The Need for Human-centered Design in Fact-checking Research

Anubrata Das*, Houjiang Liu, Venelin Kovatchev and Matthew Lease

[a] *School of Information, The University of Texas at Austin, Austin, TX, USA*

ARTICLE INFO

Keywords:
Natural Language Processing
Misinformation
Disinformation
Explainability
Human-AI Teaming

ABSTRACT

Misinformation threatens modern society by promoting distrust in science, changing narratives in public health, heightening social polarization, and disrupting democratic elections and financial markets, among a myriad of other societal harms. To address this, a growing cadre of professional fact-checkers and journalists provide high-quality investigations into purported facts. However, these largely manual efforts have struggled to match the enormous scale of the problem. In response, a growing body of Natural Language Processing (NLP) technologies have been proposed for more scalable fact-checking. Despite tremendous growth in such research, however, practical adoption of NLP technologies for fact-checking still remains in its infancy today.

We recommend that future research include collaboration with fact-checker stakeholders early on in NLP research, as well as incorporation of human-centered design practices in model development, in order to further guide technology development for human use and practical adoption. In addition, we advocate for more research on benchmark development supporting extrinsic evaluation of human-centered fact-checking technologies.

Note. This work is excerpted from a longer version in Das, Liu, Kovatchev and Lease (2023).

1. Introduction

Misinformation and related issues (disinformation, deceptive news, clickbait, rumours, and information credibility) increasingly threaten society. While professional fact-checkers and journalists provide high-quality investigations of purported facts to inform the public, human effort struggles to match the global Internet scale of the problem. To address this, a growing body of research has investigated Natural Language Processing (NLP) to automate fact-checking (Guo, Schlichtkrull and Vlachos, 2022; Nakov, Corney, Hasanain, Alam, Elsayed, Barr'on-Cedeno, Papotti, Shaar and Martino, 2021; Zeng, Abumansour and Zubiaga, 2021).

However, even state-of-the-art NLP technologies still cannot match human capabilities in many areas and remain insufficient to automate fact-checking in practice. Experts argue (Arnold, 2020; Nakov et al., 2021) that fact-checking is a complex process and requires subjective judgement and expertise. While current NLP systems are increasingly better at addressing simple fact-checking tasks, identifying false claims that are contextual and beyond simple declarative statements are yet beyond the reach for fully automated systems (Chen, Sriram, Choi and Durrett, 2022; Fan, Piktus, Petroni, Wenzek, Saeidi, Vlachos, Bordes and Riedel, 2020). Additionally, NLP tools that are integrated into the existing fact-checking workflow and assist in reducing latency are most desired by fact-checking practitioners (Nakov et al., 2021; Graves, 2018b; Alam, Shaar, Dalvi, Sajjad, Nikolov, Mubarak, Da San Martino, Abdelali, Durrani, Darwish, Al-Homaid, Zaghouani, Caselli, Danoe, Stolk, Bruntink and Nakov, 2021).

Current technological limitations have important ramifications for future research. First, practical use of NLP technologies for fact-checking is likely to come from hybrid, human-in-the-loop approaches rather than full automation. Second, as the technology matures, end-to-end evaluation becomes increasingly important to ensure practical solutions are being developed to solve the real-world use-case. To this end, new benchmarks that facilitate the extrinsic evaluation of automated fact-checking applications in practical settings may help drive progress on solutions that can be adopted for use in the wild. Finally, to craft effective human-in-the-loop systems, more cross-cutting NLP and HCI integration could strengthen design of fact-checking tools, so that they are accurate, scalable, and usable in practice. It is important to collaborate with stakeholders early in research and incorporate human-centered design in model development.

Explainable and HITL approaches leverage both human and computational intelligence from a human-centered perspective, but there is a need to provide actionable guides to utilize both methods for designing useful fact-checking tools. We propose research to identify the design spaces of applying NLP technologies to assist fact-checkers.

Note. This work is excerpted from a longer treatment in Das et al. (2023), covering a detailed account of existing NLP technologies for fact-checking, their limitations, and possible future directions.

