Chatbots as decision aids: investigating reliance in Human-Chatbot collaboration vs. Human-Human collaboration

CLÉLIE AMIOT, Université de Lorraine, CNRS, Inria, Loria, France
FRANÇOIS CHAROY, Université de Lorraine, CNRS, Inria, Loria, France
JÉRÔME DINET, Université de Lorraine, CNRS, Inria, Loria, France

Chatbots have the potential to revolutionize the way we live and work, but what impact do they have on our decision-making? We investigated how people rely on advice from chatbots compared to advice from humans. We found that participants were more likely to follow the advice of a chatbot, with 64% compliance compared to 31% with a human. Our study also showed that participants were more engaged when working with a human assistant. These findings have important implications for the implementation of conversational agents in various settings.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: AI assistant, chatbot, conversational interface, reliance, automation bias, collaboration

ACM Reference Format:
Clélie Amiot, François Charoy, and Jérôme Dinet. 2023. Chatbots as decision aids: investigating reliance in Human-Chatbot collaboration vs. Human-Human collaboration. In *Workshop on Trust and Reliance in AI-Assisted Tasks, at CHI 2023, June XX–XX, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent developments in chatbot technology, particularly the release of ChatGPT[1], have significantly increased their use as decision aids. However, it is crucial to examine whether relying on chatbots for decision-making is always justified, given the documented bias towards AI advice in the literature [7]. We seek to explore whether this bias extends to chatbots, given their more substantial social presence and greater similarity to humans than classical AI advice systems and less jarring differences than robots. The study of chatbots offers a unique opportunity to analyze natural language interactions beyond the participants' decisions. Our study aims to investigate the impact of solicited versus unsolicited advice on user behavior and decision-making, as well as identify and address any measurable differences.

2 RELATED WORK

A substantial amount of works in the literature show a bias toward AI advice [7]; when given a choice between a Human or an AI system, people tend to prefer the AI's advice. This is referred to in the literature as automation bias or algorithmic appreciation. This bias contributes to over-reliance on AI-based systems, that is, people relying more on a system than they should, considering its performance and forsaking other contradictory information sources [14].

[1] https://openai.com/blog/chatgpt/

Work on cockpit automated aids [17] showed that people over-relied on the aid, which caused commission and omission errors when the aid failed, with respective drops in accuracy of 65% on false positives and 38% on false negatives. Similarly, a study on judicial decision aid [5] showed a rate of 79% agreement with an expert system providing wrong advice for law cases ruling even though a correct attorney analysis was provided concurrently.

The defining study on the concept of algorithm appreciation [13] revealed that people adapted their reliance on a given recommendation more when it was labeled as the result of an algorithmic process compared to being from other people. This was true whether the task necessitated technical skills, like estimating someone's weight from a picture, or social skills, like forecasting the future rank of a song in a top 100 list or even forecasting attraction between two people.

A study on senior executives[9] showed that they were more likely to invest if investment advice was indicated as coming from AI than from a managerial team. The executives with the AI advisor also felt that their decision quality was higher and had greater trust in it than the group of executives advised by humans.

Finally, a study on labeling conflict resolution [2] showed that allowing participants to automate the conflict resolution leads to high rates of inappropriate reliance ranging from 73% to 84%.

Literature investigating reliance bias in AI is scarcer regarding advice delivered by a conversational interface. One study [19] showed possible automation-induced complacency with conversational interfaces. However, few works directly compare compliance with conversational AI to compliance with human assistants.

We want to fill this gap and find it pertinent as chatbots are more easily anthropomorphized and have a more social presence than other AI systems, thanks to the use of natural language for communication, and might be less likely to benefit from automation bias [16].

3 EXPERIMENTAL DESIGN

We designed an experiment with two conditions (between-group design) to investigate the difference in behavior when advised and contradicted by a chatbot or human assistant. The experiment consists of two questionnaires and 20 questions on wilderness survival asked by either a cognitive assistant or a human assistant on an instant messaging platform. The goal is to measure if the participants paired with the chatbot assistant have different behavior (following hints, types of messages, answer to questionnaires) than those with the human assistant when the only thing differentiating the two is their presentations.

