# Responsible Human-AI Teaming: Interface Designs to Promote Bias Awareness in Human-AI Fact-checking Teams

CAITLIN LANCASTER* and HEBA ALY*, Clemson University, United States
SUBHASREE SENGUPTA, Clemson University, United States
NATHAN J. MCNEESE, Clemson University, United States

Human-AI collaborative endeavors, particularly human-AI teams, are increasingly becoming prominent and widely deployed in various contexts. Given the socio-cultural impact and importance of fact-checking to curb and regulate the spread of damaging false information, human-AI partnerships are being envisioned to boost the efficacy of these endeavors. Yet, the presence of biases and the propensity for overreliance on automation poses a concern for the extended use and deployment of such collaborative ventures. Two critical considerations for this issue include understanding how bias awareness mechanisms can be augmented into the human-AI collaborative pipeline and the impact such bias awareness mechanisms may have on the trust and reliance human counterparts ascribe to their autonomous counterparts. To address these considerations, we elicit four fundamental questions that arise and are essential to be addressed. Further, drawing on synergistic research that explores the impact of transparency on crafting interactive interface designs, we propose a study inspired by participatory design that captures the perspectives of end-users to explore how affordances can boost bias awareness in collaborative human-AI fact-checking pursuits. Our aim with this exploration is to create high-fidelity interactive prototypes that can provide tangible and actionable insights to address the issue of bias awareness in human-AI teams for later testing. The long-term vision of this work is to unpack how trust and reliance in human-AI teaming are impacted by bias awareness mechanisms, paving a path toward critical engagement in human-AI collaboration.

Additional Key Words and Phrases: Human-AI Collaboration, Responsible AI, Bias Mitigation, Trust and Reliance

## 1 INTRODUCTION

Digital ecosystems in recent times have boosted collaboration and connectivity, proliferating new information-sharing practices [21, 34]. However, the freedom afforded by this information access and sharing capabilities has also been accompanied by rapid upticks in mis- and disinformation [34, 67], calling for a need to manage and curtail this issue. The spread of false information poses a major threat to human society at the collective and individual levels [12]. The proliferation of misleading information can result in sowing distrust in media and increasing anxiety levels in individuals while exacerbating polarization and eroding the basic foundation of democratic information exchange [4, 22, 47]. In light of this threat and the pressing need to manage sharing and spreading of online false information, researchers, governmental bodies, and companies producing these technologies have sought to institute fact-checking procedures and tools to curb the threats that the spread of false information poses to individuals and society [8, 28, 66].

*Both authors contributed equally to this research.

Authors' addresses: Caitlin Lancaster, cma8@clemson.edu; Heba Aly, Clemson University, Clemson, South Carolina, United States, haly@clemson.edu; Subhasree Sengupta, Clemson University, Clemson, South Carolina, United States, subhass@clemson.edu; Nathan J. McNeese, Clemson University, Clemson, South Carolina, United States, mcneese@clemson.edu.


Manuscript submitted to ACM

However, the increasing volume of information online paired with intensive human labor, cost, and emotional strain that comes with fact-checking has pushed users and organizations to automate more aspects of the fact-checking process [76].

Driven by the power and prowess of computational systems, automated fact-checking holds great promise for the future, yet any automated system is also prone to latent biases that may impair the decision-making capabilities of such systems [30]. Such limitations can have a deterring effect on the rapport created between humans and agents, in turn impacting trust and reliance [55]. Biases may result from various sources, such as the data used to train these systems or the algorithm that shapes the systems' decision-making [46]. Thus, researchers and practitioners alike must consider and address the limitations of these systems, particularly as these limitations may also affect their accuracy in fact-checking contexts, impacting how information is disseminated and curated in online forums [76]. The issue of bias also raises concerns about how AI-enabled fact-checking should be harnessed and what role it should play in the entire landscape of the fact-checking process [30]. To answer this call, researchers have envisioned collaborative partnerships between humans and AI so that human intervention and involvement can, in part, be used to overcome the drawbacks of using only AI-driven mechanisms [54]. Such visions engage humans through interactive mechanisms to increase transparency about how automation is employed so that they can make an informed decision about accepting its recommendations [40].

As prior literature suggests, the human-AI collaborative approach to fact-checking can be powerful, leveraging the computational power of AI while also addressing its limitations through human intervention [20]. Yet, there is a gap in this research as previous work fails to address the importance of identifying biases to boost collaborative outcomes between human and autonomous counterparts. Given that a growing body of literature posits several antecedents of trust and reliance that can shape the nature of collaboration between human and autonomous agents [55], it becomes crucial to understand how bias awareness can play a mediating and moderating role in shaping trust and the overall outcomes of these collaborative endeavors. Further, for any human-AI collaborative enterprise to be successful, it is essential to develop mechanisms that promote bias awareness, thereby addressing potential complacency issues that end-users may inadvertently develop towards automated systems [54].

Motivated by this premise, we propose a set of critical questions aimed to understand how bias awareness in human-AI teaming can empower end-users to collaborate and make informed decisions in human-AI joint efforts toward fact-checking. Secondly, drawing inspiration from prior studies that engage end-users using participatory design approaches to enhance fairness perceptions [73], we propose a research plan stemming from these questions engaging in our own approach to participatory design to elicit design recommendations for bias awareness mechanisms in design interfaces. Our aim with this is to create high-fidelity interactive prototypes that can provide tangible and actionable insights to address the issue of bias awareness in human-AI teams for later testing. The long-term vision of this work is to unpack how trust and reliance in human-AI teaming are impacted by bias awareness mechanisms being part of the interactive loop.

