UNIVERSITÉ DE LORRAINE | BIBLIOTHÈQUES UNIVERSITAIRES

AVERTISSEMENT
DROIT D'AUTEUR

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document. Toute contrefaçon, plagiat ou reproduction illicite encourt une poursuite pénale.

Contact bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10

UNIVERSITÉ DE LORRAINE
34 COURS LÉOPOLD — CS 25233
54052 NANCY CEDEX
TÉL : 03 54 50 54 00
CONTACT@UNIV-LORRAINE.FR
WWW.UNIV-LORRAINE.FR

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Confiance dans les Chatbots pour la Collaboration

THÈSE

présentée et soutenue publiquement le 21 Novembre 2025

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Clélie AMIOT

Composition du jury

Présidente :	Serena IVALDI	DR, Centre Inria de l'U. de Lorraine
Rapporteurs :	Gilles BAILLY	DR, Sorbonne Université
	Philippe RAUFFET	PR, Université Bretagne Sud
Examinatrice :	Agnès HELME-GUIZON	PR, Université Grenoble Alpes
Directeurs de thèse :	François CHAROY	PR, Université de Lorraine
	Jérôme DINET	PR, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Trust in Chatbots for Collaboration

Doctoral Thesis

Remerciements

Je tiens à adresser mes sincères remerciements à mes encadrants, François Charoy et Jérôme Dinet, pour leur accompagnement, leur patience et leur soutien tout au long de cette thèse.

Je remercie également les doctorants de Larsen et MFX, de Coast et enfin de Loreley, dont la présence et les nombreux moments partagés m'ont encouragée à entreprendre cette thèse puis à la mener à son terme.

Mes remerciements vont aussi à mon laboratoire, à mon université et à la communauté scientifique dans son ensemble, dont l'environnement stimulant et l'esprit de partage ont renforcé mon attachement au monde de la recherche.

Enfin, je remercie ma famille pour les moments de répit précieux qu'elle m'a apportés, ainsi que Julie, pour son écoute et nos échanges qui m'ont accompagnée tout au long de cette thèse.

Table of Contents

List of Figures **ix**

List of Tables **xi**

Chapter 1
General Introduction

1.1 Challenges of collaboration . . . 3
1.1.1 Distributed and virtual collaboration . . . 4
1.1.2 Larger collaborations . . . 4
1.2 Leveraging chatbots . . . 5
1.2.1 Real-world applications . . . 6
1.2.2 ChatGPT and other LLM-based chatbots . . . 8
1.3 Problem statement . . . 8
1.4 Contributions . . . 9
1.5 Outline . . . 10

Chapter 2
Related Work

2.1 AI assistance . . . 14
2.1.1 Trust, reliance, automation bias and algorithm appreciation . . . 15
2.1.2 Trust calibration and appropriate reliance . . . 21
2.2 Chatbot assistance . . . 22
2.2.1 Anthropomorphism and social presence . . . 22
2.2.2 Reliance on chatbots . . . 23
2.3 Chatbots and teamwork . . . 24

2.3.1 Challenges of group conversations 24
2.3.2 Chatbots integrated into work collaboration tools 25
2.3.3 Chatbots for collaborative learning and group discussions 27
2.4 Conclusion 28

Chapter 3
Chatbot vs. Human Advice

3.1 Introduction 31
3.2 Context 33
3.2.1 Related work 33
3.2.2 Design considerations 34
3.3 Experimental design 35
3.3.1 Assistant types 36
3.3.2 Questionnaires 39
3.4 Results 40
3.4.1 Messages length and response time 41
3.4.2 Messages content 41
3.4.3 Answer changes 45
3.4.4 Questionnaires 46
3.5 Discussion 51
3.5.1 Summary of results 51
3.5.2 Design implications 51
3.5.3 Limits 52
3.6 Conclusion 53

Chapter 4
A Novel Online Activity for Studying Group Interactions

4.1 Introduction 55
4.2 Existing experimental collaborative activities 56
4.3 Design of the collaborative activity 58
4.3.1 Design objectives 58
4.3.2 Activity rules 59

4.3.3 Modules . . . 61
4.4 Technical implementation . . . 63
4.4.1 Database . . . 63
4.4.2 Management interface . . . 65
4.5 Validation . . . 66
4.6 Conclusion . . . 66
4.6.1 Potential future applications . . . 67

Chapter 5
Impact of Chatbot Assistance on Collaborative Tasks

5.1 Introduction . . . 69
5.2 Methodology . . . 70
5.2.1 COCCO the chatbot . . . 71
5.2.2 Pre-activity and post-activity surveys . . . 75
5.2.3 Participants . . . 76
5.3 Analysis and Results . . . 77
5.3.1 Team performance . . . 77
5.3.2 Cognitive workload and perception . . . 79
5.3.3 Team dynamics . . . 81
5.4 Discussion . . . 83
5.4.1 Limitations and future directions . . . 84
5.5 Conclusion . . . 85

Chapter 6
Conclusion

6.1 Generative AI . . . 89
6.2 Future Directions . . . 89

Appendix A Wilderness survival questions, answers and hints **93**

Appendix B First experiment post-questionnaire **99**

Appendix C COCCO's intents **101**

Appendix D Second experiment post-questionnaire **103**

Bibliography **109**

List of Figures

3.1 Experimental design 36
3.2 Example of question, answers and hints given to the participants 37
3.3 Differences in presentation between COCCO and Hugo 38
3.4 Wizard of Oz interface 39
3.5 Messages' word count by group 41
3.6 Response time by group 42
3.7 Frequency of answer change by group and hint modality 45
3.8 Post-experiment questionnaire's answers by group 47

4.1 Example bomb from the game Keep Talking and Nobody Explodes 59
4.2 Platform interface 60
4.3 Examples of each module type 61
4.4 Database diagram 64
4.5 Experience management interface 65

5.1 Excerpt from an interaction between participants with no assistance from COCCO 72
5.2 Excerpt from two groups' interaction with assistance from COCCO in the group chat 73
5.3 Excerpt from a group interaction with assistance from COCCO in private chat 74
5.4 Database diagram including COCCO 75
5.5 Experimental design 76
5.6 Module solving times for each condition 78
5.7 Click patterns of an organized team showing temporal grouping by module types 81
5.8 Click patterns of a less organized team 82

List of Tables

2.1 Studies comparing preferences between human and algorithm advice 16

3.1 Demographic information of recruited participants 40
3.2 Labels of the assistants' messages 43
3.3 Labels of the participants' messages 44
3.4 Comparison of answers to the post-experiment questionnaire 47
3.5 Correlation matrix of the post-experiment questionnaire Likert scale and answer change rates for C-group participants 48
3.6 Correlation matrix of the post-experiment questionnaire Likert scale and answer change rates for H-group participants 48
3.7 Frequency of subjects remarked on in the post-experiment questionnaire and illustrative examples 48

5.1 Comparison of module solving times between conditions 78
5.2 Comparison of response times between conditions 79
5.3 Comparison of response times between conditions excluding COCCO . . . 80

A.1 Wilderness survival questions, answer propositions and hints (translated from French) 93

C.1 COCCO's intents 101

Chapter 1

General Introduction

Contents

1.1 Challenges of collaboration 3

1.1.1 Distributed and virtual collaboration 4

1.1.2 Larger collaborations 4

1.2 Leveraging chatbots 5

1.2.1 Real-world applications 6

1.2.2 ChatGPT and other LLM-based chatbots 8

1.3 Problem statement 8

1.4 Contributions 9

1.5 Outline 10

Chatbots are now more powerful, useful, and accessible than ever before. Modern chatbots have become remarkably more advanced at understanding and generating natural language. They can now maintain longer coherent context-aware conversations, making dialogue with them almost as natural as talking with a human interlocutor. These improvements are driven by the rapid advancement of natural language processing technology, access to more diverse and higher-quality datasets, and the development of larger and more complex language models. Coupled with increased investment in the technology, they have contributed to widespread adoption of chatbots and greater familiarity among the general public. AI usage, including tools like ChatGPT, Claude, and Google Gemini, has surged in everyday life, rising in France from a fifth of the population in 2023 to a third in just one year, with adoption particularly high among younger users, as 77% of 18- to 24-year-olds reported using these tools in 2024[1].

Chatbots are used in daily life to answer questions, provide recommendations, and offer entertainment, but they are also used in professional sectors. In customer support, they filter queries, provide instant answers, and collect preliminary information for claims.

[1] Arcep, Arcom, CGE, ANCT, *Baromètre du numérique, édition 2025*, available at labo.societenumerique.gouv.fr (accessed September 20, 2025)

In healthcare, they assist with scheduling appointments, educate the general public, and offer mental health support. They also play key roles in education by tutoring students and tracking their progress, and in corporate settings by managing administrative tasks, scheduling meetings, and tracking projects.

Beyond assisting individuals, chatbots can be valuable tools for supporting teams and enhancing collaboration thanks to their unique strengths. They are operational at all hours, near-instantaneous, and able to handle multiple simultaneous conversations, ensuring that collaborations proceed and important time-sensitive information is shared, even if team members in different time zones are unavailable. They can translate conversations in real time, enabling communication across language barriers between team members situated in different countries. They can generate high-level summaries, access knowledge bases, infer the main takeaways from discussions, and automatically aggregate information from different sources, making it easier for teams to document their progress, track decisions, and stay aligned on their goals. Additionally, chatbots can facilitate the exchange of essential information and updates between their team and external collaborators, such as other teams within or outside the organization, or higher-level management. They can keep relevant stakeholders up to date, ensuring they are aware of the latest decisions, updates, and any changes in processes, thereby reducing the possibility of miscommunications and errors arising from mismatched expectations.

However, despite the potential usefulness of chatbots for collaboration, their adoption is not straightforward. Introducing a new tool into an existing workflow can be disruptive, potentially causing more drawbacks than benefits for its users. When a tool alters established dynamics, it can lead to confusion, frustration, and resistance among team members, who must relearn processes and adapt to changes in their routine [101, 14]. This shift can create a significant cognitive burden, hindering productivity during the transition period [119].

In the context of collaboration, these problems can compound, as any obstacles or challenges faced by a collaborator will likely affect the entire team. Additionally, non-symmetric usage of the tool, where some members adopt it and others do not, can further hinder adoption or even reverse it. This is especially true when the benefits and costs of adoption are influenced by the actions of others. If the tool only provides value when used collectively, isolated adoption offers little reward and may even impose extra burdens (e.g., time or effort), while teammates passively benefit. Similarly, collaborative tools often require usage interdependence, meaning the system's usefulness is contingent on mutual engagement. As a result, adoption success depends on reaching a critical mass threshold, meaning there is a minimum number of users required before the tool becomes beneficial enough for long-term use. Without this, even early adopters may abandon the tool due to a lack of reciprocation [64, 91, 43].

Another critical factor in successful integration is how users rely on the chatbot, that is, how often its advice is followed or ignored. On the one hand, when users do not rely on it enough, they may neglect the tool altogether, missing out on its expected benefits despite the resources invested. On the other hand, when users rely too much on a tool, they may misuse it by following its advice blindly and make poor decisions. To prevent these issues, users should be encouraged to rely on chatbots appropriately, using their

assistance as a support rather than a substitute for critical judgment.

In the context of collaboration, we do not want chatbot assistance to overshadow the contributions of human collaborators. Over-reliance on chatbots can lead to self-censorship, where individuals unsure about their ideas prefer the tool's suggestions to their own. This can reduce opinion diversity, limiting the range of perspectives considered in team decision-making [28]. Additionally, excessive dependence on a chatbot may diffuse accountability, as team members may feel less responsible for outcomes when they rely heavily on automated advice [114, 92].

Considering these challenges, this thesis explores how chatbots can successfully support collaboration without disrupting existing team dynamics. It investigates key factors that contribute to reliance in chatbot assistance, specifically, the differences in reliance on human versus chatbot advice. This thesis also introduces a novel experimental platform for observing group interactions and virtual collaboration. Using this platform, we then analyze how a chatbot's communication strategies in group settings influence team performance and dynamics. This work seeks to contribute to the growing body of knowledge on human-chatbot interaction, shifting focus from traditional studies on dyadic interactions to explore the potential of chatbots in collaborative environments.

1.1 Challenges of collaboration

Through collaboration, people can achieve outcomes that are greater than the sum of their individual efforts by combining complementary skills, knowledge, and resources. Team members have access to a wider and more diverse pool of expertise, leveraging each other's strengths and compensating for individual weaknesses. Sharing the workload allows teams to focus on projects that benefit everyone while avoiding the inefficiencies of working in isolation or duplicating efforts. Despite these advantages, effective collaboration is not always easy to achieve.

Successful collaboration requires that team members share a common understanding of goals, roles, and processes [21, 27]. Maintaining this alignment as a project evolves involves sustaining consensus, keeping shared understanding up to date, and ensuring robust communication channels. Without this, teams risk misalignment that can impede progress and decision-making.

Effective collaboration also depends on coordinating shared resources and managing interdependencies between tasks. Communication challenges often arise from differences in cultural backgrounds, organizational roles, or disciplinary perspectives, where mismatched expectations, norms, or vocabularies can slow decisions and create misunderstandings [82].

When these challenges are not effectively addressed, teams risk performance breakdowns and miscommunications. Collaborators may operate based on differing or incomplete sets of information, leading to conflicting actions or duplicated work that compromises the outcome of the collaboration.

1.1.1 Distributed and virtual collaboration

A significant portion of modern collaboration is facilitated by online tools, with teams connecting through video calls, messaging platforms, and collaborative software such as Slack, Zoom, and Google Workspace.

The COVID-19 pandemic, which began in 2020, created an urgent need for companies worldwide to transition to virtual and hybrid work models [89]. As lockdowns forced offices to close, and continued social distancing and preventive isolation measures delayed a full return, teams that once worked together in the same space were left dispersed. Organizations had to rely heavily on digital platforms to maintain communication, manage projects, and support remote collaboration. This rapid shift significantly accelerated the adoption of online collaborative tools, as traditional in-person workflows became increasingly nonviable. It also sparked research interest in cognitive technologies that could enhance the coordination and management of remote teams, one of which is chatbots.

Temporal distance, caused by differences in time zones and work patterns, can lead to delays and missed tasks or information requests, as team members are often unavailable during critical periods, resulting in coordination breakdowns and project setbacks [74].

Trust is another critical factor, particularly for distributed or global teams with limited face-to-face interaction [16]. A lack of trust can exacerbate misunderstandings, reduce information sharing, and weaken communication overall.

Additionally, the physical separation and absence of face-to-face communication, along with the lack of informal, unscheduled interactions, can hinder not only communication but also trust building and amplify unresolved friction. Tensions and miscommunications that would have otherwise been smoothed over with a brief in-person conversation may go unnoticed and persist, potentially worsening over time and negatively impacting team performance and relationships [83, 74].

1.1.2 Larger collaborations

Collaborative endeavors have grown sustainably larger, bringing more people together to work on increasingly complex projects. To manage this scale, collaboration tools were developed to track tasks, timelines, resource expenditure, and productivity. At scale, collaborations often split into smaller sub-teams focused on specific sub-goals that contribute to the overall objective. However, this subdivision introduces the possibility of tension or outright competition around resource allocation.

Large-scale collaborations includes cross-organizational or cross-national collaborations, where differences in processes, training, and cultures lead to differing assumptions about roles, responsibilities, and communication styles, and can exacerbate conflicts [115]. When information is inadequately disseminated across these boundaries, errors arise that are difficult to detect, trace, and rectify.

Bigger teams also bring increased communication and coordination costs, which are among the primary burdens of collaboration, especially as team size or task complexity increases. Working with a team of three is not the same as working with a team of thirty or even three hundred collaborators. The larger the scale, the more information circulates.

These coordination costs encompass the time, effort, and organizational overhead required to align collaborators, including negotiating roles, managing task dependencies, resolving conflicts, and coordinating schedules. In larger or distributed teams, coordination becomes exponentially harder due to the increased number of communication pathways and the greater likelihood of misaligned expectations [65, 9].

Closely tied to coordination are cognitive costs, the mental effort individuals expend to manage and process collaborative information. As teams grow, members must track discussions, maintain a mental map of responsibilities, communicate status updates to stakeholders, and keep others aware of schedules and progress. They also need to understand how work is divided across teams, know who has specific knowledge or how to find it, and stay informed about management decisions affecting tasks. This increasing cognitive load can slow decision-making, cause important information to be overlooked, and lead to mistakes [85].

Managing collaboration, sharing resources, and keeping all collaborators updated can become a task in itself, consuming valuable time and resources that could otherwise be spent on the project's core goals. It becomes impossible for any single person to be aware of every detail, from each teammate's tasks and expertise to meeting updates and resource availability. To maintain manageable workloads and cognitive loads, collaborators rely on high-level summaries and filtered information relevant to their responsibilities. Consequently, individuals working in different sub-teams may remain unaware of each other's expertise, resources, or potential conflicts, further complicating coordination and collaboration.

In addition to human factors, the technical platforms that support collaboration must also scale effectively to meet the demands of large and distributed teams. This includes handling increased data volume, ensuring robust security and privacy across multiple organizations, simplifying onboarding processes for new collaborators, and managing multiple concurrent projects with diverse requirements. The complexity of these technical challenges increases in tandem with team size and project scope, necessitating collaborative tools to evolve and adapt accordingly [15].

1.2 Leveraging chatbots

Chatbots are AI agents designed to converse with human users primarily through text, providing companionship, advice, or information. Some broader definitions also encompass voice assistants, though this work focuses primarily on text. They are typically developed by companies or institutions for specific purposes, such as customer support, guiding buyers from product recommendations to closing online sales, or disseminating information. Chatbots process user prompts using methods like machine learning or pattern matching, generating responses either from scratch or through templates. Simple chatbots function as decision trees within a chat interface, guiding users through predefined options. In contrast, more advanced chatbots are more context-aware, capable of remembering long conversations, understanding implied or indirect messages, and responding to off-topic user prompts [11, 100].

They are capable of performing sentiment analysis to adjust their tone and responses to match the emotions of their interlocutor. They can be fine-tuned to specialize in specific domains, such as healthcare or finance, and can update their knowledge base without requiring full retraining. Most mainstream messaging apps and communication platforms, such as WhatsApp, Slack, and Microsoft Teams, now allow for easy chatbot integration with minimal technical effort, encouraging wider adoption. The continued development of APIs has simplified chatbot interaction with third-party applications like email services, project management software, databases, version control systems, and Customer Relationship Management software, creating fully connected app ecosystems.

Chatbots are potentially well-suited to support various aspects of collaborative work. Many collaborations already rely on virtual tools for communication, knowledge management, and task tracking, making it easy to integrate chatbots without the need to introduce completely new tools in the workflow. Their text-based interface allows for natural language interactions that are user-friendly and typically require no additional training for users. They can also be modified or expanded with new functionalities after implementation without requiring a redesign of the user interface, forcing users to relearn it.

They're available at all times and can assist multiple people simultaneously. They can aggregate information from various sources, support multiple languages, and have greater memory capacity and faster response time than humans. Additionally, they can relay information to human operators, consult and update real-time data, generate documents, and provide summaries or reports to help streamline collaboration. Furthermore, automating certain aspects of collaboration with a chatbot can help formalize processes that may not be well-defined or consistently followed by everyone, as well as standardize vocabularies, preventing confusion when collaborators from different fields use different terms for the same concept. By interacting with the chatbot in structured ways and using clearly defined terms, teams can ensure more consistent communication and understanding.

However, chatbots also have some disadvantages compared to other forms of AI assistance, particularly their lack of affordance, the ability of an interface to intuitively signal its functions and possible interactions. Unlike graphical user interfaces, which provide visual cues like buttons and menus, chatbots rely solely on text-based exchanges, making it harder for users to understand what the system can and cannot do [108]. This can lead to frustration, inefficient interactions, or underuse of the chatbot's full capabilities. As a result, studying how chatbots are adopted and integrated into collaborative workflows becomes increasingly important to ensure they are used effectively and do not become a barrier to productivity [96].

1.2.1 Real-world applications

Chatbots have already been adopted and are successfully used in real-life applications across various fields, including commerce, education, healthcare, banking, workplace settings, and civil security.

In customer support, they act as the first line of assistance, filtering customer queries

and reducing the workload of human operators by providing answers to frequently asked questions at all times and collecting information for claims and refunds. Erica, Bank of America's virtual financial assistant, has reached 2.5 billion interactions with 20 million users as of 2025 since her launch in 2018. Designed to assist with everyday financial tasks, she helps clients track their spending patterns, manage transactions, and navigate banking services. Erica has enhanced customer support efficiency by handling 98% of inquiries without escalation to a human representative[2,3].

In education, chatbots have been tested to increase student engagement and offer learning support outside of classes by providing students with additional resources and tutoring them, sometimes through participation and progress tracking or personalized study plans. Nanyang Technological University (NTU) in Singapore has integrated AI chatbots to support both students and faculty. One example is Leodar, a teaching and administrative assistant that helps lecturers by handling student queries. In 2024 alone, Leodar answered 12,000 questions, using a faculty-maintained knowledge base to provide accurate and reliable responses[4]. NTU has also introduced the NALA project, which encourages faculty members to create personalized chatbots trained on their own course materials[5].

In healthcare, chatbots are increasingly used to inform patients and ease the burden of strained resources. They support a variety of functions, such as educating the public on disease prevention, providing mental health support, and guiding patients through scheduling and treatment plans. They proved especially useful during the COVID-19 pandemic, when institutions widely deployed them to disseminate accurate health information, assist with self-triage and personal risk assessments, monitor symptoms and exposure, guide users through local lockdown policies and vaccine campaigns, and combat misinformation and fake news [2].

The American Red Cross has also been using a chatbot, Clara. Until July 2024, Clara assisted users in both English and Spanish, with a response accuracy of over 95%. During emergencies, it helped people find the nearest shelter through FEMA's database and access financial assistance, and it answered questions about volunteering and donations, reducing the workload of call centres. Clara also helped manage blood donations by providing information and scheduling appointments[6].

In corporate settings, chatbots are used to handle administrative tasks, such as onboarding new employees, automating time-off requests, planning business trips, processing expense claims, and tracking employee attendance. They are also used to automate certain aspects of project management and teamwork, such as sending reminders, scheduling meetings, generating reports, and tracking task progress updates.

[2]newsroom.bankofamerica.com (accessed September 20, 2025)
[3]newsroom.bankofamerica.com (accessed September 20, 2025)
[4]straitstimes.com (accessed September 20, 2025)
[5]ntu.edu.sg (accessed September 20, 2025)
[6]redcross.org (archived)

1.2.2 ChatGPT and other LLM-based chatbots

In recent years, LLMs (Large Language Models) such as OpenAI's ChatGPT, Google's Gemini and Anthropic's Claude, have subverted the general public's expectations of chatbots. Collective awareness and enthusiasm for chatbots are growing, as are public and private funding supporting their development.

Nowadays, chatbots touch every aspect of daily life and have lost their image of rigid FAQ widgets in the corner of websites, seemingly designed to block access to real human support. They act as a novelty and source of entertainment initially, but they can serve multiple roles depending on the user's needs. They can be a social support that acts as a confidant, who is always available and can provide comfort and life advice without judgment. They are also a productivity tool that provides writing and programming assistance, as well as an idea generation machine that helps with both creative works and mundane activities, such as deciding on a meal plan or preparing for a trip abroad. For some, they even act as search engines because they are easier to query without using parameters and provide direct, summarized results that are explained, eliminating information overload without requiring multiple website comparisons [105].

However, as chatbots become more capable and integrated into daily life, issues of reliance, how much users trust and follow their suggestions, grow increasingly important, particularly when their perceived capabilities outpace their actual limitations. Their rapid adoption has shifted the focus of research from questions of usability and technical constraints to broader concerns about ethics, trustworthiness, bias, and the societal impact of embedding AI into everyday activities. In collaborative settings, these risks are amplified: biased or misleading chatbot suggestions may propagate through entire teams, shaping collective decisions with significant consequences. At the same time, their efficiency and accessibility create strong incentives for widespread adoption across industries.

1.3 Problem statement

A well-designed chatbot can help its users by simplifying complex tasks, giving easy access to information, supporting decision-making and improving efficiency. They have the potential to enhance collaboration by facilitating conversations between teammates, encouraging creativity and becoming full-fledged teammates.

However, if we want to introduce chatbots in collaboration, we need to ensure that they are helpful, used appropriately, and do not negatively disrupt existing workflows. A chatbot has to assist and augment users' capabilities without dictating their decisions and having the final say. It should be intuitive to use and encourage users to make full use of its functionalities. It has to be visibly trustworthy to promote long-term use and reliance during critical tasks.

A chatbot that supports collaboration needs to be actively utilized to be effective and enhance existing systems. Its users need to be able to rely on it, even and especially when the task at hand is high stakes. Only when trusted can a tool be fully used and be optimally helpful. Indeed, an inappropriately designed tool will only increase the cognitive

costs of its usage, hinder performance, and increase the likelihood of mistakes. In some cases, it might even be abandoned by its users who prefer reverting to a more familiar and less taxing tool.

This thesis focuses on the impact of chatbots on decision-making and productivity in collaborative settings. Specifically, we aim to explore how chatbots are used when integrated into teamwork, assess their usefulness, and determine how they should present themselves and intervene to add value, whether by improving performance, enhancing user experience, or at the very least, avoiding disruption.

The research questions of this thesis are:

- **How can chatbots effectively support collaboration?** Is there an optimal way to deliver assistance? Can it be disruptive and detrimental to the team effort?
- **How does chatbot assistance influence user behavior?** Does it influence interaction with teammates? Are changes in behavior positive and rational?

Our research will not examine individual user factors, as these are difficult to control for and have limited implications for chatbot design beyond personalized adaptations, which remain a long-term goal. Instead, we will focus on more immediate, practical attributes of collaborative chatbot functionality, such as how the chatbot communicates, supports decision-making, and influences group dynamics. By focusing on these observable and modifiable aspects, we aim to provide actionable insights for designing chatbots that enhance teamwork, while also deepening our understanding of how chatbots shape group interactions and decision-making processes.

1.4 Contributions

In this thesis, we explore some of the key challenges and opportunities related to the integration of chatbots into collaborative group settings, with a particular focus on understanding their impact on human decision-making, team performance, and interaction dynamics. This work presents initial findings and methodological contributions that can serve as a foundation for future research.

We start by reviewing the existing body of research on AI assistance, covering both general AI systems and chatbots specifically. We examine key concepts, including appropriate reliance, automation bias, and trust calibration. We then focus on chatbot-specific factors, including anthropomorphism, social presence, and the unique challenges chatbots pose when integrated into teamwork and group decision-making contexts.

Following the observation of a lack of consensus in the literature regarding the influence of chatbot advice compared to that of human advisors, we designed a controlled experimental study. In our first contribution, we directly compare the persuasive power of a chatbot with that of a human assistant in individual decision-making tasks. Our findings indicate a higher likelihood of users changing their answers when contradicted by a chatbot, suggesting potentially greater influence, but with implications for over-reliance and trust calibration.

Since much of the literature focuses on dyadic interactions, we aimed to investigate chatbot integration in multi-user environments. This led us to design a novel online collaborative activity, tailored to study group interactions in a controlled yet adaptable setting. This platform enables researchers to observe group dynamics, role distribution, and task performance in real-time. It is a methodological contribution in its own right, enabling controlled and repeatable studies of group interactions in a customizable online environment that bridges the gap between traditional laboratory experiments and real-world collaborative settings.

Leveraging this platform, we conducted our second main study on the impact of chatbot assistance on group collaboration. We investigated how a chatbot integrated within a group discussion affects team performance, user experience, and cognitive workload. Our findings suggest that the presence of a chatbot can improve overall task performance and reduce individual cognitive load. However, different chatbot strategies (e.g., proactive vs. reactive advice) resulted in varying effects on collaboration, indicating the importance of adapting chatbot behavior to the team's characteristics and needs.

Finally, we summarize our findings and the broader implications of our work. We highlight open research questions and propose future directions for exploration.

Overall, this thesis contributes to a better and deeper understanding of how chatbots influence collaborative processes and decision-making, and offers both experimental and methodological contributions for designing more effective collaborative chatbots.

1.5 Outline

This thesis is organized in the following chapters:

- **Chapter 2 – Related Work**
  We review existing literature on AI assistance and chatbot integration, focusing on key concepts such as appropriate reliance, automation bias, trust calibration, anthropomorphism, and social presence in collaborative settings.

- **Chapter 3 – Chatbot vs. Human Advice**
  We present a controlled experimental study comparing the persuasive influence of a chatbot with that of a human assistant. Results show that chatbot advice may lead to higher answer changes, raising questions about trust and over-reliance.

- **Chapter 4 – A Novel Online Activity for Studying Group Interactions**
  We introduce a novel platform designed to study group interaction in a customizable and controlled environment, allowing for real-time observation of group dynamics and task performance.

- **Chapter 5 – Impact of Chatbot Assistance on Collaborative Tasks**
  Using the developed platform, we study the impact of chatbot assistance on group discussions, team performance, user experience, and cognitive workload. We compare proactive and reactive chatbot strategies to assess their influence on collaboration.

- **Chapter 6 – Conclusion**
  We summarize our key findings, discuss their implications, and outline open questions for future research on chatbots in collaborative settings.

Chapter 2

Related Work

Contents

2.1 AI assistance 14
2.1.1 Trust, reliance, automation bias and algorithm appreciation . . 15
2.1.2 Trust calibration and appropriate reliance 21
2.2 Chatbot assistance 22
2.2.1 Anthropomorphism and social presence 22
2.2.2 Reliance on chatbots 23
2.3 Chatbots and teamwork 24
2.3.1 Challenges of group conversations 24
2.3.2 Chatbots integrated into work collaboration tools 25
2.3.3 Chatbots for collaborative learning and group discussions . . . 27
2.4 Conclusion 28

In this thesis, we investigate how chatbots can assist group collaboration, specifically examining how they should deliver advice and how this affects group dynamics and decision-making. Our main goal is to understand how chatbots can support and facilitate collaboration without disrupting it. To achieve this, we need to foster appropriate reliance on the chatbot's advice: useful and convincing enough not to be ignored, but not too much, so that it does not overshadow the conversation or take over the decision-making process. Moreover, integrating a chatbot should avoid adding cognitive load or increasing the effort required for effective collaboration.

To inform our approach, we begin by examining the literature on AI assistance, particularly on the topic of user reliance and what makes AI advice effective. This involves exploring the conditions under which AI advice is more likely to be followed or disregarded, as well as how its reception compares to that of human advice. We will then investigate what sets chatbots apart from other AI systems and how it might affect the way users collaborate with them. Finally, we will review the existing literature on chatbots interacting with groups instead of individuals. We will examine how chatbots can

be beneficial for groups, how multi-party interactions differ from those with a single user, and the specific challenges they present.

