a powerful support.

Content. Assessing content and identifying how and why controversial topics are presented is challenging, but not without precedent. Media scholars like Puschmann et al. [59] have employed word lists to successfully identify popular conspiracy tropes and topics like antisemitism, anti-elitism, anti-immigration/Islamophobia, and anti-gender/anti-feminism. Such word lists, along with more data-driven approaches like topic modeling via Latent Dirichlet Allocation (LDA), could be used to build HCI tools that enable users to assess whether a source is covering content that is considered problematic. The word lists could also serve as a starting point for more in-depth qualitative investigations by fact-checkers.

Political Alignment. People are politically affiliated with a source if they regularly share it. Data mining of social media platforms could be used to locate different sources in the political spectrum. Politicians and what they share on X, Mastodon, or any other website could be tracked, along with the links shared by users and information on who users follow. These data mining approaches would allow people to identify sources with multipartisan approval consumed across different political parties. To reliably detect extreme content, the social media profiles of people with a reputation for sharing misinformation — e.g., Alex Jones in the U.S. or Attila Hildmann in Germany — could be included. Such HCI tools would make it comparatively easy for people to assess whether a website has multipartisan approval or whether it is only popular with a specific political group.

Authors. The author criterion could be supported by specialized search engines that make it easier to understand the authors of a website and their reputation. Thanks to social media, it has become harder to invent authors. Follower networks on platforms like LinkedIn, X, Mastodon, Instagram, or TikTok make it easier to investigate authors' connections and political affiliations.

Professional Standards. Specialized search engines could help users investigate whether a source has been criticized in the past for not adhering to professional standards. Software tools could also retrieve sections titled "Criticism" or "Controversy" in the Wikipedia article about a website.

Sources. The assessment of whether a news website provides sources for claims is another criterion that could be supported through technology. HCI researchers could build tools that leverage techniques like named entity recognition to identify the people and institutions cited in an article. These names could then be used to retrieve the authors' and institutions' social media profiles which, along with information from Wikipedia articles and similar sources, could be used to ensure that an article's statements are consistent with what the authors and institutions have said in the past. HCI tools could also apply more sophisticated approaches like argumentation mining to identify the key arguments in a text and assess their quality [67, 75]. If the key arguments are not accompanied by actors that support them, this can be highlighted as problematic. Such tools could be valuable for fact-checkers, and less experienced users could be trained in media literacy classes to use the tools.

Reputation. A data-driven approach could make it easier to assess the reputation of a website using existing resources. This approach could be similar to that of prior work that helps people fact-check claims [27]. Platforms like Wikipedia are interesting in this context because the community uses a process of deliberation to decide which sources to ban. Using the strategies of JP1 described in Section 4.12, users could also enter the name of a website and keywords like "bias" into a search engine to see whether other reliable sources have accused a website of being biased. More generally, GN4's strategy of searching social media websites to understand what kind of users post links to a news websites could be supported through HCI tools.

5.5 Political Alignment

As we have shown, political alignment has not been covered in prior work, though it was widely used in our study and it is comparatively difficult to manipulate for attackers who create misinformation. This lack of recognition is surprising, as it directly relates to why people create misinformation [77]. Political alignment connects to political influence, propaganda, and partisanship. We found that political alignment is frequently taken into account by both end users and experts when rating the reliability of news websites or other information sources. One explanation why political alignment is not included by commercial businesses like Facebook [25] and NewsGuard Criteria [45] is that it is in their business interest to appear neutral. Platforms like Facebook and X have repeatedly been criticized as being biased in the past. Asking users to take a news website's political alignment into account could, therefore, make platforms vulnerable to additional scrutiny and controversy. However, as Austrian-American communication theorist Paul Watzlawick famously argued, "One cannot not communicate." Inaction on the issue of political alignment, therefore, is also an action. By ignoring the influence of political alignment, Facebook, NewsGuard, and others could enable certain actors to push propaganda and gain political influence. We do acknowledge, however, that labeling political alignment is a challenging task related to the broader debate around democratic platforms and governance.