## 2 RELATED WORK

In this section, we synthesize three streams of related research that guide our approach and perspective: 1) human-AI collaboration and teaming 2) automated fact-checking, and 3) bias in human-AI collaboration and the need for bias awareness. The convergence of this literature brings to light our four critical questions and supports the need for our proposed research.

### 2.1 Human-AI Collaboration & Teams

Teamwork and human collaboration are fundamental aspects of human society. Teams represent coordinated collaborative endeavors, where individuals act in concert to achieve a collective goal and outcome [36]. Human-human teams have been widely explored across various contexts, and several facets of such teams (such as coordination, task delegation, and emergent roles) have been widely studied which are critical to sustaining collective activities [31]. With the widespread use and adoption of automation, researchers have begun to explore what human-human teaming concepts do and do not transcend to the growing phenomenon of human-AI teams [56].

Such teams are increasingly conceptualized and deployed across contexts, including critical operations such as search and rescue [64] or emergency medical services [61]. Yet, such teams call for a shift in conceptual understanding of how such teams function and what key elements are essential for team cohesion and overall effectiveness in such contexts [56]. Particularly, teaming with an autonomous teammate may impact how individual team members engage, collaborate, and develop a shared understanding of team goals and outcomes [45]. Furthermore, researchers have found humans hold unique expectations of their autonomous counterparts, such as a desire for human likeness, communication abilities, performance, and skills as perceived necessities for successful human-AI collaboration [84]. Such investigations highlight the need to understand factors that are critical for maintaining team cohesion and performance in human-AI teams [23, 45, 61]. Trust and factors that shape trust formation can be essential for configuring norms of exchange and reciprocity vital to human-AI teams [25, 56, 75]. A burgeoning stream of scholarship has investigated and continues to explore trust formation and associated parameters that impact trust [65].

### 2.2 A Brief Overview of Automated Fact-Checking

Fact-checking plays an important role in the modern information age for maintaining social order by checking the authenticity of claims and facts disseminated [42, 77]. Such mechanisms can have an important role in how the public consumes and perceives the current state of affairs [77]. Fact-checking involves several key decision-making processes, including gathering evidence for a claim, deciding which claim should be checked and why, and setting a procedure for evaluation [79]. However, fact-checking can be very labor intensive when done manually [30] leading to an increase in the use of automated systems in the fact-checking process [82]. The field of computational journalism has been spearheading this initiative to boost the efficiency of fact-checking, while at the same time attempting to retain the essence of investigative journalism that lies at the heart of fact-checking [13]. While computational approaches promise efficiency, they miss out on the tacit skills and expertise held by human journalists to effectively capture contextual signals and identify suitable evidence sources [1].

Due to such considerations, sophisticated algorithmic approaches have emerged that can capture several nuances and contextual signals drawn from large data corpora, leveraging the computational prowess of natural language processing (NLP) [27]. Yet, the extensive use of NLP systems is fraught with several limitations. The most critical limitation is that the opacity of such algorithms renders them hard to understand and successfully integrate into the fact-checking pipeline [30]. Thus, in parallel to computational efforts toward fact-checking, mixed-initiative approaches have begun to emerge. These mixed initiative ventures engage both humans and AI and aim to tackle the limitations of computational systems, while at the same time also addressing the inefficient, operational hurdle of manual systems [54]. While such initiatives are promising, they also beg the question of how such mixed initiatives should be designed and implemented, making further investigations into factors that impact this approach critical. Notably, trust becomes a key mediating and moderating factor. Prior investigations have indicated how the prediction capabilities of automated fact-checking systems can impact the trust in using such systems [54]. Further transparency and information sharing related to the automation process have also been key factors in trust calibration, especially in relation to fact-checking [30, 79]. Thus, further investigations are needed to elaborate on how mixed initiatives toward fact-checking can be made more robust and transparent to boost the collective efficacy of such endeavors.

### 2.3 The Need for Addressing Bias for Collaborative Human-AI Success

Bias in automated systems represents a critical hurdle to the collaborative outcomes of these fact-checking approaches. The issue of bias is incredibly complicated to understand due to the many ways it can manifest related to an AI system, such as from the data used to train the AI, to the algorithmic considerations that skew its recommendations, to the evaluation and assessment of decisions by end-users that determine whether and how AI recommendations are put into place [46]. Indeed, defining, measuring, detecting, and mitigating AI bias represents an active research agenda that is continually negotiated among AI researchers and practitioners as its prominence and application grow [71]. When left unchecked, AI biases present great potential for harm to society, mirroring the human biases that have far-reaching consequences for criminal justice and recidivism evaluations [2, 5], human resources decisions [38, 53], personal finance and credit risk assessments [78], and beyond [68]. Furthermore, the massive amount of information and decision-making that humans must process in fact-checking roles leads to high cognitive loads that contribute to overreliance (automation bias) on AI decisions, which can limit oversight in the fact-checking process [62, 86]. Indeed, research on collaborative human-AI decision-making has demonstrated that users often place either *too much* trust in the AI's decisions, encouraging complacency that can leave latent biases unchecked and unaddressed [3, 41, 54].