Table 1. Demographic informations of recruited participants (M=Male and F=Female)

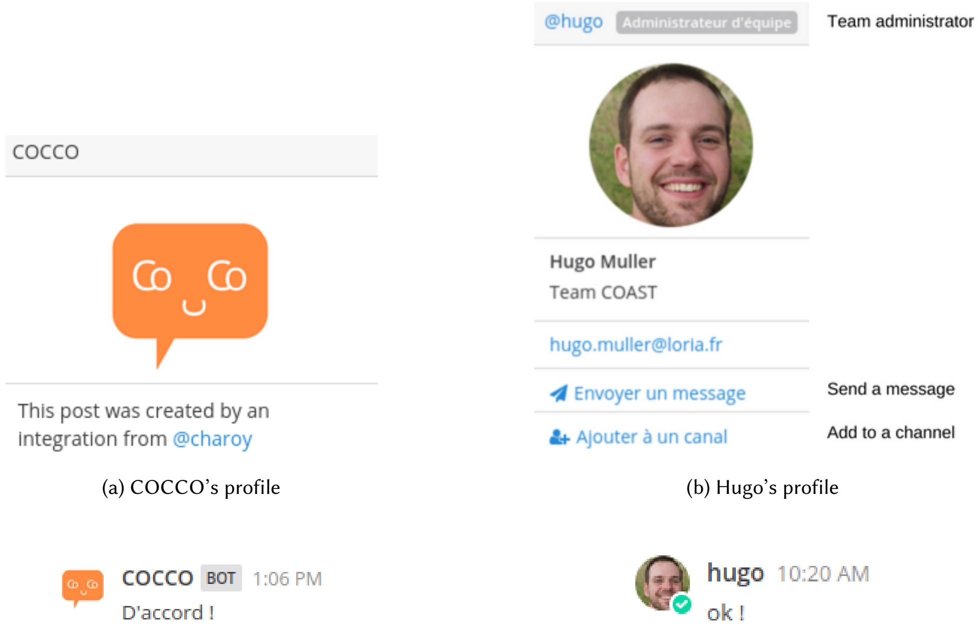
	C-group		H-group	
Age	M	F	M	F
18-25	2	3	3	1
26-35	4	4	4	5
36-45	3	4	3	4
46-55	1	1	1	2
55+	2	0	1	0
Total	12	12	12	12

We recruited 48 participants for this experiment by mailing list (see Table 1). They were compensated with a 5 euros gift check for their participation. We conducted the experiment entirely online and assigned participants a random

4-digits ID during the recruitment phase to log into the testing platforms and keep their data pseudonymized. We conducted the experiment in French, which all participants speak fluently. In this paper, all the data and labels will be translated into English. The produced data and analyzes are accessible on Zenodo [1].

This experiment protocol was approved by the Operational Committee for the Evaluation of Legal and Ethical Risks (COERLE) of the National Institute for Research in Computer Science and Automation (Inria)[2] Avis 2020-25.

3.1 Assistants



This post was created by an integration from @charoy

(a) COCCO's profile

(b) Hugo's profile

D'accord !

(c) Example message by COCCO

ok !

(d) Example message by Hugo

Fig. 1. Differences in presentation between COCCO and Hugo

Participants were asked 20 questions by an assistant, which depending on the group, was either presented as a conversational assistant or as a human experimenter. The human researcher did not use their own identity to avoid the participants having prior familiarity with them from the recruiting period [11]. No chatbot was implemented for this experiment. We used a Wizard of Oz set-up, where the experimenter sent the messages via the messaging platform API.

In the rest of the paper, we denote the Wizard of Oz assistant as the cognitive assistant or COCCO, and the human presenting one as the human assistant or Hugo. Participants paired with COCCO will be referred to as the C-group and those paired with Hugo as H-group.