2.1 AI assistance

AI assistance refers to the use of artificial intelligence systems to support, augment, or automate tasks and decision-making processes for humans. AI assistance can be delivered through various mediums depending on its specific use case. It can be physically instantiated in the form of robots or autonomous vehicles, which have the ability to perceive and act in the real world, performing tasks such as delivery, cleaning, or transportation. Because they have a physical presence, these systems can cause physical harm, resulting in safety considerations taking precedence over the quality of the interaction. Within digital environments, AI is often integrated into applications as information widgets, intelligent features, or recommender algorithms that help users navigate content, make decisions, or discover personalized suggestions. In domestic settings, voice-activated home assistants are used for hands-free control over smart home systems, managing music, lighting, temperature, and appliances through spoken commands. Finally, text-based chatbots provide conversational interfaces for customer service, technical support, mental health counseling, educational tutoring, entertainment, personal productivity and scheduling, information retrieval, and collaborative work assistance [54, 63].

As these systems become integrated into our lives, it is increasingly important to design them with a focus towards trust and reliance. Using them should be a benefit and not a cost. They should reduce friction and mental load instead of adding to it. However, although working with cooperative AI agents can boost performance, it can also introduce mental strain for users. For instance, Li et al. [61] highlight the cognitive load that may accompany human-machine collaboration. Their study on reciprocal cooperation in human-agent interactions found that people tend to reciprocate cooperative behavior from AI agents, even in complex, dynamic environments. In a hospital management microworld experiment, participants shared more resources and performed better when collaborating with highly cooperative agents, but also showed increased activation in the prefrontal cortex, indicating greater cognitive effort.

Klein et al. [50] outline ten key hurdles to effective collaborative automation. They establish the following criteria for successful human-AI teamwork: agents must understand shared goals and their roles, maintain common ground, and communicate when they cannot fulfill their responsibilities. They need a solid model of their teammates' intentions and actions, ensuring predictability on both sides. Additionally, agents should be directable and correctable through policies without the need for reprogramming, clearly signaling their status and intentions. They must also interpret teammates' signals, adapt to changes in situations, negotiate goal modifications, manage replanning in real-time, and minimize coordination costs.

Thus, the effectiveness of AI assistance depends not only on technical capabilities but also on the quality of human-AI interaction. A critical aspect of this interaction is the calibration of human attitudes toward AI systems, particularly in terms of when and how

they should be relied upon.

2.1.1 Trust, reliance, automation bias and algorithm appreciation

The paper *Trust between humans and machines, and the design of decision aids* by Muir [76] laid the theoretical groundwork for understanding how users develop trust in automated systems. It extends existing interpersonal models of trust into a multidimensional model for human-machine trust. The model factors technical reliability, expectations of competence, fiduciary responsibility, and alignment with social and moral norms. Trust is defined as not necessarily reflective of the system's objective performance, but instead as determined by the user perceptions and subjective weighting of different trust dimensions. It is conceived as a dynamic process, shaped over repeated interactions. Muir also introduced the concept of trust calibration as essential for appropriate reliance on decision aids. This involves improving users' ability to perceive a system's trustworthiness, modifying their criteria for trust, enhancing their skill in function allocation, and selectively recalibrating poorly aligned trust dimensions. Finally, she highlighted a paradox in human-machine interaction where human operators may lack control and authority over automated processes yet retain accountability for their outcomes, raising the risk of distrust or inappropriate reliance.

In 1997, Parasuraman and Riley's *Humans and Automation: Use, Misuse, Disuse, Abuse* was published [84]. This paper is a foundational work in the study of human-automation interaction, particularly influencing research on automation reliance [58]. It introduces the concepts of "use", "misuse", "disuse", and "abuse" to frame the interactions between human operators and automated systems. "Use" refers to appropriate reliance, which depends on factors such as mental workload, cognitive overhead, self-confidence, perceived risk, and trust in the system. "Misuse" arises when users over-rely on automation, leading to inadequate monitoring, omission errors or poor information discrimination. Excessive trust, cognitive shortcuts, workload levels or obscured system states can cause it. Mitigation strategies include designing salient state indicators, managing attention demands, integrating displays, and adaptive task allocation. Conversely, "disuse" involves underutilization, often through the deliberate disabling of automation or ignoring alarms due to excessive false alerts. "Abuse" reflects systemic failures, where designers or managers implement automation without adequately considering human roles, shifting potential failure from operators to flawed design logic.

Early applied research in automation has been critical in developing our understanding of these concepts, in particular, the study of aircraft autopilot systems [30]. These systems are essential to pilots, helping to reduce their cognitive load and improve the safety of routine flights by automating routine tasks (e.g., maintaining flight path, altitude, and speed), allowing pilots to instead focus on decision-making and monitoring. However, this is a critical, applied domain where any misunderstanding or miscommunication can lead to serious, potentially tragic consequences. In several studies on automated decision-making environments, specifically aircraft cockpits, Mosier and Skitka identified and defined the phenomenon of automation bias [104]. They described it as the "use of automation as a heuristic replacement for vigilant information seeking and processing", potentially

leading to both omission and commission errors in decision-making. This means that people make mistakes by overlooking important information or making wrong decisions based on incorrect information from the automated system.

In a 1999 study [102], they examined the impact of an automated monitoring aid (AMA) on participants' error rates in a flight simulation software. Participants were tasked with keeping their ship aligned with a moving target while tracking their position relative to waypoints and monitoring four gauges. Participants in the manipulation condition received automated aid that notified them when gauges were moving into the red zone or when approaching waypoints, and instructed them on the necessary action to take. The automated help improved participants' performance outside of its scripted events. However, participants with AMA made more mistakes of omission, dropping their accuracy from 97% to 59%. Additionally, the AMA did not help decrease the perceived workload of the participants. In another study [103], they investigated whether the presence of copilots reduces automation bias. They found no difference in omission errors between 1-person crews and 2-person crews. However, they did find that providing prior training on automation bias helped reduce commission errors.

Recent work on decision support systems has extended the study of automation reliance to include AI-based systems. While early automation research focused on physical systems (e.g., autopilots, monitoring aids) used by trained operators, modern AI assistants differ in being more data-driven, opaque, and adaptive, requiring users to interpret complex outputs [42]. However, over-reliance and under-reliance remain concerns. As such, concepts like automation bias have evolved into the concept of algorithmic bias or algorithm appreciation, describing users' tendency to prefer or defer to algorithmic recommendations over human ones.

Table 2.1 summarizes ten studies that have looked at reliance on AI assistance. Some of those studies use the term automation bias or algorithm appreciation. The table includes studies that compare preferences between human and algorithm advice, with columns detailing the experimental task, the conditions that were manipulated and the findings related to reliance.

Table 2.1: Studies comparing preferences between human and algorithm advice (yellow: subjective or social decisions, blue: medical decisions).

Study	Task	Conditions	Results
(Dijkstra, 1999)[19]	Deciding the correct verdict of law cases.	Concurrent choice between correct attorney analysis (not guilty) and incorrect expert system (guilty).	79% of decisions followed the expert system incorrect advice.

(Dzindolet et al, 2002)[22]	Visual discrimination task (finding a camouflaged soldier in a picture).	Advice from "a computer program designed for the task" or from "the decision made by the previous participant".	Better expected performance from the automated aid (M=6.71) than from the human aid (M=6.09). Less expected errors from the automated aid (M=24.79) than from the human aid (M=51.26).
(Logg et al, 2019)[62]	Estimating a person's weight from a photograph, before and after receiving advice.	Advice produced by either "an algorithm" or "the average estimate of [past] participants".	More conformation with an algorithm (45% on average) than with participants (30% on average).
	Forecasting the popularity of songs, before and after receiving advice.	Advice produced by either "an algorithm" or "an aggregation of 275 other participants".	More conformation with an algorithm (37% on average) than with participants (21% on average).
	Predicting how someone would judge a potential romantic partner, before and after receiving advice.	Advice produced by either "an algorithm" or an estimate by 48 people.	More conformation with an algorithm (38% on average) than with participants (26% on average).
	Guessing a person's weight from a photograph before and after receiving advice.	Advice produced by either "an algorithm" or "the average estimate of [past] participants", or a choice between both advice.	More conformation with an algorithm (50% on average) than with participants (35% on average). When given the choice, participants preferred receiving advice from an algorithm (75%) more than from another participant (25%).
	Estimating how a US state ranks in terms of airline passengers volume.	Choice between an algorithmic estimate or a human estimate (the participant's or someone else's) for the calculation of the bonus pay.	Participants preferred to base their bonus pay on the algorithmic estimate more than on the human estimate whether it was by another participant (88%) or their own (66%).

(Keding and Meissner, 2021)[44]	Reading an investment scenario and rating likelihood of investing.	Advice to invest from an "AI-based advisory system" or from a "human managerial team".	Participants receiving advice from the AI-based system rated their likelihood to invest higher (M=5.05) than those receiving advice from the managerial team (M=4.61).
(Vodrahall et al, 2022)[116]	Classification of art, city images, text containing sarcasm, or census data, before and after receiving advice.	Advice from an "average human prediction" or from an "AI prediction" both presented as having 80% accuracy.	Participants changed their answer more often and more significantly when given AI advice than human advice in the art (+6.4% activation), cities (+6.3% activation) and census (+5.7% activation) tasks, but not the sarcasm task (-5.8% activation).
	Classification of skin lesions by clinicians, before and after receiving advice.	Advice from an "average human prediction" or an "AI prediction".	There were no differences in how clinicians followed AI or human advice.
(Alon-Barkat and Busuioc, 2022)[3]	Choosing a teacher to fire based on conflicting qualitative assessments and numeric scores.	The score is either labeled as produced by an algorithm or by a human expert.	No significant differences were observed between the groups. Participants chose to follow the qualitative assessment and not the scores, regardless of the score's source.
(Castelo et al, 2019)[12]	Rating trust in an assistant for 26 task statements.	The assistant was presented as either an algorithm or a "very well qualified person".	Trust was on average higher in a qualified human (M=70.2) than in an algorithm (M=52.8), except for the tasks of predicting the stock market, predicting the weather, analyzing data and giving directions.
	Click-through rate (CTR) of Facebook ads for dating or financial advice.	The ads portrayed a human or an algorithmic advisor.	CTR for the dating advice was significantly higher with a human advisor (2.1%), than an algorithmic advisor (.5%). For the financial advice, CTR was also higher with a human advisor (1.8% vs. 1.6%).

	Rating preference for algorithm assistance relative to human (0=human preferred, 100=algorithm preferred) for 9 task statements.	Direct comparison between an algorithm and a qualified human. Half of the participants were informed that the algorithm had superior performance.	Without performance data, participants preferred human assistance (M=28.6). When given performance data, participants were indifferent between human and algorithm (M=50.4).
	Rating preference for algorithm assistance relative to human (0=human preferred, 100=algorithm preferred) for movie and romantic partner recommendation.	Direct comparison between an algorithm and a "friend" for the movie recommendation and a "professional matchmaker" for the romantic partner recommendation.	Participants trusted humans more than algorithm whether they were told to think of their emotions (M_{movie}=31.5, $M_{partner}$=36.4) or quantifiable data (M_{movie}=45.1, $M_{partner}$=43.8).
(Yeomans et al, 2019)[120]	Rating a set of three recommended jokes.	Recommendations labeled as from another user or from a recommender system.	Participants rated jokes recommendations higher when they believe the recommender was human (M=0.07 vs. M=-0.07).
(Soellner and Koenigstorfer, 2021)[107]	Estimating intention to follow medical recommendation after reading a diagnosis scenario.	Recommendation described as coming from a physician, a physician using AI or an automated AI tool.	Participants' intention to follow medical recommendation was significantly higher with a physician (assisted by AI or not) than an AI.
(Hou and Jung, 2021)[37]	Answering objective (quantifying number, length, color of geometric objects) and subjective (estimating people personnel opinions) questions, before and after receiving advice.	Suggestion described as from "another Mturker" or an "AI".	Participants were more influenced by advice from an algorithm (M=0.57) than from humans (M=0.45).

	Answering objective (quantifying number, length, color of geometric objects) and subjective (estimating people's opinions) questions, before and after receiving advice.	Agent described as either "a group of experts" or "an algorithm [...] aggregating Mturkers' responses".	Participants were more influenced by advice labeled as from humans (0.58) than from an algorithm (0.45).

The studies summarized in Table 2.1 reveal that in many tasks, particularly those involving objective or quantitative judgments, participants showed a tendency to rely more on algorithmic advice than on human recommendations. This occurred even when the algorithm was known to be incorrect or only marginally more accurate. For instance, participants in Dijkstra [19] overwhelmingly followed incorrect algorithmic advice in legal judgments, and Logg et al. [62] consistently found higher conformation with algorithmic suggestions across a variety of estimation tasks. Similarly, Dzindolet et al. [22] showed stronger perceived performance and influence from algorithmic sources compared to human ones.

However, this preference is not universal. Several studies, particularly those involving emotionally or socially complex decisions (e.g., romantic advice, medical diagnosis, humor, or personnel evaluation), found that participants either preferred human advice or were skeptical of algorithmic input. Castelo et al. [12] and Soellner and Koenigstorfer [107] highlight that trust in algorithms is sensitive to task type and framing: people are more likely to trust humans for subjective or high-stakes judgments. Hou and Jung [37] also showed that the framing of the advisor can shift users' reliance towards human or algorithms, demonstrating that algorithm appreciation or bias is highly context-dependent.

Literature investigating reliance bias in AI is scarcer regarding advice delivered by a conversational interface. One study by Zaroukian et al. [121] examined how users interacted with a conversational agent during a simulated intelligence, surveillance, and reconnaissance (ISR) task. Participants provided natural language input to the chatbot, which then generated a structured interpretation of their reports. Participants had the opportunity to accept or reject the agent's interpretation. Despite this, users frequently accepted misparsed outputs, incorrect interpretations that should have been rejected. The authors attributed this behavior to automation-induced complacency, possibly from the agent's high overall accuracy (92%), which led users to over-trust its outputs.

However, few works directly compare reliance on conversational AI with reliance on human assistants. A direct comparison is crucial to determine whether differences in reliance stem from the nature of the advisor, specifically whether it is algorithmic or human, rather than from the mere presence of advice. Additionally, conversational agents often offer more human-like interaction, making them more socially and affectively engaging than other AI systems. This increased anthropomorphism could either mitigate or exacerbate reliance biases, depending on the context.

2.1.2 Trust calibration and appropriate reliance

After establishing difference in reliance, some studies looked at how to calibrate this reliance, measuring and react to it to prevent over- or under-reliance.

According to Hancock et al.'s meta-analysis on trust in human-robot interaction, the most significant predictors of user trust are the robot itself, specifically its performance and attributes [33]. Human and environmental factors showed only limited influence, indicating that trust can be more effectively calibrated through adaptations of the robot's design.

Interestingly, research by Van der Hoorn et al. [113] suggests that robots are often trusted more when they incorrectly blame themselves rather than accurately pointing out a human collaborator's mistake. This effect aligns with the human self-serving bias, where people prefer to take credit for successes and avoid blame for failures.

Additionally, Nesset showed that transparency is only beneficial to trust if the communicated error rate of the AI system is low, whereas with a high error rate it is better for the system to have low transparency [81]. This underscores the importance of carefully designing how robots communicate errors or negative outcomes in ways that maintain trust and support effective long-term collaboration.

Herse et al. [35] explored a possible new calibration methodology to foster appropriate trust to enhance human-agent collaboration. In their study, participants first interacted with a virtual assistant in an adapted version of the Trust Game to obtain an indirect measure of trust. They were then asked to classify 40 X-ray images with assistance from the agent, which provided diagnostic suggestions with a 50% accuracy rate. Trust was measured behaviorally via participants' decisions in the Trust Game, and decision bias was quantified using signal detection theory (SDT). The study demonstrated that trust in the agent could be indirectly assessed, as behavior during the trust game correlated with participants' bias. Additionally, participants with higher initial trust exhibited less bias.

Kupfer et al. [53] examined automation bias in AI-based personnel selection, highlighting strategies to foster appropriate reliance. They observed that higher scores on verification intensity indicators, reflecting less automation bias, were correlated with higher objective decision quality. Their study also found that providing decision-makers with information about potential system errors increased verification efforts and improved subjective decision quality, whereas emphasizing personal responsibility had little effect. These findings suggest that trust calibration can be supported by educating users about AI limitations and providing less aggregated, more detailed information.

Buçinca et al. [10] explored Cognitive Forcing Functions (CFF), interventions designed to encourage analytical thinking and reduce over-reliance on AI in decision-making. In their study, CFFs included mechanisms such as requiring users to request the AI's suggestion, allowing them to revise an initial decision after seeing the AI's recommendation, or introducing a brief delay before showing the AI output. These interventions were intended to prompt users to actively engage with the AI's explanation rather than accept it blindly. The study found that participants who experienced CFFs were less likely to over-rely on AI, in contrast to traditional explainable AI methods, which often failed to reduce over-

reliance. However, while CFFs improved decision accuracy, they were associated with lower user satisfaction.

De Jong et al. [17] investigated the impact of partial explanations on over-reliance in AI-assisted decision-making in two studies, a graph-based task and a spelling and grammar task. They found that explanations generally increased reliance, with full explanations leading to greater acceptance of the AI's outputs. Their results further showed that partial explanations reduced over-reliance on incorrect suggestions but also introduced more under-reliance when the AI was correct. Finally, they observed that the degree of reliance varied depending on the complexity of the task.

In conclusion, AI assistance works best when users trust it enough to benefit from its support, but not so much that it overrides their own judgment. Approaches like clear communication of errors, prompting users to engage critically, and providing tailored explanations can help reach this balance, reducing both over- and under-reliance without adding unnecessary cognitive load. These insights set the stage for exploring chatbots, whose conversational nature brings unique opportunities and challenges for supporting collaboration and decision-making in groups.

2.2 Chatbot assistance

Chatbots are a distinct subcategory of artificial intelligence, defined by their ability to communicate through natural language text and their lack of physical embodiment. This text-based modality makes them especially effective at conveying large volumes of information efficiently. Moreover, chatbots can be easily integrated into existing chat tools and workflows, often requiring minimal modifications to current systems. For researchers, chatbots also offer the advantage of generating complete interaction logs, which can be analyzed to better understand human-agent communication dynamics.

In recent years, the number of studies investigating chatbot interactions has grown substantially, driven by technological advancements and the expanding adoption of chatbots across various domains [122]. Research has explored their applications in various areas, including customer service, mental health support, education, and team coordination, revealing diverse benefits and challenges.

Recent studies, such as those by Schmidhuber et al. [95], have demonstrated that chatbot assistance can play a significant role in alleviating users' cognitive load during complex tasks. Their findings indicate that while interaction with chatbots may lead to slightly longer task completion times, it simultaneously reduces mental effort and frustration.

2.2.1 Anthropomorphism and social presence

An essential aspect of chatbots is their anthropomorphism, which refers to their more human-like nature compared to other forms of AI. Chatbots with their anthropomorphism and use of natural language, have a greater social presence [52] and similarity to humans than other decision aid systems. This could be leveraged to reduce the likelihood of automation bias. Additionally, unlike embodied robots, chatbots are also less likely to

provoke feelings of uncanniness, since they have fewer modalities, such as motion or facial expressions, where perceptual inconsistencies can occur [13, 63].

The Computers Are Social Actors (CASA) paradigm suggests that people tend to apply social rules and behaviors to computers as if they were human [79, 59]. According to this paradigm, chatbots would benefit from anthropomorphism in contradiction to the concept of automation bias, which would instead predict that more human chatbots are less convincing [97]. In a survey, Rietz et al. [90] found that anthropomorphism was the primary determinant of the perceived usefulness of chatbots.

Even with anthropomorphism, there are differences in how people perceive and interact with chatbots compared to other humans. In Hill et al.'s study [36], a comparison of conversations with the chatbot Cleverbot and conversations between humans showed that with chatbots, people tended to have longer conversations but composed of shorter messages. Additionally, participants used a more limited vocabulary and exhibited a tendency toward more vulgar language when communicating with the chatbot.

People also do not hold chatbots to the same privacy expectations as they would human interlocutors. Sannon et al. [93] explored users' privacy expectations when interacting with conversational agents, using a 3x3 design that varied both the chatbot's personality (form-like, basic chat, or friendly) and its data-sharing behavior (no sharing, sharing with the company, or sharing with third parties). In the condition when the chatbot shared its data with its parent company, users did not perceive it as a breach of confidentiality. In fact, participants reacted similarly to this condition as they did when no data was shared at all. This suggests that users generally assume chatbots will share information with the companies that created them. However, when data was shared with third parties such as advertisers, participants reacted more negatively and perceived this as a more serious privacy violation. More socially interactive chatbots, those with friendly, human-like behaviors, were judged less harshly for the same data-sharing practices, indicating a potential privacy risk: users may be more trusting, and thus more vulnerable, when a chatbot presents itself as socially engaging.

2.2.2 Reliance on chatbots

According to Brandtzaeg and Følstad [8], productivity is identified as the primary motivation for using chatbots, followed by entertainment, social interaction, and novelty. They also noted that some people find chatbots easier to converse with than with people, especially when discussing important issues.

However, questions of trust, as well as data security, are still obstacles to their implementation [87, 118]. This point has also been raised as a weakness for chatbots in the humanitarian field [72].

One of the important questions for chatbot research is trust as it is a key factor to their proper usage. Research on chatbot has identified multiple factors for trust in chatbots.

Følstad et al. [26] identify two categories of factors for chatbot trust. They differentiate between internal attributes, such as answer relevance, humanity, limits transparency, general polish, and external attributes of the chatbot, such as brand, security, confidentiality,

task significance.

In comparison, Kerly et al. study [45] on the subject of Open Learner Models identifies a single factor more important than all others: the ability of the chatbot to provide a rationale to support its advice.

In Müller et al. [77], trust in chatbots was investigated through the lens of the HEXACO personality inventory. The authors used latent profile analysis to identify three distinct personality profiles based on users' levels of Extraversion, Agreeableness, and Honesty-Humility. These profiles, "introverted, careless, distrusting users"; "conscientious, curious, trusting users"; and "careless, dishonest, trusting users", differed significantly in how much they trusted chatbots such as Alexa. Their findings suggest that willingness to trust chatbots is strongly influenced by personality traits, with Extraversion and Agreeableness being key predictors of higher trust, and Honesty-Humility playing a more modest role.

Taken together, research on chatbot assistance highlights both the potential and the limitations of these systems. On the one hand, chatbots can reduce cognitive load, enhance efficiency, and provide socially engaging interactions through their anthropomorphic qualities. On the other hand, their effectiveness is shaped by complex factors such as trust, perceived usefulness, and user personality traits. While anthropomorphism and social presence can increase acceptance and mitigate feelings of uncanniness, they also raise concerns around privacy and data sharing, as users may lower their guard when interacting with more human-like agents.

2.3 Chatbots and teamwork

In team settings, chatbots can function as supportive collaborators that enhance communication, coordination, and overall team efficiency. By providing timely information, reminders, and structured assistance, they can help teams maintain alignment and reduce the cognitive burden associated with managing tasks and discussions. Their role may range from passive support, such as logging decisions or tracking progress, to more active involvement, like prompting discussions or facilitating knowledge sharing.

Even in dyadic interactions, that is, conversations between two people, chatbots have the potential to support collaboration. They can improve shared understanding by summarizing or clarifying points during conversations, automate team management tasks such as scheduling or task delegation, and help regularize vocabulary across participants by consistently applying terminology. These abilities can foster smoother communication and coordination in the team without intervening in more complex team interactions.

2.3.1 Challenges of group conversations

Extending the positive effects observed in dyadic interactions to larger groups presents additional challenges. Most studies are currently limited to dyadic interactions [98], which do not readily capture the complexity of multi-party conversations. In group conversations, chatbots are not the sole participants, most messages are likely between humans and not

intended for the chatbot. Chatbots in these settings primarily support the conversation, either by mediating or reacting to specific messages, and should not hinder teamwork with excessive or irrelevant messages that disrupt the conversation's flow [25, 7].

As identified by Zheng et al. [122], polyadic interactions, or group conversations, present distinct considerations for conversational agents (CAs) compared to dyadic ones. Their review of 36 studies on polyadic CAs and 135 on dyadic CAs highlights that research on agents in group settings remains relatively limited. Polyadic CAs are specifically designed to address collaboration issues across several dimensions: communication efficiency (e.g., workflow support, conversation structuring, consensus building, task management, schedule coordination, and reducing tool switching), group engagement (e.g., encouraging participation, balancing uneven contributions), relationship maintenance (e.g., fostering emotional awareness and mutual understanding, trust-building, managing privacy boundaries), and connection building (e.g., facilitating ice-breaking, supporting idea generation, enabling cross-cultural communication).

The review also highlights ongoing challenges for polyadic CAs, particularly in striking a balance between being visible (so users remain aware of their presence) and ignorable (so they do not intrude unnecessarily), as well as in determining accountability, and in being accountable in situations where users disagree. The authors conclude by emphasizing the research opportunities in this area, pointing to polyadic CAs as an underexplored design space that remains weakly integrated with existing theoretical frameworks. They suggest that future work should focus on developing the concept of boundary-awareness as a core principle for designing effective and ethical polyadic agents.

A key challenge for chatbots in group settings is determining when to speak and when to remain silent. In multi-person conversations, most messages aren't directed at the bot, and poorly timed interruptions can disrupt the flow. Good timing is essential; chatbots should step in only when they can add value, like clarifying something or nudging the group forward. Studies have shown that users get annoyed when chatbots talk too much or at the wrong moment [38]. To avoid this, bots must be aware of the context and pick up on cues that indicate whether their input is welcome or not. Designing for this kind of turn-taking is still an open problem, but it's crucial for making chatbots helpful teammates rather than distractions.

We identified two main subjects of interest for research on chatbots in polyadic interaction in the literature: chatbot-assisted collaborative learning and bots in work collaboration tools.

2.3.2 Chatbots integrated into work collaboration tools

One domain of interest in chatbot interaction with groups is the integration of chatbots into work communication tools, such as Slack or Google Chat. These bots can perform a variety of tasks, including sending reminders, managing calendars, and retrieving information from other apps without requiring users to leave the chat. They can also help with tasks like creating tickets, approving requests, notifying user about important events, help new team members by answering common questions, sharing useful documents, helping schedule meetings, or posting updates about code changes [100].

Bimbot, a chatbot integrated into BIM (Building Information Modeling) software, assists users in managing architectural and engineering tasks efficiently in teams by promoting shared understandings and providing advice [78].

Wang et al. [117] collected datasets of instant messaging usage from over 4,300 Slack group chat channels within the R&D department of a multinational IT company during the COVID-19 pandemic, aiming to understand how organizational chat practices relate to team performance. They complemented this with performance data from 117 project teams, based on whether they submitted work to an academic conference or not. Through a sample of 100 group chats, they identified nine categories: project, social group, IT support, employee support, tech enthusiasts, event, bot, test, and announcement. They then trained a machine learning classifier on meta information (message length, thread activity, reactions, emojis, ...) without looking at the messages content. A regression analysis on a subset of 57 project teams revealed a modest but notable correlation between certain communication features and successful submission outcomes. These predictors are active timespan, number of bot messages, @here mentions, and normalized counts of messages from the top contributor, threaded discussions, emojis, and code snippets. Notably, communication styles varied widely across categories: project channels tended to have few participants and short lifespans but a high volume of file sharing, while social groups emphasized emoji use and reactions, and IT support channels featured more threaded discussions, code snippets, and bot interactions. These findings show that not all group chats serve the same function or have the same dynamics, and consequently, should not be treated uniformly in analyses or when designing new systems, such as chatbots.

Laitinen et al. [56] examined the impact of a Slackbot message on a journalistic team, with interactions ranging from greetings and acknowledgments to work-related tasks such as developing story ideas or coordinating work distribution and planning. They concluded that bots can facilitate team relational communication by increasing the number of socioemotional and personal interactions.

Lebeuf et al. [57] identified areas where software developers can leverage chatbots to mitigate friction in their collaborative work environment. At the team level, these challenges include understanding team members' roles and expertise, adhering to shared procedures and agreements, aligning with team goals, coordinating activities, and maintaining trust and cooperation among team members. On the individual level, developers struggle with distracting and interruptive technologies, keeping up with emerging tools and practices, and understanding the affordances of various communication channels. At the intersection of teams and technology, additional friction points emerge, including information overload, difficulty in adopting or contextualizing tools within specific team workflows, poor awareness of ongoing project activities, inadequate collaboration tooling, and frequent miscommunication on text-based platforms.

To investigate the possibility of using chatbots for task management, Toxtli et al. [111] designed and tested a chatbot, named Taskbot, for Microsoft Teams. The chatbot is capable of handling both reactive (response to Taskbot's questions) and proactive messages (messages mentioning Taskbot or direct messages) from users. After testing it with eight teams of information workers, they observed that the chatbot was sometimes used for self-communication and was challenged when multiple conversations occurred simultaneously

or when dealing with ambiguous messages lacking context.

Stoeckli et al. [108] conducted a survey of the bots available on the work application Slack. They identified 14 functionalities grouped into four categories: message reception (aggregated information, notifications, performance indicators), setting and receiving triggers (reminders, team management), querying third-party systems (information lookup, data capture, functionality invocation), and enriching messages with links, metadata, or GIFs. Their findings revealed that most Slack chatbots lack these affordances, and users are often unaware of many functionalities unless they are directly demonstrated by others. The menu and buttons make those functionalities clearer to users.

These studies highlight the versatility and potential benefits of chatbots integrated into collaborative work environments. They show that chatbots can support task coordination, reduce friction in team workflows, facilitate information sharing, and even enhance relational communication among team members. However, those studies are mostly surveys of existing chatbot implementations, work tools, or teams, and while they provide valuable insights into current usage patterns and perceived benefits, they offer limited empirical evidence on the direct impact of chatbot interventions on team performance or dynamics. Few controlled experiments have been conducted to systematically evaluate how chatbot behaviors (timing, tone, or message frequency) affect collaboration, coordination, or trust within teams. Moreover, many of these bots are designed with narrow, task-specific functions, leaving open questions about whether more conversational or socially intelligent chatbots could play a broader role in team workflows.

2.3.3 Chatbots for collaborative learning and group discussions

One of the main research interests of studies on chatbot interactions with groups are chatbots acting as debate moderators and structuring group discussions [122, 7]. Chatbots have been shown to help with balancing participation between participants [20, 47], argument quality [110] as well as team cohesion [47, 6].

Kim et al. study [47] presents GroupfeedBot, a chatbot that facilitates group discussions by structuring the conversation, encouraging even participation between members, and organizing group participants' opinions. GroupfeedBot was designed with the capacity to manage time, elicit comments from the least contributing members, and organize and summarize members' opinions. Groups of four or five participants completed four tasks, two assisted by a simple chatbot and two assisted by GroupfeedBot. On tasks assisted by Groupfeedbot, groups showed increased opinion diversity and more evenly distributed participation. Another study of Kim et al. [49] with the chatbot DebateBot also showed an increased discussion quality and balanced participation, thanks to the chatbot facilitation.

The delivery method of the assistance is also essential. Indeed, when interacting with a group, it matters whether advice and directions are given to individuals or the group and whether only the recipient or everyone can see it. Tegos et al. pilot study [110] presents their chatbot MentorChat, which was made to support collaborative learning. They looked at differences in effectiveness between directly prompting the weaker member of a group of two or three students or addressing the group as a whole. Using discourse

analysis, they found that directly prompting a specific student resulted in considerably more explicitly reasoned responses. However, in the post-task survey, the students preferred the undirected intervention strategy to the direct one.