*Corresponding author

ORCID(s): 0000-0002-5412-6149 (A. Das); 0000-0003-1259-1541 (V. Kovatchev); 0000-0002-0056-2834 (M. Lease)

2. Distributing Work between Human and AI for Mixed-initiative Fact-checking

The practice of fact-checking has already become one type of complex and distributed computer-mediated work for human fact-checkers (Graves, 2018a). Although Graves (2017) breaks down a traditional journalist fact-checking pipeline into five steps, the real situation of fact-checking a claim is not a streamlined process (Juneja and Mitra, 2022). Various AI tools are adopted dynamically and diversely by fact-checkers to complete different fact-checking tasks (Arnold, 2020; Beers, Haughey, Melinda, Arif and Starbird, 2020; Micallef, Armacost, Memon and Patil, 2022).

Researchers and practitioners increasingly believe that future fact-checking should be a mixed-initiative practice in which humans perform specific tasks while machines take over others (Nguyen, Kharosekar, Krishnan, Krishnan, Tate, Wallace and Lease, 2018; Lease, 2020; Nakov et al., 2021). To embed such hybrid and dynamic human-machine collaborations into existing fact-checking workflow, the task arrangement between human and AI need to be articulated clearly by understanding the expected outcomes and criteria for each. Furthermore, designing a mixed-initiative tool for different fact-checking tasks requires a more fine-grained level of task definition for human and AI (Lease, 2018, 2020). Several studies highlighting the role of humans in the fact-checking workflow, e.g., a) human experts select check-worthy claims from claim detection tools (Hassan, Zhang, Arslan, Caraballo, Jimenez, Gawsane, Hasan, Joseph, Kulkarni, Nayak et al., 2017) and deliver them to fact-checkers, b) ask crowd workers to judge reliable claims sources (Shabani, Charlesworth, Sokhn and Schuldt, 2021), or c) flag potential misinformation (Roitero, Soprano, Portelli, Spina, Della Mea, Serra, Mizzaro and Demartini, 2020) to improve veracity prediction. All the above human activities are examples of micro-tasks within a mixed-initiative fact-checking process.

Prior work in crowdsourcing has shown that it is possible to effectively break down the academic research process and utilize crowd workers to partake in smaller research tasks (Vaish, Davis and Bernstein, 2015; Vaish, Gaikwad, Kovacs, Veit, Krishna, Arrieta Ibarra, Simoiu, Wilber, Belongie, Goel et al., 2017). Given this evidence, we can also break down sub-tasks of a traditional fact-checking process into more fine-grained tasks. Therefore, key research questions are a) How can we design these micro-tasks to facilitate each sub-task of fact-checking? b) What are the appropriate roles for human and AI in different micro-tasks?

Before arranging human and AI work into an order, researchers need to understand the form of influence and supervision between the role of human and AI because it will directly affect whether humans decide to take AI advice (Cimolino and Graham, 2022). Usually, if AI aims to assist high-stake decision-making tasks, such as recidivism prediction (Veale, Van Kleek and Binns, 2018) and medical treatments (Cai, Winter, Steiner, Wilcox and Terry, 2019), accessing the risk of AI decisions and building trust with AI systems become important factors for human to adopt such AI assistants (Lai, Chen, Liao, Smith-Renner and Tan, 2021). In the context of fact-checking, if AI directly predicts the verdict of a claim, fact-checkers are skeptical about how AI makes such prediction (Arnold, 2020). On the other hand, if AI only helps to filter claims that are uncheckable, such as opinions and personal experience, fact-checkers are more willingly to use such automation with less concern on how AI achieves it. Apparently, deciding whether a claim is true or false is a high-stake decision-making task for fact-checkers while filtering uncheckable claims is a less important but tedious task that fact-checkers want automation to help with. Therefore, the extent of human acceptance of AI varies according to how humans assess the task assigned to AI, resulting in different human factors, such as trust, transparency, and fairness. Researchers need to take these human factors into considerations while designing micro-tasks between human and AI. Researchers also need to specify or decompose these human factors into different key variables that can be measured during the model development process.

Given a deep understanding of the task relationship between human and AI, researchers can then ask further research questions on how to apply an explainable approach, or employ a HITL system vs. automated solutions, to conduct fact-checking. Here we list out several specific research topics that contain mixed-initiative tasks, including a) assessing claim difficulty leveraging crowd workers, b) breaking down a claim into a multi-hop reasoning task and engaging the crowd to find information relevant to the sub-claims, c) designing micro-tasks to parse a large number of documents retrieved by web search to identify sources that contain the information needed for veracity prediction.