The differences between the two conditions are mainly in the presentation of the assistant, chatbot or human. The capacities and interventions of the assistants are the same, both assistants followed the same script. We identified the following differences between the two experimental conditions :

- Profile picture: COCCO icon was a logo while Hugo had a smiling picture (IA-generated and edited)

[2] https://www.inria.fr/en/operational-committee-assesment-legal-and-ethical-risks

- Profile: We completed the profile of Hugo to show his status as a researcher of the team conducting the experiment (Figure 1b), while we kept COCCO's blank (Figure 1a)
- Status: Hugo had an online indicator next to his name (Figure 1d); COCCO had a bot badge automatically assigned by the messaging app (Figure 1c)
- Typing: Aside from the questions, which were copy-pasted, Hugo's messages were manually typed and, as such, accompanied by a typing indicator. COCCO's messages were sent instantly. This led to a difference in response speed between COCCO (2±2s) and Hugo (12±7s).
- Speech variations: Hugo's answers were more varied than COCCO's, whose messages were picked from a small pool of alternate phrasings. Minor typographic errors from Hugo were also kept, while more important ones were edited in the app.
- Structural messages: COCCO systematically announced the following question and when he was going to give a hint; Hugo only did it occasionally when it was natural.

3.2 Questions

The questions are inspired by survival scenarios team-building exercises, like the NASA moon survival problem [8] and desert survival situation [12] (Figure 2). The goal was to have questions easily understandable for any participant but on a subject where people rarely have expert knowledge. For each question, participants had to choose between three possible answers. They were instructed beforehand that they could ask for a hint before answering. Questions were asked in random order and randomly assigned to one of four modalities:

- *simple*: 5 questions were asked with no additional prompting from the assistant
- *hint*: 5 questions were asked with a hint
- *contradiction*: 5 questions were asked, followed by the assistant contradicting the answer given and allowing the participant to change his answer
- *contradiction and hint*: 5 questions were asked, followed by the assistant contradicting the answer given with a hint and allowing the participant to change his answer

We kept the same modalities regardless of the participant's answer (two hints were prepared for each question so the assistant could always contradict the participant's answer). However, they were dropped when the participant spontaneously asked for a hint before giving his first answer to avoid contradicting them when they were already following the assistant advice.

We used the open-source messaging platform Mattermost[3] for this section of the experiment. We chose it for its commonalities with most professional messaging platforms (Slack, Microsoft Teams) and easy chatbot integration.

3.3 Questionnaires

We asked participants to fill two online questionnaires, before and after the questions. The first questionnaire collected the informed consent of the participants and demographics to assign them an experimental group. In the second questionnaire, we asked participants to answer four ten-level Likert items about their estimated performance on the questions, the influence exerted by the assistant on their answers, and how helpful and tailored to their previous answers this influence was. We also asked them to provide feedback on the experiment in an open-ended question.

[3] https://mattermost.org/

You are lost in the wilderness and you are hungry. The best way to recognize which plants are safe to eat (at least the ones you don't know) is to:

(a) Question

1. Place a small piece of the plant on your lip a few minutes before eating it a little.
2. Avoid brightly colored plants.
3. Try small quantities of what the birds are eating.

(b) Possible answers

A. Birds have a different digestive system from ours.
B. Some plants can be lethal on contact.

(c) Hints

Fig. 2. Example of question, answers and hints given to the participants

4 RESULTS

4.1 Messages

After the experiments, we exported the transcripts of conversations between participants and assistants from the messaging platform. We annotated each of the 5809 messages with the role of the sender (COCCO, Hugo, C-participant, or H-participant), the number of the current question (or "x" for non-question related messages), and one label describing the nature of the message. 16 labels were used for the assistants' messages and 11 for the participants' messages (Table 2).

Some messages were given multiple labels. For example, the message *"I think maybe answer c"* sent by a participant after being contradicted was annotated with the labels "new answer" and "doubt". On the other hand, a message where the participant chose an answer in contradiction to a hint given by the assistant was annotated with "contradictory" and "answer".