Do et al. [20] investigated potential communication strategies for chatbots to strike a balance in group ideation participation. The chatbot would prompt participants either as a group or as an individual, in public or in private, or ask another participant privately to prompt the least contributing member. The private messages were displayed in the chat but were only visible to the concerned participant. The study found that participants preferred and were more reactive to being prompted individually and privately, and that the peer-mediated strategy was viewed as less intrusive and embarrassing than other strategies.

Overall, these studies suggest that chatbots can effectively foster more inclusive, structured, and higher-quality discussions. However, these studies are less relevant for collaborations that do not involve structured debates and where conversation is not the main objective. In many workplaces, balanced participation often takes a backseat to measurable performance metrics such as speed, work quality, and error rates.

2.4 Conclusion

We have reviewed the literature on AI assistance, focusing on how users rely on and interact with automated and algorithmic systems. Foundational research on automation has introduced key concepts such as appropriate reliance, automation bias, and trust calibration, highlighting the importance of designing AI systems that support human decision-making without introducing cognitive strain or overdependence. Recent studies have extended these insights to modern AI assistants, showing that algorithmic advice is often preferred over human input in objective tasks, but less so in more subjective or socially sensitive contexts. Importantly, reliance is not only shaped by system performance but also by how the advice is delivered, perceived, and framed. Despite this progress, most of the existing research focuses on non-conversational interfaces or one-on-one interactions. Few studies have explored how these dynamics unfold when AI advice is delivered by chatbots in group settings, where factors such as social presence, conversational flow, and group dynamics add complexity.

We explored how chatbots differ from other AI systems through their use of natural language, ease of integration, and ability to log interactions. Their human-like qualities give them greater social presence, which can affect how users relate to them. While factors like relevance, transparency, and brand influence trust, studies also show that users don't treat chatbots like people and may not hold them to the same standards. Overall, trust in chatbots is shaped by both their design and the context of use, making it especially important to understand how these dynamics play out in collaborative settings.

Finally, the research on chatbots in group settings reveals promising roles in facilitating teamwork, particularly through collaborative learning and structured discussions. Studies show that chatbots can help balance participation, improve the quality of group dialogue, and support team cohesion by encouraging contributions and organizing opin-

ions. However, the effectiveness of these bots depends greatly on how their interventions are delivered, whether they address individuals or the group, and whether prompts are public or private.

To address the gaps identified in the literature, the next chapter will present a confirmatory study examining how the specificities of chatbot interaction influence users' reliance on the advice provided. We will then introduce a novel group activity specifically designed to facilitate controlled experimentation on group dynamics. Using this activity, we aim to uncover both the benefits and potential challenges of chatbot assistance within collaborative teams, providing insights to guide the design of more effective and trusted group-oriented chatbots.

Chapter 3

Chatbot vs. Human Advice

Contents

3.1 Introduction 31
3.2 Context 33
3.2.1 Related work 33
3.2.2 Design considerations 34
3.3 Experimental design 35
3.3.1 Assistant types 36
3.3.2 Questionnaires 39
3.4 Results 40
3.4.1 Messages length and response time 41
3.4.2 Messages content 41
3.4.3 Answer changes 45
3.4.4 Questionnaires 46
3.5 Discussion 51
3.5.1 Summary of results 51
3.5.2 Design implications 51
3.5.3 Limits 52
3.6 Conclusion 53

3.1 Introduction

As chatbots become integrated into collaborative environments as teammates, it is essential to understand how people rely on them. Flawed reliance at the individual level can lead to errors in judgment, reduce critical thinking, create overdependence on automated systems, and compromise professional responsibility. At the team level, it can

disrupt group dynamics and hinder effective collaboration by influencing consensus and reducing diversity of thought. If a chatbot advisor is not accepted and relied upon in the workflow, it can lead to disuse, causing important updates or insights to be overlooked and potentially wasting time and resources. Conversely, over-reliance may suppress human contributors, either through uncritical acceptance of chatbot recommendations or self-censorship among team members reluctant to challenge its input. By deferring to a unified perspective shaped by the chatbot, users risk stifling creativity, diluting accountability, and diminishing their sense of ownership over final decisions [29, 114].

The literature on human-AI interaction reveals a consistent pattern of automation bias, where users are more likely to follow advice, whether correct or incorrect, when it comes from an AI system [30]. Research into automation bias has primarily focused on traditional AI systems and algorithms, with limited exploration of how this bias manifests in interactions with chatbots.

Chatbots warrant special attention because of their distinct characteristics. Unlike other AI systems, chatbots engage users in natural language conversations, creating a stronger sense of social presence. This human-like interaction might make users anthropomorphize chatbots and perceive them as more similar to human collaborators, potentially altering how they engage with chatbots compared to other AI systems [97] and reducing bias.

While prior work has explored automation bias in algorithmic or system-level AI contexts, few studies have investigated its presence in dynamic, conversational AI interactions. To better understand how chatbot assistance influences user decision-making, more specifically how automation bias manifests in interactions with chatbots, we conducted a study with the following objectives:

> *1. Determine whether automation bias occurs in interactions with chatbots by comparing user behavior when receiving advice from chatbots or human advisors.*
>
> *2. Identify the factors behind automation bias, if there is bias, in order to detect and mitigate it.*

In order to investigate these questions, we designed a between-subjects study comparing decision-making behavior when participants received advice from either a chatbot or a human assistant. Each participant was tasked with answering 20 questions, choosing from three answer options. For half of the questions, participants received advice from their assigned assistant after their initial answer, which contradicted their response and had the opportunity to change their answer. The questions focused on wilderness survival, a topic where participants were not experts and where there was no single ground truth. We used a Wizard of Oz setup to prevent differences in the assistants' abilities, so that any observed effects on behavior and decision-making were due to the presentation and expectation differences rather than differences in assistant performance.

After completing the task, participants completed a survey assessing their perceived performance, the impact of the assistant on their answers, and their perception of the

assistant's usefulness and adaptability. In total, 48 participants were evenly split across the two conditions.

Our analysis revealed that participants were twice as likely to change their answer when contradicted by a chatbot (63.7%) compared to a human assistant (31%). This difference in decision-making was also reflected in participants' attitudes: when interacting with the chatbot, they were less likely to explain their decision-making process, such as expressing hesitation or providing justifications, and less inclined to argue than when engaging with the human assistant.

These findings suggest that automation bias exists in interactions with chatbots, as evidenced not only by the higher likelihood of participants changing their answers but also by their reduced engagement in the decision-making process. Specifically, participants' reluctance to articulate their reasoning and their decreased willingness to challenge chatbot advice indicate a reduced critical stance and sense of responsibility. This pattern highlights how chatbots may inadvertently discourage reflective thinking and critical evaluation, increasing users' susceptibility to over-reliance on automated recommendations.

This chapter begins with the rationale behind our experimental design in regards to the state of the art in AI advice and automation bias research. We will then provide a detailed description of the experimental design and present the results of the experiment, our subsequent analysis, and our conclusions.

3.2 Context

3.2.1 Related work

Before conducting our study, we reviewed the existing literature on automation bias to better understand how collaborating with AI influences users' behavior. Several studies, such as the work of Skitka and Mosier [102, 75], explored how AI advice impacts decision-making and performance. However, these studies typically did not compare the effect of receiving advice from a chatbot to that of receiving advice from another source. In our study, we aimed to directly compare the influence of chatbot and human advice on participants' decision-making behavior. By holding the advice content constant and varying only the source (chatbot vs. human), we sought to isolate the effect of the advisor's identity. This design allowed us to distinguish whether differences in participants' behavior were caused by a general susceptibility to advice or by specific biases related to the perceived nature of the advisor.

In the previous chapter, Table 2.1 presents a summary of studies comparing advice acceptance and reliance based on whether the advice source is human or AI. The majority of those studies show evidence of automation bias, where participants are more likely to rely on AI advice than on human advice. However, automation bias is not always observed. Some studies found no significant algorithmic bias [116, 3], and some even found algorithm aversion [12, 120, 107], where participants showed a preference for human advice over AI. This was most commonly observed in studies involving subjective tasks or medical decision-making, where interpersonal rapport and communication are often valued more

than technical proficiency alone.

Importantly, while much of the existing research compares generic AI to human advice, there is a lack of studies specifically examining chatbots. To address this gap, we designed an experimental setup that highlighted the unique nature of the chatbot assistant by emphasizing presentation differences between the chatbot and the human assistant. Participants interacted with the assistants through a chat-based interface, where they were asked and answered questions within the same chat environment. Unlike previous studies where AI and human assistants might simply be labeled as such, our setup allowed participants to engage in real-time conversations with the assistants, providing a more natural interaction that more closely mirrors how chatbots are used in real-world settings.

3.2.2 Design considerations

We selected wilderness survival as the theme for our questions to avoid overly sensitive or high-stakes topics that could introduce external biases and are not representative of the routine scenarios in which chatbots are commonly used. Wilderness survival is a topic where participants generally have no emotional investment and minimal to no expert knowledge. Additionally, this is not a task that benefits from complex analytical reasoning, such as chess or stock market predictions, where AI's computational advantages might lead to an expectation of better performance compared to a human expert. This choice aimed to ensure that participants were unlikely to assume that AI would outperform a human assistant based on perceived superior capabilities.

Consequently, an important limitation of our study design is that the questions do not have numerical answers that would allow us to measure the strength of the advice (i.e., the percentage of change between a participant's initial answer and the recommended answer). Instead, we can only count the number of instances in which participants change their answer based on the advice they received.

As demonstrated by the study conducted by Hou and Jung [37], framing plays a crucial role in shaping participants' perceptions of the advice source and determining whether algorithm bias or algorithm aversion happens. Their findings show that how the assistant is labeled, whether as an "expert" or a "human", an "AI" or an "algorithm", can significantly influence how participants respond to its advice. To address this in our study, we attempted to keep the implied expertise and authority of the chatbot assistant and the human assistant as equal as possible. Rather than relying on a single label to establish framing, the assistants' status and nature were conveyed through multiple sources of information and visual cues (see section 3.3.1) designed to be coherent with their respective identities.

Finally, like most studies on automation bias, we chose a between-subjects design to compare the AI and human assistants. A within-subjects design, where participants would interact with both assistants or chose between their advice, could lead to distinction bias [39]. Such a design might encourage participants to make direct comparisons between the two assistants, which could undermine the goal of our study to reveal implicit biases in advice acceptance. By using a between-subjects design, we ensured that participants'

responses were not influenced by the artificial distinction between the two assistants but were instead caused by their general biases toward AI versus human advice.

In summary, while existing literature has highlighted automation bias in advice acceptance, significant gaps remain, particularly regarding chatbots. Our study contributes to this field by directly comparing advice from humans and chatbots within a conversational setting, an experimental design that emphasizes the specificities of chatbots and enables additional qualitative analysis of participants' interactions.

3.3 Experimental design

Our goal is to understand the difference in user decision-making when receiving advice from a human or a chatbot. We designed an experiment with two conditions (between-group design) comparing users answering questions and receiving advice from a chatbot or a human assistant when the only aspect differentiating the two is their presentations. The task involves answering questions on wilderness survival on an instant messaging platform, as well as two questionnaires filled out before and after the experimental task (Figure 3.1).

We chose wilderness survival as a topic to ensure questions were easily understandable for any participant but on a subject where people rarely have expert knowledge so they would need and accept suggestions more easily. The questions have no definitive answer, because we are interested in complex and more organic collaborations where there is no ground truth and the correct action to take is ambiguous. With this subject, the assistant can imply to the participant that it does not agree with their initial answer and challenge it without being overtly incorrect. We are not interested in knowing if chatbots can help participants get closer to an objective true answer or improve their performance, we want to see how a chatbot's interventions affect decision-making compared to a human, to which extent, and in which ways.

The questions are inspired by survival scenarios and team-building exercises, such as the NASA Moon Survival Problem [32] and the Desert Survival Situation [55]. In these exercises, participants are presented with a scenario of being stranded on the moon or in the desert and must rank items by their importance for survival, then compare their choices with expert rankings to evaluate their teamwork and decision-making skills.

As part of the main experimental task, participants answered 20 questions on the topic of wilderness survival. Each question offered three possible answers, and participants were informed beforehand that they could request a hint before making their choice (Figure 3.2). A complete list of the questions, answer options, and hints is provided in Appendix A.

While some answers were more well-supported than others, the study did not focus on determining whether participants' answers were correct, since the goal was to observe how easily they could be influenced by the assistant. Questions were asked in random

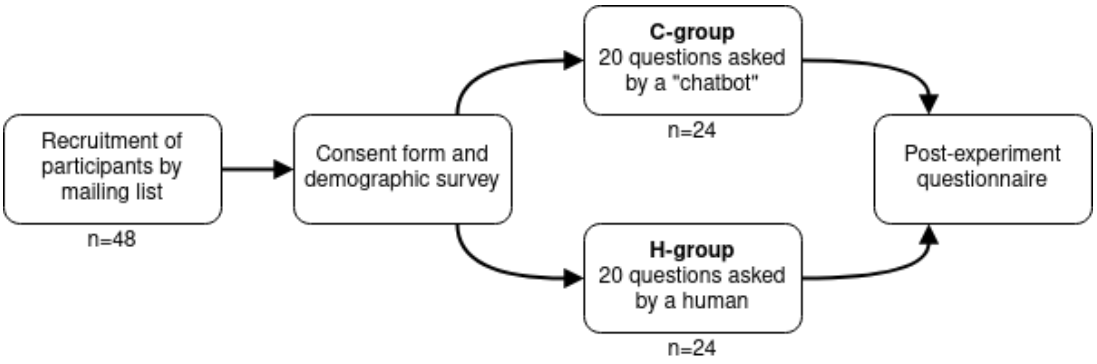


Figure 3.1: Experimental design

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 and immediately followed with a hint
- *contradiction*: 5 questions were asked and followed by the assistant contradicting the answer given and allowing the participant to change his answer
- *contradiction and hint*: 5 questions were asked and followed by the assistant contradicting the answer given with a hint and allowing the participant to change his answer

With this approach, participants' answers were challenged only half of the time. Our primary measure was the rate at which participants changed their answers, with the additional comparison of contradictions accompanied with a hint to those without. By limiting challenges to half of the questions, we wanted to make the assistant sound less confrontational and create the impression of a more natural and collaborative interaction.

We kept the same modalities regardless of which answer the participant chose. Two hints were prepared for each question so the assistant could always contradict the participant's answer and direct them to another one. However, if a participant voluntarily requested a hint before providing their initial answer, further hints and contradictions were omitted, so as not to undermine the participant's perception of having followed the assistant's advice.

We used the open-source messaging platform Mattermost[7] for the experimental task. We chose it for its commonalities with most professional messaging platforms (Slack, Microsoft Teams) and easy chatbot integration.

3.3.1 Assistant types

Participants were split in two groups: the assistant was either presented as a chatbot or as a human experimenter, respectively named COCCO and Hugo. The human researcher

[7]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

A. Place a small piece of the plant on your lip a few minutes before eating it. B. Avoid brightly colored plants. C. Try small quantities of what the birds are eating.

(b) Possible answers

1. Birds have a different digestive system from ours. 2. Some plants can be lethal on contact.

(c) Hints

Figure 3.2: Example of question, answers and hints given to the participants

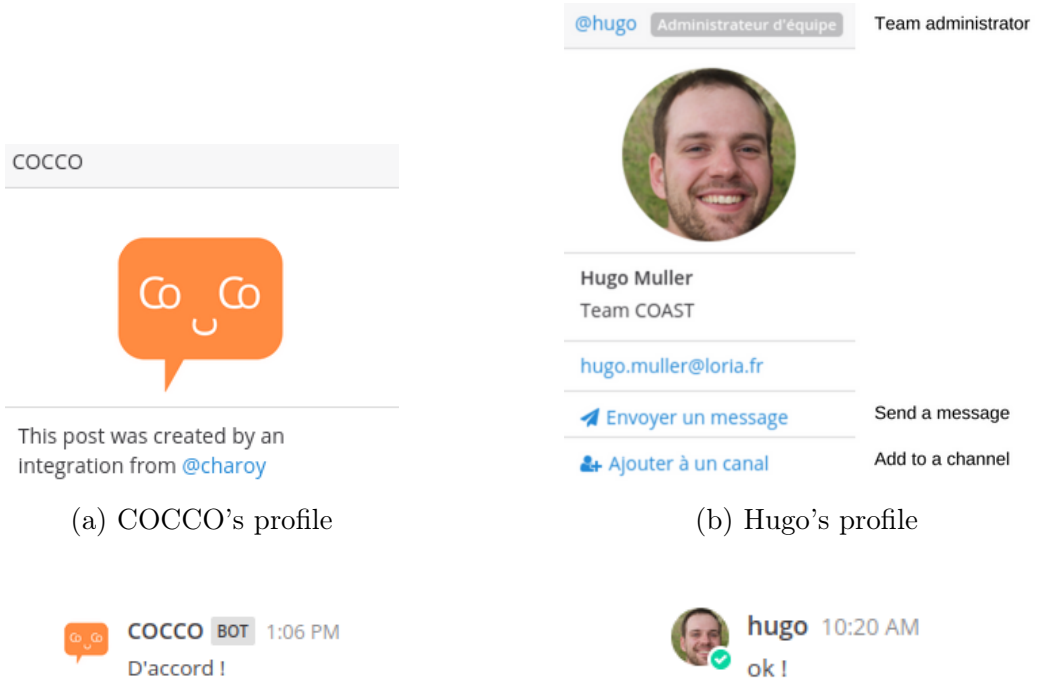
did not use their own identity to avoid the participants having prior familiarity with them from the recruiting period [51]. Instead the human assistant profile was filled using non-existing credentials and a generated and edited picture. For the chatbot condition, no chatbot was implemented for this experiment. Instead, we used a Wizard of Oz set-up in order to not introduce any capabilities differences.

The two conditions are differentiated primarily by the presentation of the assistant: chatbot or human, whose capacities and interventions are the same, as we are only interested in the difference in the participants' attitudes caused by how they perceive the assistants. Both assistants followed the same script. Importantly, neither assistant presented themselves as possessing task-specific expertise. We identified the following differences between the two experimental conditions:

- Name: The human assistant was named Hugo, a common and relatively popular masculine given name in France. The chatbot assistant was named COCCO, a synthetic name standing for COnfiance (trust), Chatbot, and COllaboration, that is short, distinctive, and easy to pronounce and read.

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

- Profile: We completed the profile of Hugo to show his status as a researcher of the team experimenting (Figure 3.3b), while COCCO's profile reflected its status as a third-party integration (Figure 3.3a).



(a) COCCO's profile

(b) Hugo's profile

(c) Example message by COCCO

(d) Example message by Hugo

Figure 3.3: Differences in presentation between COCCO and Hugo

- Status: Hugo had an online indicator next to his name (Figure 3.3d). COCCO had a bot badge automatically assigned by the messaging app (Figure 3.3c).
- Typing: Aside from the questions, which were copy-pasted, Hugo's messages were manually typed and thus 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, automatically labeling the message as "edited".
- Structural messages: COCCO systematically announced the next question or hint. Hugo only did it occasionally to keep his speech natural.

The experimenter sent COCCO's messages via the messaging platform API. We developed an interface that generated a script for the experimenter to follow with the questions randomized and assigned to one of the modalities. The interface comprised buttons to send all scripted messages necessary to communicate as COCCO (explaining the experiment, saying yes or no, asking if the participant needs a hint, ...), as well as a message field for unforeseen interactions (Figure 3.4). The interface was also used by the experimenter in the human assistant condition to consult the script and copy-paste the questions, an-

swers, and hints. In human assistant mode the buttons do not send messages via the API and instead copy the text to the clipboard.

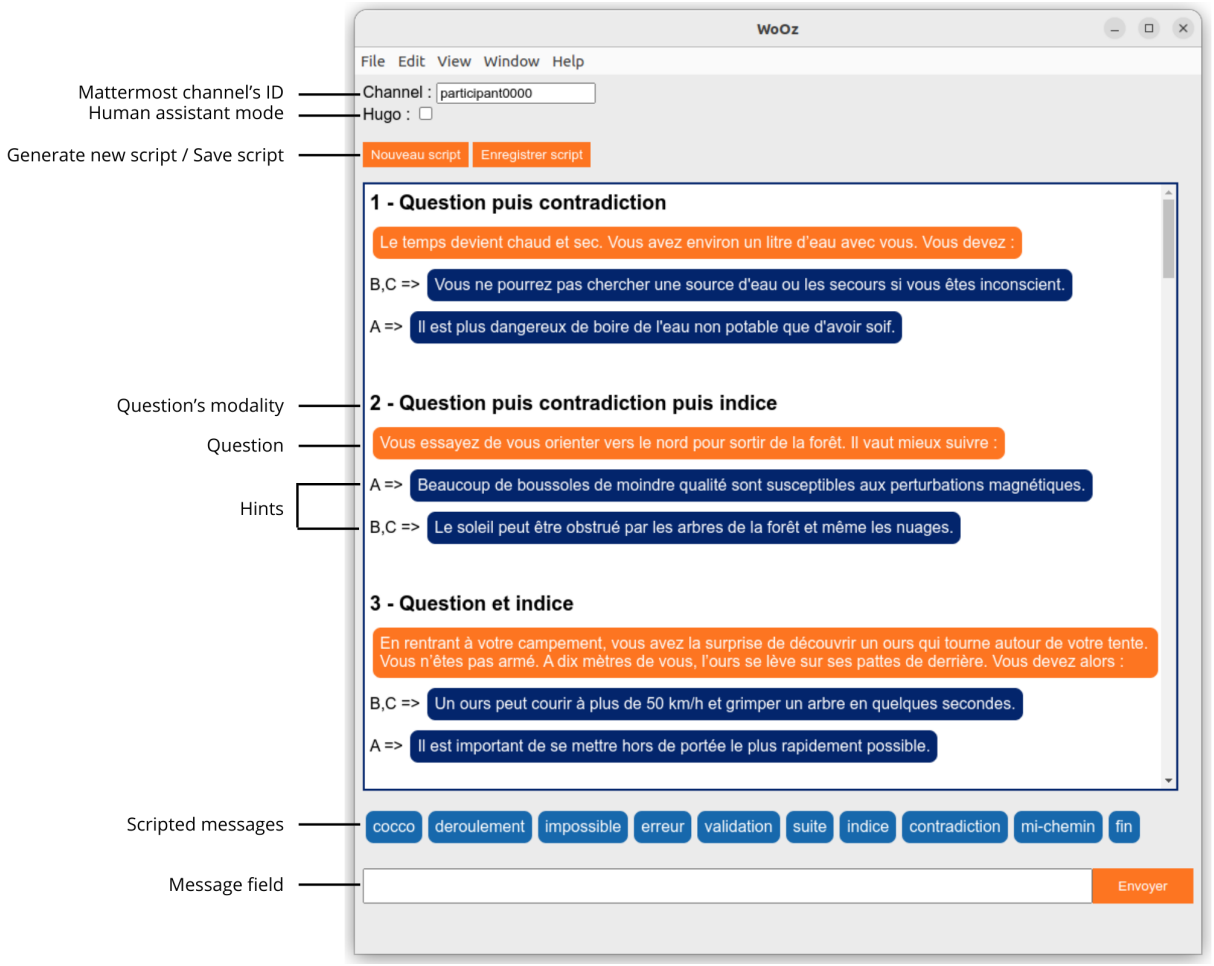


Figure 3.4: Wizard of Oz interface

In the rest of the chapter, we refer to the Wizard of Oz assistant as the chatbot 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 the H-group.

3.3.2 Questionnaires

In addition to the 20 wilderness survival questions, participants completed two online questionnaires before and after the session with the assistant. The first questionnaire collected informed consent from participants as well as demographic information to categorize them into experimental groups while maintaining a balanced gender ratio and comparable age distributions across groups.

The second questionnaire consisted of four ten-level Likert items, which required participants to rate (1) their perceived performance on the questions, (2) the extent to which

the assistant influenced their responses, and (3) the helpfulness and (4) specificity of its interventions. At the end of the questionnaire, a free-text open-ended question invited participants to provide feedback on the experiment (see Appendix B).

3.4 Results

We recruited 48 participants for this experiment through our university's staff mailing list. Participants demographics are summarized in Table 3.1. After registering for the experiment, participants received a random 4-digit ID to log into the testing platforms and keep their data pseudonymized. Their answers to the demographic survey were then used to assign them to one of the two groups in a way that would keep gender and age distribution as balanced as possible. All participants were compensated with a €5 gift voucher for their participation.

Table 3.1: Demographic information of recruited participants (M=Male and F=Female)

	C-group		H-group		Total
Age	M	F	M	F	
18-25	2	3	3	1	9
26-35	4	4	4	5	17
36-45	3	4	3	4	14
46-55	1	1	1	2	5
55+	2	0	1	0	3
Total	12	12	12	12	48

This experimental 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)[8] Avis 2020-25.

We conducted the experiment entirely online and in French, which all participants spoke fluently. In this paper, all the data and labels are translated into English. The data and analysis are accessible on Zenodo [4].

The experiments produced a lot of data, including the messages exchanged between participants and assistants along with the messages' 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 prompts. Finally, this section concludes with an analysis of the answers to the post-experiment questionnaire.

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

3.4.1 Messages length and response time

To better understand how participants interacted with the assistants, we examined the length of their messages and their response times, focusing solely on these quantitative aspects without yet considering message content.

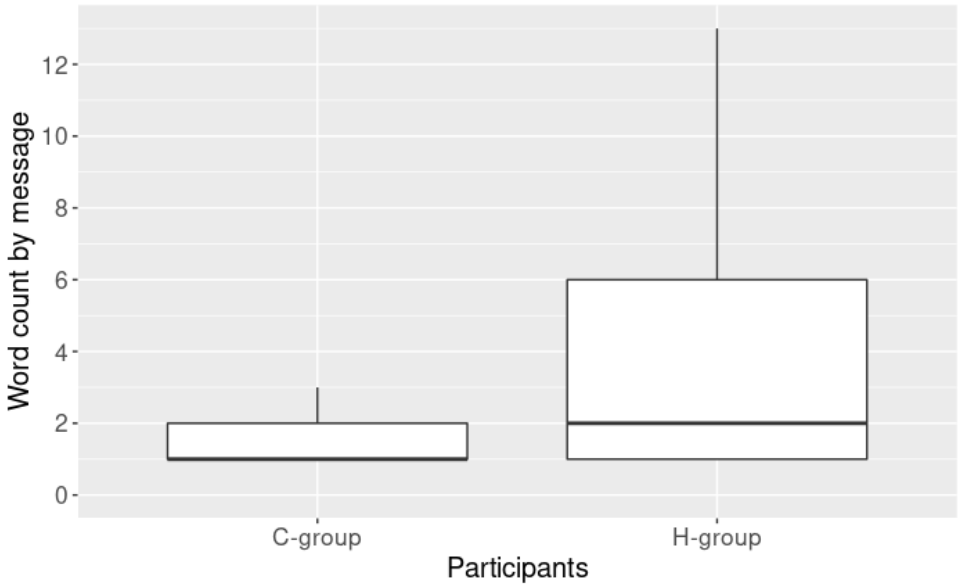


Figure 3.5: Messages' word count by group

H-group participants sent significantly longer messages than C-group participants, as determined by a Wilcoxon rank-sum test ($p = 2 \times 10^{-16}$), with average message lengths of 4.7 and 2.1 words, respectively (Figure 3.5).

H-group participants took significantly longer to respond to the assistants' messages than C-group participants, as determined by a Wilcoxon rank-sum test ($p = 3 \times 10^{-8}$), with average response times of 30 seconds and 23 seconds, respectively (Figure 3.6).

3.4.2 Messages content

We annotated each of the 5809 messages between participants and assistants. Each message was labeled with the role of the sender (COCCO, Hugo, C-participant, or H-participant), the index of the current question (or "x" for non-question related messages), and one or more labels describing the nature of the message. Sixteen labels were used for the assistants' messages and eleven for the participants' messages (Tables 3.2 and 3.3).

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". A message where the participant chose an answer contradicting a hint given by the assistant was annotated with "contradictory" and "answer".

The tables show the number of times the messages are tagged with the labels. For example, Table 3.3 does not necessarily show that 4 participants in the C-group said

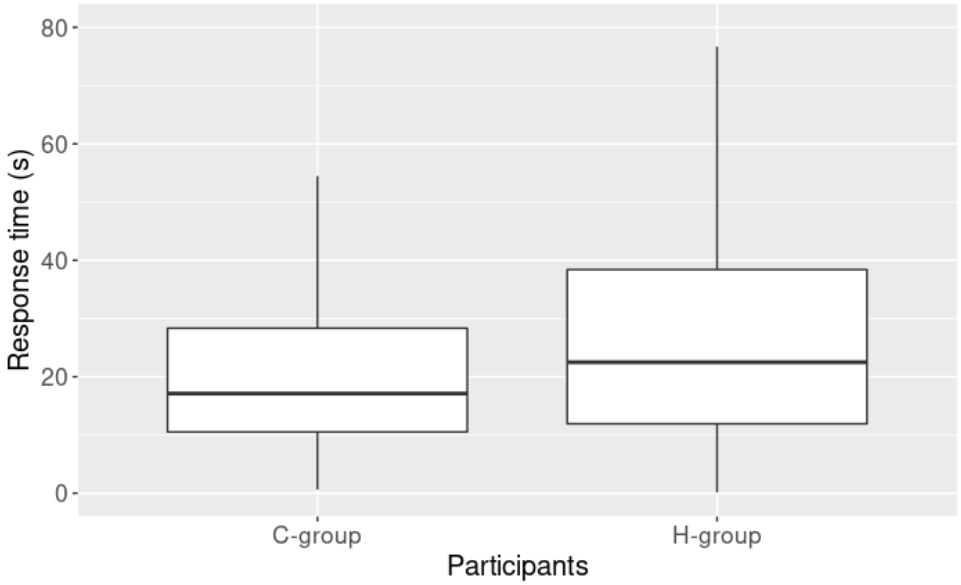


Figure 3.6: Response time by group

goodbye to COCCO: all of those messages could have been sent by the same participant. Further data processing was needed to get the number of times each interaction happened. For the statistical analysis, labels are grouped by question to avoid counting interactions divided into multiple messages more than once.

In Table 3.2, the differences in label counts highlight the distinctions in the scripts followed by the two types of assistant.

- There is no message sent by COCCO labeled "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 3.3.1, 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 labeled "technical information", "error", "validation", "hint", and "clarification" are consequences of the participants' messages and reflect 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 a 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).

Table 3.2: Description and frequency of the labels categorizing the assistants' 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 gives a hint to the participant.*	281	227
Clarification *The assistant provides clarification on the question after a request from the participant.*	2	22

Table 3.3: Description and frequency of the labels categorizing the participants' 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

We also observed several instances in the H-group where participants provided their answers across multiple messages, rephrasing, stating the answer letter followed by its content, hesitating, and then reconfirming. This resulted in a total of n = 487 messages labeled as "answer." Additionally, we observed differences in the frequency of the tags “confirmation” and “new answer,” which suggest a varying rate of answer changes following contradictions. These differences will be examined in more detail in the next section.

