Whom Do We Trust ?: A Comparative Study on Reliance between Chatbot and Human Assistance

Clélie Amiot, François Charoy, Jérôme Dinet

Whom Do We Trust ?: A Comparative Study on Reliance between Chatbot and Human Assistance

Clélie Amiot[1*], François Charoy[1] and Jérome Dinet[2]

[1*]LORIA, Université de Lorraine, CNRS, Inria, Campus Scientifique, Vandoeuvre-lès-Nancy, 54506, France.

[2]2LPN, Université de Lorraine, Nancy, 54000, France.

*Corresponding author(s). E-mail(s): clelie.amiot@loria.fr;
Contributing authors: francois.charoy@loria.fr;
jerome.dinet@univ-lorraine.fr;

Abstract

Conversational agents, such as chatbots, can potentially improve professional lives by streamlining existing workflows. In a recent study, 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 participants were more engaged when working with a human assistant. These findings have important implications for implementing conversational agents in various settings.

Keywords: AI assistant, chatbot, over-reliance, automation bias, decision-making

The rise of Large Language Model and more precisely ChatGPT has changed the way we consider chatbots. In the last seven years [1], we have witnessed a rise in the availability of these cognitive assistants on all kinds of devices and websites. They have gone from simple FAQ-style chatbots to intelligent assistants supporting decision-making by providing contextually pertinent information. They are now ubiquitous, lying in our pockets like Siri, waiting to be summoned to set up a timer or popping abruptly asking if we need help

while browsing a commercial website. More recently, they promise to help us in more critical tasks like healthcare [2, 3], civil security [4], and other collaborative endeavours [5]. Those applications have little allowance for problems, as failures can lead to damaging consequences for their users. Therefore, it is essential to understand how users rely on chatbots to adapt their designs and facilitate more complex collaborations. The aim of our research is to explore the distinctions that arise when individuals interact with a chatbot as opposed to a human agent in the context of decision-making. Specifically, we seek to elucidate the extent to which users rely on and comply with advice from each source.

This research focuses on understanding the differences in attitudes and decision-making when users interact with human advisors versus conversational assistants or chatbots. The literature has shown a pattern of bias toward AI advice in decision-making [6]. It is crucial to examine the role of social presence in this context, as chatbots have a stronger social presence than classical AI advice systems.

To address the research gap, the study will have two main objectives:

1. Evaluate the difference in user behavior and decision-making when receiving advice from a human advisor or a chatbot, whether the advice is solicited or unsolicited.

2. Investigate the reasons behind the observed differences, if any, and propose potential solutions to overcome or mitigate these differences.

To achieve these objectives, the study will involve a controlled experimental setup where participants are randomly assigned to receive advice from either a human or a chatbot. The advice provided will be either solicited or unsolicited, and the decision-making outcomes will be analyzed to determine any differences in the participants' behavior.

The experiment will incorporate various measures, such as self-report questionnaires, to assess the participants' trust, perceived expertise, and perceived social presence of the advice source. Additionally, behavioral data, such as response time, choice consistency, and natural language interactions, will be collected to understand the impact of advice source on decision-making.

Upon analyzing the results, the study aims to reveal the underlying reasons for any observed differences in attitudes and decision-making between human and chatbot advice sources. Potential factors that may contribute to these differences include the perceived trustworthiness, expertise, and social presence of the advice source, as well as any inherent biases or preferences of the participants.

Based on the findings, we will suggest solutions to overcome or mitigate the identified differences, such as enhancing the chatbots' social presence, improving their perceived trustworthiness and expertise, or addressing user biases through education and awareness initiatives.

To this end, we have designed an experiment to compare users' attitudes and reliance toward assistants based on their nature as human or conversational agents. We followed a wizard of Oz approach and a between-subjects design. We had 48 participants interact with either a human experimenter or a simulated chatbot. Only their presentation differentiated the two conditions to introduce no confounding factors like assistants' abilities. Our results show a clear bias toward chatbot assistance compared to human assistance.

