Designing Cognitive 'Copilots': Augmenting Our Minds Through Prompting Conversational Interactions

Candidate:	Leon Reicherts
Supervisors:	Prof Yvonne Rogers Prof Sam Gilbert
Examiners:	Prof Joel Fischer University of Nottingham
	Prof Nadia Bianchi-Berthouze University College London

Thesis Submitted for the Degree of
Doctor of Philosophy

University College London

Division of Psychology and Language Sciences

January 2024

Abstract

Natural language user interfaces (NLUIs) have become increasingly popular for a range of tasks and activities, given the interactive, adaptive, and contextual interactions they can offer. While the goal of many NLUIs is to make it easier for the user to request, find, or generate information or content to complete certain tasks more efficiently, it is argued here that they can also be effective at scaffolding people's thinking while performing cognitive tasks. It is hypothesised that the way this can be achieved is by proactively asking people questions to engage them in reflective thinking about a task at hand. The focus in this thesis is on open-ended tasks that benefit from such reflective thinking, as it can enable people to discover alternative perspectives, approaches, and possibilities which can help to progress with the task. The NLUIs presented in this thesis are embedded into the interfaces used to perform different types of cognitive tasks and were thus named 'cognitive co-pilots'. The tasks included a collaborative exploratory data analysis task, a complex decision-making task, and a creative three-dimensional drawing task. This PhD research explored the opportunities, as well as the challenges of 'embedding' such prompting co-pilots into these types of cognitive activities. It also examines the design parameters at the interface, such as the adequate timing, phrasing, and delivery of prompts – and how they can affect ongoing activities. The thesis reports five studies conducted on different types of proactive NLUIs and provides a novel way of conceptualising them in terms of how they can extend human cognition.

Impact Statement

This PhD research was concerned with designing Natural Language User Interfaces (NLUIs) that proactively ask questions to support human cognition. So far, this research has had an impact through academic dissemination, in particular within the Human-Computer Interaction (HCI) community at venues including the *CHI Conference on Human Factors in Computing* Systems and the *ACM Conference on Conversational User Interfaces*, or the *ACM Transactions on Computer-Human Interaction* journal.

The main contributions of this PhD research are (a) a novel way to design and conceptualise software user interfaces which support human cognition by asking the user task-specific questions through natural language, (b) a set of prototypes that were designed around this idea, and (c) empirical evidence for the ways in which these prototypes can support human cognition and in particular reflective thinking.

The studies in this thesis show that the question-asking NLUIs can be particularly effective for tasks that are open-ended[1], yet for which there are best practices, useful heuristics, techniques, or other considerations that the NLUI can proactively make a user aware of through 'reflective' questions. Such NLUIs can be most impactful for tasks and activities where people are known to frequently get stuck, make errors, forget important aspects, make biased decisions, or face any other difficulties which the NLUI can help mitigate or overcome. Examples of such tasks which were chosen in this thesis are *exploratory data analysis,* where people can benefit from being asked questions about certain patterns that can get overlooked; *complex decision-making,* where a range of aspects need to be considered and may be missed; or *creative tasks,* where it can be difficult to know how to express an idea or abstract thought without guidance. There are many other open-ended tasks with similar characteristics for which similar NLUIs could be designed, and which could be informed by the approaches, design considerations, and evidence reported in this thesis.

[1] Meaning that there is not a specific outcome that needs to be achieved nor a specific way in which it should be achieved

The last set of studies reported in this thesis also provide insights into how proactive NLUIs in the form of voice assistants could be designed for everyday settings, which can be particularly relevant for designing future voice assistants, smart speakers, or robots (Chapter 7). As NLUIs and the devices they are embedded into (e.g. smart home devices) are becoming increasingly more capable and aware of their surroundings, they will also improve their abilities to proactively intervene in certain situations. The reported studies provide evidence for people's perceptions of different types of proactive interventions in a range of settings, showing that some of them are more acceptable than others and that there are marked differences between individuals' attitudes.

In the near future, the contributions of this PhD may be of particular relevance given the increasing capabilities of AI, and specifically generative AI. As many of these tools promise to reduce the cognitive effort of many tasks – which has its benefits – they should ideally also be designed in a way that they support a person's thinking. The vision that this PhD subscribes to is that the next generations of AI tools are designed to support humans in the tasks and activities they perform with the aim to *augment* rather than *replace* human thinking. This thesis proposes that this can be achieved by designing AI tools that ask humans reflective questions about the task which they want to perform with AI, enabling them to structure and make sense of the task and its components to empower their own thinking.

Acknowledgements

As I am writing the closing words of my thesis and look back at the journey that brought me here, I am filled with deep gratitude – to everyone who I had the chance to work with, who supported me, and who contributed to this 'adventure'. This PhD thesis would not have been possible without you.

First and foremost, I would like to express my deepest appreciation to my supervisor, Yvonne Rogers, for her unwavering support throughout this journey. You inspired, challenged, and guided me – which made this an invaluable experience of learning and growth. I am so fortunate to have had you as my supervisor!

My gratitude extends to my second supervisor, Sam Gilbert. Your ideas, feedback, and guidance were truly enriching – and they helped me expand and approach my research in new ways. I am also thankful for all the thought-provoking and inspiring conversations we had together with Yvonne and Ava in our 'extended mind' book club.

Further, I would like to thank Mirco Musolesi, Hugo Spiers, and Gabriella Vigliocco for the great experience I had as part of the Ecological Brain DTP. In particular, I would like to thank Hugo Spiers for an enriching and insightful 'rotation' in his lab, which gave me valuable insights into cognitive neuroscience while working on a fascinating project on real-world navigation. My gratitude further extends to Mirco Musolesi for an equally insightful rotation that introduced me to the 'world' of *data infrastructures*, certain technologies and architectures that they involve, as well as their societal implications. I really enjoyed the conversations with you and Didem Özkul as part of this project. Furthermore, I would also like to thank you, Mirco, for your subsequent guidance and feedback during the first part of my PhD, which helped me refine the direction of my research.

A crucial part of my PhD journey has been all the enriching collaborations which I am deeply thankful for – and of which a great number were part of the Excellence Chair Program with the University of Bremen. My sincere appreciation goes to the fantastic collaboration with Nadine Wagener on the SelVReflect project (the third study in this PhD thesis). Working with you was such a joy – from the inception of the study to the moment when we both presented the paper at CHI 2023 in Hamburg. Furthermore, I would like to express my gratitude for the great collaboration with Nima Zargham on a series of studies we did on (proactive) voice assistants together with Michael Bonfert and Vino Avanesi. My sincere thanks go to the three of you for making these joint projects so enriching! Further, I would like to thank Evropi Stefanidi, Carolin Stellmacher, Thomas Mildner, Natalia Bartłomiejczyk, and Marit Bentvelzen for the delightful collaborations on various further projects and publications as part of the Excellence Chair Program. I had such a great time working with you – and I truly appreciate everything you did! Moreover, I also want to express my gratitude to Jasmin Niess and Paweł Woźniak for your feedback and support on research methods, data analyses, and reviewing the manuscripts across various Excellence Chair projects. Finally, I would like to thank Johannes Schöning and Rainer Malaka for all your support and efforts to organise gatherings and shared events which initiated and facilitated this whole series of collaborations as part of the Excellence Chair Program. I am also thankful for your input and feedback on study ideas and designs and for reviewing the resulting manuscripts across a range of projects.

My thanks further extend to all my colleagues at UCLIC – I could not think of a better group of people to be surrounded by or of a better place to work, learn, and grow! First and foremost, I would like to express my gratitude to Ava Scott for being the best 'ally' I could wish for to share my PhD journey with, for all the projects and conversations – related to and beyond our PhDs – exploring new territory and developing ideas, and for all your valuable input and feedback. I also sincerely thank you, Tu Dinh Duong, for your unwavering support, the enriching collaborations, and the many other things we have shared along the way since we got to know each other in 2017 when we both did the Human-Computer Interaction MSc! Furthermore, I am also very grateful for the great conversations, collaborations, and the help I received from Sheena Visram, Maryam Bandukda, Elahi Hossain, and Katherine Wang as part of various teaching and research efforts.

Among the colleagues from UCLIC, I would also express my special gratitude to Aneesha Singh, Nicolai Marquardt, Catherine Holloway, Anna Cox, Enrico Costanza, Amid Ayobi, and Martin Dechant for all the fruitful collaborations across teaching and research efforts, as well as for your exceptional support and mentorship. A special word of thanks also goes to Louise

Gaynor, who always provided incredible help with any administrative task – no matter how complicated! Finally, I would also like to thank Licia Capra and Neil Sebire for your support and feedback on my first study – it was a pleasure to work with you.

Further, I sincerely thank Ethan Wood for your contributions to preparing, running, and analysing my first study – your efforts were greatly appreciated! Moreover, I would also like to thank Warren Park for your support with building the prototype of the second study and with the data analysis – your contributions were truly invaluable!

My profound gratitude also extends to Tigmanshu Bhatnagar, for being such a great friend, the best flatmate, and a true companion in this PhD adventure. I am deeply grateful for all that we could share, all that I learned from you, and all the fun we had along the way!

I am, of course, also immensely thankful to both my parents for your guidance, support, encouragement, and for sharing my excitement for this journey. Furthermore, I would like to express my sincere appreciation to you, Papa, for your ideas and thoughts on my research projects, as well as the countless conversations about study designs, research methods, and relevant literature!

Special recognition also goes to all my other valued colleagues, friends, and family members who contributed to, inspired, and enriched this journey!

My sincere appreciation also goes to the participants who took part in the studies I conducted – you have been an integral part of this research! And finally, I would also like to acknowledge the Leverhulme Trust for funding my PhD.

My gratitude extends to all of you.

Publications

My research has been disseminated through the publications presented below.

Full Papers

FP4 Wagener, N., **Reicherts, L.,** Zargham, N., Bartłomiejczyk, N., Scott, A., Wang, K., Bentvelzen, M., Stefanidi, E., Mildner, T., Rogers, Y., Niess, J., (2023) SelVReflect: A Guided VR Experience Fostering Reflection on Personal Challenges. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,* 1–21. https://doi.org/10.1145/3544548.3580763

FP3 Zargham, N., **Reicherts, L.,** Bonfert, M., Voelkel, S. T., Schoening, J., Malaka, R., & Rogers, Y. (2022). Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–14. https://doi.org/10.1145/3543829.3543834

FP2 **Reicherts, L.,** Park, G. W., & Rogers, Y. (2022). Extending Chatbots to Probe Users: Enhancing Complex Decision-Making Through Probing Conversations. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–10. https://doi.org/10.1145/3543829.3543832

FP1 **Reicherts, L.,** Rogers, Y., Capra, L., Wood, E., Duong, T. D., & Sebire, N. (2022). It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. *ACM Transactions on Computer-Human Interaction*, *29*(3), 25:1-25:41. https://doi.org/10.1145/3484221

Short Papers

SP **Reicherts, L.,** Zargham, N., Bonfert, M., Rogers, Y., & Malaka, R. (2021). May I Interrupt? Diverging Opinions on Proactive Smart Speakers. *Proceedings of the 3rd Conference on Conversational User Interfaces*, 1–10. https://doi.org/10.1145/3469595.3469629

Further Publications not Included in this PhD Thesis

Below is a list of papers that were published during the course of this PhD (2019-2023), but which are not included in this PhD thesis.

Full Papers

Zargham, N., **Reicherts, L.**, Avanesi, V., Rogers, Y., & Malaka, R. (2023). Tickling Proactivity: Exploring the Use of Humor in Proactive Voice Assistants. *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*, 294–320. https://doi.org/10.1145/3626705.3627777

Short Papers

Avanesi, V., Rockstroh, J., Mildner, T., Zargham, N., **Reicherts, L.,** Friehs, M., Kontogiorgos, D., Wenig, N., Malaka, R., (2023) From C-3PO to HAL: Opening The Discourse About The Dark Side of Multi-Modal Social Agents. *Proceedings of the 5th Conference on Conversational User Interfaces*, 1-7. https://doi.org/10.1145/3571884.3597441

Zargham, N., Avanesi, V., **Reicherts, L.,** Scott, A., Rogers, Y., Malaka, R., (2023). "Funny How?" A Serious Look at Humor in Conversational Agents. *Proceedings of the 5th Conference on Conversational User Interfaces*, 1-6. https://doi.org/10.1145/3571884.3603761

Wagener, N., Stefanidi, E. and **Reicherts, L.** 2023. *Supporting Collaborative Reflection for Teenagers Through Shared Emotional Expression in Virtual Reality.* In CHI 2023 Workshop: Integrating Individual and Social Contexts into Self-Reflection Technologies. https://doi.org/10.13140/RG.2.2.10500.50564 (Workshop paper)

Reicherts, L., & Rogers, Y. (2020). Do Make me Think! How CUIs Can Support Cognitive Processes. *Proceedings of the 2nd Conference on Conversational User Interfaces,* 1–4. https://doi.org/10.1145/3405755.3406157

Bouwman, T., & **Reicherts, L.** (2020). Manypulo: A Flexible System Facilitating the Creation of Interactive Physical Prototypes. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–6. https://doi.org/10.1145/3334480.3383028

UCL Research Paper Declaration Form for Full Paper 1 [FP1]

1. For a research manuscript that has already been published:
   (a) What is the title of the manuscript?
   It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks
   (b) Please include a link to or DOI for the work:
   https://doi.org/10.1145/3484221
   (c) Where was the work published?
   ACM Transactions on Computer-Human Interaction, 29(3)
   (d) Who published the work?
   Association for Computing Machinery (ACM)
   (e) When was the work published?
   January 2022
   (f) List the manuscript's authors in the order they appear on the publication:
   Leon Reicherts, *Yvonne Rogers, Licia Capra, Ethan Wood, Tu Dinh Duong, Neil Sebire*
   (g) Was the work peer reviewed?
   Yes
   (h) Have you retained the copyright?
   No
   (i) Was an earlier form of the manuscript uploaded to a preprint server?
   No
   (j) If 'No', please seek permission from the relevant publisher and check the box next to the below statement:
   ☑ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. For a research manuscript prepared for publication but that has not yet been published: *[Excluded as not applicable.]*

3. For multi-authored work, please give a statement of contribution covering all authors:

1. author:	***Conceptualised, prepared, conducted, and wrote-up the study***
2. author:	*Assisted with the conceptualisation, preparation, and write-up of the study*
3. author:	*Assisted with the study conceptualisation and preparation and reviewed the manuscript*
4. author:	*Assisted with the preparation of the study materials, conducting the study, and analysing the data*

5. author: *Assisted with building the study materials (in particular with building the prototype used for the study)*

6. author: *Assisted with the study preparation and reviewed the manuscript*

4. In which chapter(s) of your thesis can this material be found?
Chapter 4, Chapter 8

5. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

Candidate:

Date: 19/01/2024

Supervisor/Senior Author signature:

Date: 14/01/2024

UCL Research Paper Declaration Form for Full Paper 2 [FP2]

1. For a research manuscript that has already been published:
   - (a) What is the title of the manuscript?
     Extending Chatbots to Probe Users: Enhancing Complex Decision-Making Through Probing Conversations.
   - (b) Please include a link to or DOI for the work:
     https://doi.org/10.1145/3543829.3543832
   - (c) Where was the work published?
     Proceedings of the 4th Conference on Conversational User Interfaces
   - (d) Who published the work?
     Association for Computing Machinery (ACM)
   - (e) When was the work published?
     September 2022
   - (f) List the manuscript's authors in the order they appear on the publication:
     ***Leon Reicherts**, Gun Woo (Warren) Park, Yvonne Rogers*
   - (g) Was the work peer reviewed?
     Yes
   - (h) Have you retained the copyright?
     No
   - (i) Was an earlier form of the manuscript uploaded to a preprint server?
     No
   - (j) If 'No', please seek permission from the relevant publisher and check the box next to the below statement:
     ☑ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. For a research manuscript prepared for publication but that has not yet been published: *[Excluded as not applicable.]*

3. For multi-authored work, please give a statement of contribution covering all authors:

   1. author: ***Conceptualised, prepared, conducted, and wrote-up the study***

   2. author: *Assisted with the development of the study prototype and the data analysis*

   3. author: *Assisted with the conceptualisation, preparation, and write-up of the study*

4. In which chapter(s) of your thesis can this material be found?
   Chapter 5

5. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

   Candidate:

   Date: 19/01/2024

   Supervisor/Senior Author signature:

   Date: 14/01/2024

UCL Research Paper Declaration Form for Full Paper 3 [FP3]

1. For a research manuscript that has already been published:
   (a) What is the title of the manuscript?
   Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma
   (b) Please include a link to or DOI for the work:
   https://doi.org/10.1145/3543829.3543834
   (c) Where was the work published?
   Proceedings of the 4th Conference on Conversational User Interfaces
   (d) Who published the work?
   Association for Computing Machinery (ACM)
   (e) When was the work published?
   September 2022
   (f) List the manuscript's authors in the order they appear on the publication:
   *Nima Zargham, **Leon Reicherts,** Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, Yvonne Rogers*
   (g) Was the work peer reviewed?
   Yes
   (h) Have you retained the copyright?
   No
   (i) Was an earlier form of the manuscript uploaded to a preprint server?
   No
   (j) If 'No', please seek permission from the relevant publisher and check the box next to the below statement:
   ☑ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. For a research manuscript prepared for publication but that has not yet been published: *[Excluded as not applicable.]*

3. For multi-authored work, please give a statement of contribution covering all authors:

1. author:	*Led the conceptualisation, preparation, execution, data analysis, and write-up the study*
2. author:	***Contributed to conceptualisation, preparation, execution, data analysis, and write-up of the study – in particular the write-up of the findings included in this thesis***
3. author:	*Contributed to conceptualisation, preparation, data analysis, and write-up of the study*

4. author: *Gave feedback on the study preparation and reviewed the manuscript*

5. author: *Gave feedback on the study preparation and reviewed the manuscript*

6. author: *Gave feedback on the study preparation and reviewed the manuscript*

7. author: *Gave feedback on the study preparation and reviewed the manuscript*

4. In which chapter(s) of your thesis can this material be found?
   Chapter 7

5. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

   Candidate:

   Date: 19/01/2024

   Supervisor/Senior Author signature:

   Date: 14/01/2024

UCL Research Paper Declaration Form for Full Paper 4 [FP4]

1. For a research manuscript that has already been published:
   (a) What is the title of the manuscript?
   SelVReflect: A Guided VR Experience Fostering Reflection on Personal Challenges
   (b) Please include a link to or DOI for the work:
   https://doi.org/10.1145/3544548.3580763
   (c) Where was the work published?
   Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems
   (d) Who published the work?
   Association for Computing Machinery (ACM)
   (e) When was the work published?
   January 2022
   (f) List the manuscript's authors in the order they appear on the publication:
   Nadine Wagener, ***Leon Reicherts****, Nima Zargham, Natalia Bartłomiejczyk, Ava Scott, Katherine Wang, Marit Bentvelzen, Evropi Stefanidi, Thomas Mildner, Yvonne Rogers, Jasmin Niess*
   (g) Was the work peer reviewed?
   Yes
   (h) Have you retained the copyright?
   No
   (i) Was an earlier form of the manuscript uploaded to a preprint server?
   No
   (j) If 'No', please seek permission from the relevant publisher and check the box next to the below statement:
   ☑ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. For a research manuscript prepared for publication but that has not yet been published: *[Excluded as not applicable.]*