3.4.3 Answer changes

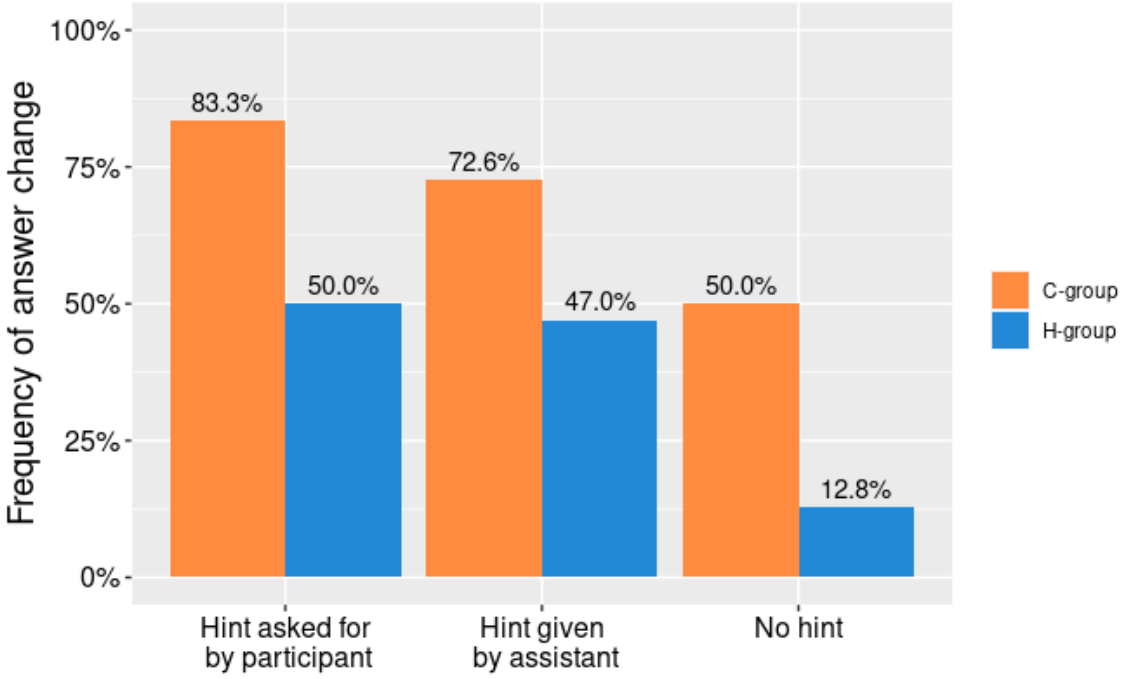


Figure 3.7: Frequency of answer change by group and hint modality

Participants answered 20 questions in total. For 10 of these, their answers were accepted immediately (the “simple” and “hint” modalities). For the other 10, participants were contradicted and then given the option to change their answers, either after receiving a hint (“contradiction and hint” modality) or with no hint provided (“contradiction” modality). For each contradicted question, we recorded whether participants changed their answer and whether a hint was provided or requested. We then computed the answer change rate for each group and hint modality (Figure 3.7).

In general, C-group participants changed their answer 63.7% of the time (130 out of 204), while H-group participants changed their answer only 31% of the time (62 out of 200). A Pearson’s Chi-square test revealed a significant association between the participants’ group and their likelihood of changing their answer, $\chi^2(1, n{=}404) = 42.07$, $p < .001$.

If participants are contradicted but given no hint ("contradiction" modality), C-group participants changed their answer half the time (43 out of 86), and H-group participants

only 12.8% of the time (12 out of 94). A Pearson's Chi-square test revealed a significant association between the participants' group and their likelihood of changing their answer after being contradicted with no hint, $\chi^2(1, n=180) = 27.62$, $p < .001$. Participants prefer following a chatbot assistant's advice to a human assistant's, even when contradicted but provided no additional hint.

In the situation where participants were contradicted and given no hint, but requested one themselves, C-group participants changed their answer 83.3% of the time (10 out of 12), while H-group participants changed their answer only half the time (3 out of 6). For this condition, a Chi-square test was not performed, as the test's assumption about the minimum expected frequency was not met.

If the assistant had given a hint automatically ("contradiction and hint" modality), C-group participants changed their answer 72.6% of the time (77 out of 106) and H-group participants 47% of the time (47 out of 100). A Pearson's Chi-square test revealed a significant association between the participants' group and their likelihood of changing their answer after being contradicted and receiving a hint, $\chi^2(1, n=206) = 13.07$, $p < .001$.

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

3.4.4 Questionnaires

We compared the answers of each group to the four ten-level Likert items of the post-experiment questionnaire (Figure 3.8):

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 3.4), 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 (m=7.21 vs. m=5.375).

We then looked at the correlation between each Likert item and the answer change rate using Spearman's rank correlation (Table 3.5 and Table 3.6). For both groups, the ratings of the assistant's helpfulness and influence and the participants' answer change rate were positively correlated (C-group $\rho = 0.72$ and H-group $\rho = 0.81$). Participants who found the assistant helpful also felt influenced by its suggestions and were more likely to change their answers accordingly.

Unlike participants in the H-group, those in the C-group showed a significant inverse correlation between answer change rate and perceived success ($\rho = -0.52$). This suggests

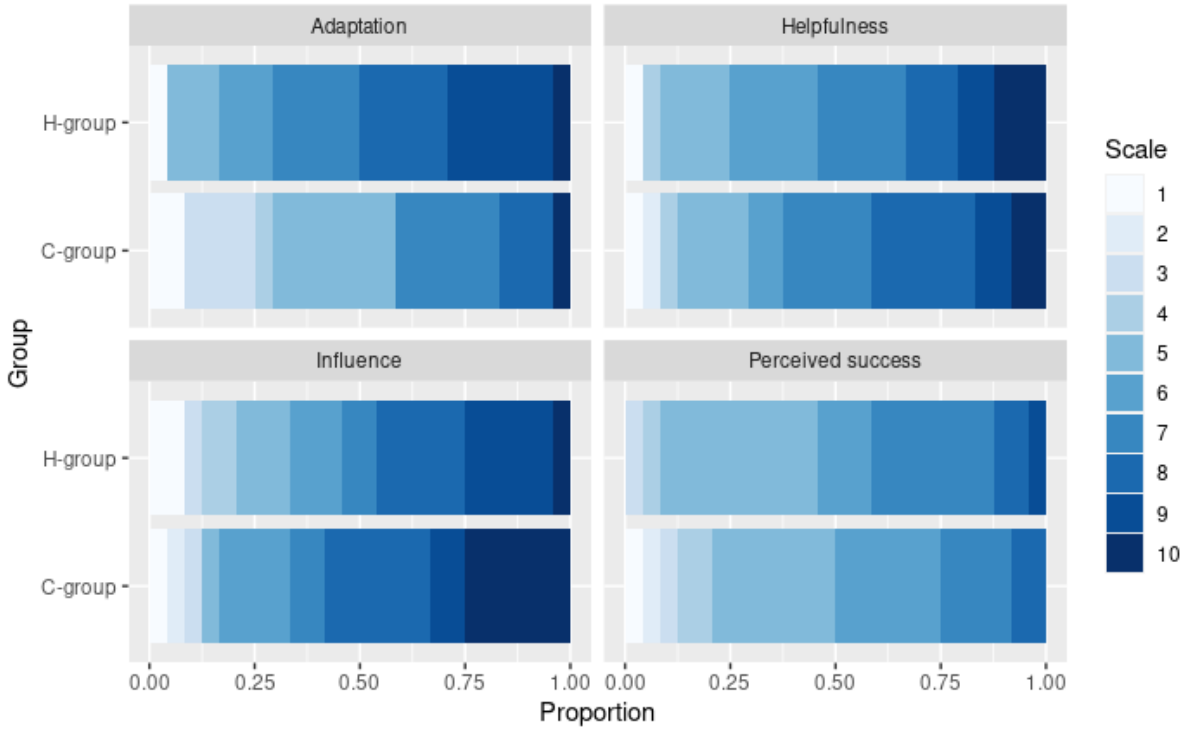


Figure 3.8: Post-experiment questionnaire's answers by group

Table 3.4: 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

that participants who more frequently changed their answers, likely under COCCO's influence, tended to view their performance less favorably. Conversely, those who believed they performed well were less inclined to change their answers in response to COCCO's contradictions.

Unlike participants in the C-group, those in the H-group showed positive correlations between perceived adaptability and three other variables. Perceived adaptability of Hugo was positively correlated with perceived helpfulness ($\rho = 0.50$), perceived influence ($\rho = 0.41$), and the rate of answer change ($\rho = 0.44$). This indicates that participants who experienced Hugo as more adaptive to their responses also tended to evaluate it as more helpful, felt more influenced by its suggestions, and adjusted their answers accordingly.

Table 3.5: Spearman correlation matrix of 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	-0.10	-			
3. Influence of COCCO	0.72*	-0.21	-		
4. Adaptation of COCCO	0.09	-0.39	0.08	-	
5. Rate of answer change	0.51*	-0.52*	0.48*	0.35	-

*p<0.05

Table 3.6: Spearman correlation matrix of 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	0.09	-			
3. Influence of Hugo	0.81*	-0.02	-		
4. Adaptation of Hugo	0.50*	0.29	0.41*	-	
5. Rate of answer change	0.64*	-0.15	0.64*	0.44*	-

*p<0.05

Thematic analysis

Finally, we performed a thematic analysis of the comments left by the participants in the questionnaire's open-ended question (Table 3.7). 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.

Table 3.7: 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

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 with its advice.
- 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 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 were women.

3.5 Discussion

3.5.1 Summary of results

Our results reveal a clear bias favoring chatbots, caused solely by their presentation and perception rather than any actual difference in capabilities. Regardless of whether hints were provided, participants guided by a chatbot assistant were significantly more likely to change their answers compared to those advised by a human assistant. In contrast, participants in the H-group demonstrated greater resistance to changing their initial responses, suggesting more confidence in their original answers or increased skepticism toward Hugo's input. These findings suggest that users may implicitly trust or rely more heavily on chatbot advice, possibly because they perceive chatbots as neutral or authoritative tools, whereas human-like assistants may trigger more critical evaluation or social skepticism.

The type of assistant also affected participants' behavior and perceptions during the experiment. Participants in the C-group associated following COCCO's advice with lower perceived performance. This suggests that they prioritized following COCCO's advice over giving correct answers.

Supporting this, participants asked for advice from the chatbot assistant three times more often than from the human assistant. This may reflect a greater willingness to follow a chatbot's directives, or conversely, a reluctance to inconvenience a human assistant.

In contrast, H-group participants relied less on Hugo's advice but engaged in deeper reflection when answering. This was evident in their longer response times, more frequent expressions of doubt, higher rates of clarification requests, and a greater tendency to justify their answers or refusals to change them. Notably, Hugo was perceived in the survey as more adaptable than COCCO, despite both assistants following the same script. This perception may have contributed to participants viewing Hugo as a better partner overall.

3.5.2 Design implications

These findings have implications for integrating chatbots into collaborative work environments. They are consistent with the literature on automation bias and, in particular, with Van Der Waa et al.'s [114] findings on artificial agent assistance for moral decisions. That study observed that human collaborators often lacked a sense of responsibility for the agent's decisions, even when they were supposed to act as supervisors. Similarly, our findings may indicate that participants in the chatbot condition did not engage in the same trust evaluation or skepticism as those in the human-assistant group. This could be because they did not perceive a sense of risk or vulnerability in failing with their answers when guided by the chatbot, potentially due to a diminished sense of responsibility [41]. In other words, they may not have felt accountable for the outcome, which could have contributed to their increased willingness to follow the chatbot's advice without critical scrutiny.

This increased reliance on conversational assistants could facilitate early adoption

and implementation. However, apparent trust in a new chatbot may not reflect genuine collaboration. Other studies have shown that over-reliance on AI can lead to decreased trust if the AI makes errors or fails to meet expectations [86, 22, 18]. Furthermore, early mistakes can lead to long-lasting decreases in user reliance [46].

Based on our experimental findings, we propose that over-reliance on chatbot advice could be monitored by measuring the frequency of messages reflecting users' decision-making processes and doubts. This could serve as an indicator of how independently users are thinking. A higher frequency of such messages may indicate a more reflective, human-like interaction and a more balanced level of reliance on the chatbot's advice.

To promote appropriate reliance on chatbots and encourage more balanced decision-making, one direction suggested in the literature is the use of cognition-forcing methods. These include negative framing (e.g., presenting the AI's limitations and possible errors before usage) [22], and techniques such as time delays or requiring a preliminary decision before providing advice [10]. However, our methodology already asked participants for their answers before presenting advice, yet still revealed significant bias. This suggests that such methods may have limitations. Further studies should explore this issue in more detail.

3.5.3 Limits

We acknowledge that having the assistant both ask questions and provide hints may have influenced how participants perceived and trusted them. However, separating these roles between two entities could have confused participants.

The difference in answer change rate between the two groups might have partially been explainable by participant response bias, more precisely, participants' reactivity to demand characteristics [40]. Although the experimental conditions were the same for all participants, those assisted by COCCO might have expected the experiment to be focused on the conversational assistant and its functioning and that the expected behavior, the "right" behavior, was to follow COCCO's advice. In contrast, the participants in the H-group would have had more reasons to believe that their behavior was the subject of study and, as such, were more cautious and less inclined to accept and trust the advice given by the human assistant. However, this interpretation conflicts with the observation that C-group participants who followed COCCO's advice reported feeling it negatively impacted their performance, which can be interpreted that they did not think following the chatbot advice was what was asked of them.

Finally, experimenting with higher stakes might have changed how invested the participants were about their performance and the validity of their answers, possibly changing how they engaged with the experiment and the assistants' advice. However, we believe that this is still representative of real-world situations where chatbot usage is typically voluntary and mainly used for routine and low-stakes tasks.

3.6 Conclusion

In this study, our goal was to determine whether automation bias occurs in interactions with chatbots, specifically examining whether decision-making differed when participants received advice from chatbots versus humans. The results revealed a clear bias in favor of chatbots: participants followed the advice of the chatbot assistant significantly more often than that of the human assistant, regardless of whether they had requested the advice. This occurred despite the fact that both assistants delivered the same messages and followed identical scripts. Our experimental design, which relied on natural language interactions, also provided valuable insights into how participants behave during the decision-making process, offering justification, expressing doubt, or seeking clarification, depending on the assistant's perceived identity.

While participants appeared to place high trust in chatbot advice, our qualitative analysis suggests that this reliance may be superficial and not indicative of genuine collaboration. Participants guided by the chatbot were more likely to change their answers, but less likely to reflect deeply or critically evaluate the assistant's input. This raises concerns about long-term effectiveness, as reliance on chatbot advice might not promote thoughtful engagement or a strong sense of responsibility for one's decisions.

These findings have practical implications for the implementation of chatbot collaborators. To foster healthy and effective human-chatbot collaboration, users must not only trust chatbot advice but also maintain a reasonable level of critical thinking and accountability, similar to how they would interact with a knowledgeable human expert. Monitoring user engagement, such as the presence of justification or doubt in responses, can help gauge the depth of collaboration and detect signs of over-reliance.

Having explored how a chatbot collaborator influences individual decision-making in one-on-one interactions, we now aim to examine the impact of its presence within group settings involving multiple human collaborators. In such contexts, the human-chatbot interaction becomes only one component of a broader collaborative dynamic. This shift opens new avenues to explore how chatbots contribute to collective reasoning and team-based decision-making, especially when individuals must integrate both the chatbot's input and the perspectives of their human teammates. Understanding how people weigh chatbot advice alongside input from other humans will be crucial to developing chatbot collaborators that enhance and do not undermine group performance, accountability, and trust.

Chapter 4

A Novel Online Activity for Studying Group Interactions

Contents

4.1 Introduction 55

4.2 Existing experimental collaborative activities 56

4.3 Design of the collaborative activity 58

4.3.1 Design objectives 58

4.3.2 Activity rules 59

4.3.3 Modules 61

4.4 Technical implementation 63

4.4.1 Database 63

4.4.2 Management interface 65

4.5 Validation 66

4.6 Conclusion 66

4.6.1 Potential future applications 67

4.1 Introduction

Following our study of chatbots interacting with a single user which showed the influence of chatbot assistant on a single person, we expanded our research to chatbots interacting with multiple users. Indeed, despite the growing use of chatbots in collaborative settings, research on their impact on group dynamics remains relatively unexplored. As recent surveys and research agendas highlight [25, 98, 7], the vast majority of chatbots are still designed for dyadic interactions, with only a small fraction functioning in multiparty or group contexts.

This gap is particularly important to address as interactions with chatbots are becoming more and more common. Chatbots are already being integrated into group chats, educational settings, and decision-making processes. As such, it is essential to investigate chatbots' impact on task-oriented group activities where their primary goal is to facilitate collaboration and performances.

Considering we did not have access to a dataset of team communications with a chatbot, we decided to approach this research question experimentally. However, an overview of the existing experimental tools used in group collaboration studies, as described in section 4.2 of this chapter, shows a limited selection, none of which met our requirements.

For this reason, we designed a new experimental methodology and online platform designed to simulate real-world collaborative scenarios, which we will present in section 4.3 of this chapter. The activity involves teams of participants working together to solve a series of puzzle modules. These modules are solvable without training through actions such as button clicks or numeric entries and are designed to require effective communication and information exchange between team members.

4.2 Existing experimental collaborative activities

Studying group interactions is an inherently complex task, involving many variables, which is often approached using observational data. In consequence, there are relatively few experimental activities designed to study group dynamics, and even fewer focusing on collaboration and using communication channels that are suitable for chatbot integration. Below, we present existing experimental activities designed to study groups of participants.

The NeoCITIES simulation [71] is an online activity simulating emergency dispatching in a virtual city space. It was developed to research decision-making, information overload and teamwork. Schelbe et al. [94] implemented the simulation to investigate team cognition within human-agent teams of three. Humans collaborating with agents perceived and had a significantly higher team performance, but a lower perceived team cognition than with human teammates. Additionally, humans with only agent teammates had significantly lower trust in them than when they had both a human and an agent teammate.

The NeoCITIES simulation engine was then used in subsequent activities like idsNETS and TeamNETS. IdsNETS [67] aims to simulate cybersecurity intrusions to study analysts' situation awareness. In this activity, players need to monitor IDS alerts (intrusion detection system), categorize and report. TeamNETS [66] is a cybersecurity simulation, where teams of three are each responsible for either policy management, intrusion detection or malicious software detection. They must identify threats and either accurately categorize it or transfer it to a better suited teammate.

Another example of an experimental activity designed to study group dynamics is the StrikeCom game [112]. It is defined as a "C3ISR (Command, Control, Communication, Intelligence, Surveillance, Reconnaissance) task designed to examine the development of shared awareness and communication in distributed groups". The game consisted of a map

divided in a grid of cells and is separated into multiple phases. Participants place assets on the cells to search for a target. After multiple search rounds, the group decides on which cells to strike. The game makes participants use different communication mediums, such as chat, audio, video or in-person.

In a different genre, Cargo is a large-scale pervasive game designed to investigate group dynamics and human-agent interaction in real-world settings [73]. In this game, teams of 6–8 players competed under time pressure to collect credits by visiting card-swiping stations scattered across a city area. A semi-autonomous, rule-based virtual agent named the Instructor guided the teams by placing real-time, text-to-speech calls to individual players delivering recommendations for which station to visit next. These recommendations often conflicted within a team, forcing players to negotiate whose instructions to follow, as not all could be acted upon simultaneously. Central to the game were issues of trust and compliance: players had to assess the Instructor's reliability, especially as it reacted to players' non-compliance. The study highlighted how perceived trustworthiness of the agent and interpretation of its motivations affected player behavior and team dynamics.

The mixed-reality game AtomicOrchid[24] was designed to study human-agent teaming using a disaster response scenario. In this game, players assume the role of field responders rescuing targets after a radioactive explosion. They navigate in real space while being guided by players in the headquarters who are the only ones able to see where the radioactive cloud is. This game was used in Ramchurn et al. study [88] which used supervised virtual agents to dispatch the players. They suggest that agents being able to adapt to users' requests is a crucial factor for their acceptability.

Another relevant effort is the experimental platform CrowdMapper developed by Mao et al. [68], in which groups of varying sizes collaborated on a crisis-mapping task based on real-world disaster data. Participants worked together online to classify and geolocate reports of damage from social media streams, a complex activity requiring coordination, verification, and integration of multiple sub-tasks. The study demonstrated how larger teams, despite reductions in individual effort, benefited from increased collaboration and produced better overall results compared to smaller groups or independent workers.

Those collaborative activities are too militaristic, competitive, or framed around high-stakes scenarios for our purpose. Instead, we want to study group interactions in more casual settings that reflect the everyday collaborative experiences of most people.

More in line with our purpose, there are a number of studies that use video games to study collaboration, such as the cooperative game "Overcook!" [23, 70]. Video games allow complete tracking of participants behaviors while being engaging and immersive for the participants. Multiplayer games also provide controlled environments that require players to communicate and coordinate to achieve shared goals, as well as encouraging delegation, planning, problem-solving, and adaptability. However, using commercial video games also have drawbacks for research, mainly a partial to complete lack of control on the rules of the activity, no possibility to adapt the activity and no access to complete logs (creating a need to annotate videos afterward) [1, 23].

For the need of our research, commercial video games are unsuitable in their current form. They offer no straightforward means of integrating a chatbot, neither in terms of gameplay, which typically does not accommodate text-based interactions without disrup-

tion, nor from a technical standpoint, as their codebases are generally inaccessible and not open to modification.

To address these limitations and create a research setting better suited to our objectives, we developed our own experimental activity. In the next section, we describe these objectives and how they informed our activity design.

4.3 Design of the collaborative activity

4.3.1 Design objectives

Considering the lack of suitable tool for research on chatbot interactions with groups, we decided to design and develop our own experimental task. We identified the following criteria that fit our needs:

- Collaborative: the activity should require multiple participants and promote collaboration between them, making it necessary to work with teammates as much as possible to achieve their goal and deter players from brute forcing the objective by themselves.
- Concrete task: the activity should present the participants with a task that is achieved by communicating. Communication should not be the goal itself (e.g., learning, debating, brainstorming), it should only be the mean by which the goal is achieved. This task should also permit us to define objective metrics and measure participants' performance.
- Text communication: all communication between teammates should be done using text, in order for a chatbot to have the same situational comprehension as the participants and be able to participate fully. (This also allows for easier analysis with complete data on all the participants' interactions.)
- No prior expertise needed: the task should not require any domain knowledge or training to be done, to simplify participant recruitment.
- Simple concepts: the concepts manipulated in the collaboration should be simple, distinct and easy to express to prevent confusion, especially for chatbots to reduce their detection error rate.
- Short: the activity should be completable (but not necessarily so) in one session without needing to split participation in multiple sessions. Indeed, participation attrition with multiple sessions are substantial [69] and would be unmanageable when working with groups that need to keep the same make-up.
- Browser platform: the activity should be accessible by browser with no installation to allow participants to do the activity remotely from their home and simplify recruiting (additionally, this was in context of the COVID pandemic where access to our laboratory was fairly restricted for visitors).

- Single page: while real-life collaboration often involves using multiple tools, this requires task switching between different windows or screens. However, this makes it likelier to miss information and to completely forget the tool. As it is less tractable and not the main focus of our research, we wanted an activity that could be done from the same screen with no switching.

Following the definition of this problem space we started thinking of rules that would satisfy those constraints.

4.3.2 Activity rules



Figure 4.1: Example bomb from the game Keep Talking and Nobody Explodes

Inspiration For the activity design, we found inspiration in the 2015's video game "Keep Talking and Nobody Explodes"[9], which fits many of our design objectives. It is a cooperative puzzle video game where two players have to defuse a bomb before a timer runs out (Figure 4.1). One of the player has access to a defusing manual and must guide their teammate using their description of the bomb's different modules. The game creates complexity out of simple symbols and collaboration is achieved with only verbal exchanges.

As some of the rules do not fit our criteria, we used the main principle of the game as a basis and adjusted its rules for our experimental activity. The two teammates roles are asymmetric, with one doing the defusing and the other accessing the necessary information. We made it so that all participants had access to information and the modules. This

[9]https://keeptalkinggame.com/ (accessed September 20, 2025)

resulted in the game being technically adaptable to any team size and not just duos. We also changed the activity so that its solving time would be measured but without ending at a set time. This considerably lessen the time pressure on participants, but instead allow us to collect data on collaborations from start to finish. Another big difference, was the usage of voice chat for communication, which we changed to text communication. To mirror modern work environments, participants can have access to a group chat for team-wide discussions and individual messaging channels for private conversations.

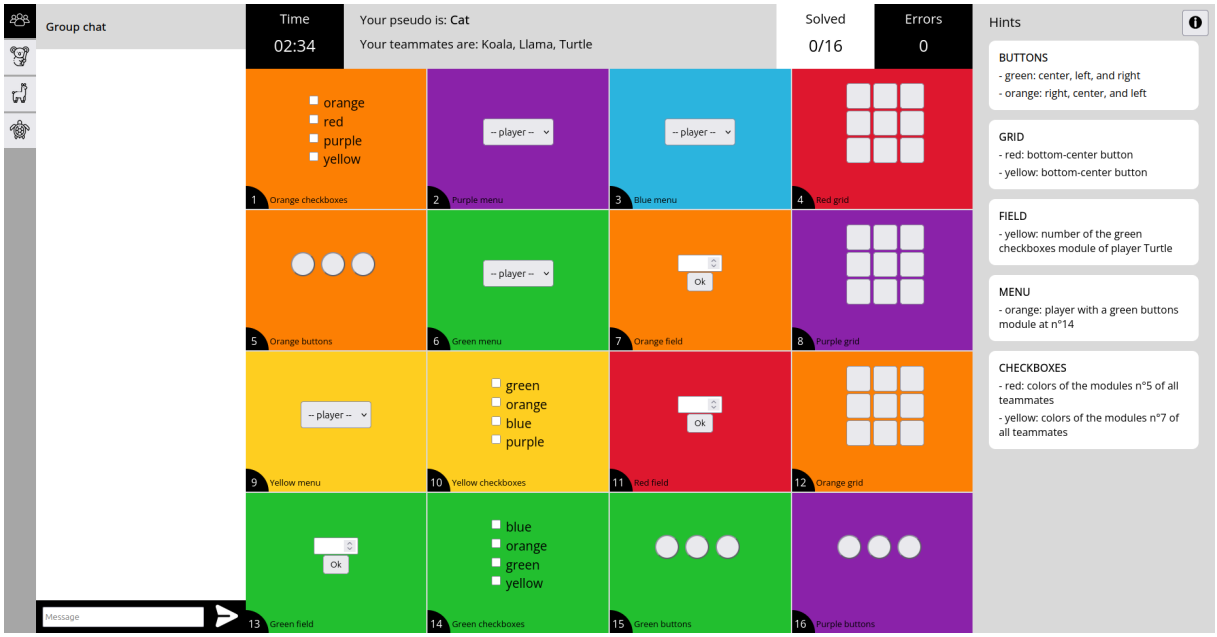
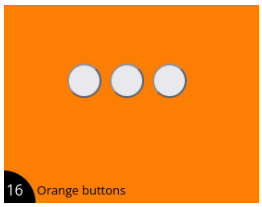


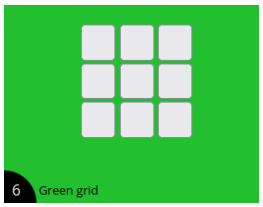
Figure 4.2: Platform interface. On the left panel: the chat interface, with different tabs for all the chats available to the participant (group chat, private chat with COCCO, and three private chats with teammates). On the middle panel: the time elapsed since the start of the experiment, the participants' pseudos, the number of modules solved, the number of errors made by the participant and the grid of 16 modules. On the right panel: the 8 hints given randomly to the participant, grouped by module type.

Following those alterations in concept from our inspiration, we refined the rules of the activity to fit all our design objectives. Multiple participants are tasked with solving their assigned 16 modules, using hints and the help of their teammates. They are instructed to do so in the shortest time possible and without making errors. To communicate, they have access to a group channel, where everyone in the team can chat and private channels where they can send direct messages to each of their teammates. The experiment ends when all participants have solved all of their modules. The platform logged every messages sent in the general chat and private chats, and every clicks on the modules.

The interface of our activity is composed of 16 distinct modules framed by a chat box to the left and textual hints to the right, as depicted in Figure 4.2.



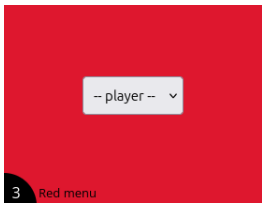
(a) Buttons module



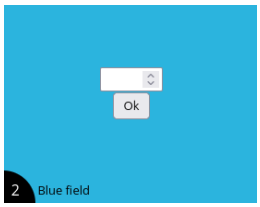
(b) Grid module



(c) Checkboxes module



(d) Menu module



(e) Field module

Figure 4.3: Examples of each module type

4.3.3 Modules

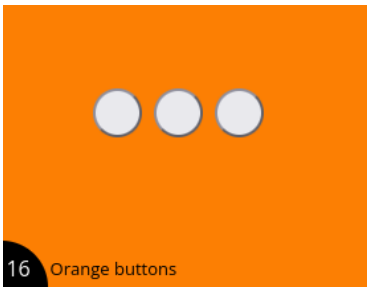
Modules are characterized by their color (red, orange, yellow, green, blue, or purple), type (checkboxes, menu, field, buttons, or grid), and associated number (ranging from 1 to 16). All modules of the same type and color have the same solution. Participants are given hints that directly or indirectly allow them to solve the modules of a specific type and color. Some hints may be useful for their own modules, but others will be useful for their teammates. Conversely, some hints necessary for a participant will be assigned to their teammates. For each team, the modules and their solutions are randomly generated using the undermentioned rules to prevent ambiguities between modules:

- No participant has two modules with the same type and color.
- No participants in the same team have a module with the same type, color and number.

The modules were designed using standard HTML elements (checkboxes, radio buttons, dropdown list, etc.) to reduce compatibility issues and make them easy for participants to interact with. The modules can be solved using only information available on the platform: the modules' types, colors and numbers and the participants' pseudos. This ensures that any participant, regardless of their background knowledge or experience, can take part in the activity. These concepts are also distinct and simple enough for a chatbot to match to predefined parameters and intents with minimal errors.

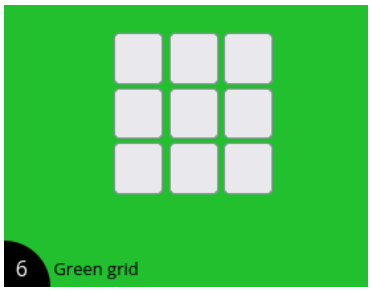
Using these simple concepts, we tried to create as much complexity in the collaboration and need for different communication strategies as possible. This was achieved by making some of the solutions dependent on additional information from one or more teammates.

Below is a breakdown of the five module types and how they are solved:



Buttons module : solved by clicking three buttons in the correct order. The hint is the solution. For example: "Purple buttons: right, center and left".