Coupled with the continued issues of system biases from flawed training data and model creation [52], there remain many questions surrounding how to address these biases in human-AI collaboration. In particular, because humans almost universally represent the final decision-maker in these collaborative endeavors, addressing the human elements, particularly their overreliance, and complacency on AI decisions, needs greater exploration to address the bias issues that lie at the core of such collaborative endeavors. [63]. This exploration is especially necessary due to the growing relevance and application of human-AI collaboration, but limited understanding of how to support appropriate human trust and reliance in the face of opaque AI decision-making [24]. As such, human-AI fact-checking becomes an important use case, given the criticality of these factors for supporting this collaborative process and overcoming the issues of distrust associated with false information. Thus, if researchers are not careful to investigate and address bias in human-AI collaborative fact-checking processes, the issues that researchers hope to address with false information will likely be met with other, potentially more damaging outcomes. This pressing issue demonstrates a need to not only understand trust and reliance related to human-AI collaboration as potential avenues for preventing biased outcomes but also to fact-checking endeavors that inherently rely on preserving the quality and sanctity of information that is shared and curated.

A related stream of research on trust and reliance has aimed to engage and support users' cognitive decision-making processes to reduce human biases and complacency by making them more aware of how they make decisions with the system itself [10], which further motivates our goal of exploring pathways to integrate bias awareness mechanisms into collaborative human-AI fact-checking. Akin to such goals, research in HCI has capitalized on design elements that promote cognitive engagement to overcome issues with heuristic reasoning, such as delaying AI recommendations to disrupt overreliance [29]. Other work on design elements to reduce overreliance capitalizing on these functions found that certain affordances can reduce, though not fully eliminate, user complacency and encourage greater engagement and reflection on an AI's decision-making process [7]. These findings motivate the need for further exploration into design interventions that raise attention to latent biases among end-users, particularly to address and assess their impact on trust calibration in human-AI collaborations.

## 3 CONSIDERATIONS FOR ELICITING INTERACTIVE BIAS AWARENESS PROCESSES IN HUMAN-AI TEAMS, WITH A FOCUS ON FACT-CHECKING

In this section, we articulate four foundational questions that arise based on the previous literature. These questions are directed toward essential considerations in collaborative human-AI fact-checking and also address the larger conceptual questions regarding the impact of bias awareness on human-AI teams.

- **RQ1: What affordances and associated interface designs help raise awareness of bias in human-AI collaborative fact-checking endeavors?**
  As discussed previously, bias represents the limitations of computational systems and mirrors the shortcomings of human thought, action, and vision [15]. The presence of bias emphasizes the need to uncover and interrogate existing autonomous agents for such biases, for which human-AI partnerships can be vital [74]. In light of such biases, being able to make informed decisions when using or collaborating with AI becomes an essential facet of such partnerships. Given how overreliance on and complacency with AI negatively impacts the efficacy of human-AI partnerships [62] and limited efficacy of proposed solutions [7], it becomes imperative to understand tools that can aid the bias perception process. From the contextual standpoint of fact-checking, bias detection becomes essential due to its impact on the decision-making procedures of fact-checking and critical socio-informatic infrastructures. For example, when using a fact-checking system to verify the spread of a deadly virus as reported through online information-sharing channels, precision in reporting can play a key role in not only the overall preparedness of vital health infrastructure but can also be important for public awareness and readiness to deal with such circumstances [59]. This framing motivates the need for crafting bias awareness interventions in collaborative human-AI fact-checking, which serves as the central basis for the above question. Such interventions can not only impact the trust fostered by the AI system used but can also impact the overall trust in the fact-checking process in the face of growing distrust in media and social structures overall.

- **RQ2: What analytical perspectives can be helpful in guiding the design approaches and directions for creating bias identification interfaces?**
  For any collaborative endeavor to succeed, transparency is an essential precursor for constituting appropriate trust and reliance among actors engaged in the collaborative process [17]. Prior investigations have indicated that transparency can nurture trust and enhance cooperative outcomes in human-AI interaction [83]. Further, boosting transparency with facets of interactivity and interpretability can be an effective guiding framework for enhancing user acceptance and empowerment in human-AI collaborative contexts [81]. Thus, we ground our approach using transparency as the guiding framework aimed at supporting critically-aware trust in AI. To foster an analytical framing toward transparency, we use the definition of agent transparency grounded on the premise of situational awareness as postulated by Chen et al. (2018), which has been empirically shown to improve team effectiveness and cooperative outcomes [11, 80]. Drawing on the tenets of situational awareness in this conceptual foundation, transparency addresses the information asymmetry between humans and autonomous agents with three fundamental goals. The first goal is to clarify goals and actions, the second is the reasoning process behind the actions, and the third is to address the uncertainty associated with the actions.

Such insights can be integral to outlining the affordances associated with eliciting bias considerations for human-AI collaborative ventures [50]. Using these three foundational underpinnings of transparency paired with a participatory design-based approach, we build and expand on this existing framework of agent transparency to address how bias awareness mechanisms can be ingrained into collaborative human-AI fact-checking endeavors.

- **RQ3: How does the role of the AI (AI as a tool or AI as a teammate) impact how affordances and associated interface designs are envisioned in the human-AI collaborative fact-checking endeavor?** As human-AI collaborative pursuits flourish, it becomes critical to understand where and how autonomous entities in the team dynamic should be situated and how their position impacts team operations and outcomes [19, 49]. The role of the autonomous agent is particularly important to understand the function and impact of AI on the overall team dynamic. This point is particularly salient given that most existing studies have investigated the use and incorporation of AI in the fact-checking process as a decision-making aid [54]. Yet, a strand of emergent scholarship also investigates the impact and influence of autonomous teammates that have the capability to think, reason, and act at par with their human counterparts [56]. The distinction of the roles an autonomous agent assumes becomes crucial as we also consider trust and reliance on autonomous agents. For example, with enhanced decision-making capabilities, how trust is envisioned for an autonomous teammate may not be the same as the trust configured for an autonomous entity acting as an aid [84]. Further, as we posit the importance of understanding bias awareness mechanisms and their impact on the collaborative dynamic, it becomes essential to understand how the role of the autonomous agent affects the type of bias identification processes necessary for end-users to engage in effective collaborative human-AI fact-checking. As such, we must investigate if having the autonomous agent act as a teammate warrants the incorporation of different types of bias identification mechanisms. The autonomous agent's role can also factor into the motivation for constituting such bias awareness pipelines to ensure that human-AI symbiosis is responsibly orchestrated. Such considerations are pivotal in the case of fact-checking, as the outcome of the mixed-initiative pipeline of fact-checking has a clear socio-cultural influence on the ethics of information exchange [26].