3. For multi-authored work, please give a statement of contribution covering all authors:

   1. author: *Co-led the conceptualisation, preparation, execution, data analysis, and write-up of the study*

   2. author: ***Co-led the conceptualisation, preparation, execution, data analysis, and write-up of the study – in particular the write-up of the findings included in this thesis***

3. author: *Contributed to parts of the literature review, helped with the preparation of the study, ran one focus group as part of the design process, contributed to the coding of the qualitative data, and reviewed the manuscript*

4. author: *Assisted with the coding of the qualitative data, creation of figures, reviewed the manuscript*

5. author: *Helped with the preparation of the study and the design of the prototype, contributed to the coding of the qualitative data, and reviewed the manuscript*

6. author: *Helped with pilot study and testing of the prototype and reviewed the manuscript*

7. author: *Contributed to parts of the literature review, helped with the qualitative data analysis, and reviewed the manuscript*

8. author: *Helped with the preparation of the study and reviewed the manuscript*

9. author: *Helped with the preparation of the study and contributed to the creation of figures and the coding of the qualitative data*

10. author: *Assisted with the preparation of the study and reviewed the manuscript*

11. author: *Assisted with the preparation of the study and qualitative analysis of the data and reviewed the manuscript*

4. In which chapter(s) of your thesis can this material be found?
   Chapter 6

5. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

   Candidate:

   Date: 19/01/2024

   Supervisor/Senior Author signature:

   Date: 14/01/2024

UCL Research Paper Declaration Form for Full Paper 4 [FP4]

1. For a research manuscript that has already been published:
   (a) What is the title of the manuscript?
   May I Interrupt? Diverging Opinions on Proactive Smart Speakers
   (b) Please include a link to or DOI for the work:
   https://doi.org/10.1145/3469595.3469629
   (c) Where was the work published?
   Proceedings of the 3rd Conference on Conversational User Interfaces
   (d) Who published the work?
   Association for Computing Machinery (ACM)
   (e) When was the work published?
   September 2022
   (f) List the manuscript's authors in the order they appear on the publication:
   Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, Rainer Malaka
   (g) Was the work peer reviewed?
   Yes
   (h) Have you retained the copyright?
   No
   (i) Was an earlier form of the manuscript uploaded to a preprint server?
   No
   (j) If 'No', please seek permission from the relevant publisher and check the box next to the below statement:
   ☑ *I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.*

2. For a research manuscript prepared for publication but that has not yet been published: *[Excluded as not applicable.]*

3. For multi-authored work, please give a statement of contribution covering all authors:

1. author:	***Co-led the conceptualisation, preparation, execution, data analysis, and write-up of the study***
2. author:	*Co-led the conceptualisation, preparation, execution, data analysis, and write-up of the study*
3. author:	*Co-led the conceptualisation, preparation, execution, data analysis, and write-up of the study*
4. author:	*Assisted with the preparation of the study and reviewed the manuscript*
5. author:	*Assisted with the preparation of the study and reviewed the manuscript*

4. In which chapter(s) of your thesis can this material be found?
   Chapter 7

5. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work):

   Candidate:

   Date: 19/01/2024

   Supervisor/Senior Author signature:

   Date: 14/01/2024

Declaration

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:

Name: Leon Reicherts

Date: 19/01/2024

Abbreviations

CA	Conversational Agent
CUI	Conversational User Interface
GenAI	Generative AI
GUI	Graphical User Interface
HCI	Human-Computer Interaction
LLM	Large Language Model
NLP	Natural Language Processing
NLUI	Natural Language User Interface
UI	User Interface
VA	Voice Assistant
VUI	Voice User Interface

Glossary

On the one hand, this glossary serves the purpose of introducing some of the domain-specific and less commonly used terms that appear in this thesis. On the other, it also intends to clarify how some familiar terms may be used (and interpreted) in a *slightly different* or *less common* way in this thesis (including *'reflection'*, for example). Furthermore, it also aims to point out relationships between some of the terms (e.g. between *'reflection'* and *'ill-defined problems'*).

Chatbot Referring to a natural language user interface which the user interacts and takes turns with through written text (e.g. question-answer turns).

(AI) Copilot Referring to AI tools released in recent years that are embedded into various applications and platforms with the aim to assist users in their activities and tasks. These tools use the metaphor of a *co-pilot* (which is usually spelled with a hyphen) to emphasise their intended role: supporting the user (i.e. the 'pilot' or 'captain') while not diminishing the user's agency, control, and their responsibility in deciding how to proceed with the specific task (i.e. having the main responsibility for as well as control over of the aircraft). The names that are given to these tools often tend to omit the hyphen in *'co-pilot'* (e.g., GitHub Copilot).

Externalisation and external representations Externalisation is understood as the *process of expressing* thoughts and ideas through external means and resources (e.g. pen and paper, a digital or analogue canvas/whiteboard, a document editor) resulting in *external representations* of these thoughts and ideas (e.g. sketches, diagrams, notes) which can then be manipulated, transformed, and revised – supporting the (collaborative) thought process as it evolves.

External cognition Understood as the *use of external tools or representations*, like notes, diagrams, or other objects, which can make information easier to access, manipulate, understand, make sense of, and to 'work with' as part of a cognitive process.

Ill-defined task/problem A task lacking clear goals, constraints, or criteria, making it ambiguous. Reflective thinking (which this PhD thesis focuses on) can help clarify objectives and develop effective strategies to proceed with the task.

Open-ended task/problem A task that often has broad goals and multiple possible solutions and approaches, thus often requiring some creativity to figure out how to proceed. Reflective thinking, among other things, can help explore diverse approaches and perspectives when progressing with the task.

Prompting NLUI Used to refer to most of the NLUIs designed and studied in this thesis. Sometimes used interchangeably with *proactive NLUI*, *question-asking NLUI*, or *probing NLUI,* depending on which aspect/characteristic a specific chapter/section focuses on.

Prompt Referring to the prompts provided by the NLUI, which were all delivered in a proactive way and generally in the form of questions. Sometimes, they are also referred to as *proactive prompts, guiding prompts, question prompts, probing questions,* or also *cognitive scaffolds* – depending on which aspect/characteristic is emphasised in a specific chapter/section.

Reflection and reflective thinking Both terms are generally used interchangeably in this thesis. Generally used to refer to the 'deep thinking' involved in trying to make sense of a subject (also referred to as critical thinking) or of oneself (also referred to as introspection or self-reflection).

Sensemaking The process of attributing meaning to things people observe or experience (e.g. through reflective thinking).

Cognitive scaffolds Techniques that guide and support people in sensemaking and problem-solving processes – here often in the form of question-based prompts. They aim to help structure thought processes, promote deeper reflection, and support critical thinking.

Software tool Here referring to any type of software used by humans to perform certain types of tasks (e.g. to analyse data, draw, write etc.).

User Generally referring to a person using/interacting with a specific interface, system, or device. However, in this thesis 'person' or 'people' are often used interchangeably with 'user' when there is no specific emphasis on system use or interaction.
(Also note that there is a difference to *'participant(s)'* which is generally used to refer to a group of people who took part in a specific study.)

List of Figures

List of Tables

Table of Contents

1. Introduction

As individuals living in 'information societies' [439], the number of things we do in our everyday lives that rely on the use of technology is ever-increasing. Many of us manage our tasks and our time, communicate with each other, engage with a plethora of content, find our way from one place to another, make decisions, purchase things, express ourselves and create – all through the use of different types of devices, services, and apps. It is thus not surprising that as part of our increasing reliance on technology, many of our cognitive activities have been 'taken on' by technology, or using more scientific parlance, have been 'offloaded' to technology (see Risko and Gilbert [346] and Rogers [351]). A pertinent example of this *cognitive offloading* are digital reminders [154], as they remove the need for us to memorise and subsequently recall what we intend to do at a specific time. As a result of this, the human mind can be seen as being increasingly *extended* with or by technology [87] – or even in a 'symbiosis' with it [125, 258] – for a wide range of things that we do in our private and professional lives.

Another way in which our minds have been extended through technology is by letting us *externalise* our cognitive processes by creating and/or interacting with external representations (e.g. Rogers and Scaife [362]), which can support cognitive processes and help us progress with our tasks [112, 113, 224, 291]. This applies to interfaces that let us create, adapt, and manipulate representations in real time while we are thinking (e.g. Zhang and Norman [469]). An early example for this is the transition from typewriters to word processors starting in the 1970s. For most people nowadays, it might be hard to imagine a world in which it is not possible to directly manipulate a document and see the changes in real time while working on it – moving sections around and iteratively refining them. However, until the 1970s, this was something most people could only dream of – among them Douglas Engelbart [125] with his ideas on how human intellect might be augmented with future technologies that would allow people to create, rearrange, and manipulate various types of representations – like text, 2D or 3D graphics, simulated environments, etc. – in *real time*. This is, of course, just one of countless examples of how technological developments enable 'external thinking' with digital tools and resources – and how they have been giving our 'extended minds' ever more capabilities.

Besides the technological tools that help us *externalise* our thinking, another 'mind extension' has come from technologies that provide us with the relevant kind of data, information, or content – whenever we might need it. One of the most prominent examples of this is, of course, the Internet [391], on which we can find just about anything with the help of search engines – allowing us to explore specific topics, discover and learn new things, solve problems, make more informed decisions about our health, finances, intended purchases, and so on. Through the use of recommender systems and machine learning, the information we receive has also become increasingly personalised and context-specific; and more recently, with generative AI (GenAI), we can even receive a wide range of content as a conversational response to nearly any request or question that we may have on our mind – overall enabling us to get what we want even faster, in the desired form and modality, and tailored to our interests and our needs in the specific context.

These are just a few of the many ways in which technologies have been designed to not only extend what we can do and how we can 'think with' them – but also *how efficiently* we can do so. Over time, this has enabled continuous increases in our productivity, which has been one of the key driving forces for humanity's progress and welfare (e.g. [120, 294, 402]). However, with the growing capabilities of all these technologies, one question is, if some of them could take more thinking 'away from us' than desirable (e.g. [377]), which might be the case, for example, when a certain technology risks to reduce our own engagement, exploration, and sensemaking of a topic. As we proceed with developing these technologies further, it is thus important to consider how they can be designed so that they remain *tools* that can extend our human abilities, which *at the same time* maintain and support our autonomy, agency, and (cognitive) engagement with whatever we do (see also Schmidt et al. [364, 366]); but what might be the most promising ways to achieve this?

1.1 Envisioning Technologies that Make Us Think Better

A long-standing vision for further augmenting our cognitive capabilities has been to develop so-called *natural language user interfaces (NLUIs)* – which are also at the centre of this PhD – that we can talk to and that talk to us to help us achieve our goals. One prominent example for this vision are the Apple 'Knowledge Navigator' and 'Project 2000' concept videos[2] released at the end of the 1980s to showcase Apple's vision of a future computer that one can talk to in order to perform a wide range of everyday and work tasks (see for a description [64, 90]). In recent years, considerable advances have been made towards this vision – particularly with the rise of GenAI. As part of these developments, NLUIs have not only become one of the main ways of interacting with the plethora of AI models that have become available – offering users a familiar 'chat-based' interaction – but they are also increasingly being embedded into a wide range of software applications and services to allow people an easy way to access and use these different 'AIs', whenever they might need them.

One popular term that has emerged for these GenAI-based 'embedded' NLUIs are so-called 'co-pilots' – examples are *GitHub Copilot*, *Microsoft Copilot*, or *Einstein Copilot* from Salesforce, among others. This *'co-pilot'* framing has introduced a new metaphor to the diversity of metaphors that have already been proposed over time for NLUIs, AI tools, and other technologies, such as *assistants*, *advisors*, or *companions*, among others (see also [302, 316]). The idea of this co-pilot framing is to underline the vision behind these new tools, which is to *contribute* to what users are doing/working on (e.g. by providing specific content) *without undermining the user's agency and control* over what they are doing (as also expected from a real co-pilot) [377]. However, the increasing availability and capability of these AI tools to take on tasks that – until recently – could only be done by humans also raises the question of **how the vision of a co-pilot that *augments rather than replaces* human cognition can be best achieved – *also in the future*.**

Steve Krug's well-known book 'Don't Make Me Think' [237] has an ironic title that epitomises the desire to increase efficiency, speed, and ease of use of most software tools and their interfaces – often to remove as much friction from the interaction as possible – which is also one of the main goals of the current AI co-pilots (e.g. [307]). Aiming for such frictionless interactions has – and will continue to have – advantages in many cases where efficiency is key. However, the question asked here is what opportunities are there to design interfaces that *make us think*? For example, by adding a certain amount of friction where this might be

[2] https://www.fastcompany.com/90913458/apples-1987-knowledge-navigator-video-depicted-a-future-thats-still-a-work-in-progress

beneficial [93, 147:141], getting us to slow down, step back, and think more systematically or deeply about something when we might be missing something or following 'automatic' or too constrained thought processes or inadequate heuristics (e.g. Tversky and Kahneman [419]), or when we might benefit of an *alternative* heuristic (e.g. [152, 153, 296, 411]), approach, or perspective for/on something. These are just a few ways in which interfaces could *make us think*, but what kind of thinking should they 'target' to achieve this?

1.2 Augmenting Cognition by Supporting Reflective Thinking

The focus in this thesis is on '*reflective thinking*' or just '*reflection*' – which can be defined as "*thinking carefully and deeply about something*" [249] and which is involved in and beneficial to a wide range of tasks and activities – in particular, if there is not a clear way in which they need to be performed. Reflection can include thinking about the reasons that led to a specific event, (historical) development, or a (personal) decision, as well as what approaches there might be to deal with a (personal) problem someone is facing – just to name a few. A common way in which two different types of reflection are often distinguished, is if someone reflects on themselves (sometimes also called *self-reflection*) or other 'material' (e.g. as part of a task) [384].

The NLUIs that were built as part of this PhD cover both types of reflection across a range of tasks. Their goal is to enable reflection to help people become aware and explore ideas that can support them in making sense of and progressing with an *open-ended* (e.g. [309]) and potentially *ill-defined* (e.g. [374, 393]) task/problem (also see the Glossary at the beginning for a definition of both terms) – either by reflecting upon the task material or upon oneself. However, people often find it difficult to engage in deep reflection that successfully transforms their understanding of something [51, 169, 232, 291, 373, 390], unless they receive guidance – but how could such guidance be provided to them?

The approach taken here is to make the NLUIs *proactive* so that they can intervene on their own initiative to ask or point something out to the user. However, this leads to another important question of this thesis, which is, what happens if, as a result of this, the conventional interaction paradigm of '*user requests, system responds*' (that currently comprises most existing

NLUIs and co-pilot tools) is shifted towards the converse of *'system requests, user responds'*. Even if many current AI (co-pilot) tools can, of course, ask questions back to the user, they are generally not yet designed to do so to actively support, guide, and structure people's thought processes. **A specific aim of my research, therefore, is to explore how to design interfaces that can proactively provide questions and prompts to the user to support them in their thought process by enabling them to reflect when needed.**

There have already been various conversational interfaces in the past that were developed to enable people to think and reflect. One of the earliest programs that could be seen as an example for this was Weizenbaum's *Eliza* developed in the early 1960s, which offered a human-like conversation by using a pattern-matching method [443]. Eliza was effective at convincing people at the time that they were having a conversation with another person, mainly by asking questions back to them based on their input, which would get them to reflect on what they said previously (for example, by asking "What makes you think that [something that the person previously said]?"). More recently, there have been various examples which have shown how conversational interfaces have the potential to help people reflect on and make sense of a subject [71], understand it better, learn or work more effectively [65, 230], and make sense of themselves and their behaviour through self-reflection [255]. Various commercial conversational agent apps have also appeared, such as *Woebot* or *Wysa*, which have been found to be successful in helping people reflect on and change their behaviour in meaningful ways [37, 141, 197, 268, 329]. While there has been some success in developing NLUIs to facilitate reflection in educational contexts [6, 71, 160, 203, 260], through asking the user questions about their thought process, they have not yet been more broadly embedded into software tools to proactively support other cognitive tasks, which is the aim of the NLUIs that were designed as part of this research.

1.3 Framing Natural Language User Interfaces as Cognitive Co-pilots

All of the aforementioned tools and apps contain some form of a conversational user interface (CUI). CUI is a widely used term to refer to these types of interfaces. However, the term that will generally be used in this thesis is *NLUI* (for natural language user interface, as introduced earlier). The reason for using NLUI is that it is somewhat more general, and it also includes interfaces that might not be categorised as a CUI. As the name CUI suggests, the interaction is generally not only based on natural language, but it is also in a *conversational* form (e.g. a question followed by an answer and so on). Examples for CUIs are chatbots (e.g. *Eliza* or *Woebot*), voice assistants (e.g. *Siri* or *Alexa*), and virtual agents (e.g. Ikea's *Anna*, which used to answer customers' questions about products), among many others. **NLUIs, on the other hand, refer to any interface that people interact with using natural language – which can (but does not have to be) through conversational interactions/turn-taking**. An example for this would be when one conversation partner asks a question but the other partner may not respond via a conversational turn but rather by performing an action. This could be a virtual tutor that guides a student in an online learning environment through an exercise and gives hints based on what they are doing while the student 'responds' to the tutor through the way they proceed with the exercise (rather than by taking a conversational turn in response to the tutor). **As not all the prototypes presented in this thesis are strictly conversational (see also [34]), NLUI was a more apt overarching term to use** – despite it being less commonly used in the field of HCI than CUIs.

An advantage of NLUIs (and CUIs) is that they can offer forms of interaction and turn-taking between user and system that other interfaces like graphical user interfaces (GUIs) cannot. In particular, NLUIs are often perceived as some form of 'social entity', which can lead to different responses and behaviours in people using them, such as enabling them to express themselves better (e.g. [459, 460]). Some of these effects can be further supported (and better targeted) by framing NLUIs in adequate ways to help a person get a better idea of how they should understand and interact with a specific NLUI. To achieve this they can make use of metaphors and be given certain *'roles'* depending on what their goals are (which might be inspired by familiar human roles and relationships). There are a variety of roles/metaphors that have been used for or 'given to' NLUIs (and CUIs), such as advisors, assistants, companions, or more recently, co-pilots. **In this thesis, they are framed as *'cognitive co-pilots'*, which aims to convey the idea of what they are trying to achieve – to support and empower people in performing cognitive tasks.** The term draws from the existing term of 'AI co-pilots' introduced earlier. The relationships between the key terms of NLUI, CUI, and

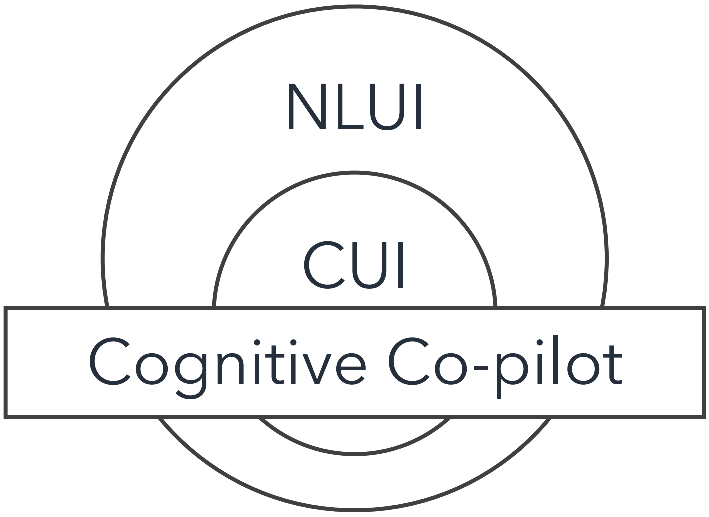


Figure 1.1: Diagram illustrating the 'relationships' between key terms in this thesis.

cognitive co-pilot are illustrated in Figure 1.1 – namely, that NLUIs encompass CUIs and that what are called *cognitive co-pilots* in this thesis can be considered an NLUI and/or a CUI. Having considered in this section these two kinds of interfaces and how they 'work' as well as the idea behind cognitive co-pilots, the next section shifts the focus to outlining the key aims and research questions that guided the design and evaluation of the cognitive co-pilots that were built as part of this research.