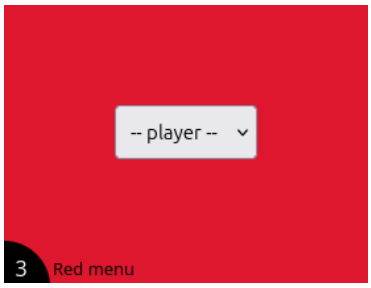
Grid module: solved by clicking the correct button in a 3x3 grid. The hint is the solution. For example: "Green grid: button on the bottom right".

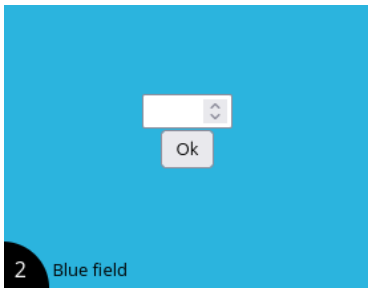




Checkboxes module : solved by checking the colors of the modules of a given number of all participants. The hint requires asking all teammates the color of one of their modules. For example: "Yellow checkboxes: colors of the modules number 9 of all players". The corresponding solution would be checking "vert" (green), "orange" and "bleu" (blue).

Menu module : solved by choosing in a dropdown menu the pseudo of the teammate with a module of a given type, color and number. The hint requires asking all teammates if they have a specific module (or deduct the answer by elimination). For example: "Red menu: player with a red grid module at number 1". The corresponding solution would be choosing "Llama" in the dropdown menu.





Field module: solved by entering the number of a specific teammate's module with a given type and color. The hint requires asking a teammate for the number of one of his modules. For example: "Blue field: player Owl's orange checkboxes module's number". The corresponding solution would be entering "14" in the field.

The different module types enable us to propose different difficulty levels regarding information sharing and encourage effective group communication. Buttons and grid modules can be solved by just knowing their corresponding hints. Checkboxes modules can be solved by polling every teammates and remembering multiple colors. Menu modules can be solved by asking some or all teammates until the solution can be deducted. Field modules can be solved by contacting a specific teammate. This results in a cooperative game where participants collaborate to solve all the puzzles from each board to achieve their common goal.

However, it is still possible to solve the modules through brute force by testing every possible combination. To discourage this approach, our implementation penalizes players by counting their mistakes, which negatively impacts their final score, as an incentive to collaborate with teammates. Alternatively, it would be possible to temporarily disables inputs after incorrect attempts, creating a brief timeout, deterring brute forcing and encouraging players to think more carefully before acting and coordinate with teammates.

4.4 Technical implementation

Our experimental web platform was developed with Javascript front-end and Python back-end using Flask, connected to a SQLite database, and hosted on a VPS.

4.4.1 Database

The activity's data is stored SQLite database with 9 tables (Figure 4.4). The data can generally be categorized into two types:

- The fixed data created before the experiment: the experiment's parameters and generated solutions.
- The data created during the experiment by the participants: the solving attempts and message logs.

The key entities of the database include the *experiment* table, which stores information about the experiment's parameters, such as its start and end times, the number of modules of each type used, and available communication channels. Each *participant* is associated with an experiment and is tracked for their number of attempts, errors, and

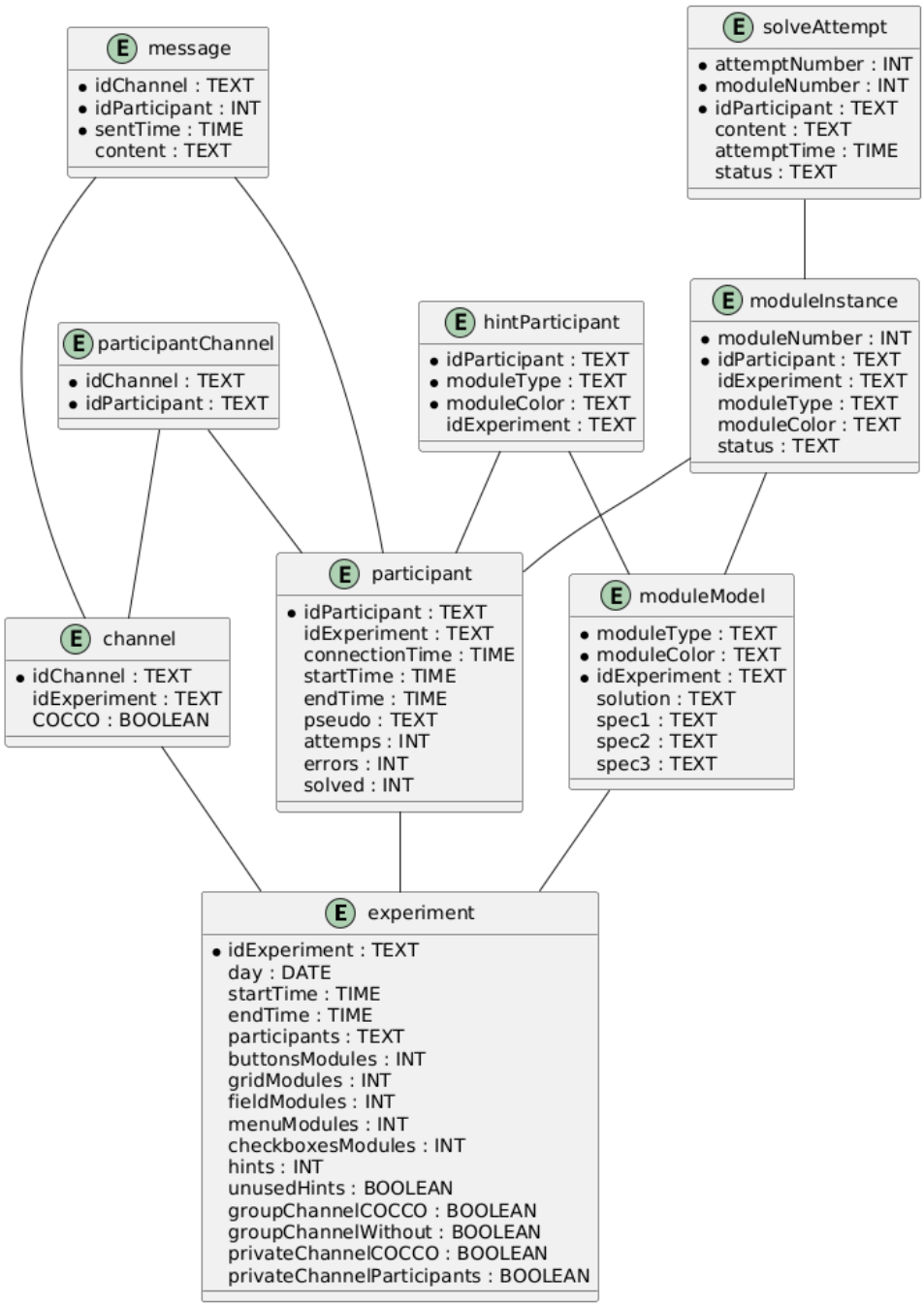


Figure 4.4: Database diagram (made using PlantUML)

solved modules. The *moduleModel* entity defines the properties and solutions for each possible module (type and color) used in the experiment, while *moduleInstance* represents the individual instances of these modules assigned to participants, including their status during the activity.

Additionally, the database includes the entities *solveAttempt* to log each participant's attempt to solve a module, *hintParticipant* to track hints given to participants for spe-

cific modules. The *channel* and *participantChannel* entities represent the communication channels used and which participants have access to them. The *message* entity logs messages sent in those channels. This structure supports both the operational aspects of the experiment (displaying the modules, checking participants attempts against the modules solutions) and the post-experiment analysis of the collaboration (logs of the messages and attempts).

4.4.2 Management interface

In addition to the experimental interface for participants, we also implemented a password protected interface to create, manage and monitor the experiments (Figure 4.5). The interface is composed of a left panel list all the experiments past and future, a central panel showing the selected experiment parameters and a right panel showing the selected experiment data (participants' data, message logs and module and hints distribution).

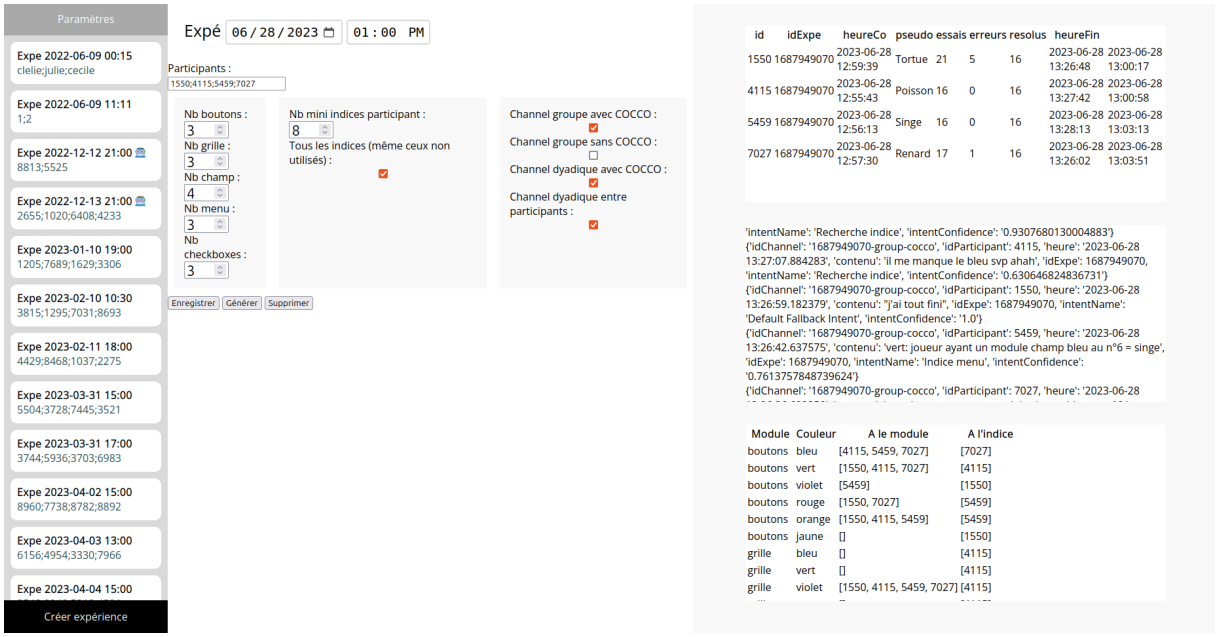


Figure 4.5: Experience management interface

Experiments are handled using the following process:

1. The experiment is created by choosing a date and an hour.

2. The experiment can then be updated with important parameters: participants ids, distribution of each module type, distribution of hints and types of communication channels available (group or private chat, with or without a chatbot).

3. Once all parameters are filled, the experiment can be generated. This creates and assigns modules and hints to participants following the parameters and the previously explained generation rules.

4. Until the scheduled start of the experiment, all parameters are editable and modules and information can be generated again.

5. During the experiment, the interface can be used to monitor the players statuses (connection time, number of module solved, number of errors, ...) and the message logs (sender, channel, content, ...).

The interface also permits the control of the following general parameters: how long before the scheduled start of the experiment can participants log onto the platform, how long before they can start the activity, and whether all participants have to confirm their presence before the activity can start (this option was mostly disabled to simplify technical tests).

4.5 Validation

Once the platform was deployed and accessible online, we organized multiple test runs of the activity. This allowed us to correct technical issues and oversights, as well as, to verify the stability of the platform. We also assessed the clarity and ease of understanding of the interface and instructions and the difficulty of the tasks.

Following the tests, we made the following noteworthy changes to the platform:

- We added a text label to all modules describing their type and color, to help color blind participants and more importantly prevent confusion between different types of modules' names.

- We added an obligatory connection check for all participants, preventing them from starting the activity without a complete team.

The test runs also provided additional material for us to refine the activity tutorial included in the platform, allowing us to clarify and improve the explanation for elements that were often misunderstood. We also added detailed information on how to troubleshoot various display problems possibly encountered by participants.

Those test runs confirmed that the modules were easy to understand and to complete. Some participants also mentioned that they found the activity entertaining and would like to play it again. This is a promising sign for completion rate of the activity and participants engagement, potentially enhancing the ecological validity of the experiment [1, 31].

4.6 Conclusion

After the tests, we decided to concentrate on teams of four solving 16 modules and to keep the rules as they were without implementing additional rule-sets, as we judged that the activity complexity was high enough for our intended study.

Participants took between 30 minutes to an hour to complete the experiment, with notable variability in their performance. We were concerned that adding further complexity could risk overwhelming some participants, especially those who already needed guidance from their teammates, or even completely halting a team's progress. For these reasons, we decided that more advanced versions of the activity should be explored in future iterations, once we had a better understanding of participant performance and could potentially recruit returning participants who are already familiar with the activity.

4.6.1 Potential future applications

During the design phase, we considered multiple ways in which the activity could be easily altered to accommodate different goals. First, parameters like team size or number of module could be adjusted. With larger teams (above a dozen people), this would require introducing new module colors and types to prevent ambiguities.

In this version, all modules can be solved at any time. We considered a version where some modules would only unlock or appear after some progression (for example, another specific module solved or the team general progression). Additionally, all players must solve their own modules, but they could instead be tasked with collectively solving the same modules (letting some players contribute more than others).

Some manipulation can also be done with no modification to the activity by manipulating the recruiting process (participants that know each others identities, expert players), the communicated goals (individual vs. team time, importance of errors) or the availability of other type of communication channels.

Another significant design possibility we did not implement was the deliberate introduction of inaccurate information, which would require participants to exercise greater vigilance. This could involve creating ambiguities within modules and hints or even providing outright false information. In such scenarios, participants would need to carefully evaluate new information, track its sources, and verify its reliability. They would also be encouraged to share failures and mistakes, as these could serve as important information for the team. A further variation could make certain information valid only for a limited time, adding an extra layer of strategic decision-making and attention management.

A larger-scale version of the project could make the activity available online at all times, with automated team making to facilitate continuous participation and data gathering. However, this would require significant scaling of the platform's technical infrastructure, as it was originally developed for small-scale, supervised experiments. In particular, the database would need to be migrated from SQLite to a more robust solution capable of handling concurrent modifications from multiple users. Additionally, important security considerations would need to be addressed. The current setup relies primarily on security through obscurity and restricted access during monitored sessions, which would likely be insufficient to withstand targeted attacks in a publicly accessible deployment.

To conclude, we have designed an activity to study group interactions and collaboration in a controlled experimental setting. For our research objectives, we chose to keep the activity relatively simple. However, the methodology we propose is highly flexible and can be easily adapted to a wide range of research protocols. This platform can serve

as a practical tool for exploring team interactions and collaboration dynamics, whether in human-only groups or in settings that include chatbot participants. With this perspective in mind, we have made the platform codebase available in a public repository: `https://gitlab.univ-lorraine.fr/amiot1/cocco-collaborative-activity` .

Chapter 5

Impact of Chatbot Assistance on Collaborative Tasks

Contents

5.1 Introduction 69
5.2 Methodology 70
5.2.1 COCCO the chatbot 71
5.2.2 Pre-activity and post-activity surveys 75
5.2.3 Participants 76
5.3 Analysis and Results 77
5.3.1 Team performance 77
5.3.2 Cognitive workload and perception 79
5.3.3 Team dynamics 81
5.4 Discussion 83
5.4.1 Limitations and future directions 84
5.5 Conclusion 85

5.1 Introduction

In today's predominantly digital work environments, effective team collaboration often relies on the ability to communicate and share information efficiently between teammates on online platforms. To better understand how group interactions are structured in those settings, we developed an innovative experimental activity designed to simulate online collaboration. In this activity, teams are tasked with solving a collection of puzzle modules on an online platform. While the modules can be completed with simple interactions, such as clicking buttons or entering numbers, successful completion of the activity relies on team coordination and efficient communication and information sharing between teammates.

Using this experimental activity, we investigated how groups engage with chatbots and how those interactions impact group dynamics, a topic that hasn't been deeply explored yet in the literature [98]. To address this gap, we designed an experimental protocol that uses our experimental activity and integrates a chatbot created to assist teams and capable of different communication strategies. It aims to answer two key research questions about the interactions between chatbots and teams:

- **How do groups engage with a chatbot?** How is it used? How is it perceived?
- **How do chatbots impact team dynamics?** Do they consistently facilitate collaboration? Can they be disruptive? What is their impact on team performance? What is their impact on cognitive load?

In this chapter, we present the design of our custom-built chatbot, COCCO, developed to be integrated into our experimental platform to provide assistance to participants. COCCO was specifically designed to understand and retain task-relevant information shared by participants through both group and private messages, allowing it to offer support and retrieve relevant details when needed.

To mimic a modern workplace environment, participants have access to both a group chat and private messaging channels. This setup allows us to observe how communication structures influence team dynamics and performance in a controlled, yet realistic, setting. Depending on the condition, COCCO was either absent (control), available only in the group chat, only in private chats, or active in both.

This chapter outlines the design and implementation of COCCO across the four conditions, and the experimental methodology used for evaluating its impact. We conclude with an analysis of the data collected and our key findings regarding chatbot-mediated team collaboration.

5.2 Methodology

For this experiment, we used the previously explained group activity we designed, played in teams of four. Team size was determined after preliminary test runs of the experiment. Teams of 3 made communication too streamlined with less cross-talk, and participants were able to deduce modules' solutions with less help from their teammates. On the other hand, teams of 5 lengthened the experiment too much, with a run lasting over one hour.

We also made the following parameter choices. Each participant has eight hints assigned randomly, where a hint for each module type and color should be assigned to at least one participant. There are 6x5=30 combinations of module types and colors and 8x4=32 hints assigned in a team of four, resulting in only two duplicated hints. Each participant has 4 field modules and 3 modules of the other types (menu, checkboxes, buttons, and grid).

5.2.1 COCCO the chatbot

Design

Our objective in designing COCCO was to create an "ideal" virtual assistant that could actively support team collaboration and facilitate interactions between teammates. We aimed to leverage the chatbot's inherent strengths, such as its memory, constant availability, and near-instant response speed.

To achieve this, COCCO was developed to parse participants' messages for task-relevant information in group and private chats. Once identified, this information was stored in its memory to be later recalled and communicated when requested. This design was intended to help COCCO support teams in maintaining their shared understanding of the activity and avoiding repeating discussions to request and provide information.

Importantly, we aimed to build COCCO in a way that reflects real-world constraints. To that end, COCCO was limited to using only the information it encountered during the course of the experiment in the participants' exchanges. It could not access backend data to provide participants with answers. This design choice ensures that its behavior more closely mirrors that of future real-world chatbot assistants operating in live, information-rich and unpredictable environments. We however included an internal validation mechanism for the information COCCO stored, to ensure that it was accurate and avoid mistakes being propagated, which could have complicated both the task and the analysis of results.

Communication strategies

To study how our chatbot interact with a group, we designed three different communication strategies: group-chat, private-chat and both chats. Those strategies are modeled on common workspace communication tools (Slack, Microsoft Teams) where some chatbots can be interacted with from any channel, some are only used to broadcast information in a public channel or some have to be contacted by direct message [100, 117].

When in the group chat, COCCO parses all the messages and only intervenes when it recognizes an information request for which it has the correct information.

When in a private chat with a single participant, COCCO intervened similarly to when it was in the group chat. Moreover, when missing a requested information, it would signal it lacked the information and could request it from other participants at the participant's behest. This was not included in the group chat condition to prevent unnecessary clutter in the conversation and context confusion for COCCO, especially as this feature would be superfluous in a group chat where other participants could already see the information request. Additionally, in the private chat, COCCO had a small contextual memory to be able to keep track of which question is being answered or which module a new message is about. Private chats involved fewer interlocutors and less visibility, but COCCO was given more functionality to help sustain the conversation.

In one of the experimental conditions, COCCO was present in both the group chat and the private chats. This setup enabled it to use the information given to it in private chat

for questions in the group chat and vice versa, making it better informed and potentially reducing repeated conversations in different channels.

We provide examples of those different communication scenarios in figures 5.1, 5.2 and 5.3.



Figure 5.1: Excerpt from an interaction between participants with no assistance from COCCO (translated from french, timestamps in minutes and seconds)

Figure 5.1 shows an interaction between participants with no intervention from COCCO, from player Fish's request for the solution of purple grid modules to the answer two minutes later. We can see that there is often cross-talk between participants and that answers are not always clear or easy to understand without context. In this case, by the time player Cat indicates that the solution to the purple grid module is the same as the yellow grid which is "upper center", the solution has already been given by player Fox.

Figure 5.2 shows two interactions from different teams where COCCO provided assistance in the group chat. In scenario 5.2a, COCCO knows the solution to the modules thanks to previous interactions. This allows it to immediately respond to one of the two requests of player Owl (i.e., red field module). The other request (orange field module) is answered later by player Llama. We can additionally see, that different participants, Owl and Turtle, ask for the same module solution in short succession.

In scenario 5.2b, player Sheep requests information for two modules (i.e., yellow and green field). COCCO only knows the hint for the yellow field module, but not its solution. However, player Monkey in suite of COCCO's message, completes the hint by giving the necessary information (i.e., the number of his blue menu module). This sequence also happens for the green field module's hint with players Owl and Bee and for the red field module's hint with COCCO and player Monkey.



(a) Simple scenario (b) Complex scenario

Figure 5.2: Excerpt from two groups' interaction with assistance from COCCO in the group chat (translated from french, timestamps in minutes and seconds)

Finally, scenario 5.3 shows an interaction where COCCO provides assistance in a private channel, but do not have the requested information (otherwise the interaction would be similar to those in group chat). Instead, after confirming with player Koala, COCCO asks the other participants for the information (i.e., whether they have a purple field module at the number 13 or not). Player Turtle states that they do not have this module and player Monkey does not respond. When player Fox confirms that they have the module, COCCO informs player Koala.

Implementation

To extract information from the messages (hints, solutions, module characteristics), we used Google Dialogflow[10]. The intents are defined and trained on the pre-tests' logs. We looked at how participants exchanged information and talked with each others when doing the activity without COCCO and created intents that would allow it to learn new information from the conversation and recognize requests for information. We then conducted additional pre-tests where COCCO was implemented in group chat and private chat, to fix problems and tune its training data. The list and descriptions of the 19 implemented intents (4 conversational, 11 for new information and 4 for requests) can be found in appendix C. During the activity, the recognized intent is used to dictate COCCO's behavior. If the message includes new information, COCCO updates its database. If the message is

[10]https://cloud.google.com/dialogflow

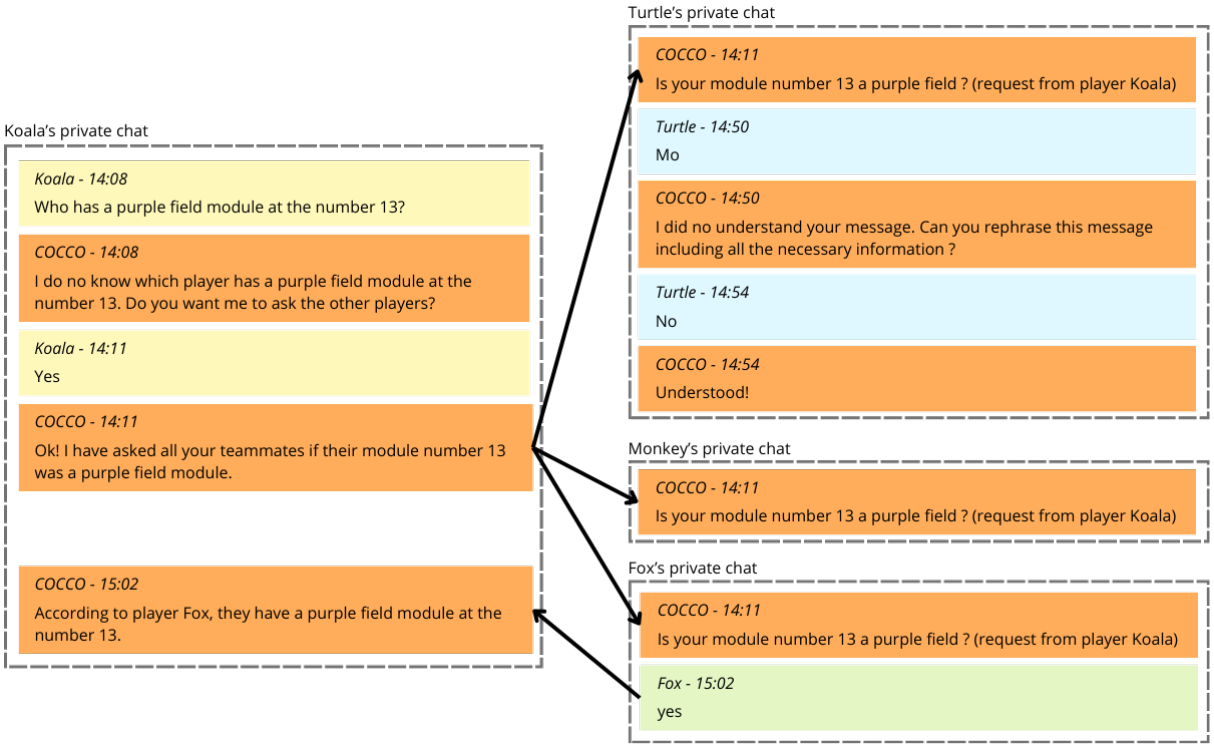


Figure 5.3: Excerpt from a group interaction with assistance from COCCO in private chat (translated from french, timestamps in minutes and seconds)

an information request, COCCO checks its database and shares the requested information if it has it. All COCCO's messages to the participants use pre-written templates, which allows for clear communication and avoid possible generated ambiguous messages, even if it trades-off against sounding natural.

Following pre-tests done with COCCO, we added a check before saving new information in COCCO's database to prevent it from repeating wrong information. This is a workaround for the need of the experiment, as we consider miscommunication repair beyond the scope of this research. However, even with the check COCCO's intent recognition is not perfect. It can still intervene at the wrong moment when it misunderstands a participant's request. This made it similar to the participants' own messages, who rarely made outright mistakes (1,78% of answers) but sometimes would make ambiguous requests or give information relating to no clear subject.

To support the representation of the chatbot COCCO's internal reasoning process during the experiment, two new entities have been introduced to the platform's database: *moduleModeleCOCCO* and *moduleInstanceCOCCO* (in orange in Figure 5.4). These tables mirror the structure of the existing *moduleModele* and *moduleInstance* entities, respectively. However, rather than storing predefined participant-assigned module data, they serve as a mental model of the activity for the chatbot. As COCCO interacts with participants and sees their messages throughout the experiment, it fills these tables with its deduced understanding of the modules configurations and their properties.

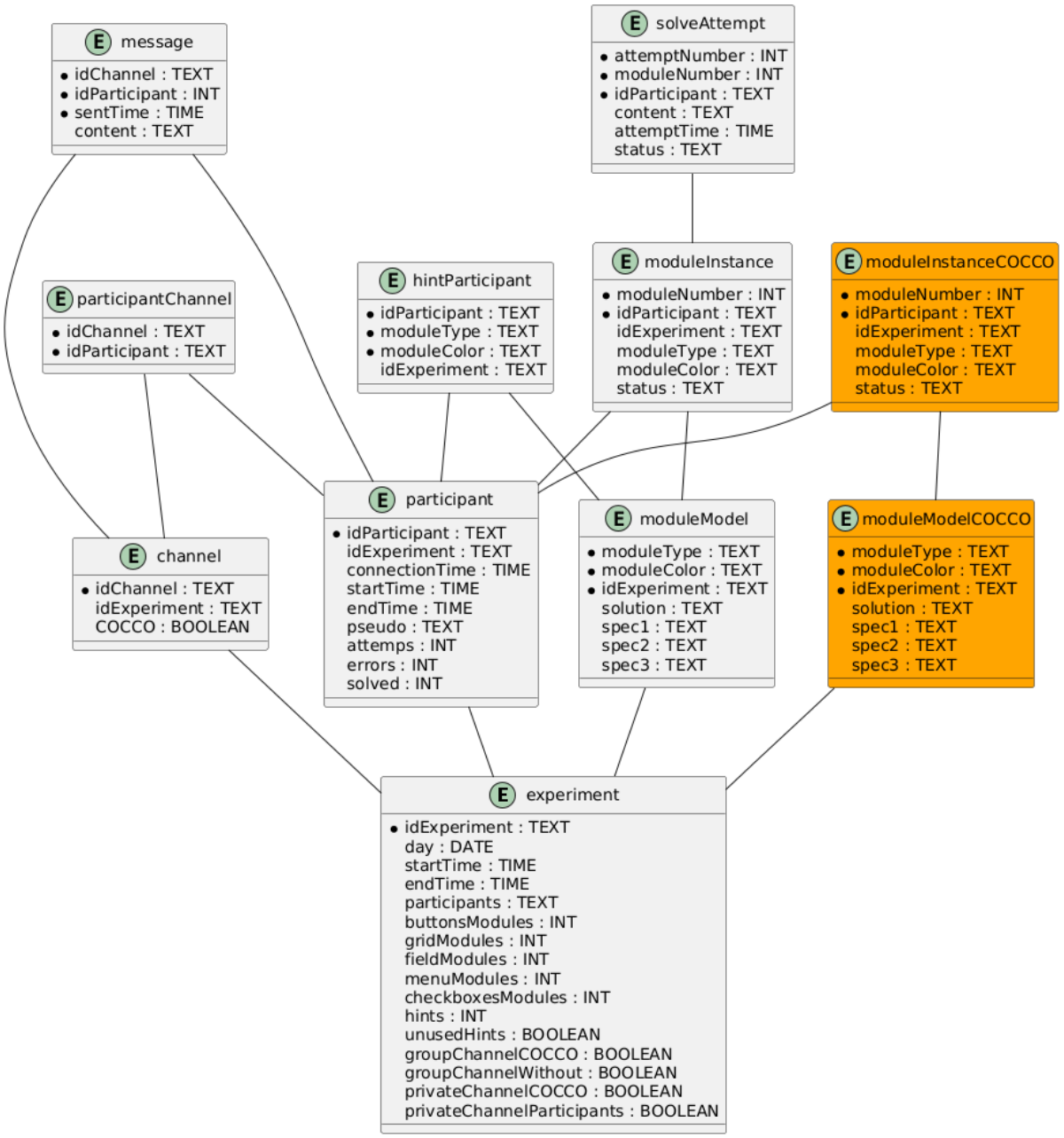


Figure 5.4: Database diagram including COCCO (made using PlantUML)

5.2.2 Pre-activity and post-activity surveys

In addition to the collaborative activity, participants answered two surveys, before and after the activity. The pre-activity survey collected participants' consent, demographic information, languages, and perceived ease with instant messaging tools. The post-task survey collected answers to the NASA Task Load Index (NASA-TLX)[34], an assessment tool for subjective workload. A general score is calculated using six weighted factors (mental demand, physical demand, temporal demand, performance, effort, and frustra-

tion). We added a Likert scale on participant's perception of the team's performance and communication quality, and trust in the team. Participants not in the control group also evaluated COCCO's communication quality and their trust in COCCO. Finally, a free-text field was included at the end of the questionnaire, allowing participants to offer general feedback or elaborate on their earlier responses (see Appendix D).