3. Human-centered Evaluation of NLP Technology for Fact-Checkers

We first highlighting four key metrics from human factors for evaluating systems (i.e., what to measure and how to measure them): accuracy, time, model understanding, and trust. Following this, we propose a template for an experimental protocol for human-centered evaluations in fact-checking.

Accuracy. Most fact-checking user studies assume task accuracy as the primary user goal (Nguyen et al., 2018; Mohseni, Yang, Pentyala, Du, Liu, Lupfer, Hu, Ji and Ragan, 2021). Whereas non-expert users (i.e., social media users or other form of content consumers) might be most interested in veracity outcome along with justification, fact-checkers often want to use automation and manual effort interchangeably in their workflow (Arnold, 2020; Nakov et al., 2021). Thus a more fine-grained approach towards measuring accuracy is essential beyond the final veracity prediction accuracy. For fine-grained accuracy evaluation, it is also crucial to capture fact-checkers accuracy, particularly for the sub-tasks for which they use the fact-checking tool.

With the assumption that "ground truth" exists for all of the sub-tasks in the fact-checking pipeline, accuracy can be computed by comparing users' answers with the ground truth. Note that measuring sub-task level accuracy is trickier than end-to-end fact-checking accuracy. Sub-task level accuracy can be captured by conducting separate experiments for each sub-task. Suppose the point of interest is to understand users' performance for detecting *claim-checkworthiness*. In that case, we will require to collect data only in the context of the *claim-checkworthiness* task.

In some cases, it is possible to merge multiple sub-tasks for evaluation purposes. For example, Miranda, Nogueira, Mendes, Vlachos, Secker, Garrett, Mitchel and Marinho (2019) evaluate the effectiveness of their tool with journalists by capturing the following two key variables: a) the relevance of retrieved evidence, and b) the accuracy of the predicted stance. This method provides essential insight into evidence retrieval, stance detection, and the final fact-checking task. Depending on the tool, the exact detail of this metric will require specific changes according to tool affordances.

Time. Time taken at each stage of the fact-checking process is important to understand how the tool helps in reducing overall latency in fact-checking. Note that both time and accuracy measures need to control for claim properties. For example, if a claim has been previously fact-checked, it will take less time to fact-check such claims. On the other hand, a new claim that is more difficult to assess would require more time.

Model Understanding. Fact-checkers express concern over understanding of the tools they use in the existing workflow. For example, Arnold (2020) pointed out that fact-checkers expressed a need for understanding CrowdTangle's algorithm for detecting viral content on various social media platforms. Similarly, Nakov et al. (2021) observed a need for increased system transparency in the fact-checking tools used by different organizations. Lease (2018) argues that transparency is equally important for non-expert users to understand the underlying system and make an informed judgement. Although this is not a key variable related to user performance, it is important for the practical adoption.

To measure understating, users could be asked to self-report their level of understanding on a Likert-scale. However, simply asking the participants if they understand the algorithm is not a sufficient metric. The true understanding of a tool will lead to simulating the tool behavior (Hase and Bansal, 2020). We suggest the following steps for measuring understanding based on prior work (Cheng, Wang, Zhang, O'Connell, Gray, Harper and Zhu, 2019).

Decision Prediction: To capture users' holistic understanding of the tool, users could be provided claims and asked the following: "What label would the tool assign to this claim?"

Alternative Prediction: Capturing how changes in the input influence the output can also measure understanding, e.g., by asking users how the tool would assign a label to a claim when input parts are changed. Imagine a tool that showed the users the evidence it has considered to arrive at a veracity conclusion. Now, if certain pieces of evidence were swapped, how would that reflect in the model prediction?