1.4 Research Questions and Aims

The overarching goal of this thesis is to address the question of **how cognitive co-pilots can be embedded into tasks to support and scaffold people's thought processes by proactively asking them questions to get them to reflect and externalise their thoughts**. What is meant here by 'scaffolding'[3] is to guide and help structure someone's thought process towards a direction that could help progress with a task and that enables them to get a new insight or idea (see, for example [341, 476]). The use of the term here is inspired by educational and learning sciences where scaffolding refers to "the process that enables a child or novice to solve a problem, carry out a task or achieve a goal which would be beyond his unassisted

[3] The scaffolding metaphor draws from building construction where the scaffolds provide both "adjustable and temporal" support to the building under construction [312].

efforts" [455]. The aim of the research conducted in this PhD was to explore different application areas and types of cognitive tasks where the 'interaction paradigm' of *'system requests, user responds'* could support people in a specific task by scaffolding their reflective thinking. The overarching research questions of this PhD thus are:

RQ1: How can 'cognitive co-pilots' be designed to proactively support people in tasks they engage in?

RQ2: How can cognitive co-pilots support reflective thinking?

RQ3: How can the findings of the studies be conceptualised and lead to a model of how scaffolding NLUIs, like cognitive co-pilots, extend people's minds?

To investigate these questions I have explored in my PhD (i) how a cognitive co-pilot can support groups of people performing an exploratory sensemaking task together [FP1][4], (ii) how cognitive co-pilots can be designed to probe and scaffold a person's decision-making processes [FP2], (iii) how cognitive co-pilots can guide people when reflecting and creatively expressing themselves [FP4], as well as (iv) the ways in which 'co-pilot devices' could (and how they should *not*) proactively prompt people and make suggestions in everyday life [SP, FP3]. See Table 1.1 for an overview of the studies/chapters and their key characteristics.

In what follows I will generally use the previously introduced term *NLUIs* to refer to interfaces which people interact with through natural language. The cognitive co-pilot framing or metaphor will only be used when the *role* of the NLUI in supporting a person's cognition is emphasised. In the remainder of this introduction, the chapters of the thesis are outlined and what NLUI each of them covered.

1.5 Contents of this Thesis

Chapter 2 provides the background to the thesis by presenting relevant work conducted in this burgeoning area of research. In the first part, it focuses on the literature on the capabilities, characteristics, and interactions provided by current NLUIs. The second part provides an

[4] Note that these are the abbreviations used in the "Publications" list at the beginning of this thesis.

overview of reflective thinking, describing how it can be defined and understood, what its benefits can be, and how it can be enabled and supported through technology. **Chapter 3** outlines the methodological approach taken to both designing and evaluating the cognitive co-pilot prototypes that were designed to support different types of cognitive tasks.

In the first study of my PhD, published in the *ACM Transactions on Computer-Human Interaction 29* [FP1] and described in **Chapter 4**, I designed an NLUI that supported groups of users exploring a dataset together. The aim was to investigate whether this set-up could enhance collaborative sensemaking. The findings from the study suggest that the provision of proactive questions triggered more reflection on the reasons for certain trends and patterns and helped people make sense of them. Furthermore, differences were found in participants' interactions among themselves and with the system depending on whether the interaction modality was text/screen or voice-based, providing important implications for designing future NLUIs that are designed to be embedded or take part in social and collaborative interactions among people.

Following on from this idea of probing users to enhance cognition, the second study of my PhD, published in the *Proceedings of the 4th Conference on Conversational User Interfaces 2022* [FP2], and presented in **Chapter 5**, explores how a chatbot can be designed to probe and scaffold complex decision-making. The aim of using probing questions at the NLUI in this context was to engage the user in reflective conversations about their reasoning as part of their decision-making. The chosen scenario was stock investing, which involves exploring and making sense of different types of data to subsequently construct an 'investment thesis' for a stock. The study showed that when experienced stock investors used the prototype they reflected on and reconsidered certain investment decisions they had made and their reasoning behind this – which they would not necessarily have done without the scaffolding from the NLUI.

The finding that an NLUI could successfully support human cognition in specific tasks led to a new direction in the PhD research, which was to explore how this might also be done in everyday contexts where it can help people to understand themselves. **Chapter 6** presents my third study published in the *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* [FP4], which was a collaborative project conducted with Nadine Wagener and other colleagues from the University of Bremen[5] and UCL. We explored how to design an NLUI to foster someone's reflection on a past personal challenge and how they overcame it. The NLUI was embedded in a VR experience that was inspired by art therapy in which the user visually represents a past challenge they went through. VR often works by immersing the person in a virtual environment, which the user can interact with, explore, and shape by themselves. In our case, we provided a virtual 3D drawing tool for participants to enable them to express themselves through creating 3D representations. An NLUI was embedded in this software environment to take the role of a talking guide, which encouraged, inspired, and prompted the user to express themselves and reflect within their VR 'world'. Our findings show that this kind of 'NLUI + VR' experience enabled the participants in our study to draw connections between different components of a past challenge and enabled them to identify new approaches for overcoming similar challenges in the future.

Building on this work investigating how to design NLUIs that take on the role of a co-pilot to support human cognition, I finally explored more broadly how to take into account individual needs, preferences, and (social) contexts that need to be considered when designing co-pilots for everyday life. This work is presented in **Chapter 7**. In this chapter, the co-pilots are not embedded into task interfaces but into everyday contexts with the aim to provide relevant information for the ongoing tasks. Again, this work was done in collaboration with colleagues from the University of Bremen as part of the Excellence Chair program. The aim was to investigate how people might react in different contexts to different types of prompts from an NLUI. We used scenario-based questionnaires and interviews to investigate in which (social) situations and for which activities different interventions by voice assistants – another form of NLUI – may be useful, appropriate, and desirable. This work was published in the *Proceedings of the 2nd and 3rd Conference on Conversational User Interfaces 2021 and 2022* [SP, FP3]. The research revealed that certain interventions are generally seen to be more acceptable than others (e.g. depending on the urgency of the intervention, the characteristics of the social setting, etc.). However, one of the main findings was that there are marked differences *between* individuals: While some participants were against having an NLUI that corrects people when

[5] As part of my supervisor Prof Yvonne Rogers' 'Excellence Chair', which she was awarded by the University of Bremen, several collaborations took place among PhD students from both Yvonne's group and the partners in Bremen, including Prof Johannes Schöning's and Prof Rainer Malaka's group.

they have a disagreement, others were in favour of this form of proactive intervention – in particular when multiple people were present, and the ongoing activity was considered to be personal or private. A conclusion from this line of research is the importance of individual differences; one size does not fit all. A challenge for future research, therefore, is to identify how to accommodate for different users in different contexts so that they can be prompted in ways that are appropriate for them.

The main contribution of this series of studies is the finding that 'cognitive co-pilots' can be developed that enable and facilitate human thinking and reflection about the task at hand, for example, to help them discover more or new things, to make them more aware of how they decide, think, or regulate their emotions, or to just help with everyday tasks and activities. The studies also showed that it is possible to design co-pilots for a range of tasks, although the final study showed that some participants found them too intrusive for certain contexts. Table 1.1 below provides an overview of the key characteristics of the chapters/studies. The **final chapter** discusses the findings from the set of studies conducted with respect to the research questions outlined earlier. Its focus is on what factors are important when designing cognitive co-pilots in order to support and augment the human mind in different (open-ended) activities and tasks which can benefit from having more reflective thinking involved. It discusses how the cognitive co-pilots were able to help people become aware of things they have not considered before, explore new approaches, and take on new perspectives. It introduces a model that can be used to conceptualise how tools like cognitive co-pilots can extend human cognition. It also discusses key design considerations as well as the challenges and limitations of building cognitive co-pilots. The chapter ends by discussing future research avenues and the ethical considerations involved in designing cognitive co-pilots.

Table 1.1: Overview of the four study chapters and the key characteristics.

	VoiceViz **Chapter 4**	**ProberBot** **Chapter 5**	**SelVReflect** **Chapter 6**	**Scenarios** **Chapter 7**
Supported activity:	Analysis and sensemaking	Decision-making	Self-expression and self-reflection	Everyday (social) activities
Study design:	Experiment comparing NLUI modalities	Qualitative user study	Mixed methods user study (pre/post)	Scenario-based questionnaire and interview study
Single- or multi-user:	Multi-user (pairs)	Single-user	Single-user	Single and multi-user (hypothetical)

In conclusion, this PhD research contributes to the following:

1. Three different types of cognitive co-pilot tools following the 'system requests, user responds' paradigm showing how they can be designed for different types of tasks.
2. Empirical evidence of how people interact with and respond to these different types of co-pilots and how the co-pilots' prompts augment their thought processes and reflection.
3. A discussion of some of the key challenges and opportunities involved in designing proactive systems and how they deliver their prompts.
4. A model to conceptualise tools like cognitive co-pilots and the ways in which they can extend human cognition.

2. Background and Related Work

This chapter is divided into two parts. The first part (2.1) covers relevant literature concerned with the design, use, and evaluation of NLUIs. As mentioned earlier, NLUIs encompass a range of interfaces, including CUIs. Although many of the interfaces and software tools covered in this literature review could also be referred to as CUIs, dialogue systems, conversational agent, smart personal assistant, chatbot, robots (with natural language capabilities), to name a few, the term NLUI will generally be used here for consistency, unless a clear specification of the interface seems meaningful. On a high level, NLUIs refer to interfaces that users interact with via *written* natural language – which includes chatbots – or *spoken* natural language – which includes voice user interfaces (VUIs), voice assistants (VAs), etc. (see also [472]). Despite differences in their instantiation, there is clearly an overlap between the different forms of user interaction of these different types of NLUIs, in so far as they have been designed to enable a user to have a conversational or 'conversation-like' interaction at the interface following a model of turn-taking.

The second part of this chapter (2.2) focuses on how cognition, in particular reflective thinking, can be understood as well as what its purposes and benefits can be. Section 2.2 also provides an overview of some existing approaches and interfaces that support reflective thinking and how this can benefit people's self-awareness, sensemaking, and decision-making among other things. To begin, the state of the art of NLUIs is discussed.

2.1 An Overview of Relevant Research on NLUIs

This section is structured as follows: Section 2.1.2 provides an overview of some of the general advantages and challenges of NLUIs and the interactions they enable. Section 2.1.3 covers how the use of different NLUI characteristics can affect the user experience and outcomes for different types of activities and tasks. Section 2.1.4 then outlines some of the NLUIs that have been embedded into software tools to support certain tasks. Section 2.1.5 describes some of the key developments in the shift from NLUIs taking a reactive role to becoming more proactive, and finally, section 2.1.6 discusses one of the key challenges of proactivity, which is to identify opportune moments for an NLUI to intervene. To begin, a brief history of the emergence and development of NLUIs is provided.

2.1.1 The Emergence of NLUIs

The foundational work for NLUIs can be traced back to 1966 when Joseph Weizenbaum introduced the Eliza program [443]. Since then, NLUIs have been developed in many different ways for a variety of tasks and have also been given voices and bodies (e.g. [278]). Driven by a number of technological advancements, there have been various 'waves' of NLUI research over time – for example, Schöbel et al. [371] identified five waves of NLUI research:

- The 1st wave, which Eliza was part of, was defined by scripted conversations, using rule-based methods,
- the 2nd wave up until the 1990s started using more basic AI methods, pattern recognition, and embodiment,
- the 3rd wave, referred to as the 'kick-off wave' and lasting until around 2010, NLUIs started receiving attention from big tech and real-world uses were developed (such as for customer support),
- the 4th wave or the 'hype wave' brought the introduction of many voice-based NLUIs and personal assistants like Siri or Alexa,
- the 5th wave started to see extensive use of AI (including LLMs), making NLUIs significantly more capable, flexible, and adaptable.

Though the study and development of NLUIs spans several decades, their practical application has only really materialised at some point during the 'hype wave' in the early 2010s [277], propelled by progress in AI domains like natural language processing (NLP), natural language understanding (NLU), as well as speech recognition and speech synthesis. With new technical capabilities and opportunities, academic and commercial interest in NLUIs has thus also accelerated throughout the hype wave (e.g. [277, 320]) and recently – in the 5th wave – experiencing an additional 'boost' with the proliferation of GenAI and, in particular, LLMs and the new possibilities they offer for building NLUIs such as ChatGPT (e.g. [138]).

With this widespread adoption and interest, chatbots have found applications and have been researched in a wide range of industries, contexts, and use cases, including e-commerce, customer service [97, 168, 333], (mental) health and wellbeing [21, 45, 226, 238, 266, 298, 313, 314, 335, 405, 416, 420], education [182, 395, 431, 437, 448, 451, 462], banking and financial advice [177, 252, 287, 383], cars [54, 215, 279, 280, 370, 454], writing support tools [151, 297, 430], collaborative tasks [11, 20, 115, 194, 406, 449], and more recently GenAI-based NLUIs supporting a range of professional/work-related tasks [61, 67, 104, 212, 213, 307, 355].

As this widespread NLUI proliferation foreshadows, the expanding domain of NLUI research is distinctly interdisciplinary, encompassing areas like (computational) linguistics, media studies, communication science, management and marketing, philosophy, psychology, sociology, informatics, engineering, design, and HCI (see also [145]). While this broad knowledge base is beneficial, it also indicates that chatbot research is dispersed among various disciplines and application fields with different research foci and agendas – which is also reflected in the increasing number of papers that are mapping out the different research streams on NLUIs [145, 271, 371, 472]. Taken together, the field of NLUIs is vast and diverse, with various aspects being researched – covering technical aspects, different ways of designing them, their psychological, social, and cultural implications and so on. One important focus within HCI has been to investigate how they can affect people's behaviour and experience – in both good and bad ways – when interacting with them, which will be described in the next section.

2.1.2 Advantages and Challenges of NLUIs (and Their Metaphors)

Most advantages *and* challenges of NLUIs are related to people perceiving them as some form of social entity with various degrees of human-likeness and anthropomorphism. The advantages are generally related to producing or 'inducing' through their human-likeness different and often more engaging, personal, and interactive experiences in people (see also [271, 472] for a review). The challenges typically are related to when the human-likeness and anthropomorphism, including the metaphors used for an NLUI, are inadequately applied or taken 'too far' (e.g. [19]). This might be the case, for example, if an NLUI is imbued with a level of human-likeness that might feel inadequate for a given use or task or uncanny (see also Desai and Twidale [108]). However, first some of the advantages of NLUIs are introduced.

Overall, NLUIs have been found to enable interactions, as well as behaviours and outcomes for users that other interfaces such as GUIs – or even human-human interactions – might not be able to provide in the same ways. This can be traced back to the theory of computers as social actors (CASA), introduced by Reeves and Nass [337], which proposes that humans interact with computers as they would with other humans, attributing social characteristics to these interfaces as well as social norms even if they display the most minimal of social cues (see also [300]). This can be amplified when technologies have more human-like characteristics, such as when offering interactions through natural language (e.g. [2]). For instance, Gnewuch, Morana, and Maedche [155] conducted a study on the effectiveness of conversational agents in service encounters, revealing how these agents can elicit more engaged and accountable responses from users. Other studies showed that if the NLUI discloses something 'personal' to the user, it can also facilitate disclosure of certain information from the user, such as on a personal experience (see for example Lee et al. [253]), also referred to as reciprocity in people's behaviour. These effects of 'higher engagement' and reciprocity when using NLUIs are likely to be related to people perceiving a form of social presence 'through' them, which might be explained using the *social presence theory* [387]. This theory has been developed in interpersonal mediated communication studies where social presence was defined as "a person is perceived as a 'real person' in mediated communication" [387:151] and which has also been applied in research on NLUIs [271] where the NLUI can convey some form of social presence of a real person. This can then lead to a phenomenon referred to as 'social facilitation' first identified/introduced by Triplett [415], where performance is affected as a consequence of the felt social presence. An example for this might also be a study which compared 'conventional' questionnaires with conversational ones and which found that participants tended to write answers to open questions that were both longer and of higher quality in the latter [459, 460]. At the same time, people were found to feel more comfortable talking about certain topics with an agent than with a human, for example, in situations that might involve judgment or any social pressure when interacting with a human [314, 335, 449].

However, there are also various challenges involved in designing NLUIs. For example, some of their human-like characteristics were found to lead to heightened expectations in people about what NLUIs are capable of which they then often cannot live up to [263]. The resulting mismatch or ‘gulf’ between expectations and experience, as it has been described by Luger and Sellen [263], can lead to frustration, disengagement [263], and sometimes even abandonment of the technology [83]. One approach to mitigate this is to clearly introduce the purpose of an NLUI and its capabilities to people before they start using it, by deploying fitting metaphors or descriptions of what roles the NLUI will play that can help people understand what they can expect from it (see for example [214]). A main reason why metaphors have become popular in computing is that they can provide grounding, couching the technologies’ capabilities and interactions with it in familiar terms for what might be otherwise described in more abstract terms [302]. For example, the desktop metaphors [302] (e.g. files, folders, bins, stacks and so on), which became broadly used in graphical user interfaces (GUIs) at the beginning of the 1990s, are thought of being an important contributor to Personal Computers (PC) becoming popular. When adequately employed, the use of metaphors can also improve people’s experience of using NLUIs. For example, when Xiao et al. [458] introduced their NLUI to participants as a “learner” or a “collaborator”, they were more willing to respond to its requests for feedback, which were also perceived as less disruptive. Similarly, Khadpe et al. [214] found that when NLUIs were introduced through metaphors suggesting lower competence (such as a child), the NLUI received better ratings on aspects like participants’ desire to adopt and to cooperate with the system after having used it. The reason for this is most likely that the metaphor suggesting lower competence narrowed the gulf between expectation and experience, adjusting participants’ expectations before interacting with the NLUI. Conversely, choosing metaphors that suggest a system is more capable than it truly is can increase this gulf [263] and lead to more frustration.

Various analyses of how technologies are embedded into social situations have shown the complexities of how their use is being interwoven into social interaction – for example, how NLUIs are used in multi-user settings and collaborative action [326, 327, 338]. It was found that sometimes when an NLUI was addressed as part of an ongoing human-human interaction, it failed to respond adequately, as it was lacking the contextual understanding, including the ongoing conversation and what was previously discussed. In this environment, an `assistant’ metaphor “fails to capture the complexities of having a natural language based ‘conversational’ system in peoples’ homes”, as Desai and Twidale [108] argue. However, as other studies suggest, people learn to adapt to technical limitations over time and, in the long term, they tend to adjust and evolve their interactions with the devices. For example, a study by Beirl et al. [39] showed how families created new social games and routines when using a smart speaker in their homes, developing new family interaction patterns.