5.2.3 Participants

We recruited 72 participants for this experiment through our university's staff mailing lists and word of mouth. Participants were required to be fluent French speakers, to prevent communication problems between teammates and to use a computer with a stable internet connection. They were between 17 to 68 years old (m=28.6 years, sd=11.3), 39 female and 33 male.

Participants were split into 18 teams of four according to their availability. Teams were then assigned to one of the four conditions resulting in 5 teams in the control, 4 teams with COCCO in private chat, 4 teams with COCCO in the group chat, and 5 teams with COCCO in both private and group chat (Figure 5.5).

Each participant was assigned a random animal name (taken from a pre-defined list of ten) as a pseudo. The pseudo prevented participants from knowing the identities of their teammates and being influenced by possible previous familiarity or biases towards their names or genders.

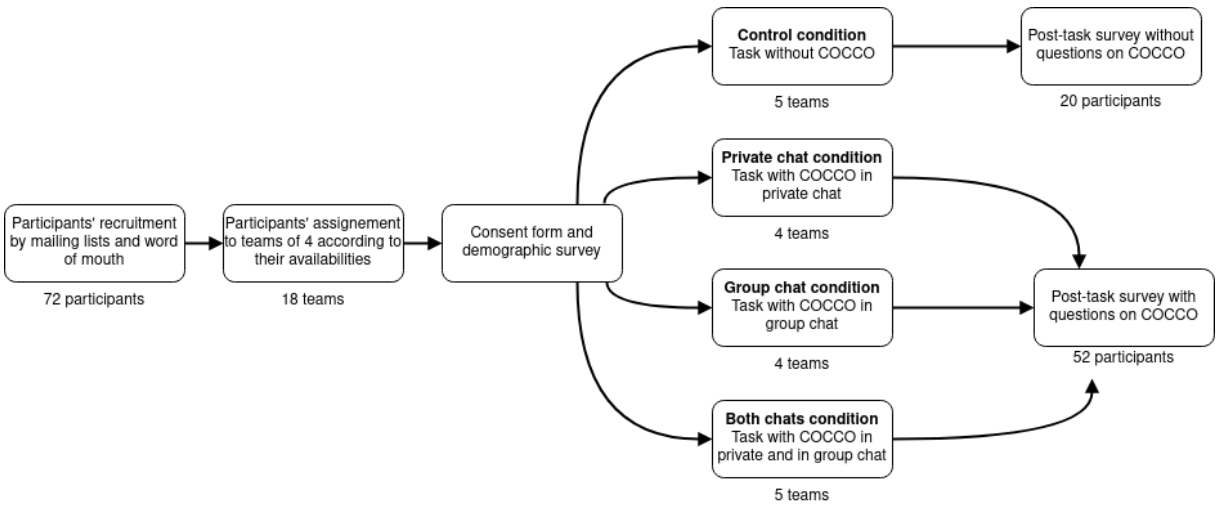


Figure 5.5: Experimental design

Participants were instructed to log onto the platform at their group's scheduled time and use a confidential code communicated beforehand in e-mail. Upon connection, participants were presented with instructions for the activity and had to confirm their presence by sending a message in the chat. Once all four participants confirmed their presence, the interface would become interactable, and the activity could start.

All participants received a €10 gift card after completing the activity. To encourage participants to perform to the best of their abilities, participants of the team with the best performance received a second gift card. For this purpose, team performance was calculated using the slowest participant's time (to encourage teamwork) and number of errors (to dissuade brute-forcing the tasks). This experimental 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)[11] Avis 2021-18.

5.3 Analysis and Results

The 18 groups generated a large amount of data [5], including the module solving attempts, the messages sent between participants and with the chatbot, and the surveys' answers. Our goal with this study was to investigate the following research questions:

- **How do groups engage with a chatbot?**
- **How do chatbots impact team dynamics?**

To address these questions, we first analyzed group performance across conditions, focusing on module-solving times and the response times to information requests. In the post-task survey, we examined differences in teammates' cognitive workload, as well as their perceptions of their own performance and that of their team. Next, we investigated how COCCO was used during the task and how it was perceived. Finally, we looked at team dynamics, analyzing how participants organized themselves and how these patterns related to other factors.

5.3.1 Team performance

Module solving time

To measure team performance, our initial metric was the time participants took to solve each of their respective modules (Figure 5.6). Participants with better performance were able to comprehend the information and solve modules faster than others. However, this metric also reflects the collective team performance, as a quicker module resolution suggests efficient information dissemination within the team.

We chose this metric over others, such as the one used to determine the winning team, or simply using each participant's final completion time, because it provides a more precise measure of performance. Additionally, we observed that errors were often due to accidental clicks or confusion about the rules and information received, rather than intentional attempts to brute-force solutions individually. Error count was not used as a performance metric because most participants made very few mistakes: 15 participants made no errors at all, and another 15 made only one error, together representing nearly

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

42% of the sample. In contrast, a small subset of participants had disproportionately high error counts, largely caused by repeatedly clicking the same incorrect answers (min = 0, Q1 = 1, median = 2, Q3 = 5.25, max = 105).

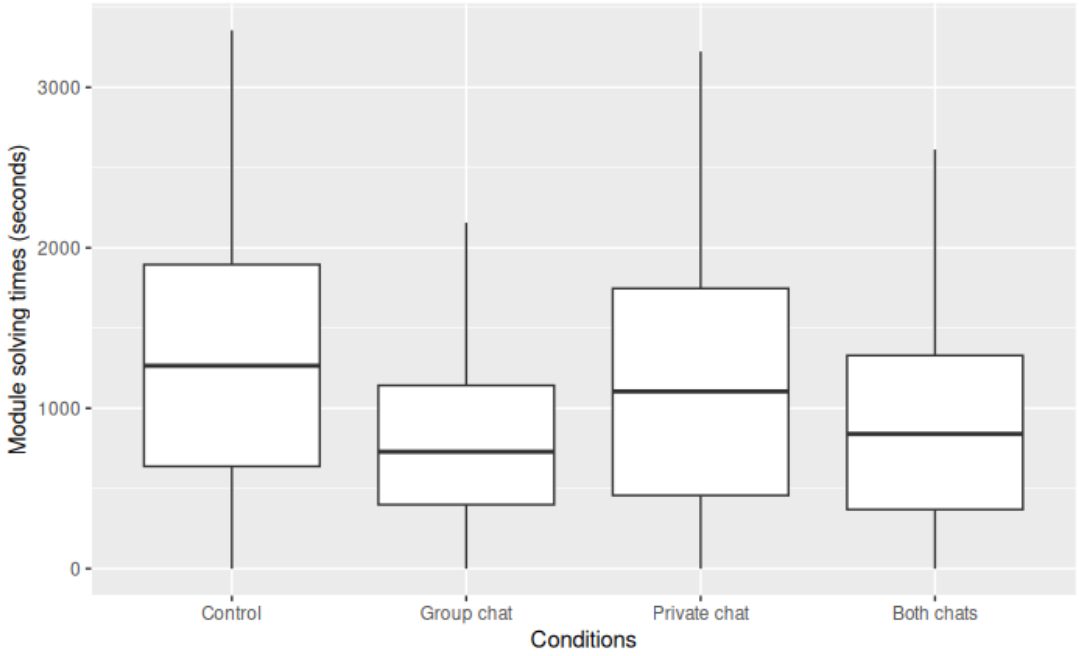


Figure 5.6: Module solving times for each condition

To compare solving times between conditions, we used a Kruskal-Wallis test. It revealed a significant difference in performance between the conditions (H=73.55, df=3, p<0.001). We conducted post-hoc pairwise comparisons using Wilcoxon rank sum test (Holm correction) and found significant differences between all conditions, as presented in Table 5.1. Participants in the control condition had the longest median module solving time (≈21min), then the private chat condition (≈18min), the both chats condition (≈14min), and the group chat condition (≈12min).

Table 5.1: Comparison of module solving times between conditions using Wilcoxon rank sum test (Holm correction)

Conditions	Control	Group chat	Private chat	Both chats
Control condition (med=1265s)	-			
Group chat condition (med=729s)	p<0.001***	-		
Private chat condition (med=1105s)	p=0.0354*	p<0.001***	-	
Both chats condition (med=840s)	p<0.001***	p=0.0354*	p=0.0032**	-

Response time

To compute the response time, we checked how long it took for a question to get an answer by marking down when someone asked for a hint or information and when they got a reply. Then, we calculated the time between the question and its answer. If a question did not get an answer, we did not include it in the data. This might occur, for example, if an information request was made again later on or if the participant tried to guess the answer on their own.

A Kruskal-Wallis test revealed a significant difference in answer delays between conditions ($H=61.71$, $df=3$, $p<0.001$). We conducted post-hoc pairwise comparisons using Wilcoxon rank sum test (Holm correction) and found significant differences between all the conditions, as presented in Table 5.2. Participants in the control condition had the longest median response delays (29s), then the private chat condition (23s), the group chat condition (17s), and the both chats condition (14s).

Table 5.2: Comparison of response times between conditions using Wilcoxon rank sum test (Holm correction)

Conditions	Control	Group chat	Private chat	Both chats
Control condition (med=29s)	-			
Group chat condition (med=17s)	$p<0.001$***	-		
Private chat condition (med=23s)	$p=0.0049$**	$p=0.0028$**	-	
Both chats condition (med=14s)	$p<0.001$***	$p=0.0408$*	$p<0.001$***	-

Response time excluding COCCO

The presence of COCCO notably influenced the variance in response times. When prompted, COCCO typically provided instant responses. To account for this, and check if participants were also sharing information quickly, we did a subsequent analysis, excluding COCCO's interventions from the measured response times.

We used a Kruskal-Wallis test, that revealed a significant difference in answer delays between conditions ($H=31.48$, $df=3$, $p<0.001$). We conducted post-hoc pairwise comparisons using Wilcoxon rank sum test (Holm correction) and found significant differences between the control group and the three manipulation groups, as presented in Table 5.3. Participants in the control condition had the longest median response delays (29s), compared to all other conditions (ranging from 18s to 22s).

5.3.2 Cognitive workload and perception

Cognitive workload

To evaluate how COCCO influenced participants' perception of the task we first calculated their cognitive workload using their answers to the NASA Task Load Index [34] in the post-task survey. To calculate the weighted score, participants first rate six dimensions (Mental

Table 5.3: Comparison of response times between conditions excluding COCCO using Wilcoxon rank sum test (Holm correction)

Conditions	Control	Group chat	Private chat	Both chats
Control condition (med=29s)	-			
Group-chat condition (med=18s)	p<0.001***	-		
Private-chat condition (med=22s)	p=0.0093**	p>0.05	-	
Both-chats condition (med=18s)	p<0.001***	p>0.05	p>0.05	-

Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration) on a scale ranging from 0 to 100. Then, they complete 15 pairwise comparisons between the dimensions to indicate which factors contribute more to their workload. The number of times each factor is selected determines its weight (ranging from 0 to 5). Each factor's rating is multiplied by its weight and divided by 15, the total sum of the weights. The final NASA-TLX score is a sum of each weighted factor, possibly ranging from 0 to 100.

Using a Kruskal-Wallis test, we found no significant differences in NASA-TLX score between conditions. However, a comparison of the weighted factors of the NASA-TLX score found a significant difference between groups (H=13.19, df=3, p<0.01) for the effort factor (*"How hard did you have to work (mentally and physically) to accomplish your level of performance?"*). We conducted post-hoc pairwise comparisons using Wilcoxon rank sum test (Holm correction) and found a significant difference (p<0.01) between the control condition (med=2.0) and the private-chat condition (med=9.0). This means that participants in the private-chat condition found the activity more effortful than participants with no chatbot assistance.

Perceived team performance, communication and trust

We looked at how the conditions affected the participants' perception of the task in terms of team performance and communication quality, trust in the team and, for non-control groups, COCCO's communication quality and trust in it. Using a Kruskal-Wallis test, we found a significant difference in the answers to the Likert-scale between conditions (H=6.13, df=2, p=0.0467) for COCCO's perceived communication quality. However, subsequent post-hoc comparisons using Wilcoxon rank sum test found no significant pairwise differences.

Cognitive workload and perceived team performance, communication and trust

We tested for correlation between the NASA-TLX score and its weighted factors, and the answer to the Likert-scale. The correlation analysis using Kendall's tau revealed a significant relationship between the NASA-TLX score and the COCCO's perceived communication quality (τ = -0.27, p = 0.006). Participants who had a good perception of COCCO's communication capabilities experienced a lower cognitive workload.

We then tested for correlation between the NASA-TLX's individual weighted factors and the answers to the Likert-scale. We found a correlation between the temporal demand factor and COCCO's perceived communication quality (τ=-0.33, p=0.0007). Participants who had a good perception of COCCO's communication capabilities experienced a lower time pressure. We additionally found negative correlations between the frustration factor and the perceived team's communication quality (τ=-0.36, p=0.0003), performance (τ=-0.26, p=0.0100) and trust in the team (τ=-0.30, p=0.0029). Participants who had a higher frustration level had an overall worse impression of their team.

5.3.3 Team dynamics

Finally, we turn to our research question on team dynamics. Within the context of the experimental task, the most prominent dynamic we observed was team organization. While some groups did not attempt to coordinate their actions and instead worked independently, with each member requesting only the information necessary for their own modules, other groups displayed a higher level of cooperation. These more organized teams proactively shared information with their teammates, monitored progress at the group level, and sometimes agreed on explicit strategies such as distributing information by module type or solving similar modules sequentially. This resulted in visually distinct timelines of group activity, where well organized teams progressed in a more structured and aligned manner compared to less organized teams (Figure 5.7).

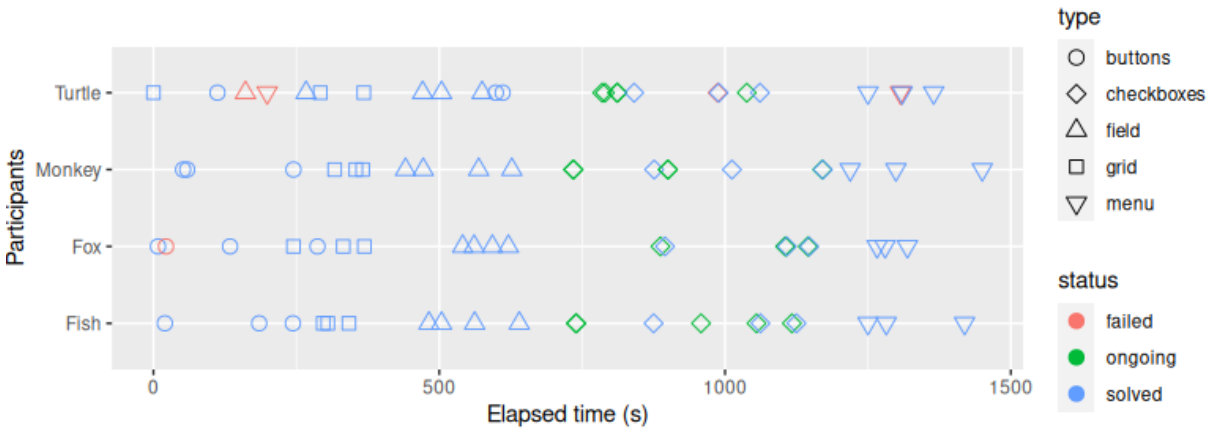


Figure 5.7: Click patterns of an organized team showing temporal grouping by module types

To move beyond visual observation, we computed a proxy measure of team organization: eta-squared (η^2) derived from the Kruskal-Wallis statistic, based on the distribution of solving attempts across module types. While more advanced models could capture the temporal complexity of the data, this measure was selected for its simplicity and interpretability as an initial assessment. The Kruskal–Wallis test indicated that 3 of the 18

groups showed no association between the timing of their clicks and the type of module clicked (Figure 5.8). For the remaining groups, we calculated eta-squared to quantify the strength of this relationship, which we refer to as the organization score. These eta-squared values are consistent with visual observations of the teams' clicks activity, suggesting that the measure provides a reasonable first approximation of team organization.

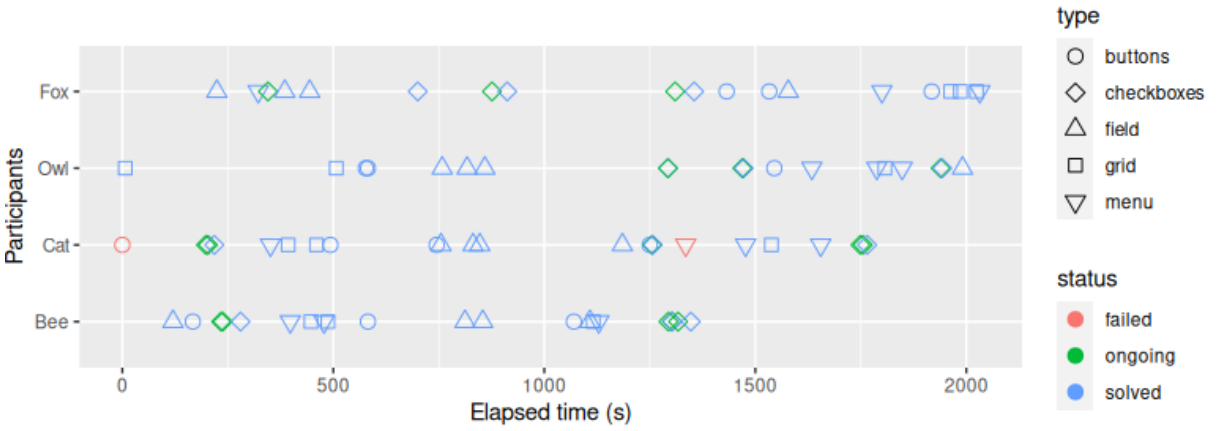


Figure 5.8: Click patterns of a less organized team

We first examined whether the degree of team organization was associated with other aspects of team performance using Spearman's rank correlation. First, we found no significant correlation between team organization and either the number of errors or task completion time.

Looking at team communication, we found a significant positive correlation between organization and the number of teamwork-related messages exchanged ($\rho = 0.753, p = 0.0012$). This result supports the interpretation that our organizational measure captures meaningful coordination processes, as more organized teams also engaged more frequently in communication about teamwork and mutual progress.

In addition, examining post-task survey data revealed a positive correlation between the organization score and the "Performance" score of the NASA-TLX ($\rho = 0.543, p = 0.039$). This suggests that more organized teams reported higher perceived performance.

Finally, we investigated whether chatbot interventions influenced the degree of organization. To do so, we correlated the organization score with the number of interactions between the teams and the chatbot COCCO. This analysis revealed no significant relationship ($\rho = 0.251, p = 0.457$), indicating that organized teamwork was not directly driven by the chatbot's prompts or interventions.

5.4 Discussion

Our exploration of the impact of a chatbot on a collaborative environment revealed how chatbot assistance can influence team dynamics and performance.

First, COCCO's interventions positively influenced teams' performances by reducing participants' module solving times in non-control groups. The median time went from 21 minutes for groups in the control condition to 18 minutes for the private chat condition, 14 minutes for the both chats condition and 12 minutes for the group chat condition.

Additionally, they improved team's communication, as can be seen by participants' faster response times to teammates' information requests. This indicates that COCCO's fast answers set a rhythm for the participants and possibly made the team's exchanges more fluid by preventing unanswered information requests from piling up. This finding reinforces previous conclusions in the literature that chatbots are effective conversational facilitators [48, 109, 99]. However, our study demonstrates this effectiveness in a context where debate structuring was not the primarily goal of the chatbot.

Finally, participants with a good impression of the chatbot's communication quality had a lower cognitive workload. This suggests that chatbot assistance is able to reduce cognitive workload. More precisely, we found a link between a participant's perceived communication quality of COCCO and experienced time pressure. This might indicate that the chatbot reduced some of the pressure and workload experienced by the participants, possibly by speeding the rate of all the answers to information requests and answering some of the requests by itself leaving more time for participant to solve their own modules.

Looking at the different communication strategies tested with COCCO, we observe that providing assistance by private chat was much less effective than by group chat. Indeed, while team performances in the private chat condition were still better than control groups, they were significantly worse than teams in the group chat condition. This is also true of teams in the both chat condition, who had worse time performances than group chat condition teams, showing that chatbot assistance by private chat is diminishing the effectiveness of the assistance in terms of speed. This results in our observation that participants interacting with COCCO in private chat felt that they had to exert more effort than control group participants.

This can be explained by the fact that this communication strategy requires participants to frequently switch channels and actively request assistance. This fits with our observation that requests to COCCO in private chat would decline with time or even stop for some teams. In private chats, COCCO needed direct engagement or prompting by participants to be optimally functional given that it had only one conversational partner. In group chats, COCCO could passively monitor the team's conversation and had more opportunities to react to information requests without needing extra effort from the participants. Additionally, non-use by one teammate would directly diminish the usefulness of COCCO's for the rest of the group and increase chances that it would be completely ignored. Indeed out of a total number of 126 requests for information by COCCO, only 48 (38%) were answered. This is mostly explained by the fact that participants would receive requests for information they did not have and chose to simply not answer COCCO. Some

other participants might also have received successive requests and chose to only answer the most recent one. However, some of them would also not answer when they were the only one with the requested information, which prevented COCCO from relaying back an answer to the original asker and rendering it ineffective.

Analysis of team dynamics revealed that groups exhibited varying levels of organization in how they approached the task. Teams with higher organization scores, reflecting structured coordination and proactive information sharing, also engaged in more teamwork-related communication and reported higher perceived performance on the NASA-TLX survey. Interestingly, this organization was not associated with objective measures of task success, such as the number of errors or module-solving time. Importantly, we did not observe evidence that COCCO's interventions directly impacted team organization. These results suggest that while organized teams felt more effective and communicated more, their objective performance was not improved. Similarly, the chatbot interventions were insufficient to meaningfully influence team organization, though they also did not appear to have any negative effects.

Following those insights, we suggest that chatbots should adopt a more dynamic role in team collaboration and proactive strategies. Instead of passively absorbing information and reacting to specific requests, chatbots should take a more proactive part in the activity by, for example, noticing missing information in their model and inferring it from other available data or requesting it from human teammates. This would help maintain the communication with the chatbot and keep it relevant throughout the collaboration, especially for longer-term collaborations.

5.4.1 Limitations and future directions

Our study offers valuable insights into the role of chatbots in collaborative environments, but it comes with limits. To solidify the generalizability of our findings, the study could benefit from replication with a more extensive participant base, especially by forming teams of varied sizes and with different objectives (time limit, competition between teammates). This would provide a more comprehensive understanding of chatbots influence on group dynamics across different collaborative scales and configurations.

While chatbot assistance in private channels showed limited benefits in this study, it may prove more valuable in larger teams. In bigger groups, where group chats tend to become busier and participants must manage more teammates and greater amounts of information, raising the risk of information overload [80], private channels could offer a quieter, more focused space for chatbot support. Exploring this possibility in future research could help us better understand how chatbots impact team collaboration depending on group size and communication complexity.

As teams familiarize themselves with chatbot assistance, an analysis of how expert usage evolves becomes imperative, especially for teams where teammates have become familiar with each other and roles have cemented themselves.

One significant limitation was the absence of other prevalent communication channels in our setup, such as phone or e-mail. As noted by Snyder and Lee-Partridge [106], these methods are substantial competitors to newly introduced collaboration tools. Our study's

exclusion of these channels may have unintentionally inflated the perceived effectiveness and influence of COCCO in certain collaborative contexts.

5.5 Conclusion

This chapter explored the role of chatbots in shaping collaborative spaces. Our study shows that chatbots can meaningfully enhance teamwork, particularly when integrated into communication platforms. COCCO's noticeable impact in the group chat condition, compared with its limited utility in private chats, illustrates how strongly deployment context influences a chatbot's effectiveness and adaptability.

One of the clearest benefits we observed was COCCO's role as a conversational facilitator. By reducing response delays, the chatbot helped teams keep discussions moving and avoid the stalls, confusion, and frustration that often come with unanswered requests. Just as importantly, COCCO directly supported performance by providing hints that guided teams toward solutions. These two forms of assistance, smoother communication and task-specific support, translated into shorter module solving times, pointing to tangible performance gains. The picture was less straightforward when it came to cognitive load. While some participants reported feeling less pressure and effort, others found the interaction more demanding.

COCCO's design choices help explain these outcomes. The system was deliberately reactive, stepping in only when prompted so as not to disrupt the natural flow of collaboration. In practice, though, this led to minimal private interactions and limited utility in that setting. This finding contrasts with earlier research [20], where participants often preferred private chatbot support. Balancing proactivity with user control therefore emerges as a key consideration for effective chatbot deployment.

Looking forward, advances in models such as ChatGPT open up new possibilities for collaborative environments. New functionalities designed specifically for group use[12] may prove better at interpreting team dynamics, supporting decision-making, assisting brainstorming, and even helping to navigate conflicts. While COCCO showed promise in our study, future systems built on more powerful models could reshape how teams interact, organize, and rely on chatbots as active collaborators.

[12] `https://openai.com/index/group-chats-in-chatgpt/` (accessed December 12, 2025)

Chapter 6

Conclusion

This thesis investigates how chatbots can support collaborative work without disrupting or introducing bias into team dynamics. As chatbots become increasingly integrated in everyday tools and workflows, understanding the impact of chatbot assistance on decision-making, collaboration, and team communication becomes essential. Current chatbot designs primarily focus on one-on-one interactions, with limited capacity to engage with multiple users simultaneously. Our work contributes to this emerging field by evaluating when and how chatbots can be effectively integrated into team environments, as well as the factors that influence trust and reliance on their assistance. Our research aims to bridge these gaps and makes contributions to the field in several important ways.

First, we reviewed the literature on AI assistance, focusing on how users rely on automated systems and the factors influencing trust, appropriate reliance, and automation bias. While algorithmic advice often outperforms human advice in objective tasks, users tend to prefer human advice for subjective or socially sensitive decisions. Chatbots differ from other systems through natural language interaction, ease of integration, and social presence, which can affect user trust and reliance. In group settings, chatbots can support collaboration by structuring discussions, balancing participation, and facilitating coordination, but their effectiveness depends on timing, delivery method, and whether prompts are directed at individuals or the group. Overall, while chatbots show potential in enhancing teamwork and reducing cognitive load, significant research gaps remain in understanding how users calibrate reliance and trust in these systems, how different interaction styles and communication modalities influence group decision-making, and how the dynamics of multi-party collaboration shape the effectiveness and acceptance of chatbot interventions over time.

Following our conclusions from the literature review, we conducted a confirmatory study to investigate automation bias in interactions with chatbots, comparing how participants responded to advice from a chatbot or a human assistant in a wilderness survival task. Using a between-subjects Wizard of Oz design, participants answered 20 questions and received advice from either a human or a chatbot assistant. Our results showed that participants were significantly more likely to change their answers when contradicted by the chatbot than by the human assistant. They also asked for more hints and engaged less in justification or reflection than with the human assistant, despite both assistants

providing identical guidance. Analyses of the post-experiment questionnaire indicated that reliance on the chatbot was higher but more superficial, with participants perceiving lower personal performance when following its advice. These findings highlight that chatbot presentation alone can increase user reliance, raising concerns about over-reliance, reduced critical thinking, and accountability.

Extending our focus from individual to group interactions, we developed a web-based experimental platform to study group interactions and collaboration with chatbots. The platform presents teams of participants with 16 puzzle modules that require communication and coordination to solve, using instant messaging through group and private chats. Modules vary in type and color, requiring different forms of information to be solved. Solutions are distributed across teammates to encourage collaboration, while errors are penalized to discourage brute-force solving. The activity is browser-based, takes less than an hour, and requires no prior knowledge, making it accessible and easy to recruit for. The concepts manipulated are simple, which also facilitates participant onboarding, data annotation, and chatbot integration. Extensive testing refined the interface, instructions, and technical stability, confirming participant engagement and task clarity. The platform is flexible, allowing adjustments in team size, module complexity, communication channels, and task structure, and could be scaled for larger studies or adapted for future research on collaborative teamwork.

Using this experimental activity, we investigated the impact of a chatbot assistant on team collaboration. The chatbot was designed to parse and store task-relevant information from participants' messages and provide assistance either in group chat, private chat, or both, depending on the experimental condition. Results showed that the chatbot assistant significantly improved team performance, reducing module-solving times and response delays to teammates' information requests, particularly when integrated into the group chat. In contrast, private chat assistance was less effective and perceived as more effortful. Participants who rated the chatbot's communication quality highly, experienced lower cognitive workload and time pressure. Team organization positively correlated with communication frequency and perceived performance, but not directly with chatbot interactions. The study highlights that reactive chatbots can enhance collaboration but suggests that more proactive, context-aware chatbots may further optimize team dynamics, especially in larger or more complex collaborative environments.

In summary, this thesis explored how chatbots can support collaborative work by examining their influence on individual reliance, group communication, and team performance. Through a combination of literature review, controlled experiments, and the development of a dedicated experimental activity, we showed that chatbots have the potential to support decision-making and teamwork. While chatbot assistance was found to improve efficiency and reduce workload in team settings, it also revealed risks of over-reliance and a decrease in critical engagement. These findings underscore that many of the key concerns highlighted in our research remain pressing. The need for appropriate reliance, critical thinking, and accountability when engaging with a chatbot is unchanged, despite the rapid advancement of chatbot systems. If anything, it becomes even more critical. This is particularly true as generative AI accelerates the integration of chatbots into professional tools and workflows.

6.1 Generative AI

When we defined our research questions, we did not consider the correctness of chatbot advice to be a central concern. At the time, it was widely expected that chatbots deployed in real-world applications primarily retrieved factual information from verified databases or knowledge sources. Any inaccuracies were likely to stem from outdated or conflicting information, or from misinterpretations of ambiguous queries.

However, the emergence of generative AI, particularly large language models such as ChatGPT and Claude, has fundamentally altered this assumption. These models do not retrieve answers from a static knowledge base; instead, they generate responses based on learned statistical patterns. While this allows them to produce fluent and contextually relevant replies, it also introduces the risk of "hallucinations": convincing but entirely fabricated content. For example, these systems may invent plausible-looking URLs to nonexistent pages of well-known websites or fabricate citations for books that were never published by real authors.

These hallucinations raise significant concerns about the trustworthiness and reliability of chatbot-generated advice, particularly in collaborative or high-stakes environments. This marks a major departure from earlier chatbot paradigms, which were largely deterministic and based on pre-validated responses. The challenge is no longer just about how chatbots should present information to avoid biasing users, but whether the information itself is grounded in reality. Compounding this issue, findings show that users often exhibit a bias toward accepting chatbot advice, even when it is incorrect, highlighting the urgent need to support users in exercising critical judgment about when to trust and act on chatbot-generated content.

At the same time, recent advances in AI offer promising strategies to mitigate these concerns and enhance the reliability of chatbot assistance. Techniques such as retrieval-augmented generation [60], grounding responses in verified databases, and implementing real-time fact-checking mechanisms can substantially reduce hallucinations and improve the accuracy of generated content. Furthermore, emerging approaches in explainable AI allow chatbots to provide users with transparent reasoning or confidence estimates, helping teams assess the trustworthiness of advice more effectively. These developments suggest that, although generative chatbots introduce new challenges, they also open opportunities for designing systems that are both highly capable and reliably supportive in collaborative environments.