The paper's first section will present the current work on AI advice reliance and automation bias. Then we will describe in detail the design of our experiment. Next, in section 4, we will present the experiment's results and our analyses. Finally, section 5 discusses the results' potential consequences for conversational agents' design.

1 Related work

A substantial amount of works in the literature show a bias toward AI advice [6]; when given a choice between their own opinion or an AI system's, people tend to prefer the AI's advice. They trust the AI system's opinion more than their own. This is referred to in the literature as automation bias or algorithmic appreciation. This bias contributes to over-reliance on AI-based systems. People rely more on a system than they should, considering its performance and forsaking other contradictory information sources [7].

Work on cockpit automated aids [8] showed that people over-relied on the aid and caused commission and omission errors when the aid failed, with respective drops in the accuracy of 65% on false positives and 38% on false negatives. Similarly, a study on judicial decision aid [9] 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 [10] revealed that people adapted their reliance on a given recommendation more when it was labelled 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[11] 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 labelling conflict resolution [12] 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 [13] showed possible

automation-induced complacency with conversational interfaces. However, few works directly compare compliance with conversational AI to compliance with human assistants.

Research has shown that chatbots, as compared to other artificial intelligence (AI) systems, possess a greater degree of anthropomorphism and social presence. By using natural language for communication, chatbots may also reduce the likelihood of producing automation bias, a phenomenon where individuals place excessive trust in automated systems. In contrast to studies examining reliance on embodied robots, the chat medium affords a greater degree of similarity to humans, potentially minimizing feelings of uncanniness.

Our study aims to investigate the degree of reliance on chatbots, particularly in situations where automation bias may occur. We designed a wilderness survival-related task, which employs a between-subjects design where participants receive assistance from either a human or a conversational AI. Both the human and AI assistants have similar capabilities, with only minor differences in presentation.

2 Experimental design

As stated before, we want to understand the difference in attitude when a user has to decide something and receive solicited or unsolicited advice from a human or a conversational agent. More precisely, we want to evaluate their influence differences on users' thought processes and decisions.

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 (Figure 1). The experiment consists of two questionnaires and 20 questions on wilderness survival asked by either a conversational 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.

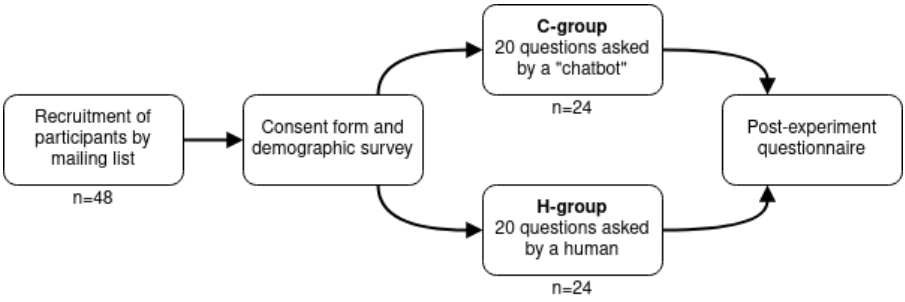


Fig. 1: Experimental design

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. We assigned participants a random 4-digit 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, we will translate all the data and labels into English.

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

This experiment protocol was approved by the legal and ethical committee of our research institute.