Taken together, NLUIs provide a 'design space' with many opportunities; however, navigating this space can be challenging and should be done with care. Given their conceptualisation as social entities or even as social actors, they can trigger and enable interactions that go beyond what other interfaces can do. However, as a result, they sometimes also get people to think that they can provide better responses than they are truly capable of. Attributing certain roles and metaphors to an NLUI can help 'calibrate' people's expectations of what it is capable of and what it should and should not be used for. Beyond their conceptualisation and which roles and metaphors they are given, a central concern is which specific design characteristics an NLUI should have so that it can enable and foster the desired interactions and behaviours in the human(s) using it – this question is addressed in the next section.

2.1.3 Using Different Interface Modalities and Characteristics

When designing NLUIs, one of the key questions is if the system should be text/screen or voice-based. While this decision might, in certain cases, be straightforward due to specific requirements or constraints for the intended usage of the system (e.g. using voice for a task where a user's hands are not free), it can be more nuanced or difficult to answer in other situations. To give an example, disclosing a personal issue might feel more natural to *talk* about, but at the same time, a person can feel uncomfortable having to say more personal things out loud compared to *typing them*. Therefore, NLUIs can be designed to provide different options for how to interact, especially where it might not be clear if one or the other modality is better, which could also depend on individual preferences or specific needs in a given situation.

Clark et al. [88] summarised empirical research comparing the effects of speech versus graphical interfaces on user performance and experience, which has shown mixed results. In some studies, the use of voice was more beneficial than in others. Voice has been found to support a range of domestic tasks [9], everyday tasks of people with impairments [328], and also collaborative tasks [449]. Le Bigot et al. conducted two studies [247, 248] investigating written text versus speech input with information retrieval systems, one of them for a restaurant search and the other for travel planning. In their first study [247], they found no difference in transfer effects when switching from one modality to the other. Subsequently, they found people were faster when working in written mode than in spoken mode, although the latter was considered to be easier. They also found spoken interaction led to more *'collaboration'* with the system – in terms of users matching their utterances to the system's utterances – while written interaction was more *efficient* – in terms of turns being required to complete the tasks [248]. In another study, Begany et al. [38] investigated users' perceptions of spoken versus written text input for a search interface. Written input was preferred compared with voice input because it was easier to learn and to use. Limerick et al. [259] studied pressing keys versus using voice commands and found that speech leads to a diminished sense of agency in users – which was defined as the experience of controlling one's own actions and their outcomes. There have also been modality comparisons of a computer tutoring system; D'Mello et al. [114] found no difference in learning outcomes if students made system input via keyboard or speech. Similarly, Litman et al. [261] found no difference in students' learning gains for spoken versus typed modalities.

These mixed findings suggest that whether speaking or typing is more effective depends on the context and task. However, most of the research on using different modalities has focused on how task completion performance varies or on exploring users' perceptions of using each modality. There is only a limited number of studies that investigated how they impact experiential and behavioural aspects, such as reflection, sensemaking, or collaboration with others. An example of a study more concerned with such aspects is that of Gonzalez and Gordon [158], which compared how using speech versus text as input affected the player experience and user understanding of their fictional role in an interactive narrative (which has similarities with playing a game). They found that in the text condition, participants were more likely to adopt the role of the narrator (speaking in past tense), and in the voice condition, they were more likely to speak directly to the narrator (or the computer), saying what was happening or what should happen (speaking in present tense). This suggests that speaking at the interface can enable users to step more into character, see themselves more as part of the story and, in doing so, change how they feel, think and experience. This resonates with the study by Kocielnik et al. [230], comparing a speech-based and a text-based agent for the workplace to support employees' activity journaling and self-learning through reflection. Their findings suggest that the voice-based system is easier to use and feels more personal, interactive, and engaging. In addition, compared to screen-based conversational interfaces, voice-based interfaces have been found to be less distracting for certain types of tasks [295, 456]. However, this research on how different modalities can engender different user experiences and behaviours beyond task performance is still limited.

Taken together, while the use of a keyboard and a graphical user interface (GUI) to type or select commands often seem preferable for many tasks because of their ease of use as well as increased agency and efficiency, voice user interfaces may be more 'natural', immediate, or engaging, enabling users to 'collaborate' more with the system and to have a different and possibly more interactive experience. In particular, users may also be able to draw upon familiar conversational practices and social norms when speaking with an NLUI.

In addition to deciding on whether to provide text/screen or voice-based interactions, another important question is what characteristics the voice should have to best accommodate user's needs for the given task (e.g. to not distract them), including how it might be chosen to match user's expectations (e.g. [54, 127]). This includes considering the gender of the agent's voice. Research has shown that the identified gender of an agent has an impact on the user's experience and perception of them [48, 53]. For example, female voices were found to be perceived as more trustworthy in a study where participants received health-related advice [171]. However, previous research suggests that designing the right voice for an assistant in a given application largely depends on its context [68, 292, 299, 404]. Whether it should sound realistic or synthesised is another design question. While real human voices have often been used [299], the preferences between synthesised and real human voices can also be context-specific [68]. Furthermore, research has shown that there are significant marked differences in people's preferences for different kinds of voices [68, 404]. In conclusion, the current evidence base does thus not provide a clear picture of when specific voice characteristics might be best. The next section provides an overview of how some NLUI have been integrated into a range of tasks and tools.

2.1.4 NLUIs Embedded Into Software Tools

Besides being standalone systems, conversational agents have been increasingly integrated with other technologies and software applications (e.g. [22, 23, 95, 131, 183, 187, 246, 363, 383, 395, 442, 448]) and even more so now with recent developments in GenAI and LLMs, giving rise to tools like *GitHub Copilot, Microsoft Copilot, Gemini for Google Workspace*, *Adobe Firefly,* the AI learning tutor *Khanmigo* from/on Khan Academy, or *Einstein Copilot* from Salesforce, etc. The benefits of doing so are that the AI co-pilot can support humans in their ongoing task within the software tool used for it, which can enable the co-pilot to directly contribute to what the person is doing.

With the prospect of future advances in AI, there is also the potential for NLUIs to become more intelligent and, as a result, capable of supporting and guiding users through ever more of their tasks at the interface. However, it is not straightforward as to how best to integrate more intelligent NLUIs with the other kinds of software tools that a human currently uses for carrying out their tasks. How can we ensure that the next generation of NLUIs do not simply automate what humans currently do but instead amplify and empower them in their activities (see also Schmidt [364], Shneiderman [386], and Bainbridge [26])? For example, should the NLUI provide the user with solutions, or instead, should it assist them in finding a solution themselves by probing them and asking them questions? Here, it is argued that it is important to conceive and design NLUIs for human cognition by embedding them into the tasks humans perform rather than 'just' conceiving them as assistants that complete tasks for them. But how? This will be discussed at the example of data analytics tools in particular.

Two different kinds of software tools that were found to be particularly promising for such NLUI embedding – that are also relevant for the studies presented here – are data analytics and visualisation tools [18, 111, 133, 150, 188, 208, 361, 379, 380, 396–398, 403, 413] and learning tools and platforms [131, 186, 432, 437, 447, 448]. One of the main reasons is that this embedding can augment the interactions with the given dataset or learning materials, either to scaffold the user's analysis or learning process through conversational turn-taking and/or to be able to ask questions. A commercial tool offering such interactions is *Tableau's Ask Data* (see [413]), which enables users to formulate queries in natural language to generate and modify visualisations. One of the earlier systems that allowed users to plot data using speech or text queries was *Articulate* by Sun et al. [403]. Their evaluation showed that when participants had to plot the same data in Microsoft Excel, they were significantly slower and found the steps required more complex and more confusing, despite the majority of them being familiar with Excel and its charting features. Another data analysis tool, *Ava*, which provided a chatbot interface, was designed to allow data scientists to assemble data analytics pipelines [201]. Computer scientists who were knowledgeable about data science were able to build machine learning models faster than when using Python. Another example is *Eviza* by Setlur et al. [379], which allowed users to interact with and modify visualisations of geospatial data via typed natural language queries. In a user study comparing *Eviza* with Tableau (without any natural language features), participants found *Eviza* to be more natural to use and completed the analysis tasks significantly faster; however, some users experienced a loss of empowerment and ownership (of the task).

In summary, this research suggests that NLUIs can help both lay and experienced users perform their tasks more efficiently when used as part of a software tool. One way it does this is by enabling them to ask questions and make requests in a more familiar way. However, most of this body of research has focused on how to speed up task completion by using natural language. Little is known as to how integrating NLUIs in software tools can also support other forms of cognitive activity, such as improving sensemaking and reflection about the task at hand. One way to do this could be by making the NLUIs more proactive, actively prompting the user in their thought process. Next, we consider what is involved in making proactive NLUIs that can achieve this.

2.1.5 Proactive NLUIs

Proactivity has been of interest for decades – covering a wide range of NLUIs and other assistants and tools (e.g. [166, 242, 244, 286, 457, 463]) or even robots (e.g. [63, 318, 318, 424]) with a significant acceleration in research and development efforts in recent years (e.g. see Deng et al. [105] for a review of proactive dialogue systems). One reason for making NLUIs proactive is to help users carry out their activities, by making suggestions or asking questions at particular times. This can be achieved by the NLUI proposing or requesting something on its own initiative, without waiting for a prompt from the user.

While sometimes the aim might be to enable more human-like interactions [195, 210, 256, 408], proactive capabilities have also been developed and researched to help provide better recommendations [234, 468, 471], to retrieve relevant information for an ongoing task or conversation with another person [11], or perform a task more efficiently, such as to make a restaurant booking [49]. This interest in proactivity has been more pronounced in recent years with the rapidly advancing capabilities of AI systems, which have increasingly better contextual awareness that is central to knowing when best to be proactive (e.g., [165]). Much has been done to address the technical challenges of building proactive conversational interfaces in the past few years [105, 368, 370, 401], especially with recent developments of LLMs (see for example [106, 257, 433]).

Despite this interest in and promise of proactivity, interactions with commercially available voice assistants are usually still highly constrained to reactive interactions [89, 92, 117, 263, 326]. A reason for little progress being made in the development of proactive commercial NLUIs is the worry that they will be perceived by users as being intrusive; proactivity typically requires understanding the context, which involves collecting data (e.g. monitoring conversations and what is happening in a space/environment) which can make people feel uncomfortable (e.g. [269, 286, 407]). More research is needed to address these user concerns while exploring how proactive interactions in such devices can open up new opportunities and potentially empower a broad range of applications [440, 441]. Besides addressing the privacy concerns when developing proactive systems, it requires knowing how to design meaningful proactive interventions that are delivered in a way that is appropriate for the user and their ongoing activity (e.g., [285, 286]).

Emerging research in HCI has begun to investigate how open people are to proactive interventions, generally suggesting that people see benefits in them – although this depends on the type of tasks and activities an individual may be engaging in as well as their preferences regarding proactivity [285, 441]. Furthermore, a range of application areas and use cases have been explored for which proactivity might be effective or adequate, such as for wellbeing [52], or in educational settings [274, 447, 448]. Others are also beginning to investigate the privacy implications that such proactive and possibly ‘always-listening’ systems could have for users and how they are perceived [269, 407].

A survey conducted by Schmidt and Braunger [367], involving 1,550 participants, showed that users generally see potential in proactivity. Similarly, a study conducted by Völkel et al. [425], which focused on users' envisioned interactions with an ideal voice assistant, revealed that many participants regarded proactive behaviour in voice assistants as desirable. Specifically, the interactions envisioned by participants indicated a preference for agents that can anticipate potential actions and provide suggestions without explicit user requests. In another survey, Chaves and Gerosa [77] identified various characteristics chatbots should be enriched with, which included proactivity. Based on their review of the literature, they found several benefits of having proactivity such as to provide additional, useful information, to recover the chatbot from a failure, to improve the productivity of a conversation, and to guide and engage users. Taken together, this line of research suggests that people generally see the promise of proactivity for certain situations and activities, but they also have some concerns for others.

Besides this research exploring people’s perceptions of proactive interactions, another line of research has investigated how people *interact with* NLUIs offering proactive interactions. For example, Andolina et al. [11] developed a proactive search agent designed to monitor user conversations and offer information based on detected entities within the dialogue. Their findings indicated that this agent effectively enriched conversations with factual information and ideas while minimally disrupting the conversational flow. One setting that has received considerable attention is car driving (e.g., [279, 368–370, 401, 454]), where proactive NLUIs can assist with driving the car, planning their route, or doing other tasks. For example, Meck et al. [280] investigated *how* proactive assistants should intervene when driving. Their findings suggest that proactive suggestions were found to be useful in specific situations. In particular, proactive interventions that were more specific to or more closely related to the participant’s driving task were accepted more frequently than ones not related to driving. For example, while proactive information on the remaining fuel range was accepted *in almost all cases*, and information on the availability of a faster route or suggestions related to the destination or where to park *in most cases*, offering the user a more relaxing mode for driving or suggestions for customising the map were accepted significantly less frequently. Hence, in general, the findings suggest that the more helpful and relevant a proactive suggestion is in a given situation, the more likely it is that it will be accepted.

Peng et al. [318] conducted a study on a robot which participants interacted with through an NLUI. The robot provided three levels of proactivity: (i) offering help directly (high proactivity), (ii) asking for permission to intervene (medium), and (iii) waiting for the participant's explicit help-seeking signal (low). Participants said that the robots with high and medium proactivity were more informative for the given task than the one with low proactivity. Overall, participants preferred the robot with the 'medium proactivity' setting even if they were interrupted more in their task compared to 'low proactivity'. This was despite the 'medium proactivity' robot taking more control over the conversation and interrupting them more than the 'low proactivity' one, and also requiring more steps of interaction than the 'high proactivity' one. Taken together, the findings show that even if there can be some less desirable aspects of proactivity, people may still prefer it if it provides meaningful information for the task at hand. Moreover, some participants appreciated that the robots provided information without having to explicitly request it: *"It's great that the robot proactively provides more information when I am hesitant. It broadens my mind as there are some points I didn't consider".*

Privacy and intrusiveness, however, are key concerns for proactive NLUIs (Tabassum et al. [407], Miksik et al. [286]). A study by Lau, Zimmerman, and Schaub [245] showed that distrust in the companies behind these devices makes many individuals hesitant to use smart speakers, especially if they are to make them proactive. Another issue is that proactive interventions, if not matching a task and what a user might need at a certain moment, can interfere with the task, negatively affect the user's agency, and be perceived as intrusive (e.g. [20, 279, 286, 457]). For example, Miksik et al. [286] found that their voice-based NLUI, which proactively intervened while participants were engaging in different tasks in a domestic environment, was often considered intrusive: P6 in their study said, *"I found it invasive [and] couldn't concentrate on the tasks"* while P13 pointed out *"there were too many updates, barely had time to think*". However, it is important to note that the NLUI, in this case, was not providing proactive suggestions for the task at hand, but instead information about incoming emails, calendar events, or other updates. The proactive interventions were thus not relevant to the tasks participants were engaging in, which might be the reason why they were perceived as distractive and intrusive. This is in line with previous research on task interruptions mentioned earlier, which found that they tend to be more disruptive when they are not relevant for the current task [58].

Research has also investigated how proactive interventions can be best delivered to make them more acceptable. For example, Dubiel et al. [118] designed a proactive voice-based NLUI and investigated its appropriateness for a food ordering decision-making task. They compared three different delivery strategies for the NLUI's proactive interventions in the interactive food ordering scenario: It would either provide (i) no feedback on the user's decisions, or provide feedback in either an (ii) unsolicited or (iii) solicited way. They found that unsolicited feedback was perceived to be more appropriate than solicited feedback, suggesting that in this scenario, it was acceptable for participants if the NLUI provided proactive feedback or suggestions without their previous approval. However, some participants questioned the desirability of having proactive interventions, with one participant asking, *"Why are you saying bad things about the food that I am going to eat?"*. Luria et al. [264] identified three thresholds of agent proactivity including reactive to user requests, proactive by providing information, and proactive by providing recommendations for a course of action. They found that users differed in their comfort levels with each threshold. In their study in which participants had to reflect on a range of storyboards that illustrated potential future scenarios of socially sophisticated agents in a domestic setting, they found that most participants were open to the idea of a proactive agent in a multi-user situation, but no one wanted the agent to enforce certain recommendations such as preventing them from ordering unhealthy food.

In sum, proactive NLUIs appear to be most promising when users might benefit from suggestions or questions during their ongoing activity. However, there are a number of concerns about a technology that takes the initiative to talk to the user, namely, privacy, agency, intrusiveness, and trust. Research on how to address these, while ensuring the putative benefits materialise, is in its infancy. *"Work has only just begun to focus on (…) for what types of experiences these proactive CUIs may be suited"*, as Cowan et al. [91] argue in their special issue on *New Theory and Design Perspectives for Conversational User Interfaces*. A central concern for designing interfaces to have proactive features is knowing what the opportune moments for proactive interventions are – to which we now turn.

2.1.6 Opportune Moments for Proactive Interventions

One line of research on when a system might best intervene and interrupt a user in what they are currently doing is to consider when it is least disruptive. This involves investigating how interruptions can affect ongoing tasks (in positive and negative ways) and when to minimise their intrusiveness and maximise their usefulness in the given moment [137, 139, 215, 418]. To achieve this, one research focus has been on how to make a system context-aware, i.e. know what is happening at a given time (e.g. [185, 317]) often with the aim to build context-aware and intelligent notification delivery systems (e.g. [15, 281–283, 317, 394]). Opportune moments for voice-based NLUI to intervene have been explored in environments like homes [72, 206, 233, 440, 441] and cars [215, 216, 370, 378].

Identifying opportune moments for a voice-based NLUI to start interacting with a user requires knowing when not to interrupt users with their current activities or social interactions [401], which is challenging to achieve given the high number of contextual cues that need to be considered [191, 347, 418]. Nevertheless, there have been a range of studies to get a better understanding of *when* NLUIs could intervene in different settings. For example, Cha et al. [72] used a smartphone camera with a wide-angle lens to detect so-called 'activity transition moments' (e.g. when people go to the kitchen after working at the desk, turn on the television after cleaning the flat, etc.), which indicate when people were more interruptible [185]. Their findings suggest that the key determinants for opportune moments are linked to personal factors such as busyness, mood, and urgency, as well as other factors related to everyday routines at home, including social context such as the presence of other people, and user mobility. Wei, Dingler, and Kostakos [441] also found participants' availability to be interrupted depended on the current activity but that their availability ratings varied strongly. However, they found that boredom and mood are significantly correlated to perceived availability, and participants tended to be more available/open to being interrupted when they were engaged in entertainment tasks rather than when studying or working. Similarly, a study by Nothdurft et al. [306] suggests the importance of the intervention for the user, their surroundings and their mental state, and the accurate placement of the interaction are key to whether proactive behaviour is acceptable or desired. Beyond the use of cameras as in the study of Cha et al. [72] or Komori et al. [233] research has explored the use of physiological sensing [74, 75] including EEG [475] to identify opportune moments to prompt the user, for example during moments of low cognitive load or emotional arousal. However, despite the value of all this research on identifying opportune moments, Fischer et al. [139] found that rather than trying to perfectly time the delivery the content of the intervention is often a more important determinant of how receptive someone will be to it – which corroborates some of the research presented in the previous section on how interventions that are relevant for the ongoing task are often more accepted [58, 280].