Trust. For practical adoption, trust in a fact-checking tool is crucial. While model understanding is often positively correlated with trust, understanding alone may not suffice to establish trust, and users may trust a model for even without it based on other properties of it. In this domain, fact-checkers and journalists may have less trust in algorithmic tools (Arnold, 2020). On the other hand, there is also the risk of over-trust, or users blindly following model predictions (Nguyen et al., 2018; Mohseni et al., 2021). To maximize the tool effectiveness, we would want users to neither dismiss all model predictions out of hand (complete skepticism) nor blindly follow all model predictions (complete faith), it is important to calibrate users' trust for the most effective tool usage. We suggest measuring a notion of *calibrated trust*: how often users abide by correct model decisions and override erroneous model decisions. This could be captured by a typical *confusion matrix*? in which we distinguish Type I errors (false positive) vs. Type II errors (false negatives). To promote effective human-AI teaming, AI tools should assist their human users in developing strong calibrated trust to appropriately trust and distrust model predictions as each case merits.

Beyond calibrated trust once could also mearue quantiative trust by adopting methodologies from human-machine trust literature (Lee and Moray, 1992). For example, Cheng et al. (2019) adopted prior work into a 7-point Likert scale. A similar scale can be reused for evaluating trust in fact-checking tool. For example, users might be asked to denote their agreement with the following statements on a Likert-scale: 1. I understand the fact-checking tool. 2. I can predict how the tool will behave. 3. I have faith that the tool would be able to cope with the different fact-checking task. 4. I trust the decisions made by the tool. 5. I can count on the tool to provide reliable fact-checking decisions.

Additional factors. Individual differences among users might result in substantial variation in experimental outcomes. For example, varying technical literacy (Cheng et al., 2019), any prior knowledge about the claims, and users' political leaning (Thornhill, Meeus, Peperkamp and Berendt, 2019) might influence user performance on the task with fact-checking tools. Thus it is valuable to capture factors such as technical literacy (familiarity with fact-checking tools and popular AI tools), media literacy (familiarity with fact-checking), and demographics in study design.

Quantitative measures alone are not sufficient as they do not capture certain nuances about how effectively a tool integrates into a fact-checker's workflow. For example, even if users understand and trust the working principle of a tool, it is unclear *why* they do so. Hence, users might be asked a few open-ended questions at the end of the study to gather qualitative insights. Such questions could include: 1. Describe your understanding of the tool. How did the tool design itself helped in understanding the tool? 2. Why do you trust or not trust the tool? 3. In what capacity do you see utilizing this tool beyond the scope of this study?

Experimental Protocol. One strategy to capture the aforementioned metrics is to design a mixed-methods study. Here we outline the template for such a study. Imagine the goal were to measure the user performance for fact-checking using a new tool (let's call it *tool A*) compared to an existing tool (*tool B*). Fact-checking tasks in real world might be influenced by users' priors about the claims being checked. Thus, a *within-subject* study protocol may be more appropriate to account for such priors (Shi, Bhattacharya, Das, Lease and Gwizdka, 2022).

Pre-task: Users would first be asked to fact-check a set of claims. A user would be asked to leverage a pre-existing *tool B* at this stage. Tool B can be replaced with different baselines, e.g., simple web-search by non-experts to proprietary tools used by fact-checkers, depending on the particular use case. Users are then asked to think aloud.

Learning: At this stage users would familiarize themselves with the new tool (*tool A*). Users would need to fact-check a different set of claims from the first one. Ground truth is also accessible to the user to form a prior about what kind of mistakes a tool might make. The claims here need to be selected at random to reflect the tools' capabilities. Moreover, tools' performance metrics needs to be given to the users as additional information. Users would be encouraged to ask questions about the tool at this stage.

Prediction: Users would now be asked to fact-check the same claims from step-1 above but this time they are asked to leverage the *tool A*. Users would be asked to think out loud through this stage. Users could simply guess the answers and achieve a high accuracy score. Thus, the claims selected for stages (1) & (3) have to be a balanced set of claims with an equal number of distributions from true positive, true negative, false positive, and false negative samples. This idea is adopted from prior work (Hase and Bansal, 2020).

Post-task survey: Users would now be asked to take a small survey for capturing trust, understanding, technical literacy, media literacy, and demographic information.

Post-task interview: Upon completion of these steps, users would be interviewed with open-ended questions to gather insights about their understanding and trust in the system.