- **RQ4: How do these affordances and associated interface design impact relationships between human and AI agents, in terms of trust and reliance?** The trust and reliance fostered in these human-AI collaborations are the most crucial dimensions of inquiry we will tackle. Trust is the fundamental factor that drives team performance and cohesion [57]. The existence of bias can play a key mediating role in trust and reliance. For example, bias can compromise the decision-making ability of autonomous agents, leading to faulty and prejudiced assessments that impair the trust and reliance humans may ascribe to their autonomous counterparts. Studies have indicated how design elements such as color and shape are fundamental interactive units embedded into the design of a user interface that can have an impact on user trust [32]. These findings motivate our goal of exploring how interface designs aimed at bias elicitation can play a critical role in trust calibration in human-AI fact-checking pursuits. Further, as stated earlier, transparency is the guiding lens upon which our approach to affordances for bias awareness is based. Studies, in this regard, have indicated how a balanced amount of transparency can factor into the trust formed [37]. This investigation highlights that the amount of information being conveyed through an interface can be critical for trust in algorithmic entities, indicating another key design consideration that can be pivotal for trust in human-AI collaborations. Given these motivating investigations, understanding how interfaces geared toward bias awareness impact trust and reliance becomes a key question not only for the contextual premise of fact-checking but also for the larger conceptual arena of human-AI collaboration.

## 4 INVESTIGATION PLAN

Each of the questions as postulated above explicate critical considerations that emerge as we consider the interplay of bias perception and overreliance, especially in the human-AI collaborative arena. Crafting critically aware human-AI collaboration will need careful consideration of each of the facets of the collaborative dynamic as articulated above. Drawing inspiration from the above questions, we propose an initial exploratory analysis as motivated and outlined below. While this investigative plan does not address in detail all the dimensions of the questions as highlighted above, our goal with the proposed exploration is to distill how the conceptual umbrella of agent transparency can provide crucial insights toward designing interactive mechanisms that can bolster critical engagement in the context of human-AI collaborative fact-checking.

While a burgeoning array of prior research details the type of biases that manifest in AI-driven systems [46], the goal of this exploration is to devise novel user-centered mechanisms to cater to ways in which latent biases manifested in AI-driven systems can be made explicit to increase the overall transparency in human-AI collaborative ventures for fact-checking. Drawing inspiration from synergistic yet conceptually different explorations investigating perceived fairness [51], we propose an inductive process drawing on participatory design methodology to address the research vision proposed. The goal of this phase is to involve end-users in the process of ideating and understanding the needs that drive how bias awareness mechanisms need to be incorporated when humans collaborate with autonomous agents [87]. The subsections below provide a detailed account of the procedure that will be followed to carry out the proposed methodological approach. Subsequently, we also provide a brief description and visual representation of the mock-ups we envision using as part of this endeavor.

### 4.1 Participatory design - Ideate, Deliberate, and Create a Proof of Concept for Bias Awareness Mechanisms

Participatory design promotes democratic visions for equitable stakeholder participation and expression, allowing them to feel involved in the ideation and conceptualization phase of technical artifacts [60]. It repositions the approach to design from being *for* users to being *with* users - ushering a new wave of design methodology rooted in the experiential wisdom of those who would directly engage with the envisioned systems. Thus, the key appeal of this approach lies in how it centers the **voice** of the users in the design pipeline [6]. Given this premise, participatory design is becoming an increasingly popular framework for conceptualizing use contexts across various domains including education, sustainability, and healthcare [14, 16, 69]. With the growth of automated systems, participatory design methods are also useful in the adoption and incorporation of conversational agents, smart assistants, and robots [18, 43]. Extending these lines of inquiry, increasingly participatory design approaches are being incorporated to address the need to reduce algorithmic opacity and engage users in creating transparent and user-centric human-AI collaborative endeavors [33, 51, 85]. Inspired by this growing research objective, we leverage participatory design as the methodological approach to our investigations. To frame our exploration, we will draw on the three foundational tenets of agent transparency. These three approaches will serve as guiding principles to shape the design activities we employ.

Participatory design is conducted using various artistic yet scientifically grounded methods and frames [70]. While this approach is highly flexible in its conceptualization and use, it involves creating a dialogue between users and designers (or researchers) to create a partnership that promotes increased awareness of the relevant use cases, and needs, and ends with prototyping, feedback, and evaluation by the users [70]. In this regard, storyboarding is a key methodology used to create this dialogue between users and researchers [44]. Storyboarding involves tapping into the user's beliefs, desires and needs by crafting a *journey* which serves as a use case and motivational guide [72]. However, the blank state of storyboarding can pose a cognitive challenge, and studies often augment storyboarding with additional probes to create a holistic experience for the user in the co-design process [48]. Thus, given the flexibility and versatility of this approach, we will use this as the main method that shapes how we conduct our workshop.