Apart from that, researchers have also examined *how* the agent should initiate a conversation, which arguably also has impacts on how receptive a person is and how adequate the timing of an intervention is perceived. This can be observed when people adjust the phrasing and tone of what they are saying when trying to interrupt someone else in an acceptable way. Drawing from these dynamics in human-human interactions, Edwards et al. [122] looked at how people interrupt another person who is engaged in a complex task, as an approach to inform the design of proactive VAs. Their results showed that the level of urgency of an intervention significantly affects how long people wait before interrupting. Furthermore, the participants balanced speed and accuracy in timing interruptions, often using cues from the ongoing task they interrupted. The participants also varied the phrasing and the delivery of interruptions to reflect urgency.

Taken together, the studies reviewed here underline that contextual understanding is key to users' acceptance of proactive NLUIs and their interventions. If they are perceived to be too disruptive or intrusive people will not want to use and interact with them. If, on the other hand, they are perceived to be useful and fit in with ongoing activities then they will be more accepted. The aim of the research presented in this thesis was to address the extent of this 'sweet spot': which tasks have the potential to benefit from proactive interventions, and how proactive interventions can be tailored to the tasks so that they scaffold people's thought processes in meaningful ways and when they are considered unacceptable and why.

A diversity of research has shown that key barriers for the design and adoption of proactive NLUIs are the amount of contextual sensing and data collection that this would require. While this is indeed a key concern, it is important to note that most scenarios covered in this thesis are constrained to people performing specific tasks using a specific software tool, rather than monitoring people's everyday lives and interactions with other people. Furthermore, prototypes designed as part of this PhD generally had relatively simple rules for their proactive interventions (based on people's task-specific behaviours), which do not need extensive data collection and processing (even if they were fully implemented) to determine when they should intervene. In other words, the proactive interventions of NLUIs presented here have somewhat different – and likely less extensive – privacy implications. Although this does, of course, not resolve the other challenges of invasiveness and distraction that proactive behaviours can involve.

To summarise this first part of the literature review, many of the studies reviewed so far have largely focused on how to improve user performance of NLUIs at the interface in terms of traditional UX measures, such as improving *efficiency* and reducing the disruptiveness of proactive interventions. The thesis looks beyond these criteria to consider wider theoretical concerns, namely how they can *extend* and *enhance* cognition. To this end, the next section considers, at a slightly more theoretical level, the aspects of cognition that NLUIs can facilitate, support or trigger to achieve this.

2.2 Extending Human Cognition Through Technology

As outlined in the introduction, throughout human history, technology – analogue and digital – has been extending and augmenting our cognitive abilities in many different ways (e.g. [62, 87, 258, 364]). Many digital technologies have been developed that enable us to perform a wide range of cognitive tasks more flexibly, with less effort, and more efficiently (e.g. [322, 357]). Examples include complex calculations (e.g. spreadsheets), retrieving information (e.g. search engines) or 'remembering' things for us (e.g. digital storage). Digital technologies have also provided us with new ways to externalise our cognition [362], allowing us to build rich, interactive external representations to work with [224], enabling us to write, draw, create, and manipulate things in real time that go beyond the possibilities of many analogue tools/technologies (e.g. digital whiteboards, media editing tools, 3D modelling tools, simulation tools).

Although technology can enhance and amplify our (cognitive) abilities in terms of the plethora of tools it provides to support many parts of our thinking and its externalisation, the question asked in this thesis is how can technology be designed to help us make sense of the tasks we are doing, facilitate our thought process, and help us have new insights and ideas? How could this be achieved by enabling, facilitating, and guiding reflective thinking through a different kind of interface that *prompts* humans while performing certain tasks?

Reflective thinking can allow people to make sense of, gain new insights, and learn new things about a topic or oneself [291]. As the next section describes, there are thus many activities that can benefit from or even require a person to engage in reflective thinking to achieve meaningful outcomes – as is often the case with open-ended and possibly ill-defined activities such as exploratory, creative/expressive, as well as learning activities. How such activities might benefit from reflective thinking and how this can be achieved is the focus of this second part of the chapter. The following sections will first provide an overview of reflective thinking (Section 2.2.1) and then outline approaches to how reflective thinking has been facilitated by existing tools/technologies while focusing in particular on NLUIs (Section 2.2.2).

2.2.1 Reflective Thinking and Its Role in Various Cognitive Tasks

Human cognition covers a wide range of cognitive activities (see for example Keane [211]) from visual perception and attention, memory, language (e.g. comprehension and production), thinking and reasoning, as well as metacognitive processes, which are there to 'monitor' and adjust the aforementioned cognitive processes. On a more general level, a well-known distinction introduced by Norman [304] is between *experiential* and *reflective* cognition. *Experiential cognition* is intuitive and effortless. It requires a certain level of expertise and familiarity. Examples could include riding a bicycle, buying groceries, driving to work, or reading an article on a familiar topic. In contrast, *reflective cognition* involves mental effort, attention, sensemaking, and decision-making. Examples could include analysing an unfamiliar dataset, writing a report, deciding which stock to invest in, which car to buy, or reading a book on a complex and unfamiliar topic. Thus, which type of cognition is involved generally depends on the task.

The cognitive process of *thinking* is a very large area of research and has been written about extensively (e.g. [348]). Unsurprisingly, it lacks a commonly agreed taxonomy. Many different 'kinds of thinking' and categorisations have been proposed, including 'fast' and 'slow' thinking as proposed by the Dual-Process Theory [205] (which has certain overlaps with 'experiential' and 'reflective' cognition mentioned before) as well as problem-solving, reasoning, concept attainment, and creative thinking [44]. For example, the Encyclopaedia Britannica [44] lists problem-solving, reasoning, concept attainment, and creative thinking for different types of thinking.

The focus in this thesis is on *reflection* also referred to as *reflective thinking*. In the following, a brief overview of some of its key theoretical 'backdrops' is provided, and how they shaped the ways in which reflection can be understood, what it involves, and what its benefits are.

The American philosopher John Dewey was one of the first to articulate the idea of reflective thinking, which he characterised as *"active, persistent, and careful consideration of any belief or form of knowledge"* [109] or, more simply, *"the kind of thinking that consists in turning a subject over in the mind and giving it serious thought"* [110]. Both framings underline the 'deep thinking' that reflective thinking involves when one tries to make sense of a subject – i.e. 'careful consideration' and 'serious thought'. About half a century later, in the 1980s, the professor of urban planning and philosopher, Donald Schön, further elaborated that reflection involves a dialogue with oneself, encompassing a 'reflective conversation' with the situation at hand. His seminal work, "The Reflective Practitioner" [373], emphasises the importance of reflective thinking in professional practice, particularly in complex problem-solving and decision-making (such as urban planning). In the same book, he also introduced the widely used framing of reflection *in* and *on* action. *Reflection-in-action* involves thinking on one's feet and adjusting actions in the midst of practice (i.e. performing certain activities/tasks), while *reflection-on-action* entails looking back at and analysing past actions to learn and improve future practice. Resonating with some of Schön's work, Jennifer Moon described in her book "Reflection in Learning and Professional Development" [291] how reflection facilitates deeper understanding and personal growth by encouraging individuals to critically analyse their experiences and integrate new insights into their existing knowledge. Moon argues that there is a range of outcomes of reflection in the literature – from learning, to having a representation of that learning and the progress being made, making a decision, new (unexpected) ideas, as well as self-development – and beyond. Besides this breadth in *outcomes*, Moon also elaborates on the differences in common *uses* of the term 'reflection', which all imply several slightly different understandings and underline the challenge of establishing a clear definition: First, reflection is often related to learning and the representation of that learning – people reflect to consider something in more detail or to (re-)represent it in oral or written form. Second, reflection implies purpose [110] – generally, people reflect for a reason, although insights can sometimes emerge without conscious reflection, which suggests an overlap with intuition. Third, reflection involves complex mental processing for issues without obvious solutions [110, 223].

However, despite these differences in understanding reflection, it can generally be described as a form of mental processing with a purpose and/or anticipated outcome, applied to complex or unstructured ideas, closely associated with learning and its representation – or in the words of Moon: ***"Reflection seems to be seen as a basic mental process with either a purpose or an outcome or both, that is applied in situations where material is ill-structured or uncertain and where there is no obvious solution. Reflection seems to be related to thinking and learning."***

These examples and definitions illustrate that reflection is often linked to **external materials** or events 'taking place' in the world around people, which they might engage with in different ways in their reflection (e.g. to understand or learn about them or make sense of what they mean to oneself). This contrasts another common use of the term, which refers to the more **introspective** forms of reflection – also known as self-reflection [384] (which can be defined as ***"the activity of thinking about your own feelings and behaviour, and the reasons that may lie behind them."***[6]). This introspective form of reflection is also often focused on within HCI – see for example Bentveltzen et al. [43].

On the other hand, 'reflection on external material' [384], also referred to as critical reflection, involves analysing and evaluating external information, ideas, or experiences. Critical reflection involves questioning assumptions, identifying biases, and considering multiple

[6] https://dictionary.cambridge.org/dictionary/english/self-reflection

perspectives, with the aim to achieve a deeper understanding of complex issues. This type of reflection is closely related to and overlaps with critical thinking, which Ennis [126] defines as "reasonable, reflective thinking focused on deciding what to believe or do" and which is considered to be one of the key skills within education for sustainable development in higher education [102].

Here, reflection is understood and conceptualised as covering this entire spectrum – from introspection to critical thinking. Namely, as **a mental process involving the consideration and examination of external materials as well as personal knowledge, ideas, experiences, and emotions to gain deeper understanding and insights into them.** Hereafter, the terms reflection and reflective thinking will generally be used interchangeably.

Reflection and Metacognition. Some of the more introspective forms of reflection (i.e. to reflect on oneself) are connected to metacognition, a term coined by Flavell [142] to describe the awareness and regulation of one's cognitive processes or, more simply, 'thinking about thinking'. Metacognition has been extensively studied in the domain of educational and developmental psychology but also in various other psychological disciplines. There are differences in conceptualisations and approaches to metacognition within different sub-fields of psychology [305]. However, more broadly, there is an agreement on the general definition of metacognition, which is the *knowledge*, *monitoring*, and *control* of one's own cognitive activity [94, 142]. Metacognitive knowledge and awareness can help people decide how to best perform certain cognitive tasks, for example, when they are deciding on the strategies or heuristics to use for it. In the acquisition of metacognitive knowledge – which is relevant for making these decisions – self-reflection can play an important role, as it can foster awareness of one's own thoughts and actions and how they impact behaviour [164, 473], which can then help form new metacognitive knowledge.

Benefits of Reflection. Reflective thinking has various benefits for experiential learning and personal development. Kolb [232] highlights its role in transforming experiences into abstract concepts, enhancing understanding and critical thinking. Moon [291] argues that reflection is essential for deepening understanding, fostering personal growth, and enhancing problem-solving. Furthermore, reflection aids emotional intelligence [157], promoting self-awareness, better decision-making, and improved relationships. Reflective practice also supports personal growth, and it can enable lifelong learning. It fosters metacognitive skills, which are crucial for self-regulated learning [142, 474]. And finally, in the form of critical thinking, it enables the evaluation of one's assumptions and beliefs [126]. In line with Moon, **the *learning* that reflective thinking can enable is here generally understood as a dynamic process of transforming information and experiences into meaningful insights, knowledge, or skills**. It encompasses experiential and lifelong learning, thereby covering those forms of learning that **extend beyond formal education.**

Challenges of Reflection. However, an important consideration is that people may often not be able to engage in deep reflection. In some situations, it might simply *not be possible*, given the context or time constraints of completing a task. Certain tasks also do *not require* deep reflective thinking (and extensive metacognitive processes) – as it might be the case for many familiar everyday activities. Yet, in other situations when someone needs to work on an open-ended and possibly ill-defined task that would require or clearly benefit from deeper reflection, people might just not be able to do so, as it can be challenging to engage in meaningful, sufficiently 'deep', and 'productive' reflection [143, 390]. Schön [373] underscores that reflective practice requires a supportive environment and structured opportunities for reflection – which might not always be available. Furthermore, without any support, individuals tend to find it difficult to engage in deep reflection that successfully transforms their understanding of something [51, 169, 232, 291, 373, 390]. For example, it cannot be assumed that people just engage in reflection by presenting them with some data – such as about themselves and their behaviour [36, 390]. Reflection often requires guidance, which can be provided through various means. One example to provide this guidance is through scaffolding prompts that encourage individuals to (i) consider different aspects of their experiences or the material they work with as part of a task and (ii) get them to think about what they are doing and how they could proceed [291]. If reflection-in-action is to be encouraged, Schön [372:102] notes that the **scaffolds need to be provided "in the midst of a task" when people might be stuck, for example**. Several technologies have been designed to support and enhance reflection by offering such guidance as part of/in a task. NLUIs, in particular, seem to have promise to engage users in reflective dialogues, prompting them to think critically about their experiences and thoughts (e.g. [37, 141, 197, 268, 329]). The next section provides an overview of the literature on NLUIs that have attempted to achieve this.

2.2.2 Facilitating and Supporting Reflective Thinking With NLUIs

One way to facilitate reflection is to ask questions that can scaffold someone's reflective thought process [143]. Within educational and learning sciences, for example, a range of techniques and approaches have been proposed that teachers/tutors can use to scaffold students' (metacognitive) reflection and sensemaking, which are often based on question prompts (e.g. [40, 324, 393]). One aim of these prompts is to help students reflect on their strategies and explore different approaches to perform an open-ended task. Inspired by these approaches, a range of NLUIs have been developed, which intend to prompt people in similar ways. These NLUIs generally aim to support different forms of reflection – covering both introspection and critical thinking about the (learning) materials a person engages with. First, NLUIs that intend to facilitate reflection to support learning and understanding are presented, followed by NLUIs that target specific forms of reflection that are involved in metacognitive processes.

NLUIs Facilitating Reflection for Learning

NLUIs have received particular attention within the realms of learning and education [107, 148, 160–162, 260, 356, 356, 385, 392, 392, 395, 449, 451–453]. For example, conversational tutoring systems like *AutoTutor* were found to produce significant learning gains [159, 310, 422]. They consist of an avatar (the 'tutor') that speaks, a graphical interface in which the user completes a learning task, and a chatbot-like interface that shows what the avatar has said and where the users can provide their input. Some of the key 'moves' which the agent supports are the following: asking questions about the topic at hand, providing hints (until the learner provides a correct or acceptable answer), correcting students' answers, and providing feedback. The AutoTutor systems have been mostly designed to support individual users' learning of topics like computer literacy or physics. The studies of it being used by students have shown how it can help them learn about a specific topic by motivating and guiding them [159].

Asking learners questions is key to sparking curiosity and scaffolding reflection and sensemaking [40, 412]. Furthermore, enabling students to formulate their own questions can increase their learning performance [101, 221, 222]. Research has thus also investigated how question-asking agents could support learners. For example, Alaimi et al. [6] investigated how different types of agents can encourage children to formulate questions and Ceha et al. [71] found that question-asking robots can be successful at enhancing students' curiosity about a topic. Questions asked included: *"I am curious. Do the holes form when gas bubbles get trapped when the lava cools? Do you have any idea?"*; The findings of their study also showed that curiosity can be 'contagious', as the robot which verbally expressed curiosity was able to influence the participants' curiosity and got them to ask more questions themselves.

There has also been research on using chatbots for different learning contexts and activities as part of online learning (e.g., [260]). Educational chatbots have been found to improve communication while simplifying learning interactions (e.g., [353]). Winkler et al. [448] showed that in the context of online lectures, a conversational agent, which scaffolded learners' sensemaking, had more positive effects on learning compared with an agent that did not. The scaffolding NLUI asked questions at specific points about the content learned in the lecture. If a learner's answer to the agent's question was incorrect, it would follow up with sub-dialogues that stated the problem and question in different ways, scaffolding the learner's thinking and guiding them to the correct answer. Jung et al. [203] designed an NLUI that facilitated children's reflection as they designed mechatronics systems. The agent asked open-ended questions that intended to stimulate a dialogue between the child and their mechatronic artefact, which was found to be successful at guiding the process. Wambsganss et al. [431] developed an NLUI that provides feedback on students' argumentation while doing a persuasive writing exercise. It was found to help produce texts with higher *formal* and with a higher *perceived* quality of argumentation compared to traditional non-conversational tools.

Taken together, this research suggests that NLUIs are effective at guiding and scaffolding learning and sensemaking activities by giving the student/user feedback at specific points during an activity for things they could consider, try out, reflect on, and how they could adjust their approach to completing a task. These scaffolds from the NLUI supported students in proceeding with their task, employing more effective strategies to perform it, approaching possible solutions to a problem or exercise, and gaining new insights.

NLUIs Facilitating Reflection for Metacognitive Processes

NLUIs have also been designed to foster metacognitive skills – in computer-based learning environments and learning platforms – where they are used to support self-regulated learning through metacognitive strategies [360, 382], which was found to have positive effects on learning outcomes [243, 288, 360]. Similarly, Song et al. [395] developed an NLUI that was successful at getting learners to reflect on their progress – a key metacognitive activity while learning. Ramachandran et al. [336] found in their study with a tutoring robot that engaging in think-aloud – which can be used as a metacognitive strategy – can lead to improved learning outcomes (in solving selected problems from the mathematics syllabus). Interestingly, however, they found that those students who were *prompted* by the robot to think aloud outperformed the control group who engaged in think-aloud *on their own*. The authors hypothesised that this was caused by the social presence of the robot, which stimulated natural engagement in thinking aloud. Another example within an educational setting is *Muse* by Cabales [65], which prompted students to monitor and reflect on their learning strategies while working on research projects. In a user study, the agent's scaffolding of metacognitive reflection appeared to have helped students think more deeply about their studying process and apply beneficial learning strategies, indicating possible metacognitive behaviour change. Kim et al. [219] developed an agent that asks users to engage in metacognitive activities to decide if a post on social media is to be trusted or if it might be fake news. To get the user to stop and think, they used a 'pause and reflect' strategy [24, 134]. However, Kim et al. did not observe any statistically significant differences in terms of participants' activities and their accuracy in identifying fake news. Another finding was that certain metacognitive strategies can lead to forced cognitive engagement and be perceived as tiresome.

This body of research shows the diverse uses of NLUIs for learning and metacognition, where NLUIs have been successfully used to guide learners, support question-asking and argumentation skills and facilitate metacognitive processes, for example, to help them adhere to more beneficial (learning) strategies. However, the research also highlights that despite the promise of NLUIs to support reflection, there are still many open questions for how to best design them for a wider range of cognitive tasks.

2.2.3 Summary

As this second subsection of the literature review has discussed, there are many ways in which reflective thinking can benefit cognitive tasks, in particular, if they involve or rely on sensemaking to make better decisions or build an understanding of a situation, a problem, or oneself. NLUIs seem to be particularly promising to support reflection, as they can ask scaffolding questions to the human user – similar to how a tutor or counsellor might do. Beneficial for this also that people tend to perceive an NLUI as some form of social entity which they may often feel more compelled to respond to. At the same time, they might feel less pressure and thus be more exploratory in their reflection compared to being 'questioned' by another human who might judge them. The question this leads to is whether NLUIs could thus augment human cognition in new ways. While technological tools often intend to achieve this augmentation by enabling people to externalise and offload their cognition, what further opportunities are there to do so by enabling people to engage in reflective thinking that benefits their ongoing activities? Although research has shown that NLUIs can support reflective thinking in certain tasks, such as learning, there is less evidence for its efficacy for other sensemaking or decision-making tasks, which the research presented in this thesis focuses on. The next section describes the methods and approaches that were used to conduct this research.