5.6 Limitations

The goal of our user study was to elicit criteria that can be used to determine whether a news website is reliable. The framing of our study represents a special setting in which users take time to pause and reflect on a website's reliability. As discussed, prior work indicates that misinformation is sometimes shared because people do not pay attention [51, 53]. Further work is needed to understand how representative our study setting is for users who encounter unreliable news websites in their everyday life. Regardless of the outcome of such studies, the news reliability criteria presented in this paper can be used to sensitize users to pay attention to online websites and train them to evaluate reliability.

Participants' views on news reliability are shaped by their countries, cultures, and the political systems in which they live. It is also important to note that notions of "left" and "right" are always relative to a specific country.

Due to the methodological complexity and time limitations when consulting many end users and experts, we could only examine three news websites. We thus focused on unreliable websites so as to maximize our findings' impact on helping people identify misinformation. We encourage future work that applies the methodology to reliable websites.

For end users, we recruited elected politicians of a state parliament so we could include the expertise and perspectives of people from all political alignments. While these end users were elected in free and open elections and, in theory, representative of the voting body, we found that they were highly educated. This level of education corresponds to prior work showing that highly educated people are overrepresented in legislative bodies around the world [19]. Still, the end users in our investigation were laypeople regarding how news is produced. Our results are, therefore, helpful in understanding end users without expertise in news production. That said, future work is needed to better understand the influence of education in this regard.

6 CONCLUSIONS

A politically informed citizenry is the backbone of a well-functioning democracy [15]. We believe that the news reliability criteria described in this paper are an important stepping stone toward a more informed citizenry. These criteria are empirically grounded in that end users and experts apply them in practice. Such criteria have three benefits: 1) solutions based on these criteria are scalable since the assessment only has to be done for

each news source, not each article, 2) the assessment of a news source can be done independently of individual claims and news articles, thus minimizing the risk of potential backfire effects, and 3) many criteria may also be applicable to individual news articles and other kinds of content such as videos. Another important benefit of news reliability criteria is that they empower users to make their own decisions. History has shown, numerous times, that censorship and media control are central instruments of tyranny and dictatorship. Therefore, the most critical design goal of news reliability criteria is freedom of speech. None of the criteria in this paper tell people what to think or do, and all of them can help people make better and more informed decisions about the content they encounter online. We make concrete technical proposals that can empower people to determine whether a news website is reliable. With our criteria, we believe that platforms like YouTube, TikTok, Facebook, X, Mastodon, NewsGuard, and others can improve their services. Unlike other approaches towards misinformation connected to the risk of censorship, the criteria presented in this paper provide people with strategies for distinguishing reliable from unreliable sources in a way independent of political ideology. We hope that social media platforms and providers of reliability criteria make these criteria available to their users.

ACKNOWLEDGMENTS

This research has been supported by the German Academic Exchange Service (DAAD), the German Research Foundation, Grant No. 374666841 (SFB 1342), and the NSF under Grant No. 2107391. We thank all participants for their time and their insights. We also thank the reviewers for their exceptionally helpful and constructive feedback.

REFERENCES

[1] John H Aldrich, John L Sullivan, and Eugene Borgida. 1989. Foreign affairs and issue voting: Do presidential candidates "waltz before a blind audience?". *American Political Science Review* 83, 1 (1989), 123–141.

[2] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 245, 19 pages. https://doi.org/10.1145/3491102.3502040

[3] Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. 2023. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review* 4, 4 (2023).

[4] Sacha Altay, Emma de Araujo, and Hugo Mercier. 2021. "If This account is True, It is Most Enormously Wonderful": Interestingness-If-True and the Sharing of True and False News. *Digital Journalism* 0, 0 (2021), 1–22. https://doi.org/10.1080/21670811.2021.1941163 arXiv:https://doi.org/10.1080/21670811.2021.1941163

[5] Antonio Alonso Arechar, Jennifer Nancy Lee Allen, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael Stagnaro, Jerry Zhang, et al. 2022. Understanding and reducing online misinformation across 16 countries on six continents. PsyArXiv.