The table only shows the number of messages tagged with the labels, further data processing was needed to get the number of times each interaction happened. For example, Table 2 does not necessarily show that 4 participants in the C-group said goodbye to COCCO: all of those messages could have been sent by the same participant. For the statistical analysis, labels are grouped by question to avoid counting interactions divided into multiple messages more than once.

In Table 2, we can see that H-group participants were more likely than C-group participants to say goodbye to the assistant (39 vs. 4), ask about the experiment (24 vs. 4), ask for clarification on the question (23 vs. 8), give a justification to their answer (75 vs. 4) and express doubt on their answer (61 vs. 15). On the other hand, C-group participants asked more frequently for hints (56 vs. 18).

We can also see differences in frequency for the tags "confirmation" and "new answer". They indicate a different rate of answer change after being contradicted and will be examined in more detail in section 4.3.

Table 2. Description and frequency of each label describing the participant's messages (C=C-group and H=H-group)

Labels	C	H
Ready *The participant signals that they are ready to start the experiment.*	25	30
Farewell *The participant says goodbye to the assistant.*	4	39
Technical question *The participant asks a question on the experiment.*	4	24
Answer *The participant chooses one of the three answer propositions.*	480	487
Contradictory *The participant gives an answer in contradiction with a previously given hint. (Supplementary label used with "answer" or "new answer".)*	21	22
Demand for clarification *The participant asks for clarification on a question.*	8	23
Demand for hint *The participant asks for a hint.*	56	18
Confirmation *The participant confirms their answer.*	76	156
Justification *The participant provides a justification for their chosen answer.*	4	75
Doubt *The participant expresses doubt on their answer.*	15	61
New answer *The participant gives a new answer.*	131	66

4.2 Messages length and response time

Messages sent by H-group participants were on average significantly longer than messages sent by C-group participants (4.7±5.6 words vs. 2.1±2.6 words) according to a Wilcoxen rank-sum test ($p=2.10^{-16}$).

We found a statistically significant difference in the response time to the assistants' messages between the two groups (Wilcoxen rank-sum test, $p=3.10^{-8}$). H-group and C-group participants took an average of 30 seconds (sd=29) and 23 seconds (sd=25) respectively to answer the assistants' messages.

4.3 Answer changes

For every question where a participant was contradicted, we whether the participant changed their answer and whether they asked for a hint, were given one by the assistant, or had none. We then computed the answer change rate for each group and hint modality (Figure 3).

C-group participants are likelier than H-group participants to change their answer after being contradicted by their assistant whether they ask for a hint, are automatically given one by the assistant, or receive no hint at all.

In general, C-group participants change their answer 63.7% of the time (n=204), while H-group participants change their answer only 31% of the time (n=200).

After asking for a hint, C-group participants change their answer 83.3% of the time (n=12), while H-group participants change their answer only half the time (n=6). If the hint is given automatically by the assistant, C-group participants change their answer 72.6% of the time (n=106) and H-group participants 47% of the time (n=100). If participants are contradicted but given no hint, C-group participants change their answer half the time (n=86) and H-group participants only 12.8% of the time (n=94).

In short, participants prefer following a chatbot assistant's advice to a human assistant's, even when they have explicitly solicited this advice. Additionally, participants will still follow COCCO's guidance half the time when they