3. Methodology

The goal of the research was to investigate how NLUIs can augment people's cognition in the tasks they perform – in particular, by enabling reflective thinking. The studies conducted were designed specifically to answer the research questions set in the first chapter by examining how NLUIs could be designed to facilitate different kinds of reflective thinking in a range of different scenarios and cognitive tasks. The reason for exploring a large space rather than focusing on just one kind of cognitive task was to determine more broadly how reflective thinking could be facilitated by NLUIs across different contexts. Demonstrating how reflection can be achieved across different settings using different kinds of prompts enabled empirical evidence to be obtained that can arguably be more generalisable.

The first study focused on how proactive interventions can contribute to collaborative sensemaking, as well as on the different effects different interaction modalities can have on people's reflective thinking. The setting was a controlled lab study that employed an experimental design with mixed methods for data analysis. The main variables were the number of requests made to the NLUI, the number of available visualisations explored, as well as the frequency of turn-taking in participants' conversations, and the number of questions they asked each other about the dataset. The qualitative analysis of the interviews following the task investigated people's experiences of being prompted by the NLUI. Exploratory data analysis was chosen as a task, as being able to make sense of a dataset and formulate questions and hypotheses is a key activity and skill for many people who need to work with and analyse data (e.g. a large number of knowledge workers). Thus, if such sensemaking tasks could be effectively supported through 'cognitive prompts', designing similar interfaces could potentially benefit many people when performing comparable analytical activities.

The second study focused on how the capabilities of *cognitive prompting* could be extended to more complex decision-making activities – for which the domain of investment decision-making was chosen. The reason for choosing this task was that it shares many characteristics of other complex decision-making tasks where different types of information need to be compared and weighed against each other, for which both heuristics as well as more sophisticated and structured decision-making approaches may be used. Another benefit of focusing on investment decision-making is that there is a large body of research available on the heuristics that they often involve, as well as the potential undesirable outcomes and biases they can lead to. This existing research was used to inform the design of the NLUI and the questions it asked. Given the exploratory nature of this study, which was mainly interested in *how* this can be achieved, an important contribution was to explore how an NLUI and its prompts could be best designed for this purpose, which was then qualitatively evaluated. The evaluation followed a technology probe approach [192] where participants interacted with a simulated stock trading platform in which the NLUI was integrated which was then followed by an interview.

Building upon the insights of the second study on using proactive NLUIs to get people to reflect on their decision-making, the third study explored how NLUIs can support people in reflecting on a past personal challenge. The research conducted focused on how the NLUI could be embedded in a software tool used to perform a creative task, which in this case was a VR tool. After having explored the integration of NLUIs into a shared interface in the first study and a more common graphical user interface of a computer application in the second study, VR seemed to be a promising interface to explore, particularly for the present creative task. The main reason was that VR has been found to be effective for supporting similar tasks of self-expression, as it provides a personal and immersive environment with various means and possibilities for creation (e.g. [428, 429]). The focus of this study was on how the experience with the NLUI's scaffolding prompts can support people in performing the activity of expressing and reflecting on a past challenge and what could be learned from it (about themselves). The study was conducted as a controlled lab study using an 'interventional' design with pre and post-measurements concluded by an interview. The key metrics of interest were participants' self-efficacy and affect before and after using the 'NLUI+VR' tool, as well as their experience of using it and the extent to which it got them to reflect.

A range of different methods were used for the different designs of the prototypes (Section 3.1) and the study designs (Section 3.2), which will be further elaborated in the following sections.

3.1 NLUI Design

To design the different NLUIs a range of methods were used. The methods differed depending on the characteristics of the task supported by the NLUI and the users performing it. The aim was to provide proactive 'co-piloting' specific to each task by asking the user questions about what they are currently doing and thinking and how they might proceed. Of course, for many tasks, it may not be needed or desired to have a system ask users questions about their thought process as this might introduce too much friction and disruption – for example, for more well-defined or familiar/routine tasks. However, as elaborated in the literature review, in particular in Sections 2.2.2, there are many other tasks where some of this 'friction' can provoke 'deeper thinking' and can lead to more deliberate, reflected, or motivated actions and decisions, as well as the exploration of new perspectives, possibilities, or alternatives [93] – for example in open-ended and possibly ill-defined analytical, decision-making, or creative tasks.

What all the chosen tasks of the three main studies had in common was that they all involved sensemaking in different forms. Sensemaking can be defined as "the process of forming and working with meaningful representations in order to facilitate insight and subsequent intelligent action" [323]. In the present studies, this involved making sense of a dataset, making sense of various information in order to make a decision, making sense of how a decision is made, and making sense of a past personal experience by visually representing it – all with the aim to gain new insights into the task material or oneself. To make sense of something and gain insights, reflective thinking is generally beneficial or even required – which all the NLUIs were thus designed to support. In what follows, key considerations in the design of the NLUIs will be described, including single versus multi-user interaction, when to intervene, and which tasks to support.

Single or multi-user activities. The first relevant categorisation of tasks is if they are performed by one person only (single user) or by multiple people (multi-user). For example, someone can make a decision alone or in collaboration with others. Both pose their own challenges and opportunities for being facilitated through proactive NLUIs. For example, when designing for a multi-user scenario, there is the advantage that cognitive processes and the progression of the task might be more 'observable' from the outside (i.e. because people verbalise their thoughts, see for example [13]), and there is already an ongoing conversation which the NLUI can ‘join’. However, there is also the risk that the intervention could disrupt an ongoing conversation or collaborative interaction. On the other hand, when designing for a single-user scenario, there is the challenge that most of the thoughts might be less ‘observable’ (unless the person engages in think-aloud), but there is at least less of a risk of disrupting other ongoing collaborative activities. In the first study presented here, an NLUI was developed for a collaborative task. In the subsequent studies, NLUIs were developed for single users so that both aspects could be explored.

When to intervene. An NLUI may not only intervene at *different moments,* but it may also do so in *different ways* depending on the person’s needs. Concerning the former, the prompts that were developed for the studies were either delivered (i) based on people’s activity or inactivity during the task or (ii) when specific actions of the task were performed by them (e.g. buying or selling a stock). With the aim that the NLUIs would work for a range of people who may have different needs during the task (e.g. depending on how comfortable or experienced they are with expressing themselves), the interventions were designed and phrased in a way that they could be useful for different skill levels and/or giving people the possibility to request or skip/ignore certain prompts.

Which tasks to support. To identify tasks that may be augmented by an NLUI intervening proactively several aspects were considered. An aim of this was to identify tasks where there might be specific needs which the NLUI could address by proactively providing cognitive scaffolds. However, as mentioned previously, one of the main challenges of proactive interventions is that they may interfere with ongoing tasks – depending on the task and/or how open a user is to receive questions. Therefore, several criteria were used based on which the ‘suitability’ of a cognitive task would be determined to make it more likely that proactive interventions could successfully complement people’s thought processes when performing it. These characteristics were based on the idea introduced earlier of proactive NLUIs supporting people by scaffolding their thought processes rather than giving specific recommendations or solutions, as the latter might be more suitable for well-defined or closed-ended tasks.

The criteria for the task characteristics were:

1. It is an open-ended task that benefits from or requires reflection to get the desired insights.
2. Related to that, the task might generally benefit from exploring different approaches, possibilities, and perspectives to lead to a wider range of insights and a broader understanding.
3. There are specific points in the task where people might have difficulties deciding what to do (e.g. due to being overwhelmed with what they should look for, which option to choose, what to do next, etc.).
4. There might be certain best practices for completing the task that are worth considering for the user independent of their level of familiarity/expertise or confidence with the task.

The tasks chosen for the studies of this PhD for which all the above criteria were satisfied in different ways were the following:

1. **Exploratory data analysis**, specifically to speculate on what might be possible reasons for certain patterns in a dataset. The task is characterised by its open-endedness, and that it can sometimes be difficult to draw connections and reflect on what the reasons for certain patterns might be. Furthermore, there can be more nuanced or less salient but equally important patterns that may not be easy to notice. The NLUI/co-pilot prototype that was built for this task was named *VoiceViz*.
2. **Investment decision-making**: Investment decision-making is characterised by its complex nature and how many different types of data (e.g. quantitative metrics versus qualitative reports) from different sources (e.g. company, analyst, or market data) need to be considered, ideally carefully reflected on, and put together into a so-called 'investment thesis' that aligns with one's strategy. The prototype was named *ProberBot*.
3. **Self-expression for self-insight**: This involves creating representations of personal experiences to gain new understandings of them. The creation of such representations was done here through creating virtual 3D representations in VR. A challenge of such a task is to reflect on what the key components are, how these could be expressed, and how one can get insights from the created representation. The prototype was named *SelVReflect.*

The next question was how exactly and at which points the different tasks could be supported. Identifying when which prompts could be needed can be done using many approaches – the following methods were applied in the three studies reported in this thesis:

1. By using existing literature for that type of task – for example, what are the common challenges, pitfalls, and biases that people face according to existing evidence.
2. Involving experts when the literature can only partially inform when and how delivering interventions would be meaningful.
3. Running user studies in which people need to perform the given tasks to identify more specifically where the main challenges of the task are in practice (e.g. in an observation study).
4. Conducting enactment studies with participants to 'emulate' some of the NLUI interactions to explore which proactive interventions might work and which might not.
5. Evaluating different prototypes at different levels of fidelity with participants to adjust when and how prompts are delivered (e.g. in pilot/pre-studies).

How these methods were applied in the specific cases will be described in each chapter in more detail.

The designed cognitive co-pilot prototypes were generally embedded into another tool or interface used to perform the respective task – similar to some of the 'AI co-pilots' introduced earlier. The reason for this was three-fold. Firstly, this would enable the NLUI to directly intervene 'where' the task is performed. Secondly, it would also allow the NLUI to more directly act upon people's activities whilst performing the task (e.g. in the case of the second study, it could intervene when a person is about to make a buy/sell trade). Thirdly, through this kind of 'interface embedding', the participants in the studies would be able to respond to or 'act on' the NLUI's questions by performing certain actions within the interface rather than having to respond through natural language (this was the case for SelVReflect, for example). However, the embedding also meant that further aspects had to be considered regarding how the NLUI would be integrated. For example, how the NLUI should appear, and if it should be in a way that the user needs to respond to it or would it be intervening in a more 'peripheral' way, allowing the user to continue with the task for some time until they might decide to engage with the prompt. How these questions and considerations were addressed will be described in more detail in each chapter.

An important aspect of many NLUIs is that they can enable people to express their thoughts and or feelings through asking questions (see Section 2.2.2). Various NLUIs were built based on this idea, most of which were for educational or wellbeing domains, capitalising on learning requiring such processes of externalisation to make sense of and understand certain topics (e.g. by writing about them) (e.g. [291]) and in the context of wellbeing to express and make sense of certain feelings and experiences to give them (a) meaning (e.g. [267]). Here, the idea of externalising one's thoughts to support sensemaking is applied to a range of tasks. It is assumed that by externalising thoughts in the given tasks, the participants would be able to develop them further and that the NLUI could facilitate this process of externalisation through its questions, which aim to get people to consider and reflect on specific aspects.

Finally, it is important to mention that the prototypes were developed to be partially functioning to enable human behaviour to be observed in the studies and the hypotheses to be tested. VoiceViz used a Wizard-of-Oz paradigm to enable more control of aspects of the interface to be explored, and thus, the NLUI's interventions were controlled by a human experimenter. ProberBot and SelVReflect were functional prototypes that did not have to be controlled by a human, but their functionality was constrained to the given study task (i.e. specific decision-making contexts or the expression of a past challenging experience).

Table 3.1 provides an overview of the main characteristics and features of the three tasks presented in Chapters 4-6, how the NLUIs were designed to support them, and the research methods used. Next, I will cover the methodological approaches used to evaluate the prototypes once the above design steps were completed.

Table 3.1: Overview of the studies and NLUIs designed to support reflective thinking.

	VoiceViz **Chapter 4**	**ProberBot** **Chapter 5**	**SelVReflect** **Chapter 6**
What users reflect on:	The reasons behind certain patterns in data	The relevance of certain information for one's decision-making and the way one makes decisions	A challenging personal experience and how one deals with such challenges & how to represent it
What users express or 'externalise' (in response to the NLUI):	Hypotheses, 'research questions'	Investment theses, motivations, and reasoning behind decisions	A personal experience and its components
Modality of externalisation:	Speech	Text	Visual
Role of the NLUI:	Inspiring new questions and hypotheses to reflect on and explore	Probing person's reasoning and decision-making, eliciting metacognitive reflection	Facilitating self-expression and reflection
Task interface:	Collaborative interface with shared display	Web-interface	VR environment
Interaction:	Multi-user (pairs)	Single-user	Single-user
Research methods:	Mixed methods (experimental)	Qualitative	Mixed methods ('interventional')

3.2 Study Design

As elaborated in the previous section, the prototypes that were developed were novel and hence no comparable tools were more widely available yet, which also had implications for the methodological choices. Rather than evaluating specific features of these NLUIs and how they affect the task performance, the focus here was more on exploring the opportunities of these tools by evaluating the experiences, interactions, and sensemaking they can foster, which are considerations that generally necessitate the application of qualitative methods. More specifically, the studies were designed to answer the following: (i) which of the NLUI's design/interactional characteristics were most relevant for how it was used by participants, (ii) how participants interacted with and responded to the interface encompassing the NLUI with its cognitive prompts, and (iii) how the NLUI supported their reflection process as part of the given task.

Given the challenges of measuring reflective thinking per se [143, 349], mainly qualitative analyses were chosen, focusing on aspects such as participants' exploration and discovery of new perspectives, questions, or ideas, and whether they gathered new insights. For this, participants' deliberations in the interviews that were conducted after completing the specific study task were generally used. Furthermore, their conversations and interactions while completing the task were analysed (e.g. VoiceViz). For this, a form of conversation analysis was used to investigate how the interactions between participants and the NLUI were structured. The interviews were semi-structured, covering questions such as: what they thought of having an NLUI embedded into the interface they were using, their views on the questions it asked them, how the prompts affected the way they completed the task, when and to what extent certain prompts were intrusive and disruptive, and what they took away from the tasks in terms of insights. The interviews of each study were transcribed and then thematically analysed.

Study 1 (VoiceViz) and Study 3 (SelVReflect) also included quantitative metrics so that aspects of user performance could be analysed, and in Study 1, they were compared across conditions. Study 1 investigated the number of questions generated as part of the data exploration (i.e. a proxy for participants' reflective thinking and speculation) and the number of turns between participants (a proxy for participants' engagement in the collaborative exploration). In Study 3 (SelVReflect), in which participants expressed and reflected on a past experience (and what it might mean for the future), aspects like self-efficacy and participants' affective experience were relevant which were measured using validated questionnaires before and after the task to investigate the effects of the SelVReflect experience (using PANAS [435], GSE [375], and DOE-20 [340], and TSRI [42] scales). The quantitative data in VoiceViz and SelVReflect were analysed using descriptive or inference statistics, depending on which was more adequate for a given part of the analysis. Where inference statistics were employed, parametric and non-parametric tests were used depending on the characteristics of the collected data.

Following these three studies, two more studies are reported in Chapter 7, which explored the opportunities and challenges of 'unleashing' the cognitive co-pilots from being embedded into a software tool and specific task to instead 'embedding' them in a range of situations in everyday life. For these final two studies, a *speculative design/design fiction* approach [17] was employed, inspired by vignette studies [4] in psychology. This was used to explore the appropriateness and acceptability of situating proactive agents in a diversity of hypothetical settings. Specifically, a set of fictional scenarios were developed in an iterative design process, for which then storyboards were created. These storyboards were subsequently used in an online survey (Section 7.3) and an interview study (Section 7.4). The main focus of the studies was to explore both quantitatively and qualitatively how people would perceive proactive interventions by NLUIs in terms of their usefulness as well as their desirability in the given context and the ongoing (social) activities. This set of studies was intended to complement the three 'main studies', as it explores a counter-perspective or 'counter-approach' to what activities could be supported by 'cognitive co-pilots' and how they could be designed.

All the studies did not require specific participant samples to take part. Only for Study 2 (ProberBot) was it required that people with sufficient stock investing experience participate, as it would have otherwise been difficult for them to understand the NLUI's prompts. Participants' informed consent was collected for all studies, and they were approved by the UCL Research Ethics Committee, project number: UCLIC/1819/008/RogersProgrammeEthics

In sum, the methodology used for the research and the rationale for the choices made have been presented in this chapter. It has introduced the range of methods that were used for both the design and evaluation of the NLUIs to accommodate for the specific tasks performed by participants and what the tasks involved. The specific methods used and how they were conducted will be described in more detail in the respective chapters – the first one is the VoiceViz study, which is covered in the next chapter.

4. VoiceViz: Fostering Reflective Thinking in a Collaborative Sensemaking Task

The chapter describes a user study conducted with a team of researchers at UCL[7] that was conducted to investigate how proactive prompting through a cognitive co-pilot can support reflective thinking as part of a collaborative exploratory data analysis task by asking people questions. The rationale was that this kind of scaffolding could trigger more reflection (in action) and discussion about the reasons for the trends and patterns in a dataset. Many people can find it difficult to discover and make sense of such trends and patterns – in particular, less obvious ones. Here, the goal was to see how an NLUI could help scaffold and guide their exploration and sensemaking.

NLUI interventions can be presented via voice or text – which can make a difference as to how effective they are. To examine whether the modality in which the NLUI interacts with people makes a difference, the research question posed here was: Does a voice versus screen-based NLUI interaction affect people's behaviour and thought processes? It was hypothesised that voice would be more immediate and emotive, resulting in more 'fluid' conversations and interactions. This is because voice-based NLUI could get 'integrated' into the ongoing conversation while facilitating collaboration – in a different way than a screen-based NLUI can.

[7] Tu Dinh Duong helped with the development of the web app built for the Wizard-of-Oz prototype. Ethen Wood (intern at UCLIC at that time) helped further develop the web app for the study, contributed to running the study (by controlling the prototype), and helped with the preparation of the data for further analysis (e.g. implementing a solution to diarise the transcription). The other members of the project team helped conceptualise and prepare the study and contributed to its write-up.

4.1 Introduction

Many activities people engage in are collaborative and co-located. For co-located activities in professional settings, often shared displays are used to present and/or collaboratively make sense of spreadsheets, diagrams, data visualisations, or other resources and documents. The present study explores how an NLUI that is designed to facilitate the process of sensemaking in an exploratory task could be embedded in a shared interface used to perform the task, in particular with respect to the modality it uses, thereby addressing the following questions: How should an NLUI deliver its prompts and how can they support the sensemaking process? While talking with other humans and using the shared interface, is *talking* with an NLUI more effective compared with interacting with a chatbot/text-based NLUI? If so, is it because speaking aloud can enhance the flow of conversation while facilitating collaboration?

While there has been considerable research investigating the effects on human performance of using speech versus text/screen-based input at the interface (e.g. [38, 100, 114, 179, 247, 248, 259, 470]), and the use of speech versus text/screen-based output (e.g. [46, 250, 301, 334, 434], little is known as to whether using voice or screen-based interactions, when conducting a cognitive task, impacts upon (i) the way human-human interactions progress and (ii) how the humans interact with the NLUI. Furthermore, most of this research has focused on single-user scenarios. Here, it is proposed, from the users' perspective, that voice can be more engaging, sparking curiosity and interactivity, by triggering more questioning and hypothesising during (collaborative) sensemaking.