*4.1.1 Workshop Modalities:* Given time and resource constraints, we will only conduct a single workshop. Our workshop will involve approximately 10 participants and last for approximately 1.5 to 2 hours. The workshop will be conducted online to overcome restrictions imposed by the COVID-19 pandemic and also increase our geographic reach [51]. Participants will be recruited via the Prolific online recruitment tool, given its flexibility to draw participants using a wide set of filtering criteria. Given our focus on end-users, we will employ purposive sampling to induct a set of participants who have some knowledge of social media content (to create a premise for fact-checking) and have familiarity with technology (to situate the context of AI) [58] based on the filtering options available to us on the Prolific platform.

Our workshop will be guided by a set of scenarios. While fact-checking will be used as a sensitizing device to provide the contextual basis for the proposed investigation, the critical concepts here include providing deeper insights into the type of latent biases and different layers of interactivity that are motivated by the three layers of agent transparency. For this exploration, we focus on two key types of biases - (1) Data bias - the type of bias that stems from the way in which data was collected and processed (such as oversampling a particular population), and (2) Algorithmic bias (these manifest in the way in which the algorithmic processes use the data, which include the different weights and manner in which different model parameters are calibrated) [46, 76].

The workshop will be conducted as follows. First, in the introductory portion, the researchers will introduce the motivations to familiarize the participants with the context of the study. Next, in teams of 2-3, each group will work with a researcher to collaboratively ideate on the different storyboards, augmenting, critiquing, and complementing the existing probes and associated questions. The probes (i.e. the design mock-ups) will be developed in Figma (a high-fidelity prototyping software) and grounded in the three fundamental rationales associated with agent transparency [80]. In the process of engaging with the mock-ups, the researcher will ask questions to understand how the different types of biases call the different interpretive and interactive needs. Further, the participants will also be probed on how the role of the AI (tool vs teammate) affects the type of affordances desired. After this constructive deliberative process, there will be a combined final session, where the participants will be asked to share any final remarks, thoughts, and apprehensions about such systems, which will serve as the closing of the session. The workshop will be recorded using audio and video recording software. The data collected will be analyzed using thematic coding, rooted in a grounded theory approach [9]. Using the insights and inferences garnered, the researchers will embark on creating different interactive dashboards for further experimental validation.

### 4.2 Preliminary visions for design mockups

Our investigation will build on previous studies on designing a prototype that caters to the needs of end-users who interact with autonomous entities, with the goal of facilitating bias awareness. Akin to the visions of agent transparency, our focus when developing these initial prototypes is to provide users with foundational framing related to the functionality and decision-making process of automated systems comprising the critical parameters used and resources involved. Motivated by the association between transparency and interactivity, we envision these interfaces to afford users the ability to manipulate, edit and visualize changes in the underlying decision-making models to provide a holistic understanding of autonomous entities acting in unison with their human counterparts [3, 81]. Our design approach focuses on empowering users to better comprehend the different dimensions of bias and their effects in order to better calibrate their engagement with autonomous entities [35, 51, 54]. Drawing inspiration from prior investigations, our design perspectives are centered on being sound, iterative, complete, simple, and non-overwhelming [39]. We also ensure that user feedback can be actionable, reversible, and presented through incremental changes [51]. To cater to the vision of empowering the user to understand the ramifications of potential mispredictions, our prototype is equipped with a tightly coupled cycle of interactions with a user interface that explains how the machine learning model makes decisions and provides users with opportunities to make necessary corrections [39].

Our prototype's main page is designed to provide end-users with transparency in the claim identification process by displaying information such as the overall decision, accuracy, and a global explanation of the decisions made by the system. To bolster interactivity we created an iterative, sound, and complete interface as shown in Figure 1 section (A) [39, 51]. Visual aids such as pie charts, bar charts, and scatter plots are implemented to help users understand the