[6] Fatemeh Torabi Asr and Maite Taboada. 2019. Big Data and quality data for fake news and misinformation detection. *Big Data & Society* 6, 1 (2019), 2053951719843310. https://doi.org/10.1177/2053951719843310 arXiv:https://doi.org/10.1177/2053951719843310

[7] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 93 (Oct. 2020), 26 pages. https://doi.org/10.1145/3415164

[8] danah boyd. 2018. You Think You Want Media Literacy… Do You? https://points.datasociety.net/you-think-you-want-media-literacy-do-you-7cad6af18ec2

[9] Samantha Bradshaw, Philip N Howard, Bence Kollanyi, and Lisa-Maria Neudert. 2020. Sourcing and automation of political news and information over social media in the United States, 2016-2018. *Political Communication* 37, 2 (2020), 173–193.

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

[11] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806 arXiv:https://doi.org/10.1080/2159676X.2019.1628806

[12] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad, India) *(WWW '11)*. Association for Computing Machinery, New York, NY, USA, 675–684.

https://doi.org/10.1145/1963405.1963500
[13] Jan Chołoniewski, Julian Sienkiewicz, Naum Dretnik, Gregor Leban, Mike Thelwall, and Janusz A Hołyst. 2020. A calibrated measure to compare fluctuations of different entities across timescales. *Scientific reports* 10, 1 (2020), 1–16.
[14] Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Sage publications.
[15] M.X. Delli Carpini and S. Keeter. 1997. *What Americans Know About Politics and Why it Matters.* Yale University Press, New Haven.
[16] Carroll Doherty, Jocelyn Kiley, Nida Asheer, and Calvin Jordan. 2021. Beyond Red vs. Blue: The Political Typology.
[17] Judith Donath. 2007. Signals, cues and meaning. *Signals, Truth and Design* (2007).
[18] Judith Donath. 2007. Signals in Social Supernets. *Journal of Computer-Mediated Communication* 13, 1 (10 2007), 231–251. https://doi.org/10.1111/j.1083-6101.2007.00394.x arXiv:https://academic.oup.com/jcmc/article-pdf/13/1/231/22317022/jjcmcom0231.pdf
[19] Josefina Erikson and Cecilia Josefsson. 2019. Does higher education matter for MPs in their parliamentary work? Evidence from the Swedish Parliament. *Representation* 55, 1 (2019), 65–80.
[20] Laura Faragó, Anna Kende, and Péter Krekó. 2019. We only believe in news that we doctored ourselves. *Social Psychology* (2019).
[21] Martin Flintham, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. 2018. Falling for Fake News: Investigating the Consumption of News via Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173950
[22] B. J. Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, and Marissa Treinen. 2001. What Makes Web Sites Credible? A Report on a Large Quantitative Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 61–68. https://doi.org/10.1145/365024.365037
[23] B. J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. 2003. How Do Users Evaluate the Credibility of Web Sites? A Study with over 2,500 Participants. In *Proceedings of the 2003 Conference on Designing for User Experiences* (San Francisco, California) *(DUX '03)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/997078.997097
[24] Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. 2020. NELA-GT-2019: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles. arXiv:2003.08444 [cs.CY]
[25] Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15536–15545. https://doi.org/10.1073/pnas.1920498117 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1920498117
[26] Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15536–15545.
[27] Hendrik Heuer and Elena Leah Glassman. 2022. A Comparative Evaluation of Interventions Against Misinformation: Augmenting the WHO Checklist. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 241, 21 pages. https://doi.org/10.1145/3491102.3517717
[28] Hendrik Heuer, Juliane Jarke, and Andreas Breiter. 2021. Machine learning in tutorials – Universal applicability, underinformed application, and other misconceptions. *Big Data & Society* 8, 1 (2021), 20539517211017593. https://doi.org/10.1177/20539517211017593 arXiv:https://doi.org/10.1177/20539517211017593
[29] Edda Humprecht, Frank Esser, and Peter Van Aelst. 2020. Resilience to Online Disinformation: A Framework for Cross-National Comparative Research. *The International Journal of Press/Politics* 25, 3 (2020), 493–516. https://doi.org/10.1177/1940161219900126 arXiv:https://doi.org/10.1177/1940161219900126
[30] Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. 2021. Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 18 (apr 2021), 42 pages. https://doi.org/10.1145/3449092
[31] Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 82 (Nov. 2018), 23 pages. https://doi.org/10.1145/3274351
[32] Dan M Kahan. 2013. Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision making* 8, 4 (2013), 407–424.
[33] Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 140 (Oct. 2020), 27 pages. https://doi.org/10.1145/3415211
[34] Mike Kuniavsky. 2003. *Observing the user experience: a practitioner's guide to user research.* Elsevier.
[35] Hannu Kuusela and Pallab Paul. 2000. A comparison of concurrent and retrospective verbal protocol analysis. *American journal of psychology* 113, 3 (2000), 387–404.
[36] Airi Lampinen. 2015. Deceptively Simple: Unpacking the Notion of "Sharing". *Social Media + Society* 1, 1 (2015), 2056305115578135. https://doi.org/10.1177/2056305115578135 arXiv:https://doi.org/10.1177/2056305115578135