The aim of the study presented here is to explore how people interact with either voice or screen-based NLUIs when carrying out a collaborative activity during a meeting – where there is more than one person present who will interact with it. Clearly, there are different affordances of interacting with graphical/screen-based interfaces compared with voice interfaces. Thus, the focus here is less on the difference in the screen versus voice modality *per se*, but rather on the *type of interactions* these interfaces require or enable. A screen-based interface requires reading information (e.g. the NLUI's text messages/prompts) and performing manual interactions (using a mouse or a touchscreen or typing on a keyboard). It has the advantage that users can decide when they want to process/read the NLUI's prompt. On the other hand, if the interface is voice-based, people in the meeting do not have to switch modalities when addressing the NLUI or being prompted by the NLUI during an ongoing conversation with other people. With a voice-based interaction, requests to the NLUI can be made as part of the conversation, and NLUI prompts will need to be processed/listened to immediately when they are provided, since speech requires immediate attention. Furthermore, talking and listening are well-honed skills that we employ when holding a conversation in the company of others. In contrast, selecting from menus or typing text in at an interface, via a touchscreen or keyboard, is more indirect and something usually done by one user. The text as it appears on the screen may or may not be read and may require being read aloud by one in a meeting to let the other know they are reading it or for them to listen. Hence, it could be that a voice-based NLUI is a better match when the NLUI is intended to be part of a group setting, as it may be more effectively embedded into an ongoing conversation between groups of people.

In co-located group settings, users can be engaged with the system in different ways. There are situations where only one person interacts with a device (e.g. a computer) and the others are only observing the device's output, such as in a presentation, talk or demo. Then there are settings where users are co-located around a device and interact with it simultaneously; depending on the interface, the interactions can be in parallel (e.g. tabletop) or need to be coordinated sequentially (e.g. computer). In the present study, the focus is on co-located simultaneous interactions with an NLUI-enabled data analytics interface, comparing two modalities – screen and voice-based interactions. While there is a variety of ways the human-computer interaction could be configured (e.g. the NLUI's output is provided through voice while the users' input is screen-based or vice versa), the focus of this study was to compare two combinations that were considered most natural and appropriate for use with an ongoing collaborative activity – voice-based input *and* output or screen-based input *and* output.

A further aim was to investigate the differences in the way conversations unfold and progress in a group setting when having voice versus screen-based interactions with a software tool which 'incorporates' an NLUI. The NLUI was designed to play a particular role which was to 'scaffold' and facilitate the participants' activities, prompting them and giving them ideas for what to do next when progressing a task. The objective was to determine how pairs of users interacted with and responded to an NLUI for the two conditions when trying to understand

and make inferences in a data analysis task where they had to explore and make sense of a set of time series graphs. Exploratory data analysis was chosen as the task, not only because it is an important task (for knowledge workers) across domains and industries but also because many people find it difficult to make sense of the various trends and patterns as to what they signify (e.g. [343, 344]). Prompting could provide them with a way of focussing on specific features in the data to consider in their exploration and sensemaking.

While the modality of interacting with the NLUI-enabled system varied between the conditions, it was kept constant for the main task materials, so that in both conditions participants analysed time series graphs. The reasons for this were three-fold, (i) the focus was the interactions with the NLUI and not the task material itself, (ii) it was considered more ecologically valid as most computer-supported tasks, including data analysis, are usually screen-based, and (iii) it appears challenging to design a task for which the modality could also be varied (in addition to varying the modality of NLUI interaction) without introducing additional confounds.

Separating the input/output modality from the ongoing activity taking place at the graphical user interface (i.e. voice-based input and output combined with a visual data analysis task), could also enable a more natural and free-flowing conversation about the activity at hand. However, there is also the possibility that the users may ignore (or miss) what a voice NLUI has said or find it irritating when it interrupts their interactions (e.g. [280, 286, 306, 440]). Conversely, interacting with an NLUI in the same modality as the ongoing sensemaking task (i.e. using visual screen-based input and output combined with a visual data analysis task) means that the users can decide when to read the NLUI's messages, although it is a more indirect form of interaction that may not integrate as well with ongoing human conversations.

The pairs of participants were asked to discuss and make inferences about a series of data visualisations that were presented on a shared digital display. For this purpose, a prototype called *Vizzy Analytics* was developed, which uses a Wizard of Oz paradigm [325, 352], to 'mimic' an NLUI that proactively intervenes and makes suggestions. It does this by prompting the pairs at certain times, as to what they can look at in the visualisations, thus playing a role similar to a facilitator. The prompts are in the form of questions about certain trends and patterns in the data, which pairs could then reflect on while discussing and making sense of the visualisations – thereby aiming to enable *reflection-in-action* [373]. The NLUI also acts upon their voice or screen-based requests/commands to select different visualisations for them to look at and compare. To assess the differences in the interactions that took place between the participants and between participants and the NLUI, both quantitative and qualitative data were collected and analysed. Quantitative analyses included the amount of interaction with the system as well as interactions between participants themselves, where the latter included conversational turn-taking and questions participants asked each other. Qualitative analyses involved investigating how participants responded to NLUI prompts, which was done via conversation analysis of transcribed segments (Section 4.5.3), as well as an analysis of participants' reflections on the experience of doing the analysis task with the tool/NLUI, which were collected in post-study interviews (Section 4.5.4). The findings showed that there were differences between the two conditions in certain kinds of task-related behaviours and interactions.

4.2 Background and Related Work

As described in Section 2.1.4, NLUIs are beginning to be integrated with other technologies and software applications [22, 23, 95, 183, 246, 383, 395, 442, 448] – recently most notably in the form of AI co-pilots introduced into a range of products from Microsoft, Google, Adobe, Salesforce, or SAP, among others. The aim is to help with searching (e.g. Bing Chat), and with everyday work-related tasks, such as analysing data in a spreadsheet, writing or summarising texts or presentations (e.g. *Microsoft Copilot* or *Gemini for Google Workspace*). Other possibilities include realising creative ideas or providing design inspirations (e.g. *Adobe Firefly*) and providing insights into business-related data (e.g. Salesforce's *Einstein Copilot*).

Hence, NLUIs/co-pilots have much potential to guide and support users in conducting complex tasks at the user interface. One such task that is receiving increasingly more attention is data analytics – examples are *Microsoft Copilot* combined with *Power Query* in Excel, Tableau's *Ask Data*, SAP's *Analytics Cloud Conversational Analytics* or *Just Ask* features, or even data analysis libraries containing generative AI capabilities such as *PandasAI*[8]. A benefit of

[8] https://github.com/gventuri/pandas-ai

being able to interact with data analytics tools using natural language is to make it easier for users to understand what the outputs and visualisations mean or to intuitively express data-related questions [379]. For example, Jordan et al. designed a voice-based NLUI through which people could answer questions about the National Council elections with text or graphs [202], finding that a *"human communication style with a bit of wit motivates users to continue asking questions"*). Another example is *DataTone* [150] by Gao et al., which allows the user to type or speak queries when exploring databases or spreadsheets. Tabalba [406] et al. also presented an always-listening NLUI as part of a data visualisation system that aims to support collaborative exploration of datasets. Besides generating visualisations which the users request, the system would also proactively generate visualisations based on what users are discussing. This feature was considered helpful by participants also because it would generate chart types that the participants had never seen before, showing them new or different ways to visualise the data. The systems show some of the opportunities there are in designing NLUIs for data analytics – encouraging users to ask questions about the data and discovering things they might not have seen without it.

When considering how an NLUI should appear and how it should interact with users, it raises the question of what might be the optimal way for users to interact with it (see Section 2.1.3), especially when it is intended to be used in a particular setting, such as a meeting, and in combination with using a software tool to perform a given task. What parts of human communication should the NLUI be programmed to mediate and how best should an NLUI reply – in terms of suggesting, augmenting, or other ways of responding that help towards achieving an outcome (e.g. [176])? Another important question is how best to support multi-user situations, where these challenges can become even more pronounced, as the NLUI might interfere with other ongoing human-human interactions. Yet, most research has focused on single-party scenarios, and it is unclear how the findings would apply to multi-user scenarios.

4.2.1 NLUIs in Multi-User and Collaborative Settings

Research into how voice-based NLUIs are appropriated in multi-user domestic settings – where they have become popular in the form of smart speakers – has shown how they can mediate social situations, facilitating various kinds of social bonding and family interactions

[39]. For example, Porcheron et al. [326] observed how interactions with an *Amazon Echo* in a family setting were seamlessly interwoven with other ongoing activities at family mealtimes where parents were at the same time trying to get their child to eat their food. They also point out how our conversations with each other and voice-assisted technologies interleave in nuanced ways, rather than being separate conversations within the family or between the family and the device, that switch smoothly from one to another. However, another study investigating how voice assistants were used in multi-party conversations showed that sometimes interactions with Alexa in a social setting can be awkward, disrupting the flow of normal human social interaction [331]. Family members sometimes needed to repeat and refine queries, which were not understood by the voice assistants; other times, they had to enforce silence so that the assistants could better understand their queries. While some of these challenges may be due to limitations of the devices' NLP capabilities, which continuously improve with technological progress (e.g. in particular with LLMs more recently), the challenge of what *role* the NLUI can or should play in a multi-user setting is (and will likely continue to be) an important research challenge.

Beyond the domestic settings, several studies show that there are various opportunities for embedding NLUIs into multi-party interactions [20, 115, 401, 449]. For example, Kim et al. [218] found that adding chatbots to group chats as facilitators can promote diverse discussions. Their NLUI aimed to encourage members to participate evenly and organise members' opinions among others. Their findings revealed that their bot enabled more diversity in opinions and that it could encourage more even participation. Similarly, Tegos and Demetriadis [410] found that NLUIs can trigger dialogues between students in online discussions by intervening in a conversation, which substantially improved both individual and group learning outcomes. More specifically, the group of participants using the NLUI outperformed the control group in understanding and illustrating conceptual domain knowledge following the learning activity. Goda et al. [156] designed a chatbot which learners of English as a foreign language interacted with before joining a group discussion. It would ask Socratic questions like *"Has your opinion been influenced by something or someone?"* or *"What caused you to feel that way?"* or *"Why do you think that?"* to get people to reflect. The findings showed that the chatbot led to more conversations and increased students' awareness of critical thinking. Skov et al. [389] designed *Susa*, an NLUI that took part in workshops and in which it acted as a facilitator. Despite being somewhat limited in its interactional capabilities, such as only providing predetermined answers, Susa was found to help the teams stay on track, and some participants appreciated how it made them reflect before making decisions.

This research suggests that designing NLUIs to take the role of the facilitator to support the human-human interaction, for example, by asking scaffolding questions, can have positive effects on collaboration, promote equal contributions, encourage reflection, and lead to more desirable collaborative outcomes. However, as most studies discussed, the NLUIs also had several limitations; for example, that they could not take an active part in the ongoing conversations and sometimes also disrupted them. The next section will examine further some of the mechanisms of (co-located) collaboration and how they can be scaffolded and supported.

4.2.2 Supporting Conversation and Collaboration

One of the benefits of having voice-based NLUIs playing a role in a co-located group setting is their potential to support human collaboration – through all being able to speak and listen to it (e.g. [449]). To understand which role voice assistants might play needs an understanding of how collaboration takes place between humans and how other technologies have been designed to enhance this. Collaboration usually involves two or more people working together to carry out a task, often being co-located and engaging in conversation with each other, mostly in the form of verbal and non-verbal communication. When talking, each person usually takes turns, with each turn consisting of one or more *turn-constructional unit*s, which is a 'complete' utterance that possibly leads to a *transition relevance place* [359] following which another speaker may take a turn. For example, such *transition relevance places* can be found after questions, which often indicate an 'invitation' to another speaker to respond or to take a turn. Thus, questions are an important 'motor' of conversational interaction and turn-taking – relevant for a wide range of collaborative tasks. One way to promote collaboration is to design a system that supports turn-taking in a group setting, providing invitations for the different parties to take a turn. In particular, the question here is if it is possible to design a system that actively promotes turn-taking (e.g. by question prompts), which can, in turn, encourage more collaboration, for example, enabling ideas to be generated, a problem to be solved, or to learn new forms of cooperative play [465]. Increased turn-taking in a conversation can be related to a higher degree of interactivity in the conversation [79, 119, 175, 284] which can be desirable for tasks that benefit from faster-flowing interactions.

Yuill and Rogers [464] discuss a variety of methods and design considerations that can be used to promote collaboration, including constraining multi-user interactions through the design of the software. Marshall et al. [273] demonstrated that constraining the number of users at a shared interface that can interact at the same time can lead to more turn-taking, collaboration, and articulation of ideas for solving problems. Voice interfaces may, likewise, be designed to encourage turn-taking behaviour, where an NLUI asks a user a question and vice-versa. Hence, similar to how the voice modality can be more immediate in the way it encourages people to step into an interactive narrative role [158], a voice-based NLUI may have a direct impact on how turn-taking takes place in a group: It can take a proactive turn in the ongoing conversation, which a text-based NLUI displaying prompts on a screen cannot. However, little is known about the effect of this kind of directness and immediacy at the interface. On the one hand, they could facilitate and encourage new forms of collaboration by prompting or encouraging users to take turns. On the other, they could interrupt the flow of an ongoing conversation (e.g. [275, 327]) in ways that using a screen-based interface does not. The focus here is on what the effects are on human collaboration when groups interact with an NLUI, using screen-based versus voice interactions, when engaged in an exploratory sensemaking task using a software tool. Can voice interactions with the system also result in an increase in turn-taking between users compared to screen-based interaction? The challenge is deciding how many turns (of question asking) the NLUI should take in order to promote more turn-taking among the human group members and in which way and when it should do so to not interfere and negatively affect the human-human interaction.

4.2.3 Summary and Focus of This Research

To summarise the relevant literature – including the literature covered in Chapter 2 – voice and screen/text-based NLUIs have been found to be effective at stimulating users' reflection as well as guiding and supporting them in different types of tasks, such as collaborative learning or problem-solving tasks (Section 2.2.2). Another line of research has shown (so far, mostly for single-user scenarios) that combining natural language interfaces with software tools can improve the efficiency and experience when completing specific tasks using the tools, such as making it easier for users to express queries in data analytics tools (Section 2.1.4). For the design of NLUIs, an important question is, how the experience and efficiency are affected by the modality of human-NLUI communication, which has been investigated in a range of single-user contexts for different types of tasks with mixed findings (Section 2.1.3). There is a paucity of research that addresses this question in situations of collaborative action, where an NLUI 'takes part in' human-human conversations. In such collaborative settings, previous research has mainly investigated how the interfaces of collaborative systems can be designed to promote turn-taking, without the use of NLUIs (Section 4.2.1). However, there seems to be much potential in using NLUIs to this end (Section 4.2.2), in particular, because of their capability to actively prompt users by intervening in the ongoing conversation and asking them questions.

Building upon and bringing these strands of research together, the present study addresses the following questions: (i) whether a system containing a question-asking NLUI can support and facilitate reflective and exploratory thinking, getting people to speculate and hypothesise in a collaborative analysis and sensemaking task, and (ii) whether the modality of interaction, and more specifically, the modality through which the NLUI's prompts are provided to users – voice or screen-based – makes a difference to the users' question-asking and turn-taking behaviour as well their engagement with the system and the task.

4.3 Aims and Hypotheses

The aim of this study is to investigate the effects of voice versus screen-based modality on interactions between the users and the system as well as between users themselves. The open-ended collaborative task involved exploring and making sense of a set of data visualisations. The visualisations were presented on a large display in front of the participants, who had to reflect on and speculate about the reasons behind the trends and patterns that they show.

The two conditions compared were: (i) using voice requests alongside an NLUI that prompts participants by speaking and (ii) using a screen-based menu to make requests (providing familiar GUI elements, such as checkboxes, buttons, etc.) alongside an NLUI that prompts participants through text messages (being displayed in a 'chatbot-like' way). In other words, the two conditions differ in how participants request visualisations and how they receive NLUI prompts (see Figure 4.1): In the first condition, both system input and output are based on voice, while in the second condition, they are both screen-based. Thus, the second condition (the 'screen condition') mimics more common GUI-based systems. The way data visualisations were presented was the same in both conditions (also see Figure 4.1).

Making voice requests may integrate better with the participants' ongoing conversation and be more fluid than switching to making requests through the graphical user interface. It is thus expected that voice interaction leads to users interacting more with the interface and exploring more of the data, which results in the first hypothesis:

H1 – human-computer interactions: The voice condition will encourage (a) more interactions with the software tool and (b) more of the available data visualisations being looked at.

The metrics of (a) how much the interface is interacted with and (b) to what extent the task material (the available visualisations) is explored are proxy metrics for the users' engagement with both the interface and the task. These engagement variables are important for this type of exploratory task, even if they do not allow for any conclusions concerning the sensemaking and reflection that would occur.

Regarding the impact of the modalities on *human-human interactions*, we expect that certain mechanisms of *social presence* [387] (also see Section 2.1.2) – which could be somewhat stronger for an interface that uses voice due to higher human-likeness (see Section 2.1.3) – may be 'motivating' or 'triggering' participants to also proactively speak and take turns in the conversation by asking questions like the NLUI does. Therefore, it is predicted that when a speaking NLUI prompts users with questions, it is likely to get users to also ask more speculative questions themselves. Furthermore, the increased question-asking may also be due to users' increased curiosity when speaking with the NLUI compared with reading its prompts from a screen (as might be concluded from Ceha et al. [71] and Gonzalez and Gordon [158]). This leads to our second hypothesis:

H2 – human-human interactions: The voice condition will encourage (a) more turn-taking and (b) question-asking between participants.

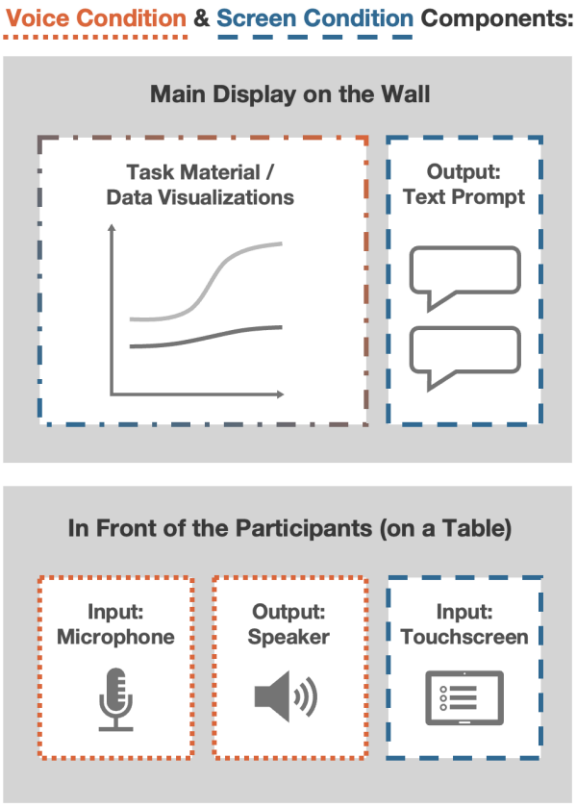

Voice Condition & Screen Condition Components:
Main Display on the Wall
Task Material / Data Visualizations
Output: Text Prompt
In Front of the Participants (on a Table)
Input: Microphone
Output: Speaker
Input: Touchscreen

Figure 4.1: An overview of the modalities used for user input and system/NLUI output in the two conditions; the task material is presented in the same way in both conditions.