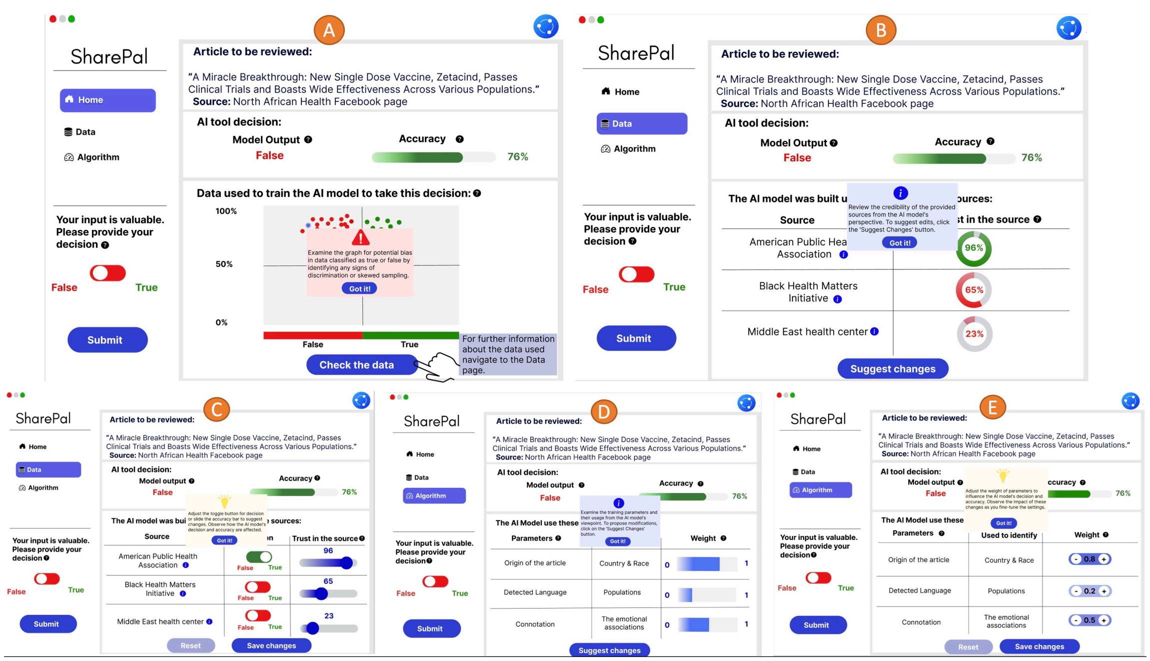


Fig. 1. The interface used in the workshop. Section (A) is the main page that contains the data, AI decision, accuracy, data used to train AI model, and icons to show more information about difficult terminologies. Section (B) outlines the data sources used for training, along with their level of trust. Section (C) represents the interface where the user can suggest changes and see the updates on the decision and accuracy. Section (D) includes the parameters and weights used in the training of the AI model. Section (E) allows users to adjust parameter weights and view real-time updates to decisions and accuracy.

data presented in the prototype better. This also caters to the vision of engaging the user with the evidence or the data used to build the models that run in the background of the automated fact-checking process. Reducing uncertainty is a fundamental pillar of agent transparency, thus components, such as tooltips, will be implemented to provide the opportunity to gain a more detailed understanding of associated decisions and actions to end-users as and when necessary as shown in Figure 1 section (B) and (D) [11, 54]. Users will have the ability to make changes to the inputs through sliders, which result in immediate model updates of decision and accuracy, and any edits can be reversed or overwritten as shown in Figure 1 section (C) and (E). The importance of resources and value distributions will also be shown, elucidating potential types of bias in the decision-making process. Additionally, users will have the option to view model predictions about relevant information and can compare information in a scatterplot based on the similarity to the currently selected data and attribute. Such visual and cognitive elements are envisioned to enhance the user's capacity to identify and interrogate the autonomous system and calibrate the way in which they engage and trust the system.

According to extant literature, these design features make the decisions made by autonomous entities more concise, interactive, and tailored to what users can process and understand [35, 51, 54]. Further, making interactive corrections enables integrating user feedback into the model and enables the user to perceive the effects of their adjustments. However, our aim with providing these initial, foundational designs is not to lock users into these set elements but to create a template aimed at igniting conversations with the goal of incorporating user perspectives more deeply and astutely into the frameworks associated with nurturing responsible human-AI teams, especially in the contextual arena of fact-checking.

## 5 CONCLUSION & POTENTIAL CONTRIBUTIONS

In the face of growing mis- and disinformation online, human-AI collaborative fact-checking offers a potentially fruitful avenue to capitalize on the powerful computational abilities of AI and the contextually-attuned competencies of humans [54]. However, these promising collaborations are potentially hindered by the presence of biases, affecting the decision-making outcomes of both parties [41]. Building on the growing body of literature around trust and reliance that influences these collaborations, we proposed a set of critical research questions that are highly relevant and contextually important to the current research landscape. Indeed, through these questions, we can better understand how bias awareness in human-AI teaming can empower end-users to collaborate and make informed decisions in human-AI joint efforts toward fact-checking with special attention paid to appropriate levels of transparency. As such, our proposed research plan capitalizes on participatory design to elicit design recommendations for bias awareness mechanisms in design interfaces. From this research, we will create high-fidelity interactive prototypes that can provide tangible and actionable insights to address the issue of bias awareness in human-AI teams for later empirical testing. In addressing our research questions and completing this proposed empirical exploration, we believe we can provide the following essential contributions:

(1) Provide a deeper understanding of the way in which transparency can be added as a design feature toward bias mitigation.
(2) Contextually, the aim will be to provide insights into the way in which bias awareness can be addressed for automated fact-checking systems.
(3) Elucidate empirically driven insights regarding how interactive interface designs aimed at bias awareness can impact human-AI team dynamics, particularly trust, and reliance.

## REFERENCES

[1] Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward "the holy grail": The continued quest to automate fact-checking. In *Computation+ Journalism Symposium,(September)*.

[2] Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. 2019. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 120–129. ISSN: 2640-3498.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[4] W Lance Bennett and Steven Livingston. 2020. *The disinformation age.* Cambridge University Press.

[5] Arpita Biswas, Marta Kolczynska, Saana Rantanen, and Polina Rozenshtein. 2020. The role of in-group bias and balanced data: A comparison of human and machine recidivism risk predictions. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies*. 97–104.

[6] Erling Björgvinsson, Pelle Ehn, and Per-Anders Hillgren. 2010. Participatory design and" democratizing innovation". In *Proceedings of the 11th Biennial participatory design conference*. 41–50.

[7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[8] Assunta Cerone, Elham Naghizade, Falk Scholer, Devi Mallal, Russell Skelton, and Damiano Spina. 2020. Watch'n'Check: Towards a social media monitoring tool to assist fact-checking experts. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 607–613.

[9] AL Chapman, M Hadfield, and CJ Chapman. 2015. Qualitative research in healthcare: an introduction to grounded theory using thematic analysis. *Journal of the Royal College of Physicians of Edinburgh* 45, 3 (2015), 201–205.

[10] Jim Q Chen and Sang M Lee. 2003. An exploratory cognitive DSS for strategic decision making. *Decision support systems* 36, 2 (2003), 147–160.