6.2 Future Directions

Looking forward, our research lays a foundation for continued exploration of chatbot-assisted collaboration. This work could also delve deeper into refining chatbot intervention strategies that balance helpfulness with unobtrusiveness, ensuring that future generations of chatbots are trusted teammates in collaborative environments rather than passive participants or sources of distraction.

In future work, our platform could be extended to explore more complex collaborative

scenarios. One possibility is to conduct repeated problem-solving sessions to investigate whether experienced teams continue to benefit from the assistant over time. Another direction is to study larger group sizes, where managing the group chat itself may become increasingly challenging.

On the chatbot side, future studies could explore making the assistant more proactive, for example, by detecting when participants fall behind and offering targeted support, autonomously querying information, or even correcting mistaken contributions. Investigating these capabilities could help reveal when and how team communication begins to shift from being primarily human driven to increasingly mediated by the chatbot, clarifying the point at which reliance on the assistant fundamentally changes the nature of collaboration.

By building on the platform and findings from our study, future research can guide the development of chatbots that not only enhance teamwork but also foster appropriate reliance and maintain users' trust in increasingly complex and collaborative digital environments. Designing future collaborative chatbots should go beyond basic functionality to address long-term engagement, adaptability, and scalability, ensuring they remain effective, trustworthy, and context-aware partners in human collaboration.

We identify the following research problems to explore in the study of collaborative chatbots:

- Long-term chatbot collaboration: Investigate whether automation bias can be mitigated through repeated interactions over time. As users become more familiar with a chatbot's capabilities and limitations, their reliance patterns may shift. Future research should examine whether such familiarity fosters more critical engagement or reinforces dependence, and how chatbots can be designed to encourage calibrated trust.

- Adaptive chatbot behavior: Explore how chatbots can dynamically assess a user's cognitive engagement with the task and adapt their behavior accordingly. This could involve adjusting the frequency, complexity, or assertiveness of interventions, which could even include intentionally obstructive actions, based on real-time signals of user attention, workload, or decision confidence.

- Scalability to larger teams: Evaluate the usefulness and impact of chatbots in larger groups with higher coordination costs. This includes examining whether the role of the chatbot changes in larger teams and whether its communication strategy must adapt to manage group dynamics, resolve conflicts, or ensure equal participation.

In conclusion, while this thesis has contributed valuable insights into chatbot-assisted collaboration, much remains to be done. Our work established empirical evidence of automation bias in chatbot interactions, introduced a novel methodological framework for studying group collaboration, and identified both opportunities and risks in integrating chatbots into collaborative environments. Together, these contributions provide a foundation for future research and design, offering both conceptual and practical tools to better understand and shape how chatbots influence teamwork. As chatbots become increasingly

embedded in professional and social contexts, the importance of ensuring appropriate reliance, fostering critical thinking, and supporting effective collaboration will only grow. This thesis thus not only advances academic knowledge but also provides timely guidance for the responsible development of chatbots as trustworthy and effective partners in human collaboration.

Appendix A

Wilderness survival questions, answers and hints

Table A.1: Wilderness survival questions, answer propositions and hints (translated from French)

Question	Answer propositions	Hints
You have strayed from your group and are now lost. You have no special equipment to signal your presence. The best way to alert your friends is to:	A. Whistle as high and loud as possible	1. Low sounds go farther.
	B. Scream the loudest you can	
	C. Call for help as loud as possible with a deep voice	2. Whistled languages exist for a reason
You are in an area infested with snakes. The best way to avoid getting bitten is to:	A. Move at night	1. Snakes prefer to flee from animals larger than themselves.
	B. Walk smoothly and quietly	
	C. Make as much noise as possible with your feet	2. Snakes attack when disturbed.
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. Place a small piece of the plant on your lip a few minutes before eating it a little	1. Some plants can be lethal on contact.
	B. Avoid brightly colored plants	2. Birds have a different digestive system from ours.
	C. Try small quantities of what the birds are eating	

The weather is getting hot and dry. You have about a liter of water with you. You have to:	A. Drink as soon as you are thirsty	1. It is more dangerous to drink unsafe water than to be thirsty
	B. Do not drink before you stop for the night; at this time drink what you think you need	2. You will not be able to look for a water source or help if you are unconscious.
	C. Ration the water and drink only about a cup a day	
You have run out of water and you are very thirsty. You finally arrive at a creek to find it completely dry. The best way to find water is to:	A. Dig in the hollow of a curve	1. Plant roots retain water.
	C. Dig anywhere in the creek bed	
	B. Dig under plants and tree roots on the banks of the stream	2. The creek bed is the last place to dry up.
You are in a desert area, along ravines where water can be found. At night, the best place to camp is:	A. Halfway down	1. The walls of ravines tend to crumble.
	B. High on an overhang	2. There may be thunderstorms overnight.
	C. Near the pocket of water in the ravine gorge	
You have moved away from your camp and it is getting dark quickly. As you return to your camp, the intensity of your flashlight begins to drop. You must then:	A. Switch on and off the light for a few seconds in order to know the path and move forward	1. Night falls quickly, you will quickly be plunged into darkness if you wait.
	C. Head to your camp as quickly as possible, keeping the torch lit	
	B. Warm the batteries with your hands before putting them back into the lamp	2. Batteries perform worse in the cold.
A snowstorm keeps you confined to your tent. To sleep, you have left your camping gas stove on. It is preferable for the flame to be:	A. Red	1. A non-blue flame means combustion is incomplete.
	C. Yellow	
	B. Blue	2. A blue flame can indicate the presence of copper sulfate.

You must cross a very strong current river, the bed of which is strewn with large stones, and the water is cloudy. After carefully choosing where you will be passing, you must:	A. Take off your boots and throw your backpack on the other bank	1. More weight allows better stability.
	B. Throw your backpack on the other bank, but cross with your boots on	
	C. Cross with your boots and backpack on	2. A backpack is a source of weight and an additional current surface.
When you cross a river that has a very strong current and you have water up to your waist, you must cross by turning:	A. Facing the current	1. Tipping over in the river is dangerous.
	C. Back to the current	
	B. Towards the other bank	2. We are more stable from the front.
You find yourself on a rocky slope on which you have to climb to pass. The rocks are slippery and covered with moss. You have to climb:	A. In socks	1. It is easier to feel the rock with bare feet.
	B. Keeping your boots on	
	C. Barefoot	2. There is always a risk of injury.
When you return to your camp, you are surprised to discover a bear circling around your tent. You are not armed. Ten meters from you, the bear rises on its hind legs. You must then:	A. Stand still but ready to slowly back up	1. It is important to get out of range as quickly as possible.
	B. Climb to the nearest tree	2. A bear can run over 50 km/h and climb a tree in seconds.
	C. Run away	
You try to head north out of the forest. It is better to follow:	A. Your keychain compass	1. Most low-quality compasses are susceptible to magnetic disturbance.
	B. The sun	2. The sun can be obstructed by forest trees and even clouds.
	C. Moss on trees	
While walking you fall badly on a branch. You realize that part of the branch has pierced your leg. What should you do?	B. Gently remove the branch	1. Once the leg has been perforated, it is best to limit movement in the injury.

	A. Break off the branch as close as possible to the injury	2. Branches can be covered with a lot of things causing infections.
	C. Don't touch the branch	
While walking through tall grass, you get bitten by a snake that gets away before you can identify it. You have to:	A. Open the wound and let some blood flow	1. Less than 20% of snakes are poisonous. It is best to avoid causing further injury.
	B. Make a tourniquet over the wound	
	C. Clean and dress the wound	2. Some venoms can be fatal within minutes. It is important to prevent the venom from spreading.
Night falls and the air cools. What is your best choice of blanket?	A. A cotton blanket still damp after crossing the river	1. Cotton is a poor insulator contrary to wool.
	B. An assembly of foliage	2. A survival blanket weighs between 60 and 65 grams with a thickness of 13 micrometers.
	C. A survival blanket	
To warm up, you try to start a fire. You try to:	A. Focus the rays of the sun with a magnifying glass	2. The presence of the sun is never guaranteed.
	B. Strike a flint with iron to obtain sparks	1. The friction method requires dry wood and a lot of energy, especially for non-experts.
	C. Rub two pieces of wood until it ignites	
To light your fire, it is best to use:	A. Packaging and other trash	1. It is not recommended to breathe plastic fumes.
	B. Freshly cut branches	2. Leaves have a very short burning time.
	C. Dead leaves	
You found some water, but you want to purify it before you drink it. It's better to:	A. Filter the water	1. Many microbes do not survive above certain temperatures.

	C. Collect the water condensation	
	B. Boil the water	2. It is easy to create a filter with cloth and sand.
You build a shelter for the night using branches and foliage gathered around for the structure. You start by placing:	A. A large branch leaning against the trunk of a tree	2. The tipi structure keeps a fire inside the shelter without too much risk.
	B. A long branch placed high between two trees	
	C. Several long branches crisscross	1. Building a sturdy and stable tipi is difficult for beginners.

Appendix B

First experiment post-questionnaire

- [Likert scale] Sur une échelle de 1 à 10 (1: Pas du tout d'accord à 10: Tout à fait d'accord), indiquez votre accord avec les propositions suivantes.
  On a scale from 1 to 10 (1: Strongly disagree to 10: Strongly agree), indicate your level of agreement with the following statements.

  – J'ai bien répondu aux questions posées et obtenu un bon score.
  I answered the questions well and obtained a good score.

1	2	3	4	5	6	7	8	9	10

  – Les interventions de l'assistant ont influencé mes réponses.
  The assistant's interventions influenced my answers.

1	2	3	4	5	6	7	8	9	10

  – L'aide apportée par l'assistant m'a permis d'obtenir un meilleur score.
  The help provided by the assistant allowed me to obtain a better score.

1	2	3	4	5	6	7	8	9	10

  – L'assistant a adapté ses interventions au cours des questions en fonction de mes réponses et de mon score.
  The assistant adapted its interventions during the questions according to my answers and my score.

1	2	3	4	5	6	7	8	9	10

- [Free text] Avez-vous des remarques vis-à-vis du déroulement de l'expérimentation, de votre ressenti ou de vos interactions avec l'assistant ?
  Do you have any comments about the course of the experiment, your experience, or your interactions with the assistant?

Appendix C

COCCO's intents

Table C.1: COCCO's intents (Dialogflow)

Name and use case	Matched parameters
Conversational intents	
Welcome *Participant greets the chatbot*	/
Help *Participant asks for help*	/
Confirmation *Participant confirms or agrees*	/
Contradiction *Participant contradicts or disagrees*	/
Information given: COCCO's knowledge base is updated.	
Buttons hint *Participant shares the hint for a button module of a specific color*	The module's color and the correct solution composed of a sequence (left, center, right).
Grid hint *Participant shares the hint for a field module of a specific color*	The module's color and the correct solution composed of an expression describing a position in a 3x3 grid (e.g., "upper right", "center left").
Checkboxes hint *Participant gives the hint for a field module of a specific color*	The module's color and the correct solution composed of a module's number.
Checkboxes solution *Participant shares a checkboxes module's solution*	The module's color and its solution which is a list of colors.
Menu hint *Participant shares the hint for a field module of a specific color*	The module's color and the correct solution composed of a type of module, a color, and a module's number.

Name and use case	Matched parameters
Menu solution *Participant shares a menu module's solution*	The module's color and its solution which is a player's pseudo.
Field hint *Participant shares the hint for a field module of a specific color*	The module's color and the correct solution composed of a type of module, a color, and one of the players' pseudo.
Field solution *Participant shares a field module's solution*	The module's color and its solution which is a number.
Module color *Participant shares the color of one of their modules*	The module's number and color.
Module number *Participant shares the number of one of their modules*	The module's number, and optionally, the module's type and color.
Module information *Participant confirms that they have a specific module*	The module's type, color, and number.
Requests: COCCO checks its knowledge base and shares the requested information if it has it.	
Color request *Participant requests the other participants' module colors*	The color of the relevant module.
Hint request *Participant requests the hint for a module*	The module's type and color.
Number request *Participant requests another player's module color*	The module's type and color, and optionally, depending on the context, the other player's pseudo.
Participant request *Participant requests to know which player has a specific module*	The module's type, color, and number.

Appendix D

Second experiment post-questionnaire

- [Slider questions] Pour chaque question, indiquez votre réponse avec le slider.
  For each question, indicate your answer using the slider.

  - **Pression temporelle:** le rythme de la tâche vous a-t-il pressé ?
    Time pressure: Did the pace of the task make you feel rushed?

    Très peu _____ Beaucoup

  - **Pression mentale:** la tâche était-elle exigeante mentalement ?
    Mental pressure: Was the task mentally demanding?

    Très peu _____ Beaucoup

  - **Performance:** avez-vous réussi à accomplir ce qu'il vous était demandé ?
    Performance: Were you able to accomplish what was asked of you?

    Très peu _____ Beaucoup

  - **Effort:** avez-vous dû réaliser des efforts pour accomplir cette tâche ?
    Effort: Did you have to put in effort to complete this task?

    Très peu _____ Beaucoup

  - **Frustration:** avez-vous été stressé, agacé ou découragé durant la tâche ?
    Frustration: Were you stressed, annoyed, or discouraged during the task?

    Très peu _____ Beaucoup

- [Explanation] Quinze paires de facteurs pouvant affecter la charge de travail vont vous être présentés. Pour chaque paire, choisissez le facteur ayant le plus affecté votre charge de travail durant la tâche.
  Fifteen pairs of factors that may affect workload will be presented to you. For each

pair, choose the factor that affected your workload the most during the task.

Les facteurs sont définis comme suit:
The factors are defined as follows:

- **Performance**: Dans quelle mesure pensez-vous avoir réussi à atteindre les objectifs de la tâche ? Dans quelle mesure étiez-vous satisfait de votre performance dans la réalisation de ces objectifs ?
  Performance: To what extent do you think you achieved the task objectives? How satisfied were you with your performance in accomplishing these objectives?

- **Pression mentale**: Quel niveau d'effort mental était nécessaire (i.e. réfléchir, se rappeler, décider, chercher, ...) ? La tâche était-elle facile ou exigeante, simple ou complexe ?
  Mental pressure: What level of mental effort was required (e.g., thinking, remembering, deciding, searching, ...)? Was the task easy or demanding, simple or complex?

- **Pression physique**: Quelle quantité d'activité physique était nécessaire (i.e. pousser, tirer, tourner, contrôler, activer, ...) ? La tâche était-elle calme ou intense, reposante ou épuisante ?
  Physical pressure: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, ...)? Was the task calm or intense, relaxing or exhausting?

- **Pression temporelle**: Quelle pression de temps avez-vous ressentie en raison du rythme des tâches ? Le rythme était-il lent et tranquille ou rapide et frénétique ?
  Time pressure: What time pressure did you feel due to the pace of the tasks? Was the pace slow and relaxed or fast and hectic?

- **Frustration**: Dans quelle mesure vous êtes-vous senti anxieux, découragé, irrité, stressé et agacé au lieu de satisfait, assuré et détendu pendant la tâche ?
  Frustration: To what extent did you feel anxious, discouraged, irritated, stressed, or annoyed rather than satisfied, confident, and relaxed during the task?

- **Effort**: À quel point avez-vous dû travailler pour atteindre votre niveau de performance ?
  Effort: How hard did you have to work to reach your level of performance?

- [Pairwise comparisons] Choisissez le facteur ayant le plus affecté votre charge de

travail.
Choose the factor that most affected your workload.

Pression physique	–	Effort
Pression temporelle	–	Performance
Frustration	–	Effort
Pression mentale	–	Pression temporelle
Performance	–	Frustration
Effort	–	Pression mentale
Pression physique	–	Performance
Pression temporelle	–	Frustration
Pression physique	–	Pression mentale
Pression physique	–	Frustration
Performance	–	Pression mentale
Performance	–	Effort
Pression temporelle	–	Effort
Pression mentale	–	Frustration
Pression temporelle	–	Pression physique

- [Likert scale] Sur une échelle de 1 à 10 (1: Pas du tout d'accord à 10: Tout à fait d'accord), indiquez votre accord avec les propositions suivantes concernant vous et vos coéquipiers.
  On a scale from 1 to 10 (1: Strongly disagree, 10: Strongly agree), indicate your agreement with the following statements concerning yourself and your teammates.

  – Mes coéquipiers ont communiqué de façon fluide et efficace.
  My teammates communicated smoothly and effectively.

1	2	3	4	5	6	7	8	9	10

  – Mes coéquipiers m'ont communiqué des informations sans erreur.
  My teammates communicated error-free information to me.

1	2	3	4	5	6	7	8	9	10

  – Les messages envoyés par mes coéquipiers étaient faciles à comprendre.
  The messages sent by my teammates were easy to understand.

1	2	3	4	5	6	7	8	9	10

– Les messages envoyés par COCCO étaient faciles à comprendre.
The messages sent by COCCO were easy to understand.

1	2	3	4	5	6	7	8	9	10

– Notre performance est moins bonne que celle de la plupart des autres équipes.
Our performance is worse than that of most other teams.

1	2	3	4	5	6	7	8	9	10

– COCCO a communiqué de façon fluide et efficace.
COCCO communicated smoothly and effectively.

1	2	3	4	5	6	7	8	9	10

– J'ai préféré vérifier les informations transmises par mes coéquipiers avant de les utiliser.
I preferred to double-check the information transmitted by my teammates before using it.

1	2	3	4	5	6	7	8	9	10

– J'avais confiance dans les informations que COCCO a partagé avec moi.
I trusted the information COCCO shared with me.

1	2	3	4	5	6	7	8	9	10

– Les interventions de COCCO ont provoqué beaucoup de malentendus.
COCCO's interventions caused many misunderstandings.

1	2	3	4	5	6	7	8	9	10

– Mon équipe a réalisé une bonne performance.
My team achieved a good performance.

1	2	3	4	5	6	7	8	9	10

– J'avais confiance dans les informations que mes coéquipiers ont partagé avec moi.
I trusted the information my teammates shared with me.

1	2	3	4	5	6	7	8	9	10

- Le temps utilisé par mon équipe et le nombre d'erreurs réalisées sont aussi bas que possible.
  The time used by my team and the number of errors made were as low as possible.

1	2	3	4	5	6	7	8	9	10

- COCCO m'a communiqué des informations sans erreur.
  COCCO communicated error-free information to me.

1	2	3	4	5	6	7	8	9	10

- Il y a eu beaucoup de malentendus dans nos échanges entre coéquipiers.
  There were many misunderstandings in our exchanges between teammates.

1	2	3	4	5	6	7	8	9	10

- [Free text] Avez-vous des remarques vis-à-vis du déroulement de l'expérimentation, de votre ressenti ou de vos interactions avec vos coéquipiers et COCCO ?
  Do you have any comments regarding the conduct of the experiment, your experience, or your interactions with your teammates and COCCO?

Bibliography

[1] Kelsey Allen, Franziska Brändle, Matthew Botvinick, Judith E Fan, Samuel J Gershman, Alison Gopnik, Thomas L Griffiths, Joshua K Hartshorne, Tobias U Hauser, Mark K Ho, et al. Using games to understand the mind. *Nature human behaviour*, 8(6):1035–1043, 2024.

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

[3] Saar Alon-Barkat and Madalina Busuioc. Human–ai interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1):153–169, 2023.

[4] Clélie Amiot. Cocco dataset and analyses, July 2022.

[5] Clélie Amiot. Cocco 2 dataset and analyses, July 2025.

[6] Ivo Benke, Michael Thomas Knierim, and Alexander Maedche. Chatbot-based emotion management for distributed teams: A participatory design study. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–30, 2020.

[7] Eva AC Bittner, Sarah Oeste-Reiß, and Jan Marco Leimeister. Where is the bot in our team? toward a taxonomy of design option combinations for conversational agents in collaborative work. In *Hawaii International Conference on System Sciences (HICSS)*, 2019.

[8] Petter Bae Brandtzaeg and Asbjørn Følstad. Why people use chatbots. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*, pages 377–392. Springer, 2017.

[9] Frederick P Brooks Jr. *The mythical man-month: essays on software engineering*. Pearson Education, 1995.

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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.

[11] Avyay Casheekar, Archit Lahiri, Kanishk Rath, Kaushik Sanjay Prabhakar, and Kathiravan Srinivasan. A contemporary review on chatbots, ai-powered virtual conversational agents, chatgpt: Applications, open challenges and future research directions. *Computer Science Review*, 52:100632, 2024.

[12] Noah Castelo, Maarten W Bos, and Donald R Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, 2019.

[13] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*, 92:539–548, 2019.

[14] Veronika Cieslak and Carmen Valor. Moving beyond conventional resistance and resistors: an integrative review of employee resistance to digital transformation. *Cogent Business & Management*, 12(1):2442550, 2025.

[15] Jean M Costa, Marcelo Cataldo, and Cleidson R de Souza. The scale and evolution of coordination needs in large-scale distributed projects: implications for the future generation of collaborative tools. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3151–3160, 2011.

[16] Tugrul U Daim, Anita Ha, Shawn Reutiman, Brennan Hughes, Ujjal Pathak, Wayne Bynum, and Ashok Bhatla. Exploring the communication breakdown in global virtual teams. *International journal of project management*, 30(2):199–212, 2012.

[17] Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. Cognitive forcing for better decision-making: Reducing overreliance on ai systems through partial explanations. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–30, 2025.

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

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

[20] Hyo Jin Do, Ha-Kyung Kong, Jaewook Lee, and Brian P Bailey. How should the agent communicate to the group? communication strategies of a conversational agent in group chat discussions. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–23, 2022.

[21] James E Driskell, Eduardo Salas, and Tripp Driskell. Foundations of teamwork and collaboration. *American Psychologist*, 73(4):334, 2018.

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

[23] Yvonne A Farah, Antonio Banuelos-Moriel, and Michael C Dorneich. Evaluating the consistency of cooperative video games in inducing teamwork behaviors. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 67, pages 104–110. SAGE Publications Sage CA: Los Angeles, CA, 2023.

[24] Joel E Fischer, Wenchao Jiang, and Stuart Moran. Atomicorchid: A mixed reality game to investigate coordination in disaster response. In *International Conference on Entertainment Computing*, pages 572–577. Springer, 2012.

[25] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. Future directions for chatbot research: an interdisciplinary research agenda. *Computing*, 103(12):2915–2942, 2021.

[26] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. What makes users trust a chatbot for customer service? an exploratory interview study. In *International conference on internet science*, pages 194–208. Springer, 2018.

[27] Jos Fransen, Paul A Kirschner, and Gijsbert Erkens. Mediating team effectiveness in the context of collaborative learning: The importance of team and task awareness. *Computers in human Behavior*, 27(3):1103–1113, 2011.

[28] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. Will humans-in-the-loop become borgs? merits and pitfalls of working with ai. *Management Information Systems Quarterly (MISQ)-Vol*, 45, 2021.

[29] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. Will humans-in-the-loop become borgs? merits and pitfalls of working with ai. *Management Information Systems Quarterly (MISQ)-Vol*, 45, 2021.

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

[31] David Gundry and Sebastian Deterding. Validity threats in quantitative data collection with games: A narrative survey. *Simulation & Gaming*, 50(3):302–328, 2019.

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

[33] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527, 2011.

[34] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[35] Sarita Herse, Jonathan Vitale, Benjamin Johnston, and Mary-Anne Williams. Using trust to determine user decision making & task outcome during a human-agent collaborative task. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, pages 73–82, 2021.

[36] Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior*, 49:245–250, 2015.

[37] Yoyo Tsung-Yu Hou and Malte F. Jung. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), October 2021.

[38] Stephanie Houde, Kristina Brimijoin, Michael Muller, Steven I Ross, Dario Andres Silva Moran, Gabriel Enrique Gonzalez, Siya Kunde, Morgan A Foreman, and Justin D Weisz. Controlling ai agent participation in group conversations: A human-centered approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 390–408, 2025.

[39] Christopher K Hsee and Jiao Zhang. Distinction bias: misprediction and mischoice due to joint evaluation. *Journal of personality and social psychology*, 86(5):680, 2004.

[40] Olga Iarygina, Kasper Hornbæk, and Aske Mottelson. Demand characteristics in human–computer experiments. *International Journal of Human-Computer Studies*, 193:103379, 2025.

[41] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

[42] Christian P Janssen, Stella F Donker, Duncan P Brumby, and Andrew L Kun. History and future of human-automation interaction. *International journal of human-computer studies*, 131:99–107, 2019.

[43] Laddawan Kaewkitipong. Diffusion of an online collaboration tool: The case of google wave adoption failure. In *2012 45th Hawaii International Conference on System Sciences*, pages 3990–3999. IEEE, 2012.

[44] Christoph Keding and Philip Meissner. 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.

[45] Alice Kerly, Norasnita Ahmad, and Susan Bull. Investigating learner trust in open learner models using a'wizard of oz'approach. In *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings 9*, pages 722–724. Springer, 2008.

[46] Antino Kim, Mochen Yang, and Jingjng Zhang. 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.

[47] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[48] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[49] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26, 2021.

[50] Glen Klein, David D Woods, Jeffrey M Bradshaw, Robert R Hoffman, and Paul J Feltovich. Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95, 2004.

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

[52] Elisa Konya-Baumbach, Miriam Biller, and Sergej von Janda. Someone out there? a study on the social presence of anthropomorphized chatbots. *Computers in Human Behavior*, 139:107513, 2023.

[53] Cordula Kupfer, Rita Prassl, Jürgen Fleiß, Christine Malin, Stefan Thalmann, and Bettina Kubicek. Check the box! how to deal with automation bias in ai-based personnel selection. *Frontiers in Psychology*, 14:1118723, 2023.

[54] Ali Ladak, Matti Wilks, Steve Loughnan, and Jacy Reese Anthis. Robots, chatbots, self-driving cars: Perceptions of mind and morality across artificial intelligences. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2025.

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

[56] Kaisa Laitinen, Salla-Maaria Laaksonen, and Minna Koivula. Slacking with the bot: Programmable social bot in virtual team interaction. *Journal of Computer-Mediated Communication*, 26(6):343–361, 2021.

[57] Carlene Lebeuf, Margaret-Anne Storey, and Alexey Zagalsky. How software developers mitigate collaboration friction with chatbots. *arXiv preprint arXiv:1702.07011*, 2017.

[58] John D Lee. Review of a pivotal human factors article:"humans and automation: use, misuse, disuse, abuse". *Human Factors*, 50(3):404–410, 2008.

[59] Jong-Eun Roselyn Lee and Clifford I Nass. Trust in computers: The computers-are-social-actors (casa) paradigm and trustworthiness perception in human-computer communication. In *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives*, pages 1–15. IGI Global Scientific Publishing, 2010.

[60] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[61] Jiahao Li, Shen Dong, Erin K Chiou, and Jie Xu. Reciprocity and its neurological correlates in human-agent cooperation. *IEEE Transactions on Human-Machine Systems*, 50(5):384–394, 2020.

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

[63] Albert Łukasik and Arkadiusz Gut. From robots to chatbots: unveiling the dynamics of human-ai interaction. *Frontiers in Psychology*, 16:1569277, 2025.

[64] M Lynne Markus and Terry Connolly. Why cscw applications fail: Problems in the adoption of interdependent work tools. In *Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, pages 371–380, 1990.

[65] Thomas W Malone and Kevin Crowston. The interdisciplinary study of coordination. *ACM Computing Surveys (CSUR)*, 26(1):87–119, 1994.

[66] Vincent F Mancuso and Michael McNeese. Teamnets: Scaled world simulation for distributed cyber teams. In *HCI International 2013-Posters' Extended Abstracts: International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part II 15*, pages 509–513. Springer, 2013.

[67] Vincent F Mancuso, Dev Minotra, Nicklaus Giacobe, Michael McNeese, and Michael Tyworth. idsnets: An experimental platform to study situation awareness for intrusion detection analysts. In *2012 IEEE international multi-disciplinary conference on cognitive methods in situation awareness and decision support*, pages 73–79. IEEE, 2012.

[68] Andrew Mao, Winter Mason, Siddharth Suri, and Duncan J Watts. An experimental study of team size and performance on a complex task. *PloS one*, 11(4):e0153048, 2016.

[69] Lenora Marcellus. Are we missing anything? Pursuing research on attrition. *The Canadian journal of nursing research = Revue canadienne de recherche en sciences infirmieres*, 36(3):82—98, September 2004.

[70] Igor Mayer. Assessment of teams in a digital game environment. *Simulation Gaming: An international journal of theory, design and research*, 49(6):602–619, 2018.

[71] Michael D. McNeese, Priya Bains, Isaac Brewer, Cliff Brown, Erik S. Connors, Tyrone Jefferson Jr., Rashaad E.T. Jones Jr., and Ivanna Terrell Jr. The NeoCITIES simulation: Understanding the design and experimental methodology used to develop a team emergency management simulation. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting, HFES 2005*, pages 591–594, December 2005.

[72] Joanna Misiura and Andrej Verity. Chatbots in the humanitarian field: Concepts, uses and shortfalls. Technical report, United Nations Office for the Coordination of Humanitarian Affairs (OCHA), May 2019. Digital Humanitarian Network Report.

[73] Stuart Moran, Nadia Pantidi, Khaled Bachour, Joel E Fischer, Martin Flintham, Tom Rodden, Simon Evans, and Simon Johnson. Team reactions to voiced agent instructions in a pervasive game. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 371–382, 2013.

[74] Sarah Morrison-Smith and Jaime Ruiz. Challenges and barriers in virtual teams: a literature review. *SN Applied Sciences*, 2(6):1096, 2020.

[75] Kathleen L Mosier, Melisa Dunbar, Lori McDonnell, Linda J Skitka, Mark Burdick, and Bonnie Rosenblatt. Automation bias and errors: Are teams better than individuals? *Proceedings of the human factors and ergonomics society annual meeting*, 42(3):201–205, 1998.

[76] Bonnie M Muir. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6):527–539, 1987.

[77] Lea Müller, Jens Mattke, Christian Maier, Tim Weitzel, and Heinrich Graser. Chatbot acceptance: A latent profile analysis on individuals' trust in conversational agents. In *Proceedings of the 2019 on Computers and People Research Conference*, SIGMIS-CPR '19, page 35–42, New York, NY, USA, 2019. Association for Computing Machinery.

[78] Ivan Mutis and Adithya Ramachandran. The bimbot: Mediating technology for enacting coordination in teamwork collaboration. *Journal of Information Technology in Construction*, 26, 2021.

[79] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, 1994.

[80] Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Yong-Yeol Ahn, and Alessandro Flammini. Information overload in group communication: from conversation to cacophony in the twitch chat. *Royal Society open science*, 6(10):191412, 2019.

[81] Birthe Nesset, David A Robb, José Lopes, and Helen Hastie. Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 313–317, 2021.

[82] Margaret Oertig and Thomas Buergi. The challenges of managing cross-cultural virtual project teams. *Team Performance Management: An International Journal*, 12(1/2):23–30, 2006.

[83] Gary M Olson and Judith S Olson. Distance matters. *Human–computer interaction*, 15(2-3):139–178, 2000.