The measures and study protocol could be useful in the context of evaluating any new fact-checking system compared to an existing system or practices. Specifics might vary depending on the target user group and the tool's intended purpose. For example, let us assume a new claim detection tool has been proposed that takes claims from a tipline (Kazemi, Garimella, Shahi, Gaffney and Hale, 2021). Currently, fact-checkers use an existing claim-matching algorithm to filter out the already fact-checking claim. Now, if we replace *tool B* above with the existing claim-matching algorithm and *tool A* with the proposed claim detection tool, we can utilize the protocol mentioned above. One could thus evaluate how users perform for claim detection tasks using the new tool compared to the existing ones wrt. accuracy, time, understanding, and trust.

4. Conclusion

In contrast with fully-automated systems, hybrid systems involve humans-in-the-loop and facilitate human-AI teaming (Bansal, Wu, Zhou, Fok, Nushi, Kamar, Ribeiro and Weld, 2021) in order to: a) scale-up human decision making; b) augment machine learning capabilities with human accuracy; and c) mitigate unintended consequences from machine errors. We recommend that future research include collaboration with fact-checker stakeholders early on in NLP research, as well as incorporation of human-centered design practices in model development, in order to further guide technology development for human use and practical adoption. In addition, we advocate for more research on benchmark development supporting extrinsic evaluation of human-centered fact-checking technologies.

Note. This work is excerpted from a longer treatment in Das et al. (2023), covering a detailed account of existing NLP technologies for fact-checking, their limitations, and possible future directions.

Acknowledgements. This research was supported in part by the Knight Foundation, the Micron Foundation, Wipro, and by Good Systems[1], a UT Austin Grand Challenge to develop responsible AI technologies. The statements made herein are solely the opinions of the authors and do not reflect the views of the sponsoring agencies.

References

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., Al-Homaid, A., Zaghouani, W., Caselli, T., Danoe, G., Stolk, F., Bruntink, B., Nakov, P., 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic. pp. 611–649. URL: https://aclanthology.org/2021.findings-emnlp.56, doi:10.18653/v1/2021.findings-emnlp.56.

Arnold, P., 2020. The challenges of online fact checking: how technology can (and can't) help - full fact. https://fullfact.org/blog/2020/dec/the-challenges-of-online-fact-checking-how-technology-can-and-cant-help/. (Accessed on 09/11/2021).

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D., 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16.

Beers, A., Haughey, Melinda, M., Arif, A., Starbird, K., 2020. Examining the digital toolsets of journalists reporting on disinformation, in: Proceedings of Computation + Journalism 2020 (C+J '20). ACM, New York, NY, USA,., p. 5. URL: https://cpb-us-w2.wpmucdn.com/express.northeastern.edu/dist/d/53/files/2020/02/CJ_2020_paper_50.pdf.

Cai, C.J., Winter, S., Steiner, D., Wilcox, L., Terry, M., 2019. " hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. Proceedings of the ACM on Human-computer Interaction 3, 1–24.

Chen, J., Sriram, A., Choi, E., Durrett, G., 2022. Generating Literal and Implied Subquestions to Fact-check Complex Claims URL: http://arxiv.org/abs/2205.06938, arXiv:2205.06938.

Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., Zhu, H., 2019. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders, in: Proceedings of the 2019 chi conference on human factors in computing systems, pp. 1–12.

Cimolino, G., Graham, T.N., 2022. Two heads are better than one: A dimension space for unifying human and artificial intelligence in shared control, in: CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3491102.3517610, doi:10.1145/3491102.3517610.

Das, A., Liu, H., Kovatchev, V., Lease, M., 2023. The state of human-centered nlp technology for fact-checking. Information Processing Management 60. URL: https://www.sciencedirect.com/science/article/pii/S030645732200320X, doi:https://doi.org/10.1016/j.ipm.2022.103219.

Fan, A., Piktus, A., Petroni, F., Wenzek, G., Saeidi, M., Vlachos, A., Bordes, A., Riedel, S., 2020. Generating fact checking briefs, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 7147–7161. URL: https://aclanthology.org/2020.emnlp-main.580, doi:10.18653/v1/2020.emnlp-main.580.

Graves, D., 2018a. Understanding the promise and limits of automated fact-checking .

Graves, L., 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. Communication, Culture & Critique 10, 518–537.

Graves, L., 2018b. Boundaries not drawn: Mapping the institutional roots of the global fact-checking movement. Journalism Studies 19, 613–631.