[11] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science* 19, 3 (2018), 259–282.

[12] Xinran Chen, Sei-Ching Joanna Sin, Yin-Leng Theng, and Chei Sian Lee. 2015. Why students share misinformation on social media: Motivation, gender, and study-level differences. *The journal of academic librarianship* 41, 5 (2015), 583–592.

[13] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015), e0128193.

[14] Jane Clemensen, Mette J Rothmann, Anthony C Smith, Liam J Caffery, and Dorthe B Danbjorg. 2017. Participatory design methods in telemedicine research. *Journal of telemedicine and telecare* 23, 9 (2017), 780–785.

[15] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press.

[16] Bronwyn Cumbo and Neil Selwyn. 2022. Using participatory design approaches in educational research. *International Journal of Research & Method in Education* 45, 1 (2022), 60–72.

[17] Christian Dalsgaard and Morten Flate Paulsen. 2009. Transparency in cooperative online education. *International Review of Research in Open and Distributed Learning* 10, 3 (2009).

[18] Aduén Darriba Frederiks, Johanna Renny Octavia, Cesar Vandevelde, and Jelle Saldien. 2019. Towards participatory design of social robots. In *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part II 17*. Springer, 527–535.

[19] David De Cremer and Garry Kasparov. 2021. AI should augment human intelligence, not replace it. *Harvard Business Review* 18 (2021).

[20] Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Eng. Bull.* 43, 3 (2020), 65–74.

[21] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. 2012. Why rumors spread so quickly in social networks. *Commun. ACM* 55, 6 (2012), 70–75.

[22] Oluwole Olumide Durodolu and Samuel Kelechukwu Ibenne. 2020. The fake news infodemic vs information literacy. *Library Hi Tech News* 37, 7 (2020), 13–14.

[23] Mica R Endsley. 2017. From here to autonomy: lessons learned from human–automation research. *Human factors* 59, 1 (2017), 5–27.

[24] Mica R Endsley. 2023. Supporting Human-AI Teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior* 140 (2023), 107574.

[25] Neta Ezer, Sylvain Bruni, Yang Cai, Sam J Hepenstal, Christopher A Miller, and Dylan D Schmorrow. 2019. Trust engineering for human-AI teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 322–326.

[26] Thomas J Froehlich. 2017. A Not-So-Brief Account of Current Information Ethics: The Ethics of Ignorance, Missing Information, Misinformation, Disinformation and Other Forms of Deception or Incompetence. *BiD* 39 (2017).

[27] D Graves. 2018. Understanding the promise and limits of automated fact-checking. (2018).

[28] Lucas Graves and Alexios Mantzarlis. 2020. Amid political spin and online misinformation, fact checking adapts. *The Political Quarterly* 91, 3 (2020), 585–591.

[29] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[30] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[31] Richard A Guzzo and Marcus W Dickson. 1996. Teams in organizations: Recent research on performance and effectiveness. *Annual review of psychology* 47, 1 (1996), 307–338.

[32] Florian Hawlitschek, Lars-Erik Jansen, Ewa Lux, Timm Teubner, and Christof Weinhardt. 2016. Colors and trust: The influence of user interface design on trust and reciprocity. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 590–599.

[33] Kenneth Holstein and Vincent Aleven. 2022. Designing for human–AI complementarity in K-12 education. *AI Magazine* 43, 2 (2022), 239–248.

[34] Y Linlin Huang, Kate Starbird, Mania Orand, Stephanie A Stanek, and Heather T Pedersen. 2015. Connected through crisis: Emotional proximity and the spread of misinformation online. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 969–980.

[35] Alireza Karduni, Ryan Wesslen, Sashank Santhanam, Isaac Cho, Svitlana Volkova, Dustin Arendt, Samira Shaikh, and Wenwen Dou. 2018. Can you verifi this? studying uncertainty and decision-making about misinformation using visual analytics. In *Twelfth international AAAI conference on web and social media*.

[36] Jon R Katzenbach and Douglas K Smith. 2008. *The discipline of teams*. Harvard Business Press.

[37] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2390–2395.

[38] Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* 13, 3 (2020), 795–848.

[39] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[40] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*. 1–10.

[41] David La Barbera, Kevin Roitero, and Stefano Mizzaro. 2022. A hybrid human-in-the-loop framework for fact checking. In *Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2022)*.

[42] Michael Lachmann, Guy Sella, and Eva Jablonka. 2000. On the advantages of information sharing. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 267, 1450 (2000), 1287–1293.

[43] Naja A Mack, Dekita G Moon Rembert, Robert Cummings, and Juan E Gilbert. 2019. Co-designing an intelligent conversational history tutor with children. In *Proceedings of the 18th ACM international conference on interaction design and children*. 482–487.

[44] Kim Halskov Madsen and Peter H Aiken. 1993. Experiences using cooperative interactive storyboard prototyping. *Commun. ACM* 36, 6 (1993), 57–64.

[45] Nathan J McNeese, Beau G Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 2021. Who/what is my teammate? team composition considerations in human–ai teaming. *IEEE Transactions on Human-Machine Systems* 51, 4 (2021), 288–299.

[46] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[47] Paul Mihailidis and Samantha Viotty. 2017. Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American behavioral scientist* 61, 4 (2017), 441–454.

[48] Robb Mitchell and Mie Nørgaard. 2011. Using DIY cartoon storyboards, live sketching and co-sketching to involve young and older users in participatory design. In *Proceedings of IASDR2011, the 4th World Conference on Design Research*, Vol. 31.