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

[85] Christopher Poile, Andrew Begel, Nachiappan Nagappan, and Lucas Layman. Coordination in large-scale software development: Helpful and unhelpful behaviors. *Dept. of Management Sciences University of Waterloo Waterloo, ON, Canada cpoile@ uwaterloo. ca and Microsoft Research Redmond, WA, fandrew. begel, naching@ microsoft. com. and Dept. of Computer Science, North Carolina State University, Raleigh, NC, lucas. layman@ ncsu. edu*, 2009.

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

[87] John Powell. Trust me, I'm a chatbot: how artificial intelligence in health care fails the turing test. *Journal of medical Internet research*, 21(10):e16222, 2019.

[88] Sarvapali D Ramchurn, Feng Wu, Wenchao Jiang, Joel E Fischer, Steve Reece, Stephen Roberts, Tom Rodden, Chris Greenhalgh, and Nicholas R Jennings. Human–agent collaboration for disaster response. *Autonomous Agents and Multi-Agent Systems*, 30:82–111, 2016.

[89] Andreas J Reuschl, Maximilian K Deist, and Adnane Maalaoui. Digital transformation during a pandemic: Stretching the organizational elasticity. *Journal of Business Research*, 144:1320–1332, 2022.

[90] Tim Rietz, Ivo Benke, and Alexander Maedche. The impact of anthropomorphic and functional chatbot design features in enterprise collaboration systems on user acceptance. In *Proceedings of the 14th International Conference on Wirtschaftsinformatik (WI2019)*, Siegen, Germany, February 2019.

[91] Everett M. Rogers. *Diffusion of Innovations*. Free Press, New York, NY, 5th edition, 2003.

[92] Adriana Salatino, Arthur Prével, Emilie Caspar, and Salvatore Lo Bue. Influence of ai behavior on human moral decisions, agency, and responsibility. *Scientific Reports*, 15(1):12329, 2025.

[93] Shruti Sannon, Brett Stoll, Dominic DiFranzo, Malte F Jung, and Natalya N Bazarova. "i just shared your responses" extending communication privacy management theory to interactions with conversational agents. *Proceedings of the ACM on Human-Computer Interaction*, 4(GROUP):1–18, 2020.

[94] Beau G Schelble, Christopher Flathmann, Nathan J McNeese, Guo Freeman, and Rohit Mallick. 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):1–29, 2022.

[95] Johanna Schmidhuber, Stephan Schlögl, and Christian Ploder. Cognitive load and productivity implications in human-chatbot interaction. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2021.

[96] Isabella Seeber, Eva Bittner, Robert O. Briggs, Gert-Jan de Vreede, Triparna de Vreede, Doug Druckenmiller, Ronald Maier, Alexander B. Merz, Sarah Oeste-Reiß, and Nils Randrup. Machines as teammates: A collaboration research agenda. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, pages 486–495, 2018.

[97] Anna-Maria Seeger, Jella Pfeiffer, and Armin Heinzl. When do we need a human? anthropomorphic design and trustworthiness of conversational agents. In Gabe Lee, editor, *Proceedings of the Sixteenth Annual Pre-ICIS Workshop on HCI Research in MIS : Seoul , Korea , December 10, 2017*, pages Paper 15, 1–6, Atlanta, GA, 2017. AISeL.

[98] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[99] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. It takes a village: integrating an adaptive chatbot into an online gaming community. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020.

[100] Amir Shevat. *Designing bots: Creating conversational experiences*. O'Reilly Media, 2017.

[101] Anuragini Shirish and Leslie Batuekueno. Technology renewal, user resistance, user adoption: status quo bias theory revisited. *Journal of Organizational Change Management*, 34(5):874–893, 2021.

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

[103] Linda J Skitka, Kathleen L Mosier, Mark Burdick, and Bonnie Rosenblatt. Automation bias and errors: are crews better than individuals? *The International journal of aviation psychology*, 10(1):85–97, 2000.

[104] LJ Skitka and KL Mosier. Automation bias: when, where and why. In *Judgment and Decision-Making Society Annual Convention, St. Louis, MO*, 1994.

[105] Marita Skjuve, Petter Bae Brandtzaeg, and Asbjørn Følstad. Why do people use ChatGPT? exploring user motivations for generative conversational AI. *First Monday*, 29(1), 2024.

[106] Jason Snyder and Joo Eng Lee-Partridge. Understanding communication channel choices in team knowledge sharing. *Corporate Communications: An International Journal*, 18(4):417–431, 2013.

[107] Michaela Soellner and Joerg Koenigstorfer. Compliance with medical recommendations depending on the use of artificial intelligence as a diagnostic method. *BMC medical informatics and decision making*, 21:1–11, 2021.

[108] Emanuel Stoeckli, Falk Uebernickel, and Walter Brenner. Exploring affordances of slack integrations and their actualization within enterprises – towards an understanding of how chatbots create value. In *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS)*, pages 2016–2025, 2018.

[109] Stergios Tegos, Stavros Demetriadis, and Anastasios Karakostas. Mentorchat: Introducing a configurable conversational agent as a tool for adaptive online collaboration support. In *2011 15th panhellenic conference on informatics*, pages 13–17. IEEE, 2011.

[110] Stergios Tegos, Stavros Demetriadis, and Thrasyvoulos Tsiatsos. A configurable conversational agent to trigger students' productive dialogue: a pilot study in the call domain. *International Journal of Artificial Intelligence in Education*, 24:62–91, 2014.

[111] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. Understanding chatbot-mediated task management. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–6, 2018.

[112] Douglas P Twitchell, Karl Wiers, Mark Adkins, Judee K Burgoon, and Jay F Nunamaker. Strikecom: A multi-player online strategy game for researching and teaching group dynamics. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 45b–45b. IEEE, 2005.

[113] Diede PM Van der Hoorn, Anouk Neerincx, and Maartje MA de Graaf. " i think you are doing a bad job!" the effect of blame attribution by a robot in human-robot collaboration. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, pages 140–148, 2021.

[114] Jasper Van Der Waa, Sabine Verdult, Karel Van Den Bosch, Jurriaan Van Diggelen, Tjalling Haije, Birgit Van Der Stigchel, and Ioana Cocu. Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI*, 8, 2021.

[115] Siv Vangen and Nik Winchester. Managing cultural diversity in collaborations: A focus on management tensions. *Public Management Review*, 16(5):686–707, 2014.

[116] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 763–777, 2022.

[117] Dakuo Wang, Haoyu Wang, Mo Yu, Zahra Ashktorab, and Ming Tan. Group chat ecology in enterprise instant messaging: How employees collaborate through multi-user chat channels on slack. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–14, 2022.

[118] Weiyu Wang and Keng Siau. Living with artificial intelligence–developing a theory on trust in health chatbots. In *Proceedings of the sixteenth annual pre-ICIS workshop on HCI research in MIS*, pages 1–5. Association for Information Systems San Francisco, CA, 2018.

[119] David D Woods and Emily S Patterson. How unexpected events produce an escalation of cognitive and coordinative demands. *PA Hancock, & PA Desmond, Stress, workload, and fatigue. Mahwah, NJ: L. Erlbaum*, 2001.

[120] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414, 2019.

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

[122] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. Ux research on conversational human-ai interaction: A literature review of the acm digital library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2022.

Abstract

This thesis examines how chatbots can support collaborative work. Their growing integration in professional contexts raises questions about their influence on decision-making, communication, and coordination. Most chatbots target individual interactions, and their potential to enhance collaboration remains largely unexplored. This work focuses on user trust, reliance, and the impact of chatbot assistance on collective performance.

The literature review addresses AI-assisted decision-making, trust, and automation bias. Algorithmic advice is more reliable for objective tasks, while human opinions are preferred for subjective or sensitive issues. Through anthropomorphism and natural language interaction, chatbots increase engagement and social presence, which can make users more trusting and socially influenced. In a collective context, they can structure discussions and balance participation, but their effectiveness depends on both the context and the form of their interventions.

An experiment investigated automation bias with a chatbot compared to a human assistant. Forty-eight participants answered 20 questions about wilderness survival, accompanied by advice and contradictions. Measures included answer changes, message length, and response time, complemented by a post-experiment questionnaire. Results show a strong bias toward the chatbot: participants changed their answers in 63.7% of cases after a chatbot contradiction, versus 31% for the human. Participants assisted by the chatbot requested help more often, wrote shorter messages, and provided fewer justifications. Qualitative feedback indicated that the human was perceived as more adaptable. These results suggest that the mere presence of a chatbot can induce dependency and reduce critical thinking.

To study collaboration, a web platform was developed where teams solved modules by sharing information via group chats and private channels. A chatbot was tested with four-member teams under three conditions: group, private, or both. The chatbot memorized relevant information and returned it on request. Performance was measured by resolution time, response delays, and cognitive load. Results show that integration into the group chat significantly improved performance, reducing median resolution time from 21 to 12 minutes. Responses to information requests were also faster. Private chat was less effective, as it required soliciting help and switching channels, increasing perceived effort. Teams that shared information proactively perceived better performance, although this did not translate into objective success.

This thesis demonstrates that chatbots can enhance collaborative performance and reduce cognitive load while introducing risks of bias and decreased critical engagement. Contributions include: (1) empirically highlighting chatbot-specific automation bias, (2) developing a methodological framework for studying group collaboration, and (3) identifying key factors for effective assistance, notably the mode of integration. Future work

should explore long-term interactions to determine whether familiarity fosters critical thinking or strengthens dependency. Studying larger teams could clarify how chatbot communication strategies should evolve to maintain effectiveness and trust. These results provide theoretical and practical guidance for designing collaborative chatbots capable of supporting decision-making and teamwork.

Keywords: chatbot, collaboration, trust.

Résumé

Cette thèse examine comment les chatbots peuvent soutenir le travail collaboratif. Leur intégration croissante dans les contextes professionnels soulève la question de leur influence sur la prise de décision, la communication et la coordination. La plupart des chatbots ciblent les interactions individuelles, et leur potentiel pour améliorer la collaboration reste peu exploré. Ce travail se concentre sur la confiance, la dépendance des utilisateurs et l'impact de l'assistance des chatbots sur la performance collective.

La revue de littérature porte sur la prise de décision assistée par l'IA, la confiance et le biais d'automatisation. Les conseils algorithmiques sont plus fiables pour les tâches objectives, tandis que les avis humains sont privilégiés pour les questions subjectives ou sensibles. Les chatbots, par leur anthropomorphisme et leur langage naturel, augmentent l'engagement et la réceptivité sociale des utilisateurs. Dans un contexte collectif, ils peuvent structurer les discussions et équilibrer la participation, mais leur efficacité dépend du contexte et de la forme de leurs interventions.

Une expérience a étudié le biais d'automatisation avec un chatbot comparé à un assistant humain. Quarante-huit participants ont répondu à 20 questions sur la survie en pleine nature, accompagnées de conseils et de contradictions. Les mesures incluent les changements de réponse, la longueur des messages et le temps de réaction, complétés par un questionnaire post-expérimental. Les résultats montrent un biais marqué envers le chatbot : les participants ont modifié leurs réponses dans 63,7% des cas après une contradiction du chatbot, contre 31% pour l'humain. Les participants assistés par le chatbot demandent plus souvent de l'aide, rédigent des messages plus courts et fournissent moins de justifications. Les retours qualitatifs indiquent que l'humain est perçu comme plus adaptable. Ces résultats montrent que la simple présence d'un chatbot peut induire une dépendance et réduire l'esprit critique.

Pour étudier la collaboration, une plateforme web a été conçue où des équipes résolvent des modules en partageant des informations via chat de groupe et canaux privés. Un chatbot a été testé avec des équipes de quatre participants dans trois conditions : groupe, privé ou les deux. Le chatbot mémorisait les informations pertinentes et les restituait à la demande. La performance a été mesurée par le temps de résolution, les délais de réponse et la charge cognitive. Les résultats montrent que l'intégration au chat de groupe améliore significativement la performance, réduisant le temps médian de résolution de 21 à 12 minutes. Les réponses aux demandes d'information sont également plus rapides. Le chat privé est moins efficace, car il oblige à solliciter l'aide et à changer de canal, augmentant l'effort perçu. Les équipes partageant l'information de manière proactive perçoivent de meilleures performances, sans que cela se traduise par un succès objectif.

Cette thèse démontre que les chatbots peuvent améliorer la performance collaborative et alléger la charge cognitive, tout en introduisant des risques de biais et de réduction de l'engagement critique. Les contributions incluent : (1) la mise en évidence empirique d'un biais d'automatisation spécifique aux chatbots, (2) le développement d'un cadre méthodologique pour l'étude de la collaboration en groupe, et (3) l'identification des facteurs clés d'une assistance efficace, notamment le mode d'intégration. Il reste à explorer les interactions à long terme pour déterminer si la familiarité favorise l'esprit critique ou

renforce la dépendance. L'étude d'équipes plus larges permettrait de préciser comment les stratégies de communication du chatbot doivent évoluer pour maintenir efficacité et confiance. Ces résultats offrent des orientations théoriques et pratiques pour concevoir des chatbots collaboratifs capables de soutenir la prise de décision et le travail en équipe.

Mots-clés: chatbot, collaboration, confiance.

Résumé long

Introduction

Les chatbots sont devenus plus puissants, accessibles et polyvalents que jamais. Ces outils d'intelligence artificielle permettent des interactions en langage naturel et peuvent maintenir une conversation cohérente au fil de plusieurs échanges. Ils sont désormais intégrés dans de nombreux aspects de la vie quotidienne et professionnelle. Leur adoption a fortement augmenté, dont en France, où un tiers de la population a utilisé des outils d'IA en 2024, avec une adoption particulièrement élevée chez les jeunes adultes. Les chatbots sont employés pour répondre à des questions, fournir des recommandations, divertir, mais également pour des tâches professionnelles comme le support client, l'éducation, la santé ou la gestion administrative.

Dans un contexte collaboratif, les chatbots présentent un potentiel notable : ils peuvent faciliter la communication entre équipes, traduire en temps réel, générer des synthèses et centraliser l'information. Ils permettent de réduire les charges cognitives, d'améliorer le suivi des projets et de tenir les parties prenantes informées. Toutefois, leur intégration n'est pas triviale. L'introduction d'un nouvel outil peut perturber les dynamiques existantes, entraîner de la résistance, et augmenter la charge cognitive de ses utilisateurs. De plus, leur efficacité dépend de l'adoption collective : un usage asymétrique ou isolé peut limiter leurs bénéfices.

La dépendance envers les chatbots représente un autre enjeu critique. Un usage insuffisant empêche de tirer parti de leurs avantages, tandis qu'une surdépendance peut conduire à des décisions biaisées et à une dilution de la responsabilité au sein des équipes. Il est donc essentiel que les utilisateurs les considèrent comme un soutien et non comme un substitut au jugement humain.

Cette thèse examine comment les chatbots peuvent soutenir efficacement la collaboration sans perturber les dynamiques existantes. Elle étudie notamment les facteurs influençant la confiance et la dépendance des utilisateurs vis-à-vis des conseils des chatbots, et compare l'impact des recommandations d'un chatbot à celles d'un assistant humain. Une plateforme expérimentale innovante a également été développée, permettant l'observation des interactions de groupe et de la collaboration virtuelle de manière contrôlée et reproductible. Enfin, un étude utilisant cette plateforme a permis de montrer que la présence de chatbots peut améliorer la performance collective et réduire la charge cognitive individuelle. Cependant, les stratégies de communication du chatbot, qu'il envoie des messages individuels ou publics, influencent différemment la dynamique de l'équipe.

Ces travaux apportent ainsi des contributions méthodologiques et conceptuelles pour la conception de chatbots collaboratifs efficaces, et ouvrent des perspectives pour mieux comprendre leur impact sur la prise de décision et la collaboration en groupe.

Travaux associés

Dans cette thèse, nous nous intéressons à la manière dont les chatbots peuvent soutenir la collaboration de groupe, en particulier à la façon dont ils délivrent des conseils et à

l'impact que cela a sur les dynamiques collectives et la prise de décision. L'objectif est de favoriser un usage appropriée du chatbot : assez convaincant pour ne pas être ignoré, mais pas au point d'éclipser la discussion ou de prendre le contrôle des décisions. De plus, l'introduction d'un chatbot ne doit pas accroître la charge cognitive ou le coût de coordination.

La littérature sur l'assistance par l'IA met en avant les notions de confiance, d'usage approprié et de biais d'automatisation. Les travaux fondateurs ont montré que l'automatisation peut entraîner des erreurs d'omission ou de commission, lorsque les utilisateurs se fient trop ou trop peu aux systèmes. Avec les systèmes modernes, opaques et adaptatifs, ces risques subsistent, et s'ajoutent les phénomènes de préférence pour les recommandations algorithmiques ou, au contraire, de scepticisme selon le contexte. Les études comparant les conseils humains et algorithmiques montrent que les algorithmes sont davantage suivis dans les tâches objectives et quantitatives (estimations, prévisions), mais que dans les domaines subjectifs, sociaux ou sensibles (relations amoureuses, humour, médecine), les utilisateurs font davantage confiance aux humains. Ainsi, l'usage dépend fortement du type de tâche et du contexte de l'assistance.

Une autre question est celle de la calibration de la confiance. Les recherches soulignent que la confiance repose surtout sur les performances et attributs perçus du système, et qu'elle peut être modulée par la transparence, l'auto-critique, ou encore le type de retours fait à l'utilisateur. Le défi est de trouver un équilibre qui évite la sur-confiance comme la sous-utilisation.

Les chatbots, en tant que sous-catégorie d'IA, se distinguent par leur communication en langage naturel et leur intégration facile aux outils existants. Ils génèrent aussi des artefacts d'interactions utiles à la recherche. Leur anthropomorphisme leur confère une forte présence sociale, ce qui peut influencer la confiance et l'engagement. Contrairement aux robots incarnés, ils suscitent moins d'effet de type "vallée de l'étrange". Toutefois, cet aspect humain soulève des enjeux de vie privée : les utilisateurs tendant à moins exiger de confidentialité vis-à-vis d'un chatbot que d'un humain.

La confiance dans les chatbots est influencée par des facteurs internes (pertinence, transparence, humanité) et externes (marque, sécurité, importance de la tâche). La capacité du chatbot à justifier ses conseils est particulièrement cruciale. La personnalité des utilisateurs joue également un rôle, certains profils se montrant plus enclins à la confiance.

Concernant les interactions entre chatbot et groupes, la recherche reste limitée. Les interactions polyadiques présentent des défis spécifiques : savoir quand intervenir, comment structurer les échanges sans perturber le flux, et comment maintenir un juste équilibre entre visibilité et discrétion. Les chatbots intégrés aux outils collaboratifs (Slack, Teams) peuvent aider à gérer les tâches, réduire les frictions, partager l'information et même renforcer la communication socio-émotionnelle. Toutefois, la plupart des études sont descriptives et peu d'expériences contrôlées évaluent leur impact direct sur la performance ou la confiance. Dans le domaine de l'apprentissage collaboratif, les chatbots se sont révélés efficaces pour équilibrer la participation, améliorer la qualité des discussions et renforcer la cohésion, à condition d'adapter la manière dont ils adressent les individus ou le groupe.

En conclusion, si l'assistance par chatbot offre un potentiel considérable pour soutenir

le travail collectif, elle exige un design attentif à la confiance, à la charge cognitive et aux dynamiques sociales. Les travaux futurs devront examiner expérimentalement comment ces spécificités influencent l'usage et la collaboration afin de concevoir des chatbots collectifs à la fois utiles, fiables et bien acceptés.

Comparaison assistant chatbot ou d'humain

Dans une première étude, nous nous intéressons à la manière dont les utilisateurs interagissent avec des chatbots présentés comme coéquipiers, en particulier à travers le prisme du biais d'automatisation. Ce biais décrit la tendance des individus à suivre les recommandations d'un système automatisé, qu'elles soient correctes ou non. Bien que la littérature ait largement étudié ce phénomène pour des algorithmes ou des systèmes d'IA classiques, peu de recherches ont porté sur les chatbots, qui se distinguent par leur interaction conversationnelle et leur plus grande dimension sociale.

Les travaux antérieurs montrent que l'IA peut influencer la prise de décision humaine, parfois en générant un excès de confiance ou une dépendance excessive. Cependant, les comparaisons entre conseils donnés par un humain et ceux donnés par un chatbot restent rares. Nous avons donc souhaité déterminer si le biais d'automatisation apparaît dans les interactions avec des chatbots, en comparant directement l'influence d'un chatbot et d'un conseiller humain, et identifier quels facteurs expliquent ce biais.

L'expérience a été menée auprès de 48 participants, répartis en deux groupes. Chaque participant devait répondre à 20 questions liées à la survie en milieu sauvage (sans réponses uniques ou expertes) dans un environnement de messagerie instantanée. La moitié des questions étaient accompagnées de conseils fournis par un assistant : soit un chatbot (COCCO), soit un humain fictif (Hugo). Afin de contrôler les différences de compétences, les deux assistants ont suivi exactement le même script, via une approche de type magicien d'Oz (l'expérimentateur envoyait manuellement les messages). Les distinctions reposaient uniquement sur la présentation : nom, photo de profil, statut en ligne, vitesse de réponse, etc.

L'expérimentation prévoie quatre modalités : question simple, question avec indice, contradiction de la réponse du participant, contradiction avec indice. Les mesures principales concernent le taux de changement de réponse après contradiction, ainsi que la longueur, le contenu et le temps des messages. Un questionnaire post-expérience a également recueilli la perception des participants (utilité, influence, adaptabilité de l'assistant, performance perçue).

Les résultats montrent que les participants assistés par le chatbot changent leurs réponses 63,7% du temps après contradiction, contre seulement 31% pour ceux suivis par l'assistant humain. Ce biais s'observe même sans indice supplémentaire : la simple contradiction du chatbot suffit à faire changer la moitié des réponses, contre 12,8% pour l'humain. Les participants du groupe humain produisent des messages plus longs, prenent davantage de temps pour répondre, expriment plus souvent leurs doutes, demandent des clarifications et justifient plus volontiers leurs choix. À l'inverse, les participants assistés par le chatbot demandent trois fois plus souvent des indices. Dans le questionnaire, Hugo (l'humain) est perçu comme plus adaptable que COCCO, malgré un script identique. En-

fin, une analyse qualitative des commentaires montre que les participants guidés par le chatbot donnent plus de retours critiques et perçoivent souvent son aide comme pouvant nuire à leurs performances.

Ces résultats confirment l'existence d'un biais d'automatisation envers les chatbots, généré uniquement par leur identité perçue. Les participants leur font plus confiance, mais cette confiance s'accompagne d'une réduction de la pensée critique et d'un moindre sentiment de responsabilité. À l'inverse, face à un conseiller humain, les participants se montrent plus sceptiques et plus réfléchis, engageant un processus de justification et d'argumentation. Cela pose un risque pour l'intégration des chatbots dans les environnements collaboratifs : une adoption rapide peut masquer un manque de collaboration réelle. L'excès de confiance pourrait mener à une dépendance superficielle et fragiliser la qualité des décisions. Une contre-mesure serait de surveiller la fréquence des messages exprimant du doute ou des justifications comme indicateurs d'un engagement critique, en parallèle à l'implémentation de techniques de forçage cognitif pour encourager une réflexion indépendante (par exemple rappeler les limites de l'IA ou introduire des délais de réponse).

Parmi les limites, le rôle unique de l'assistant (poser les questions et conseiller) a pu influencer sa perception. De plus, l'effet de biais de demande (attentes des participants sur ce qui est attendu d'eux par l'expérience) n'est pas à exclure. Enfin, l'expérience portait sur une tâche à faible enjeu, des contextes plus critiques pourraient produire d'autres résultats.

En conclusion, l'étude démontre que les chatbots génèrent un biais d'automatisation fort, en incitant les utilisateurs à suivre leurs conseils plus souvent que ceux d'un humain. Mais cette apparente confiance s'accompagne d'un engagement cognitif réduit. Comprendre ces mécanismes est essentiel pour concevoir des chatbots collaborateurs qui soutiennent la prise de décision humaine sans diminuer la responsabilité ni la diversité des points de vue.

Activité en ligne innovante pour l'étude des interactions de groupe

Alors que les chatbots se multiplient dans des environnements collaboratifs, messageries de groupe, éducation ou prise de décision, leur effet sur les dynamiques de groupe reste peu étudié. La majorité des chatbots sont encore conçus pour des interactions dyadiques, il est donc important de développer des méthodes expérimentales adaptées aux contextes collaboratifs multipersonnes.

L'examen des activités expérimentales existantes révèle que les simulations comme NeoCITIES, StrikeCom, Cargo ou AtomicOrchid, bien qu'utilisées pour étudier la collaboration et l'interaction humain-agent, sont souvent militarisées, compétitives ou à enjeux élevés, ce qui les rend peu adaptées à l'étude de situations collaboratives ordinaires. Les jeux vidéo coopératifs offrent un cadre engageant et mesurable, mais présentent des limites méthodologiques : impossibilité de modifier les règles, difficulté d'intégrer un chatbot et accès restreint aux données.

Pour pallier ces limitations, nous avons développé une activité expérimentale originale. Inspirée du jeu « Keep Talking and Nobody Explodes », cette activité implique des équipes

de participants résolvant 16 modules chacun de manière collaborative, avec communication textuelle uniquement, via un canal de groupe et des canaux privés. Chaque module a une couleur (rouge, orange, jaune, vert, bleu ou violet), un type (boutons, grille, cases à cocher, menu ou champ) et un numéro. Les participants reçoivent des indices qui peuvent être utiles pour eux ou pour leurs coéquipiers, ce qui encourage l'échange d'informations et la coopération.

Les boutons se résolvent en cliquant dans le bon ordre. Les grilles nécessitent de cliquer sur le bon bouton dans une matrice 3x3. Les cases à cocher demandent de collecter les couleurs de modules précis de tous les coéquipiers. Les menus impliquent d'identifier le pseudo du coéquipier possédant un module spécifique. Enfin, les champs requièrent de saisir le numéro d'un module particulier d'un coéquipier. Cette diversité crée différents niveaux de difficulté et de dépendance entre participants, favorisant différentes stratégies de communication.

La plateforme web développée (JavaScript/Flask, base de données SQLite) enregistre chaque interaction et tentative de résolution, tout en offrant une interface de gestion pour paramétrer et suivre les expériences. Des tests pilotes ont permis d'améliorer l'accessibilité, la clarté des instructions et la stabilité technique. L'activité est réalisable en une demi-heure à une heure pour des équipes de quatre, avec un équilibre entre complexité et faisabilité, tout en restant engageante pour les participants.

Cette méthodologie permet de simuler des situations collaboratives contrôlées et de collecter des données complètes sur les interactions humain-humain et humain-chatbot. Elle offre également une base flexible pour adapter l'expérience à différentes tailles d'équipes, configurations de modules ou scénarios expérimentaux, constituant un outil pertinent pour l'étude de la collaboration en groupe dans des contextes variés.

Impact de l'assistance par chatbot sur les tâches collaboratives

Ce chapitre explore l'impact des chatbots sur la collaboration en ligne, à travers l'activité collaborative précedemment présentée. Les équipes, composées de quatre participants, sont chargées de résoudre des modules nécessitant coordination et partage d'informations. L'objectif principal de l'étude est d'étudier comment les groupes interagissent avec un chatbot et l'influence de celui-ci sur la dynamique et la performance des équipes.

Pour cela, nous avons développé COCCO, un chatbot conçu pour assister les participants en mémorisant les informations pertinentes issues des messages échangés et en les restituant sur demande. Le chatbot ne disposait que des informations partagées par les participants, garantissant un comportement proche de celui d'un assistant réel. Trois stratégies de communication ont été testées : COCCO dans le chat de groupe, dans les messages privés, ou dans les deux canaux simultanément. Les participants ont également rempli des questionnaires pré- et post-activité, incluant l'indice de charge de travail subjective NASA-TLX et des évaluations de la performance, de la communication et de la confiance dans l'équipe et dans le chatbot.

L'analyse des performances a montré que l'assistance du chatbot réduisait significativement le temps de résolution des modules et les délais de réponse aux demandes d'informations. Les équipes sans assistance (groupe contrôle) avaient un temps médian

de résolution de module d'environ 21 minutes, contre 12 minutes pour le groupe avec COCCO uniquement dans le chat de groupe. Les réponses aux demandes d'information des coéquipiers étaient également plus rapides, indiquant que le chatbot facilite la fluidité des échanges. Cependant, l'assistance en messages privés est moins efficace, car elle nécessite que les participants sollicitent activement le chatbot et change de canal, augmentant l'effort perçu.

En termes de charge cognitive, aucun effet global significatif sur le score NASA-TLX n'a été observé, mais les participants de la condition chat privé ont perçu l'activité comme plus exigeante. Par ailleurs, une bonne perception de la qualité de communication de COCCO est corrélée à une charge cognitive et une pression temporelle réduites.

L'étude a également examiné la dynamique des équipes. Certaines équipes étaient très organisées, partageant l'information de manière proactive et coordonnant leurs actions, tandis que d'autres travaillaient de manière indépendante. L'organisation des équipes était positivement corrélée avec le nombre de messages relatifs au travail d'équipe et la perception de la performance, mais pas avec les mesures objectives de succès. Les interventions de COCCO n'ont pas directement influencé cette organisation.

En conclusion, l'assistance du chatbot a amélioré la performance et la communication des équipes, notamment dans le chat de groupe, et a réduit la charge cognitive. Les résultats montrent que la visibilité et la stratégie de communication du chatbot influencent fortement son efficacité. Pour l'avenir, il serait pertinent de rendre le chatbot plus proactif, capable de détecter et compléter les informations manquantes, afin de maintenir son utilité et favoriser l'engagement des participants, surtout dans des collaborations à long terme ou avec des équipes plus importantes.

Conclusion

Cette thèse explore le rôle des chatbots dans le travail collaboratif et apporte plusieurs contributions clés. Elle établit empiriquement l'existence du biais d'automatisation pour les interactions avec chatbot, montrant que les utilisateurs se fient davantage aux conseils d'un chatbot qu'à ceux d'un assistant humain, mais de manière plus superficielle. Elle propose également un cadre méthodologique innovant pour étudier la collaboration en groupe facilitée par chatbot, incluant une plateforme expérimentale flexible permettant d'évaluer différentes configurations de tâches et de communication. Enfin, elle identifie les facteurs qui favorisent une assistance efficace, comme l'intégration dans le chat de groupe, la qualité de la communication et l'organisation de l'équipe, tout en soulignant les risques de surdépendance et de diminution de l'engagement critique.

Les résultats ouvrent plusieurs pistes pour la recherche future. Il s'agit notamment d'étudier l'évolution de la collaboration avec le chatbot sur le long terme et de déterminer si la familiarité favorise un engagement critique plutôt qu'une dépendance excessive. Les chatbots pourraient devenir plus proactifs, en adaptant leur comportement à l'engagement cognitif des participants ou en corrigeant les erreurs et assistant les membres d'équipe ayant des difficultés. Enfin, l'étude de l'adaptation du chatbot à des équipes plus grandes et complexes permettra de préciser comment les stratégies de communication et le rôle du chatbot doivent s'ajuster pour soutenir efficacement la collaboration et renforcer la

confiance des utilisateurs.