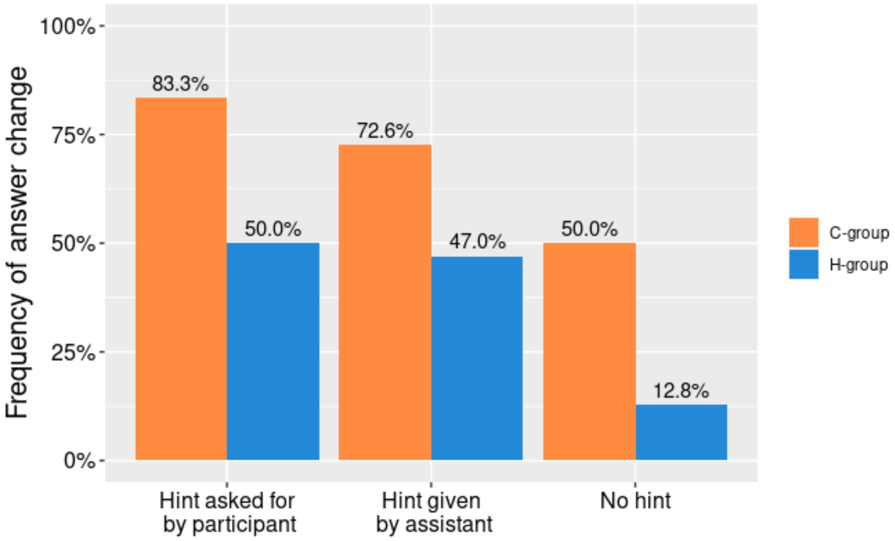


Fig. 3. Frequency of answer change by group and hint modality

are contradicted but provided no hint. This shows the participants' willingness to follow a chatbot's advice even when it is imprecise.

4.4 Questionnaires

We compared the answers of each group to the four ten-level Likert items of the post-experiment questionnaire :

1. Helpfulness of assistant: *The assistant's help allowed me to obtain a better score.*
2. Perceived success: *I answered the questions well and obtained a good score.*
3. Influence of assistant: *The assistant's interventions influenced my answers.*
4. Adaptation of assistant: *The assistant adapted his interventions according to my answers and my score.*

Using Pearson's Chi-squared test (Table 3), we found no statistically significant difference between how C-group participants and H-group participants rated their success, the influence of the assistant, and its helpfulness. However, participants in the H-group rated the adaptability of their assistant significantly higher than C-group participants (by almost 2 points).

Table 3. Comparison of answers to the questionnaire between C-group and H-group using Pearson's Chi-squared test

Question	C-group	H-group	p-value
1. Helpfulness of assistant	m=6.67	m=6.75	p=.90
2. Perceived success	m=5.375	m=6	p=.71
3. Influence of assistant	m=7.29	m=6.46	p=.40
4. Adaptation of assistant	m=5.375	m=7.21	p=.035

Table 4. Pearson correlation matrix for the post-experimental questionnaire Likert scale and answer change rates for C-group participants (n=24)

	1	2	3	4	5
1. Helpfulness of COCCO	-				
2. Perceived success	-0.28	-			
3. Influence of COCCO	0.77*	-0.32	-		
4. Adaptation of COCCO	0.33	-0.35	0.22	-	
5. Rate of answer change	0.53*	-0.56*	0.60*	0.38	-

*p<0.05

Table 5. Pearson correlation matrix for the post-experimental questionnaire Likert scale and answer change rates for H-group participants (n=24)

	1	2	3	4	5
1. Helpfulness of Hugo	-				
2. Perceived success	0.01	-			
3. Influence of Hugo	0.80*	-0.05	-		
4. Adaptation of Hugo	0.39	0.25	0.33	-	
5. Rate of answer change	0.66*	-0.16	0.65*	0.39	-

*p<0.05

We then looked at the correlation between each Likert item and the answer change rate using Pearson correlation (Table 4 and Table 5). For both groups, the ratings of the assistant's helpfulness and influence and the participants' answer change rate were positively correlated together. Participants who thought the assistant was helpful also thought they were influenced by its interventions and changed their answers to follow the assistant's advice.

Contrary to H-group participants, C-group participants' answer change rate was inversely correlated with perceived performance. Thus, participants that followed COCCO's advice more often felt that it negatively impacted their score at the end of the experiment, but H-group participants did not.

Finally, we analyzed thematically the comments left by the participants in the open-ended question. Five participants chose to leave no comment; the remaining comments were segmented into 105 items for each distinct idea present. Similar ideas were regrouped into 19 sub-themes, themselves regrouped into 5 main themes: the experiment, the questions, the participant's performance, the assistant, and the assistant's interventions.