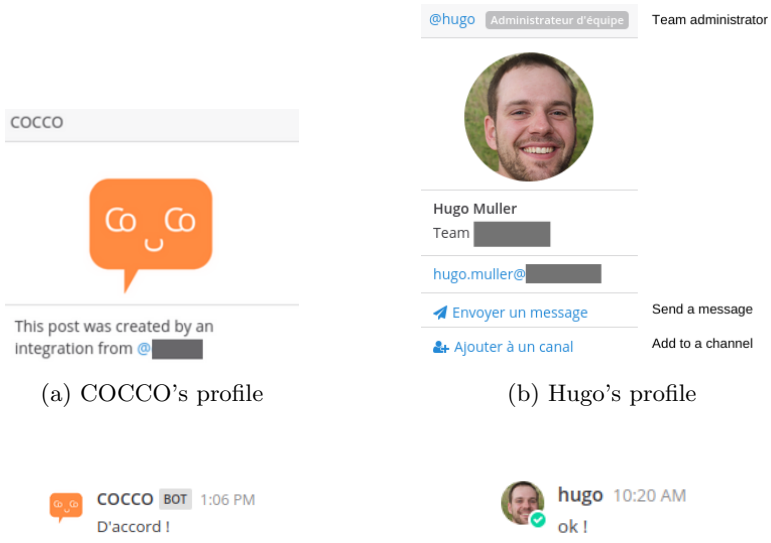
2.1 Assistants

To evaluate behavior differences, we had participants answer 20 questions asked by an assistant. Depending on the group, the assistant 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 [14]. 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 conversational 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, as we are only interested in the difference in the participants' attitudes when the assistants' presentations vary. 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)



(a) COCCO's profile

(b) Hugo's profile

(c) Example message by COCCO

(d) Example message by Hugo

Fig. 2: Differences in presentation between COCCO and Hugo

- Profile: We completed the profile of Hugo to show his status as a researcher of the team experimenting (Figure 2b), while we kept COCCO's blank (Figure 2a)
- Status: Hugo had an online indicator next to his name (Figure 2d); COCCO had a bot badge automatically assigned by the messaging app (Figure 2c)
- 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 would give a hint; Hugo only did it occasionally when it was natural.

2.2 Questions

The questions are inspired by survival scenarios team-building exercises, like the NASA moon survival problem [15] and desert survival situation [16]. 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 (Figure 3). 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[1] for this section of the experiment. We chose it for its commonalities with most professional messaging platforms (Slack, Microsoft Teams) and easy chatbot integration.

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. 3: Example of question, answers and hints given to the participants

[1] https://mattermost.org/

2.3 Questionnaires

In addition to the 20 questions, we requested that participants complete two online questionnaires both prior to and after the questioning session. The first questionnaire aimed to collect informed consent from participants as well as demographic information to categorize them into experimental groups. The second questionnaire consisted of four ten-level Likert items, which required participants to rate their perceived performance on the questions, the extent to which the assistant influenced their responses, and the helpfulness and specificity of this influence based on their prior answers. An open-ended question also prompted participants to provide feedback on the experiment.

3 Results

The experiments produced a lot of data, including the messages exchanged along with metadata (timestamps and the answers to the questionnaires). In the following, we analyze the different ways in which participants' behaviors vary between the two experimental groups. We annotated and analyzed the nature of the messages, as well as their timing and length. We also measured the number of changes in the decision of the participants in reaction to the assistant's inputs. Then we analysed the answers to the post-experiment questionnaire.

3.1 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 (Tables 2 and 3) 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.

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 contradicting a hint given by the assistant was annotated with "contradictory" and "answer".

The tables only show 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 3 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 the difference in the protocol followed by the two types of assistant.

Table 2: Description and frequency of each label describing the assistant's messages (C=COCCO and H=Hugo)