[37] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, and others. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
[38] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131. https://doi.org/10.1177/1529100612451018 arXiv:https://doi.org/10.1177/1529100612451018 PMID: 26173286.
[39] Philipp Mayring. 2021. Qualitative content analysis: A step-by-step guide. *Qualitative Content Analysis* (2021), 1–100.
[40] Solomon Messing, Christina DeGregorio, Bennett Hillenbrand, Gary King, Saurav Mahanti, Zagreb Mukerjee, Chaya Nayak, Nate Persily, Bogdan State, and Arjun Wilkins. 2020. Facebook Privacy-Protected Full URLs Data Set. https://doi.org/10.7910/DVN/TDOAPG
[41] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Nami Sumida. 2018. Can Americans tell factual from opinion statements in the news. *Pew Research Center's Journalism Project.* (2018). https://www.journalism.org/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news
[42] Amy Mitchell, Jeffrey Gottfried, Galen Stocking, Mason Walker, and Sophia Fedeli. 2019. Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed. https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/
[43] Nic Newman. 2020. Overview and Key Findings of the 2020 Digital News Report. https://www.digitalnewsreport.org/survey/2020/overview-key-findings-2020/
[44] Nic Newman, Richard Fletcher, Craig T. Robertson, Kirsten Eddy, and Rasmus Kleis Nielsen. 2022. Reuters Institute Digital News Report 2022. (2022). https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022
[45] NewsGuard. 2022. Rating process and criteria. https://www.newsguardtech.com/ratings/rating-process-criteria/
[46] David Lynn Painter and Juliana Fernandes. 2022. "The Big Lie": How Fact Checking Influences Support for Insurrection. *American Behavioral Scientist* (2022), 00027642221103179.
[47] Irene V Pasquetto, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. 2020. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review* (2020).
[48] Jarutas Pattanaphanchai, Kieron O'Hara, and Wendy Hall. 2013. Trustworthiness Criteria for Supporting Users to Assess the Credibility of Web Information. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13 Companion)*. Association for Computing Machinery, New York, NY, USA, 1123–1130. https://doi.org/10.1145/2487788.2488132
[49] Cyrus Peikari and Anton Chuvakin. 2004. *Security Warrior: Know Your Enemy.* " O'Reilly Media, Inc.".
[50] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.
[51] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (01 Apr 2021), 590–595. https://doi.org/10.1038/s41586-021-03344-2
[52] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.
[53] Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.
[54] Gordon Pennycook and David G Rand. 2020. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality* 88, 2 (2020), 185–200.
[55] Andrea Pereira, Elizabeth Harris, and Jay J Van Bavel. 2023. Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes & Intergroup Relations* 26, 1 (2023), 24–47.
[56] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3391–3401. https://www.aclweb.org/anthology/C18-1287
[57] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 231–240. https://doi.org/10.18653/v1/P18-1022
[58] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait Detection. In *Advances in Information Retrieval*, Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello (Eds.). Springer International Publishing, Cham, 810–817.
[59] Cornelius Puschmann, Hevin Karakurt, Carolin Amlinger, Nicola Gess, and Oliver Nachtwey. 0. RPC-Lex: A dictionary to measure German right-wing populist conspiracy discourse online. *Convergence* 0, 0 (0), 13548565221109440. https://doi.org/10.1177/13548565221109440 arXiv:https://doi.org/10.1177/13548565221109440
[60] George Rabinowitz and Stuart Elaine Macdonald. 1989. A directional theory of issue voting. *American political science review* 83, 1 (1989), 93–121.