The principal takeaways from this thematic analysis are :

- A third of the participants (equally in C-group and H-group) observed that the assistant might not be leading them to the correct answers.
- C-group participants made more remarks on the problems they encountered and what they thought could be improved (8 vs. 2).
- More C-group participants commented on the assistant, its interventions, and how they were perceived than H-group participants (29 vs. 13).
- Overall, C-group participants gave more feedback on the experiment and addressed more points than H-group participants (65 vs. 40).
- C-group participants commented on the quality of the software used, while H-group participants commented on the quality of the conversation.

- We also note that of the 11 participants indicating that they thought they lacked the necessary knowledge for the questions, 9 of them are women.

5 DISCUSSION

We observe a significative difference in the likelihood of being persuaded by an assistant to change one's answer depending on the assistant's nature: it is more likely for someone to change their answer when prompted by a conversational assistant than by a human. Even when given no new information, participants will still follow the guidance of a chatbot half the time. This reliance is also evidenced by C-group participants asking for hints three times as much as H-group participants, even though they felt it negatively impacted their performance.

H-group participants have been less willing to accept the advice given by the assistant and have instead chosen to put more thought into their answers. This is supported by their longer response time and a higher rate of clarification requests. While H-group participants were less likely to follow Hugo's advice, they also often provided justifications for their responses, a behavior that the C-group did not reflect.

We also note that participants from both groups felt equally manipulated by the assistant according to the questionnaire ending remarks. This supports that C-group participants felt more complacent about the validity of their answers as they followed the assistant's guidance despite feeling that he was manipulating them. In contrast, H-group participants integrated the human assistant more into their reflection (4 times more expressions of doubt) even if they did not trust his advice.

Finally, we saw that participants felt that Hugo tended to adapt his advice more than COCCO. Both assistants followed the same script and did not deviate from it. This shows the participants' expectation that chatbots act rigidly, an expectation often informed by past interactions with other chatbots. It suggests that when chatbots can act more flexibly in the future, users might need to be familiarized with the full extent of chatbots' capacities.

Those findings have implications for integrating conversational agents in a collaborative work environment. They are consistent with the literature on automation bias and especially [18] findings on artificial agent assistance for moral decisions. They observed that human collaborators lacked a sense of responsibility for the agent's decisions, even when they were supposed to act as supervisors.

This increased reliance on conversational assistants could benefit early-on implementation and adoption. However, apparent trust in a new conversational assistant may not mean there is a genuine collaboration with humans. Other studies have shown that overreliance on AI can lead to decreased trust if the AI makes an error and fails to meet too high expectations [4, 6, 15] and that early-on errors could lead to permanent decreases in reliance [10]. Methods to prevent overreliance that could benefit conversational assistants are negative framing, presenting the AI's limits and possible errors before usage [6], and cognitive forcing methods, such as time delays or demanding a preliminary decision before providing advice [3].

Based on this experiment's findings, we propose that overreliance could be monitored by measuring the rate of messages concerning the collaborator's decision process and doubts. Additionally, appropriate reliance might be improved with negotiation-forcing methods that would incite collaborators to exchange with the chatbot on its advice by finding a compromise or asking for more information. In contrast to other systems, those measures depend on the usage of natural language, which is specific to chatbots.

Limits

We acknowledge that having the assistant both ask questions and provide hints might affect how participants perceived and trusted them. However, having two separate entities asking questions and providing help might have confused the participants.

We recognize that the difference in answer change rate between both groups might have partially been explainable by participant response bias, more precisely, demand characteristics. Though the experimental conditions for every participant were the same, the ones assisted by COCCO might have expected the experiment to be about the cognitive assistant and its functioning and that the expected behavior, the "right" behavior, was to follow COCCO's advice. However, this hypothesis can be disproven because C-group participants who followed COCCO's advice felt that it negatively impacted their performance. C-group participants were also more expressive in the open-ended question, providing feedback on the assistant and experiment. This could be interpreted as taking a more complacent role akin to an outside observer whose goal is to evaluate COCCO. In contrast, the participants in H-group had more reasons to believe that their behavior was the subject of study and, as such, were more reluctant to accept and trust the advice given by the human assistant.