Guo, Z., Schlichtkrull, M., Vlachos, A., 2022. A Survey on Automated Fact-Checking. Transactions of the Association for Computational Linguistics 10, 178–206. URL: https://doi.org/10.1162/tacl_a_00454, doi:10.1162/tacl_a_00454, arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00454/1987018/tacl_a_00454.pdf.

Hase, P., Bansal, M., 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior?, in: ACL.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., Hasan, S., Joseph, M., Kulkarni, A., Nayak, A.K., et al., 2017. Claimbuster: The first-ever end-to-end fact-checking system. Proceedings of the VLDB Endowment 10, 1945–1948.

Juneja, P., Mitra, T., 2022. Human and technological infrastructures of fact-checking. arXiv preprint arXiv:2205.10894 .

Kazemi, A., Garimella, K., Shahi, G.K., Gaffney, D., Hale, S.A., 2021. Tiplines to combat misinformation on encrypted platforms: A case study of the 2019 indian election on whatsapp. ArXiv abs/2106.04726.

Lai, V., Chen, C., Liao, Q.V., Smith-Renner, A., Tan, C., 2021. Towards a science of human-ai decision making: A survey of empirical studies. arXiv preprint arXiv:2112.11471 .

Lease, M., 2018. Fact checking and information retrieval., in: DESIRES, pp. 97–98.

Lease, M., 2020. Designing human-ai partnerships to combat misinfomation .

Lee, J., Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35, 1243–1270.

Micallef, N., Armacost, V., Memon, N., Patil, S., 2022. True or False: Studying the Work Practices of Professional Fact-Checkers. Proceedings of the ACM on Human-Computer Interaction 6, 1–44. URL: https://doi.org/10.1145/3512974, doi:10.1145/3512974.

Miranda, S., Nogueira, D., Mendes, A., Vlachos, A., Secker, A., Garrett, R., Mitchel, J., Marinho, Z., 2019. Automated fact checking in the news room, in: The World Wide Web Conference, pp. 3579–3583.

Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., Hu, X., Ji, S., Ragan, E., 2021. Machine learning explanations to prevent overtrust in fake news detection, in: Proceedings of the International AAAI Conference on Web and Social Media, pp. 421–431.

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barr'on-Cedeno, A., Papotti, P., Shaar, S., Martino, G.D.S., 2021. Automated fact-checking for assisting human fact-checkers, in: IJCAI.

[1] http://goodsystems.utexas.edu/

Nguyen, A.T., Kharosekar, A., Krishnan, S., Krishnan, S., Tate, E., Wallace, B.C., Lease, M., 2018. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking, in: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, pp. 189–199.

Roitero, K., Soprano, M., Portelli, B., Spina, D., Della Mea, V., Serra, G., Mizzaro, S., Demartini, G., 2020. The covid-19 infodemic: Can the crowd judge recent misinformation objectively?, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1305–1314.

Shabani, S., Charlesworth, Z., Sokhn, M., Schuldt, H., 2021. Sams: Human-in-the-loop approach to combat the sharing of digital misinformation, in: CEUR Workshop Proc.

Shi, L., Bhattacharya, N., Das, A., Lease, M., Gwizdka, J., 2022. The Effects of Interactive AI Design on User Behavior: An Eye-tracking Study of Fact-checking COVID-19 Claims, in: Proceedings of the 7th ACM SIGIR Conference on Human Information, Interaction and Retrieval (CHIIR). URL: https://utexas.box.com/v/shi-chiir2022.

Thornhill, C., Meeus, Q., Peperkamp, J., Berendt, B., 2019. A digital nudge to counter confirmation bias. Frontiers in big data 2, 11.

Vaish, R., Davis, J., Bernstein, M., 2015. Crowdsourcing the research process. Collective Intelligence 3.

Vaish, R., Gaikwad, S.N.S., Kovacs, G., Veit, A., Krishna, R., Arrieta Ibarra, I., Simoiu, C., Wilber, M., Belongie, S., Goel, S., et al., 2017. Crowd research: Open and scalable university laboratories, in: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, pp. 829–843.

Veale, M., Van Kleek, M., Binns, R., 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making, in: Proceedings of the 2018 chi conference on human factors in computing systems, pp. 1–14.

Zeng, X., Abumansour, A.S., Zubiaga, A., 2021. Automated fact-checking: A survey. Language and Linguistics Compass 15, e12438.