Labels	C	H
Introduction *The assistant presents himself/itself.*	24	24
Explanation *The assistant explains the experiment.*	24	24
Waiting *The assistant signals that he/it is waiting for a message from the participant to start the experiment.*	0	28
Transition *The assistant signals that he/it is moving on to the next question.*	457	103
Half *The assistant signals that half of the questions are done.*	24	19
End *The assistant announces that all the questions are done and gives the link for the post-experiment questionnaire.*	24	48
Farewell *The assistant says goodbye to the participant.*	0	24
Technical information *The assistant helps the participant connect to the right messaging channel and access the questionnaires.*	1	27
Error *The assistant informs the participant that he cannot answer their question and/or ask them to reformulate their message.*	17	1
Question *The assistant asks one of the 20 questions and gives three possible answers.*	481	480
Contradiction *The assistant challenges the answer given by the participant and asks them if they want to confirm it.*	204	208
Validation *The assistant confirms that he/it understood the participant's message and/or registered their answer.*	545	384
Hint proposition *The assistant asks the participant if they want a hint. (Given after the participant has been inactive for more than one minute.)*	1	13
Hint notification *The assistant signals that he/it is going to give a hint.*	265	2
Hint *The assistant give a hint to the participant.*	281	227
Clarification *The assistant provides clarification on the question after a request from the participant.*	2	22

Table 3: 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

- There is no message sent by COCCO labelled "waiting" because those were sent in the same message as those explaining the experiment. Likewise, COCCO's "farewell" message was grouped with its "end" message, while Hugo's "end" message was split in two.
- As mentioned in section 2.1, COCCO used transitions more often than Hugo (respectively 457 and 103 times), as well as validation messages (545 and 384).
- A network error caused COCCO to send one question twice, for 481 questions, instead of the theoretical 480 (20 questions for 24 participants).
- The differences in the number of messages labelled "technical information", "error", "validation", "hint", and "clarification" are consequences of the participants' messages reflecting their attitudes.

Regarding the participants, 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 3.3.

3.2 Messages length and response time

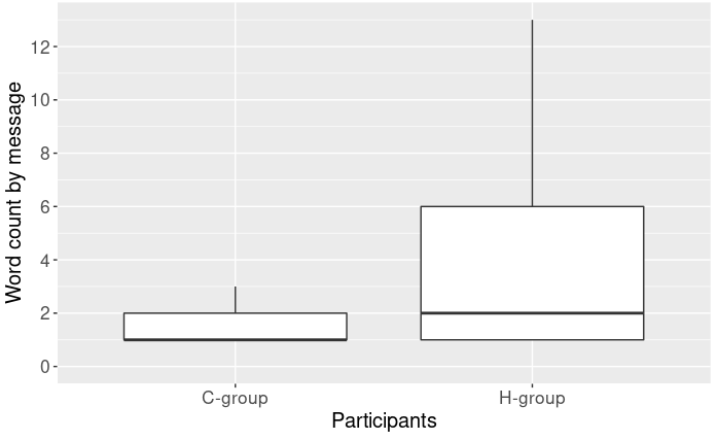


Fig. 4: Messages’ word count by group

According to a Wilcoxen rank-sum test ($p=2.10^{-16}$), 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) (Figure 4).

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 (Figure 5).

3.3 Answer changes

For every question where a participant was contradicted, we counted how many times 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 6).

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.