Additionally, experimenting with more important stakes might have changed how the participants engaged with the experiment. However, we believe that this is still representative of real-world situations where the implementation of a new tool is only tested voluntarily before moving on to a more large-scale distribution if the feedback is positive.

6 CONCLUSION

We set out to identify the differences between how people relied on chatbots and humans. Our results show a clear bias toward AI: people advised by a conversational assistant followed its advice much more than a human assistant's, whether requested or not. However, our qualitative analysis also highlighted other measures indicating that this apparent trust might be superficial. Therefore, we recommend the usage of additional indicators when looking at the reliance on a new conversational assistant.

As chatbots get more sophisticated and adaptable, the question is how do we ensure that the user both makes use of the most advanced services of a cognitive service but at the same time trusts it only reasonably as they would a comparable human expert. One salient example is ChatGPT. This chatbot has gained massive traction in the last few months thanks to its verbal fluency and ability to handle various topics. However, its high fluency allows it to eloquently present false and nonsensical claims that are difficult to notice on a surface reading. Future works should look at how conversational agents can be designed to elicit more in-depth reasoning from their users while still maintaining appropriate reliance.

REFERENCES

[1] Clélie Amiot. 2022. *COCCO dataset and analyses.* https://doi.org/10.5281/zenodo.6669635

[2] Michelle Brachman, Zahra Ashktorab, Michael Desmond, Evelyn Duesterwald, Casey Dugan, Narendra Nath Joshi, Qian Pan, and Aabhas Sharma. 2022. Reliance and Automation for Human-AI Collaborative Data Labeling Conflict Resolution. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.

[3] 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.

[4] Ewart J de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman. 2012. The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56. Sage Publications Sage CA: Los Angeles, CA, 263–267.

[5] Jaap J Dijkstra. 1999. User agreement with incorrect expert system advice. *Behaviour & Information Technology* 18, 6 (1999), 399–411.

[6] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human factors* 44, 1 (2002), 79–94.

[7] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.

[8] Jay Hall and Wilfred Harvey Watson. 1970. The effects of a normative intervention on group decision-making performance. *Human relations* 23, 4 (1970), 299–317.

[9] Christoph Keding and Philip Meissner. 2021. Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. *Technological Forecasting and Social Change* 171 (2021), 120970.

[10] Antino Kim, Mochen Yang, and Jingjng Zhang. 2020. When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms. *Late Errors on Users' Reliance on Algorithms (July 2020)* (2020).

[11] Sherrie YX Komiak and Izak Benbasat. 2006. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS quarterly* (2006), 941–960.

[12] J Clayton Lafferty and Alonzo W Pond. 1974. *The desert survival situation: A group decision making experience for examining and increasing individual and team effectiveness*. Human Synergistics.

[13] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[14] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[15] Vlad L Pop, Alex Shrewsbury, and Francis T Durso. 2015. Individual differences in the calibration of trust in automation. *Human factors* 57, 4 (2015), 545–556.

[16] Anna-Maria Seeger and Armin Heinzl. 2018. Human versus machine: Contingency factors of anthropomorphism as a trust-inducing design strategy for conversational agents. In *Information systems and neuroscience*. Springer, 129–139.

[17] Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006.

[18] Jasper Van Der Waa, Sabine Verdult, Karel Van Den Bosch, Jurriaan Van Diggelen, Tjalling Haije, Birgit Van Der Stigchel, and Ioana Cocu. 2021. Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations. *Frontiers in Robotics and AI* 8 (2021).

[19] Erin Zaroukian, Jonathan Z Bakdash, Alun Preece, and Will Webberley. 2017. Automation bias with a conversational interface: User confirmation of misparsed information. In *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 1–3.

Received 23 February 2023