[61] Kim Salazar. 2020. Contextual inquiry: inspire design by observing and interviewing users in their context. *Nielsen Norman Group* 6 (2020).
[62] Dietram A. Scheufele and Nicole M. Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7662–7669. https://doi.org/10.1073/pnas.1805871115 arXiv:https://www.pnas.org/content/116/16/7662.full.pdf
[63] Farhana Shahid, Srujana Kamath, Annie Sidotam, Vivian Jiang, Alexa Batino, and Aditya Vashistha. 2022. "It Matches My Worldview": Examining Perceptions and Attitudes Around Fake Videos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 255, 15 pages. https://doi.org/10.1145/3491102.3517646
[64] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.* 10, 3, Article 21 (April 2019), 42 pages. https://doi.org/10.1145/3305260
[65] Kai Shu, H. Russell Bernard, and Huan Liu. 2019. Studying Fake News via Network Analysis: Detection and Mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, Nitin Agarwal, Nima Dokoohaki, and Serpil Tokdemir (Eds.). Springer International Publishing, Cham, 43–65. https://doi.org/10.1007/978-3-319-94105-9_3
[66] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1 (Sept. 2017), 22–36. https://doi.org/10.1145/3137597.3137600
[67] Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1718–1729. https://doi.org/10.18653/v1/2021.eacl-main.147
[68] S. Sloman and P. Fernbach. 2017. *The Knowledge Illusion: Why We Never Think Alone*. Penguin Publishing Group. https://books.google.dk/books?id=2xuMDAAAQBAJ
[69] Michael Spence. 1978. Job market signaling. In *Uncertainty in economics*. Elsevier, 281–306.
[70] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 127 (Nov. 2019), 26 pages. https://doi.org/10.1145/3359229
[71] The Trust Project. 2017. The 8 Trust Indicators. https://thetrustproject.org/trust-indicators/
[72] Stuart A. Thompson and Davey Alba. 2022. Fact and mythmaking blend in Ukraine's Information War. https://www.nytimes.com/2022/03/03/technology/ukraine-war-misinfo.html?action=click&module=RelatedLinks&pgtype=Article
[73] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[74] W3C Community. 2020. Reviewed Credibility Signals. https://credweb.org/reviewed-signals-20200224
[75] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 176–187. https://aclanthology.org/E17-1017
[76] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067
[77] Claire Wardle. 2018. Fake news. it's complicated. https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79
[78] Claire Wardle, Hossein Derakhshan, et al. 2018. Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. *Ireton, Cherilyn; Posetti, Julie. Journalism,'fake news'& disinformation. Paris: Unesco* (2018), 43–54.
[79] Wikipedia contributors. 2021. List of political parties in Germany — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_political_parties_in_Germany&oldid=1030256103. [Online; accessed 1-July-2021].
[80] Wikipedia contributors. 2022. Five whys — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Five_whys&oldid=1102936260 [Online; accessed 23-August-2022].
[81] Wikipedia contributors. 2022. Impressum — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Impressum&oldid=1092341470 [Online; accessed 1-September-2022].
[82] Wikipedia contributors. 2022. List of newspapers in the United States — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_newspapers_in_the_United_States&oldid=1101330486 [Online; accessed 4-August-2022].
[83] Sam Wineburg, Joel Breakstone, Sarah McGrew, Mark D Smith, and Teresa Ortega. 2022. Lateral reading on the open Internet: A district-wide field study in high school government classes. *Journal of Educational Psychology* (2022).
[84] Sam Wineburg and Sarah McGrew. 2019. Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information. *Teachers College Record* 121, 11 (2019), 1–40. https://doi.org/10.1177/016146811912101102 arXiv:https://doi.org/10.1177/016146811912101102
[85] Thomas Wood and Ethan Porter. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41, 1 (2019), 135–163.