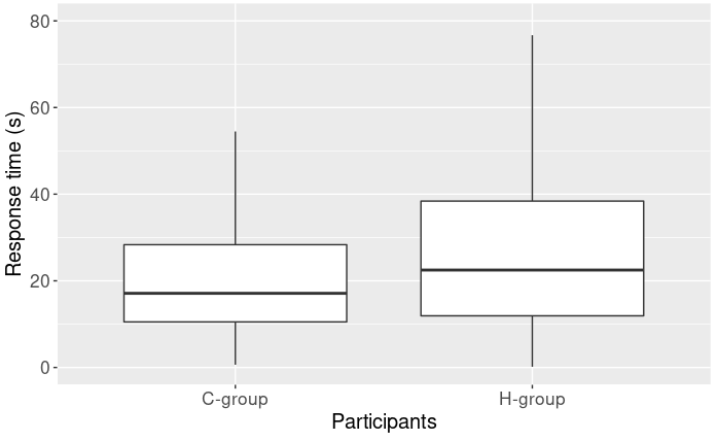


Fig. 5: Response time by group

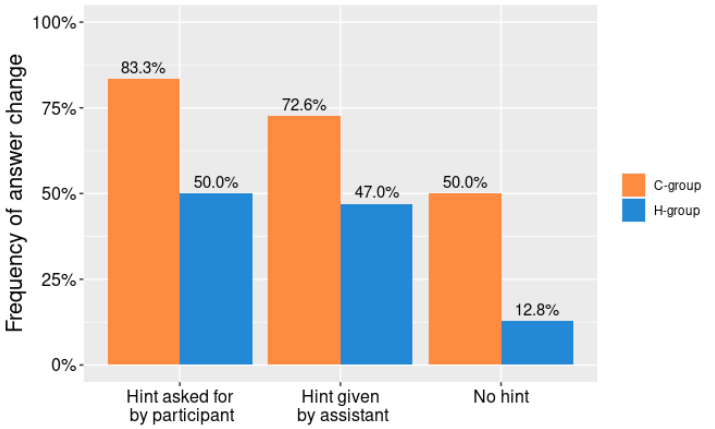


Fig. 6: Frequency of answer change by group and hint modality

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

3.4 Questionnaires

We compared the answers of each group to the four ten-level Likert items of the post-experiment questionnaire (Figure 7):

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.*

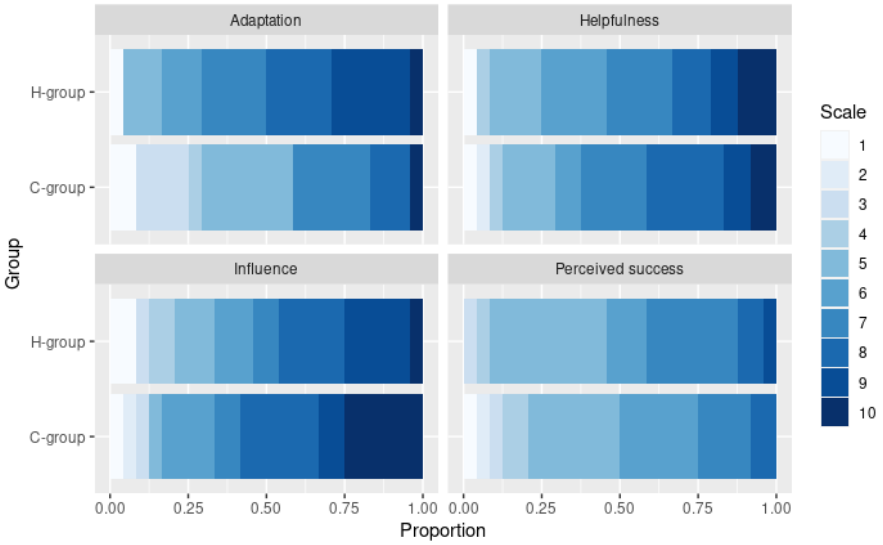


Fig. 7: Post-experiment questionnaire's answers by group

Using Pearson's Chi-squared test (Table 4), 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).

We then looked at the correlation between each Likert item and the answer change rate using Pearson correlation (Table 5 and Table 6). For both groups,

Table 4: Comparison of answers to the post-experiment 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 5: Pearson correlation matrix for the post-experiment 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 6: Pearson correlation matrix for the post-experiment 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

the ratings of the assistant's helpfulness and influence and the participants' answer change rate were positively correlated. 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.

Unlike 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 (Table 7). 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.

4 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 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.