[49] Geoff Musick, Thomas A O'Neill, Beau G Schelble, Nathan J McNeese, and Jonn B Henke. 2021. What happens when humans believe their teammate is an AI? An investigation into humans teaming with autonomy. *Computers in Human Behavior* 122 (2021), 106852.

[50] Morten Johan Mygland, Morten Schibbye, Ilias O Pappas, and Polyxeni Vassilakopoulou. 2021. Affordances in human-chatbot interaction: a review of the literature. In *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society: 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society, I3E 2021, Galway, Ireland, September 1–3, 2021, Proceedings 20*. Springer, 3–17.

[51] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 3 (2022), 1–30.

[52] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769* (2021).

[53] David T Newman, Nathanael J Fast, and Derek J Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160 (2020), 149–167.

[54] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 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*. 189–199.

[55] Nessrine Omrani, Giorgia Rivieccio, Ugo Fiore, Francesco Schiavone, and Sergio Garcia Agreda. 2022. To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change* 181 (2022), 121763.

[56] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors* 64, 5 (2022), 904–938.

[57] Ravi Paul, John R Drake, and Huigang Liang. 2016. Global virtual team performance: The effect of coordination effectiveness, trust, and team cohesion. *IEEE Transactions on Professional Communication* 59, 3 (2016), 186–202.

[58] Neetij Rai and Bikash Thapa. 2015. A study on purposive sampling method in research. *Kathmandu: Kathmandu School of Law* 5 (2015).

[59] Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *Journal of Public Health* (2021), 1–10.

[60] Elizabeth B-N Sanders. 2002. From user-centered to participatory design approaches. In *Design and the social sciences*. CRC Press, 18–25.

[61] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. 2022. Let's Think Together! Assessing Shared Mental Models, Performance, and Trust in Human-Agent Teams. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–29.

[62] Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance, Explainable AI, Human-AI Collaboration, Human-AI Complementarity. *arXiv preprint arXiv:2302.02187* (2023).

[63] Philipp Schmidt and Felix Biessmann. 2020. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*. Springer, 431–449.

[64] Tjeerd AJ Schoonderwoerd, Emma M Van Zoelen, Karel van den Bosch, and Mark A Neerincx. 2022. Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task. *International Journal of Human-Computer Studies* 164 (2022), 102831.

[65] Arathi Sethumadhavan. 2019. Trust in artificial intelligence. *Ergonomics in Design* 27, 2 (2019), 34–34.

[66] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*. 745–750.

[67] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[68] Jake Silberg and James Manyika. 2019. Notes from the AI frontier: Tackling bias in AI (and in humans). *McKinsey Global Institute* 1, 6 (2019).

[69] Rachel Charlotte Smith and Ole Sejer Iversen. 2018. Participatory design for sustainable social change. *Design Studies* 59 (2018), 9–36.

[70] Clay Spinuzzi. 2005. The methodology of participatory design. *Technical communication* 52, 2 (2005), 163–174.

[71] Ramya Srinivasan and Ajay Chander. 2021. Biases in AI systems. *Commun. ACM* 64, 8 (2021), 44–49.

[72] Elizabeth Stowell, Teresa K O'Leary, Everlyne Kimani, Michael K Paasche-Orlow, Timothy Bickmore, and Andrea G Parker. 2020. Investigating opportunities for crowdsourcing in church-based health interventions: A participatory design study. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[73] Simone Stumpf, Lorenzo Strappelli, Subeida Ahmed, Yuri Nakao, Aisha Naseer, Giulia Del Gamba, and Daniele Regoli. 2021. Design methods for artificial intelligence fairness and transparency. (2021).

[74] Mike HM Teodorescu, Lily Morse, Yazeed Awwad, and Gerald C Kane. 2021. Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *MIS Quarterly* 45, 3 (2021).

[75] Claire Textor, Rui Zhang, Jeremy Lopez, Beau G Schelble, Nathan J McNeese, Guo Freeman, Richard Pak, Chad Tossell, and Ewart J de Visser. 2022. Exploring the Relationship Between Ethics and Trust in Human–Artificial Intelligence Teaming: A Mixed Methods Approach. *Journal of cognitive engineering and decision making* 16, 4 (2022), 252–281.

[76] James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687* (2018).

[77] Joseph E Uscinski and Ryden W Butler. 2013. The epistemology of fact checking. *Critical Review* 25, 2 (2013), 162–180.

[78] Diederick van Thiel and Willem Frederik (Fred) van Raaij. 2019. Artificial intelligence credit risk prediction: An empirical study of analytical artificial intelligence tools for credit risk prediction in a digital era. *Journal of Risk Management in Financial Institutions* 12, 3 (2019), 268–286. Publisher: Henry Stewart Publications LLP.

[79] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*. 18–22.

[80] Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. 2022. Designing transparency for effective human-AI collaboration. *Information Systems Frontiers* 24, 3 (2022), 877–895.

[81] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

[82] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. *Proceedings of the VLDB Endowment* 7, 7 (2014), 589–600.

[83] John Zerilli, Umang Bhatt, and Adrian Weller. 2022. How transparency modulates trust in artificial intelligence. *Patterns* (2022), 100455.

[84] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[85] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

[86] Xianglin Zhao, Li Chen, Yucheng Jin, and Xinzhi Zhang. 2023. Comparing button-based chatbots with webpages for presenting fact-checking results: A case study of health information. *Information Processing & Management* 60, 2 (2023), 103203.

[87] Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric PS Baumer, and Min Kyung Lee. 2022. Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–4.