[86] World Health Organization. 2020. Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation
[87] Amotz Zahavi. 1975. Mate selection—a selection for a handicap. *Journal of theoretical Biology* 53, 1 (1975), 205–214.
[88] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*. 603–612.

A APPENDIX

Table 5. The politically diverse end users from Germany that participated. The table shows their party, gender, whether they have personally used any of the websites they rated, and how they rated them. Rating scale: very unreliable (−−), unreliable (−), neither reliable or unreliable (○), reliable (+), very reliable (++).

Participant		Personal Usage			Reliability Ratings		
ID	Gender	DE1	DE2	DE3	DE1	DE2	DE3
AFD1	m	**Sometimes**	Never	*Rarely*	+		
AFD2	m	**Frequently**	Never	*Rarely*	○	○	○
AFD3	m	Never	Never	Never	○	−−	○
CDU1	m	Never	Never	Never	−−	−−	○
CDU2	f	**Sometimes**	Never	Never		○	−
CDU3	m	Never	Never	Never	○	○	○
CDU4	m	Never	Never	*Rarely*	○	+	−
CDU5	m	Never	Never	Never	−	○	−
CDU6	f	Never	Never	Never	○	−	−
CDU7	f	Never	Never	Never	−	−−	−−
FDP1	m	Never	Never	Never	−	○	−
GN1	m	Never	Never	Never	−−	−−	−−
GN2	m	Never	Never	Never	−	−	−
GN3	f	Never	Never	Never	○	+	−
GN4	m	Never	Never	Never	−−	−−	−−
LT1	f	Never	*Rarely*	Never	−	−	−
LT2	f	Never	Never	Never	−	−−	−−
SPD1	m	Never	**Frequently**	Never	−	+	−
SPD2	f	*Rarely*	Never	Never		−	−
SPD3	m	*Rarely*	*Rarely*	Never	○	−	−
SPD4	m	Never	Never	Never	−−	−−	−−
SPD5	f	Never	Never	Never	−	−−	−
SPD6	m	Never	*Rarely*	Never	○	○	−−

Table 6. The professional journalists from the U.S. that participated as experts. The table shows their gender, whether they have personally used any of the websites they rated, and how they rated them. The ID indicates what kind of newspaper employs which expert. For experts from the Top 10 most popular US-based newspapers, the participant ID begins with JP. For experts recruited from newspapers with a political alignment, JL stands for left, JC for center, and JR for right. Rating scale: very unreliable (−−), unreliable (−), neither reliable or unreliable (○), reliable (+), very reliable (++).

Participant		Personal Usage			Reliability Ratings		
ID	Gender	EN1	EN2	EN3	EN1	EN2	EN3
JP1	f	Never	Never	Never	−	−	−−
JP2	f	Never	Never	Never	−−	−−	○
JP3	m	Never	Never	Never	−−	−	−−
JP4	m	Never	Never	Never	−−	−−	−
JP5	m	Never	Never	Never	+	−	−−
JP6	m	Never	Never	Never	−	−−	−
JP7	m	*Rarely*	Never	Never	○	−	−
JL1	m	Never	Never	Never	−−	−−	−−
JL2	m	*Rarely*	Never	Never	−	−−	−−
JL3	f	Never	Never	Never	−−	−	○
JL4	m	Never	Never	Never	−−	−−	−−
JL5	f	Never	Never	Never	−	−−	−−
JL6	m	Never	Never	Never	−−	−−	−−
JC1	f	Never	Never	Never	○	+	○
JC2	f	Never	Never	Never	−	○	−
JC3	f	Never	Never	Never	−−	−−	−−
JR1	f	Never	Never	Never	−−	−−	−−
JR2	m	*Rarely*	Never	Never	○	−	−−
JR3	f	Never	Never	Never	○	−−	−−
JR4	m	Never	Never	Never	○	○	