Table 7: Frequency of subjects remarked on in the post-experiment questionnaire and illustrative examples (C=C-group and H=H-group)

Theme	Sub-theme	C	H
Experiment	The experiment went well *"The experiment went smoothly"; "everything went well"*	8	5
	The conversation was fluid *"The discussion seemed really fluid to me."; "The exchanges were quick and clear."*	2	1
	The conversation was pleasant *"[The assistant was] very likeable"; "I was comfortable and confident throughout the questionnaire."*	0	2
	The software worked well *"The software used works wonderfully"; "no bug, clear interface "*	2	0
	Remarks and problems with experiment *"One of the questions was asked twice instead of once"; "An indication of the question number would be useful"; "It's frustrating to give answers without knowing whether the answer is correct or not."*	8	2
	Questioning on the nature of the experiment *"I am curious to know what research idea is behind this experiment."; "I did not really understand the meaning of this experiment, and I was very surprised by the questions."*	4	5
Questions	The questions were interesting *"The experiment was interesting."; "Surprising but interesting questions"*	2	2
	The context for the questions was insufficient *"I have already practised a few situations and there are a lot of independent elements that come into play. "; "the elements of context are reasonably questionable."*	2	1
	Multiple answers could be right *"The questions were very subtle, I think there could be several possible answers."; "I sometimes had the impression that there was not necessarily a "right" answer."*	2	3
Participant's performance	Lack of knowledge on the subject *"I think I am lacking knowledge in this field" ; "[this] reinforces my belief that I am not made for adventure :)"; "Being no survivalism expert"*	6	5
	Good perceived performance *"I think I got pretty good answers"*	0	1
Assistant	The assistant was useful *"Fortunately he was there!" ; "The assistant is helpful when you do not know how to answer"*	4	1
	Lacking interaction with the assistant *"I did not feel any real rational interaction between the assistant and me"; "I think it might have been possible for me to ask the assistant questions"*	2	1
Assistant's interventions	Perceived manipulation by the assistant *"We do not know if the assistant is always trying to help us or deceive us" ; "I had the impression the assistant tried to influence me dishonestly, that is, towards wrong answers."*	9	7
	Interventions are followed *"I gladly let myself be influenced"; "I was very influenced by the answers of the assistant"*	3	0
	Interventions created doubt *"we tend to believe that we must follow the clues"; "When the assistant asked us if we were sure, even if I was 100% sure of my answer, I had the feeling that I was wrong and that my answer had to be changed"*	4	2
	Interventions are voluntarily ignored *"It's pretty hard not to be influenced by the assistant!"; "I rather had the will not to take into account their remarks"*	2	1
	Interventions are followed less with time *"So, I finally adapted, after 1/3 or half of the answers, by trying to follow my first opinion."; "I rather trusted the assistant at first"*	3	0
	Interventions are followed depending on certainty in answer *"[the assistant] has a great influence in case of hesitation ... "; "my degree of disturbance by the assistant is correlated with my degree of ignorance"*	2	1

Those findings have implications for integrating conversational agents in a collaborative work environment. They are consistent with the literature on automation bias and especially [17] 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 genuine collaboration with humans. Other studies have shown that over-reliance on AI can lead to decreased trust if the AI makes an error and fails to meet too high expectations [18–20] and that early-on errors could lead to permanent decreases in reliance [21]. Methods to prevent overreliance that could benefit conversational assistants are negative framing, presenting the AI's limits and possible errors before usage [19], and cognitive forcing methods, such as time delays or demanding a preliminary decision before providing advice [22].

Based on the findings from the experiment, the study proposes that over-reliance on chatbot advice can be monitored by measuring the frequency of messages concerning the user's decision process and doubts. This could serve as an indicator of the extent to which users rely on the chatbot for decision-making support. A higher frequency of messages discussing decision processes and doubts may suggest a higher level of reliance on the chatbot's advice.

To improve appropriate reliance on chatbots and promote more balanced decision-making, the study suggests implementing negotiation-forcing methods. These methods encourage users to engage in active discussions with chatbots about their advice, seeking clarifications or more information, or even finding a compromise. This interactive approach takes advantage of the natural language capabilities unique to chatbots and aims to enhance users' understanding and critical evaluation of the advice provided.

Negotiation-forcing methods could include:

- Encouraging users to ask follow-up questions: Design chatbots to prompt users to ask questions or seek clarifications about the advice given. This can be achieved by incorporating cues in the chatbot's responses, such as "Feel free to ask any questions you may have" or "Is there anything you'd like me to clarify?"
- Incorporating active learning: Develop chatbots that can engage in active learning, where they ask users for feedback on their advice or adapt their recommendations based on user preferences and responses. This can foster a more collaborative decision-making process.
- Facilitating compromise: Design chatbots to be open to compromise and explore alternative solutions when users express doubts or disagreement with the advice provided. This can involve presenting multiple options or guiding users through a step-by-step evaluation of their decision.
- Providing evidence-based advice: Ensure that chatbots offer well-reasoned, evidence-based advice, including sources or explanations to support their

recommendations. This can help users better understand and evaluate the advice given.

By implementing these negotiation-forcing methods, the study aims to promote more appropriate reliance on chatbot advice, fostering a more balanced and informed user decision-making process. This approach leverages the unique strengths of chatbots, particularly their natural language capabilities, to facilitate more effective and interactive human-chatbot collaborations.

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 conversational 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 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.

5 Conclusion

In this study, the goal was to identify the differences in reliance between conversational assistants (chatbots) and humans when providing advice to users. The results demonstrated a clear bias toward AI, with participants following the advice of conversational assistants more often than that of human assistants, regardless of whether the advice was requested or not. This suggests that users might be more inclined to trust AI advice over human advice in decision-making scenarios.

However, the qualitative analysis of the study revealed that this apparent trust in AI might be superficial. Participants' interactions with the chatbots and their decision-making processes indicated that their reliance on AI advice was not as deep-rooted as the quantitative results suggested. This highlights the importance of considering multiple measures and indicators when assessing reliance on conversational assistants.

Based on these findings, we recommend the following additional indicators to evaluate reliance on new conversational assistants:

- User engagement: Analyze the depth and quality of user interactions with the conversational assistant, including the number of follow-up questions and clarifications sought by users.
- Confidence in decisions: Assess users' self-reported confidence in their decision-making when advised by a conversational assistant compared to a human advisor.
- Decision quality: Evaluate the objective quality of users' decisions when relying on advice from conversational assistants versus human advisors. This can be measured by comparing decision outcomes against established benchmarks or expert evaluations.
- Decision consistency: Investigate the consistency of users' decisions across different scenarios and advice sources, exploring whether their reliance on conversational assistants remains stable or varies depending on the context.
- User feedback: Collect user feedback on their experiences interacting with conversational assistants, including their perceptions of trustworthiness, expertise, and helpfulness.

By incorporating these additional indicators, researchers and developers can obtain a more comprehensive understanding of user reliance on conversational assistants. This will enable them to design and implement more effective and user-friendly AI advice systems that foster appropriate trust and reliance while promoting balanced decision-making.

To address the challenge of ensuring that users make the most of advanced cognitive services like ChatGPT while maintaining appropriate trust and reliance, future research should also focus on the design and development of conversational agents that encourage in-depth reasoning and critical evaluation. The following recommendations can help guide this research:

- Transparency: Develop chatbots that provide transparent explanations for their advice, making it clear how they arrive at their recommendations. This can help users assess the quality and reliability of the information provided, and ensure that they are not blindly following AI-generated advice.
- Encouraging critical thinking: Design chatbots that prompt users to think critically about the advice they receive by asking questions or providing alternative viewpoints. This can foster a more interactive and engaging decision-making process, promoting more balanced reliance on AI advice.

- Evidence-based advice: Ensure that chatbots offer well-reasoned, evidence-based advice, including sources or explanations to support their recommendations. This can help users better understand and evaluate the advice given.
- Adaptive learning: Implement adaptive learning techniques that allow chatbots to refine their recommendations based on user preferences, feedback, and decision outcomes. This can help establish a more collaborative and trust-based relationship between users and AI assistants.
- User education and awareness: Develop user education and awareness initiatives to help users understand the capabilities and limitations of AI systems like ChatGPT[2]. This can promote more informed decision-making and appropriate reliance on AI advice.
- Detecting and addressing misinformation: Incorporate mechanisms for chatbots to detect and address false or nonsensical claims. These mechanisms can prevent the chatbot from generating misleading information or alert users when the information provided is inaccurate or unreliable.

By focusing on these recommendations, future research can contribute to developing more effective and user-centred conversational agents that encourage in-depth reasoning and appropriate reliance. This will ultimately lead to a more balanced and informed decision-making process, ensuring users make the most of advanced cognitive services while maintaining trust and scepticism.

References

[1] Dale, R.: The return of the chatbots. Natural Language Engineering **22**(5), 811–817 (2016). https://doi.org/10.1017/S1351324916000243

[2] Almalki, M., Azeez, F.: Health chatbots for fighting covid-19: a scoping review. Acta Informatica Medica **28**(4), 241 (2020)

[3] Nadarzynski, T., Miles, O., Cowie, A., Ridge, D.: Acceptability of artificial intelligence (ai)-led chatbot services in healthcare: A mixed-methods study. Digital health **5**, 2055207619871808 (2019)

[4] Ahmady, S.E., Uchida, O.: Telegram-based chatbot application for foreign people in japan to share disaster-related information in real-time. In: 2020 5th International Conference on Computer and Communication Systems (ICCCS), pp. 177–181 (2020). IEEE

[5] Mutis, I., Ramachandran, A.: The bimbot: mediating technology for enacting coordination in teamwork collaboration. J. Inf. Technol. Constr. **26**, 144–158 (2021)

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

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

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

[8] Skitka, L.J., Mosier, K.L., Burdick, M.: Does automation bias decision-making? International Journal of Human-Computer Studies **51**(5), 991–1006 (1999)

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

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

[11] Keding, C., Meissner, P.: 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**, 120970 (2021)

[12] Brachman, M., Ashktorab, Z., Desmond, M., Duesterwald, E., Dugan, C., Joshi, N.N., Pan, Q., Sharma, A.: Reliance and automation for human-ai collaborative data labeling conflict resolution. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–27 (2022)

[13] Zaroukian, E., Bakdash, J.Z., Preece, A., Webberley, W.: Automation bias with a conversational interface: User confirmation of misparsed information. In: 2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), pp. 1–3 (2017). IEEE

[14] Komiak, S.Y., Benbasat, I.: The effects of personalization and familiarity on trust and adoption of recommendation agents. MIS quarterly, 941–960 (2006)

[15] Hall, J., Watson, W.H.: The effects of a normative intervention on group decision-making performance. Human relations **23**(4), 299–317 (1970)

[16] Lafferty, J.C., Pond, A.W.: The Desert Survival Situation: A Group Decision Making Experience for Examining and Increasing Individual and Team Effectiveness. Human Synergistics, ??? (1974)

[17] Van Der Waa, J., Verdult, S., Van Den Bosch, K., Van Diggelen, J., Haije, T., Van Der Stigchel, B., Cocu, I.: Moral decision making in human-agent teams: Human control and the role of explanations. Frontiers in Robotics

and AI **8** (2021)

[18] Pop, V.L., Shrewsbury, A., Durso, F.T.: Individual differences in the calibration of trust in automation. Human factors **57**(4), 545–556 (2015)

[19] Dzindolet, M.T., Pierce, L.G., Beck, H.P., Dawe, L.A.: The perceived utility of human and automated aids in a visual detection task. Human factors **44**(1), 79–94 (2002)

[20] de Visser, E.J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., Parasuraman, R.: The world is not enough: Trust in cognitive agents. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 56, pp. 263–267 (2012). Sage Publications Sage CA: Los Angeles, CA

[21] Kim, A., Yang, M., Zhang, J.: 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)

[22] Buçinca, Z., Malaya, M.B., Gajos, K.Z.: 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), 1–21 (2021)