How frequently conversation partners take turns (or the speaker alternation rate) is a relevant characteristic of a conversation. It can be considered a proxy measure for the *interactivity* of a conversation [79, 119, 175, 284], and it thus represents a relevant metric in the present collaborative sensemaking scenario. The reason why the number of questions (related to the task/the data visualisations) participants ask each other is used as a metric is not only because of its relevance in conversation and can reflect curiosity but also because it is key for an exploratory analytical task in which speculative hypotheses should be generated [163, 209, 417]. And even more generally, beyond data analysis, question-asking is fundamental for scientific inquiry, reflective and critical thinking, and intellectual exploration [82, 110, 144, 184, 421, 467]. The number of questions thus also serves as a proxy for the extent to which participants try to make sense of the data and for how exploratory and curious they are [6, 40, 71]. Although these metrics can be considered as being related to people's sensemaking and reflection on a meta-level, they will be contextualised and triangulated with a conversation analysis focusing on the patterns in participants' collaborative reflection.

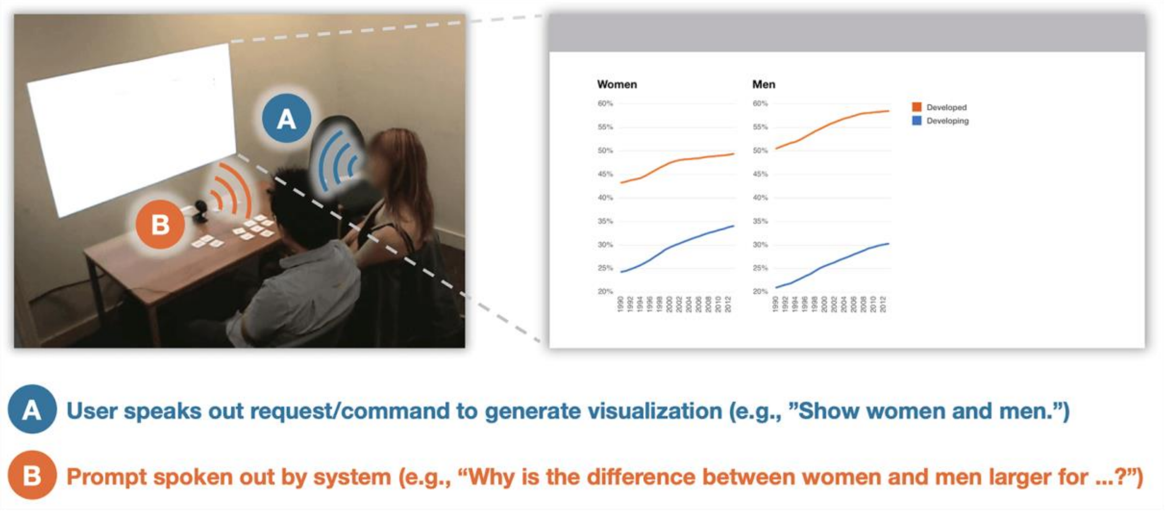
4.4 User Study Design

The study was designed so that pairs of participants could take part together. This enabled the analysis of the conversations and turn-taking that took place between the participants themselves, as well as between the participants and the NLUI. Previous research on 'pair analytics' by Arias-Hernandez et al. [13] has shown how this approach also offers a natural way of making explicit and capturing reasoning processes (in contrast to single-user scenarios) while also enabling a variety of metrics to be used to assess collaboration.

A Wizard of Oz [352:428] paradigm was used to test the two hypotheses. This set-up allows us to both simulate and control the NLUI interventions for both conditions. Wizard of Oz studies have frequently been used to simulate and test novel systems with users, in particular, 'intelligent interfaces' (e.g. Dahlbäck et al. [99] or Porcheron et al. [325]). Our NLUI, which we called Vizzy, was simulated by a human experimenter, who was tasked with triggering the prompts at certain times during the study.

In the voice condition, participants were asked to change the visualisations through voice requests. In the screen condition, they could make the same requests (i.e. through selecting variables/filters) from a menu shown on a tablet. Vizzy was designed to provide prompts in the form of questions in both conditions, either through synthesised speech (see Figure 4.2) or through chatbot-like text messages that appeared on the main screen next to the area where visualisations were displayed (see Figure 4.3). The questions were phrased in a way that they could give participants ideas for what they might also consider or explore next and, by that, aimed to scaffold their reflection process while performing the data analysis task.

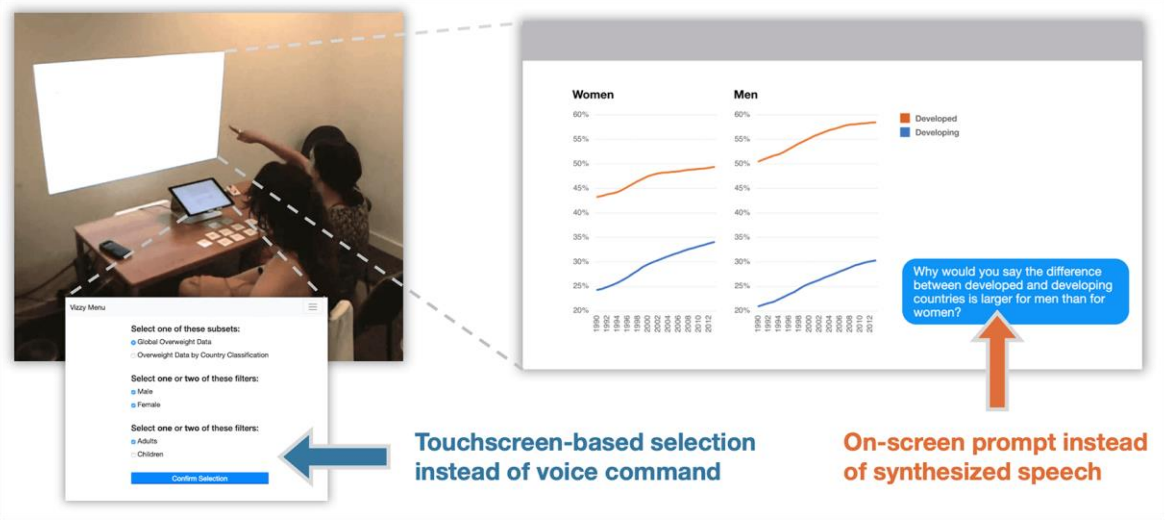
Thus, both conditions were based on examining data visualisations on a large screen; the only difference between conditions was in how requests were given to the system and how the system provided prompts (also see Figure 4.1), which were both either voice-based (voice requests & synthesised speech prompts) or screen-based (requests based on menu selection & text message prompts). The same set of visualisations was available in both conditions. To control for equivalence across the two conditions, Vizzy's prompts spoken aloud in the voice condition or displayed as text messages in the screen condition were selected from the same set of prompts.



Ⓐ User speaks out request/command to generate visualization (e.g., ”Show women and men.”)

Ⓑ Prompt spoken out by system (e.g., “Why is the difference between women and men larger for ...?”)

Figure 4.2: The interfaces for the voice condition with a microphone positioned in the middle of the table. The cards on the table show the set of available visualisations that could be generated.



Select one of these subsets:

● Global Overweight Data

○ Overweight Data by Country Classification

Select one or two of these filters:

☑ Male

☑ Female

Select one or two of these filters:

☑ Adults

□ Children

Why would you say the difference between developed and developing countries is larger for men than for women?

Touchscreen-based selection instead of voice command

On-screen prompt instead of synthesized speech

Figure 4.3: The interfaces for the screen condition with a tablet situated in the middle of the table for user input. The cards on the table show the set of available visualisations that could be generated.

The task. In an initial pilot study, well-defined tasks were tested with specific target outcomes, asking participants to find specific patterns in the data. However, participants’ conversations were short; they focused on searching for the patterns they were asked to discover. For the main study, a more open-ended and exploratory task was chosen that would require people to reflect and speculate. Thus, the sensemaking task we designed involved exploring and interpreting data visualisations based on a set of time series to then speculate on what trends and patterns they show and what might be possible reasons for them – which reflect common activities of exploratory (time series) data analysis. The aim was to enable participants to try to make sense of a set of visualisations by hypothesising about and questioning the underlying data without the need to have a data analytics background. The domain chosen was health, in particular, the prevalence of obesity throughout time, a topic that participants would have some understanding and familiarity with. Specifically, the data represented how the prevalence of obesity has increased in recent decades for different populations. The data used for the visualisations was derived from Marinez [272]. It is publicly available and comes from the Global Burden of Disease Study 2013 [303]. We chose a dataset that covers 24 years from 1990-2013. This period was considered sufficiently historical to enable participants to discuss past developments that could have led to the trends and patterns in the graphs. The visualisations could be generated from the obesity data by combining time series graphs by age (children/adults), gender (male/female), and 'country type' (developed/developing and global). 22 visualisations could be generated, which showed the graphs for adults, children, adults *and* children, men, women, boys, girls, men *and* women, men *and* boys, women *and* girls, boys *and* girls, which could all be displayed as global average or split up into developed and developing countries; see for example Figure 4.4 which shows the averages of developed and developing countries for boys and girls. In the voice condition, this visualisation is generated by saying (*"Vizzy, show boys and girls, developed and developing countries."* or *"Vizzy, show developed and developing for boys and girls."* or similar) and in the screen condition by using the menu on the tablet. The set of available time series to display meant that visualisations which could be generated were simple enough to understand but also sufficiently complex to show different interactions and trends. They comprised a range of level differences as well as changing growth rates that could be explored at a general level (e.g. the overall increase in the time series for boys and girls) or a more detailed level (e.g. how the speed of growth/growth rates changed throughout time) as can be seen in Figure 4.4. A larger set of variables and visualisations (based on the chosen dataset [303]) were tested in a pilot study. The final set of visualisations was limited to 22 so as to make sure that there would be sufficient overlap in the data the pairs look at together and to reduce possible confounds. In a subsequent pilot study, the chosen set of 22 available visualisations proved to be sufficient to allow participants to explore and discover different aspects. The possible interactions between the different time series made the problem space sufficiently complex while not being too overwhelming.

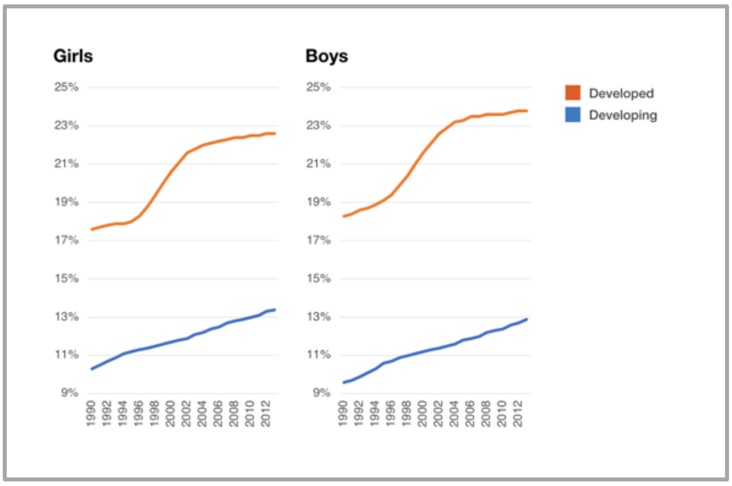


Figure 4.4: An example visualisation of two graphs generated from the task dataset showing the change in the prevalence of obese girls and boys for developed and developing countries from 1990 to 2013.

NLUI prompts. In total 19 prompts were composed for Vizzy consisting of open-ended and more well-defined questions, many of which were applicable to more than one of the available visualisations. They were based on some of the patterns and trends in the data and were aimed at helping participants discover them or take a closer or different look at them. They were inspired by techniques used in teaching to scaffold students' reflection (see also [84, 144, 217, 341, 354, 393] as well as Section 2.2.2). The questions thus generally aimed to 'probe' participants' thinking and give them ideas for certain trends and patterns they could consider and investigate further while making sense of the data visualisations. Examples of the question prompts include:

- *If I would say one of them is slowing down in recent years, which one would you say it is?*
- *Is the increase of one more significant than the other?*
- *What might have caused the sudden spike?*
- *So, if you look at this, would you say that the increase is slowing down in the number of overweight people for all four groups?*
- *Why would you say the difference between developed and developing countries is larger for men than for women?*

4.4.1 Participants

A between-subjects design was used with two conditions: *screen* versus *voice*. 36 participants took part in the study; 9 pairs for each condition. The pairs were randomly assigned to either the voice or screen condition. Instead of matching them up as stranger pairs, we asked the 18 participants we initially recruited to bring someone they knew and who they felt comfortable doing a collaborative task with. This enabled the pairs to feel at ease collaborating with each other during the study. Participants were recruited from UCL and were between 18 and 35 years of age. 18 were female (10 in voice, 8 in screen). In 8 pairs, genders were mixed (4 in voice, 4 in screen). All participants were fluent in English and had normal or corrected-to normal vision and hearing capabilities.

4.4.2 Experimental Set-Up

Physical Room Set-up. The visualisations were projected onto a screen on a wall. A desk and two chairs were positioned in front of it (also see Figure 4.2 and Figure 4.3). This enabled each pair to be able to readily see the projected visualisations while also being able to face and speak to each other. On the desk was placed a set of small cardboard cards (5x5cm) indicating the label or 'title' of each visualisation that could be created (e.g. *"boys and girls, developed and developing countries")*. The cards could be used by participants in both conditions to help them keep track of what they had already looked at or what to look at next (also see Figure 4.2, where participants grouped the cards). In the voice condition, the cards could also be used by participants to help them formulate their requests to Vizzy to show a visualisation (e.g. *"Vizzy, show [card label]")*. In the screen condition, a touchscreen display was placed on the desk, showing a menu to generate the visualisations. In the voice condition, two loudspeakers were positioned behind the desk that Vizzy could be heard through. The voice used for speech output of Vizzy was based on Amazon Polly's[9] voice *Joanna*.

Two webcams were installed in the room to record the participants: one facing down from the ceiling, the other one located behind. The former allowed to capture interactions with the touchscreen/tablet (in the screen condition) and cardboard cards, and the latter allowed to capture participants along with the main screen to see what they point and look at. Each

[9] https://aws.amazon.com/polly/

participant was also asked to wear a Lavalier microphone to record what they said. The video of the participants and the conversations were recorded using *OBS Studio*.

Wizard of Oz Set-up. The visualisations and the NLUI's prompts were controlled through a dedicated computer in an adjacent room, which was connected to the projector in the experimental room. During the study, the Wizard (a second experimenter) listened to the audio stream and observed the video feeds from the two webcams. The Wizard controlled the interface for the two conditions to present the requested visualisations on the screen and select the Vizzy prompts to be played (voice condition) or displayed on the screen next to the visualisation (screen condition).

Vizzy Interface: User Control and System Prompting. In the voice condition, a conference microphone was positioned in the centre of the desk (described to participants to be the microphone through which Vizzy 'listens to' their requests). In the screen condition, the microphone was replaced with a tablet showing the GUI from which they were to select their choices. In order to produce consistent recordings between the conditions, the conference microphone was also present in this condition but hidden under the tablet, not visible to participants. Vizzy was designed to occasionally prompt the participants, intended to encourage them to explore further what was causing the trends and the rise in different obesity levels in the displayed visualisation. For each visualisation prompts could be selected by the experimenter from a predefined set which were applicable to the specific visualisation. Each visualisation was assigned between 1-3 applicable prompts; some prompts were applicable to more than one visualisation. For example, Vizzy (or rather the wizard) could select the prompt *"Would you say that the increase is slowing down for all four groups?"* for all visualisations where there were four time series appearing (e.g. "adults and children" of "developed and developing countries"). Prompts were only provided if (i) participants had not yet discussed the pattern/trend/difference the specific prompt referred to and if then (ii) there was a silence of approximately 3 seconds or more in their conversation to avoid interruptions of the pair's discussion. This threshold was set based on the iterative design process of the system, where after testing different durations with test participants, approximately 3 seconds was considered most appropriate. The reasons for this were two-fold; it was (a) long enough that, in most cases, there was no direct interference with an ongoing conversational turn by one of the participants and (b) short enough that there were still sufficient opportunities for the NLUI to intervene. However, both conditions mentioned above (i and ii) had to be met for a prompt to be provided; a moment of silence itself did not lead to a prompt being triggered. Hence, which prompts could be provided for which visualisation depended on what pairs discussed until there was the first silence in their conversation about a specific visualisation. Furthermore, providing prompts only every two minutes on average was found to be optimal in the pilot studies, as when the intervals between the prompts are too short, they can become annoying and disruptive to the flow of the discussion. The frequency of NLUI prompts was kept as similar as possible in both conditions. In the screen condition, there was a 'clicking' notification sound (similar to sounds used in messaging apps) so that participants would not miss a prompt. To ensure consistency across the two conditions, the Wizard spent considerable time familiarising themselves with the set of prompts, practising selecting different ones for the different stages of the task and the types of visualisations being looked at before commencing the study, following the above rules and guidelines.

4.4.3 Procedure

Ethics approval was obtained from UCL (UCLIC/1819/008/RogersProgrammeEthics) prior to the study. Pairs of participants were informed about the purpose of the study and asked to fill in a consent form agreeing to being audio and video recorded during the study for subsequent analysis.

The participants were informed that they would be asked to collaborate in an exploratory data analysis task. They were told that there was no right or wrong way to do the task and that they should just try to discover and reflect on interesting trends and patterns and speculate on what might have caused them. They were further instructed that they could press a button on a remote control on the table when they had completed the task or if they had a problem during the task. They were informed that Vizzy would prompt them at certain times during the study. They were further told that they could decide for themselves if they wanted to respond to Vizzy's prompts and to use them as prompts to guide their thinking. They were then given the set of cards that showed the possible visualisations they could generate. They were told that cards have the purpose of giving them an overview of the available visualisations that could be requested. In the screen condition, the interface on the tablet was introduced to them and how they could use it to request visualisations. In the voice condition, they were shown how they could request a visualisation (by saying "Vizzy" followed by the visualisation request) and how they could also just use the cards to help them formulate their requests ("Vizzy" followed by the label of the card). Finally, they were informed that the task would normally take about 15-20 minutes.

The instructions were identical for both conditions except from when describing how to interact with the voice and screen-based interfaces. After the introduction, the experimenter left the room, and the participants commenced the task. After completing the task, the experimenter returned to the room and conducted a semi-structured interview with the pair, asking them to reflect on their experiences during the study. Participants were then debriefed about the aim of the study and that it was a Wizard of Oz design. Then, the Wizard, who was controlling Vizzy, came into the experimental room and introduced himself. The participants were each compensated with a £15 Amazon voucher.

4.4.4 Data Analysis

The video and audio data collected were analysed using a combination of *automated* speaker diarisation and transcription as well as *manual* transcription and analysis of conversations, questions asked and turns taken. Both quantitative and qualitative analyses were conducted to test the hypotheses and provide insights into the nature of the discussions that took place in the two conditions.

Quantitative Analysis: Interactions Between Participants and the System

To test H1 regarding the human-computer interactions (The voice condition will encourage (a) more interactions with the software tool and (b) more of the available data visualisations being looked at), the corresponding user interactions were broken down into (a) visualisation requests, measuring the total number of requests that were made for visualisations, (b) visualisations explored, which was determined by how many unique visualisations (of the set of available visualisations) the participants looked at together. For example, out of four possible visualisations (A, B, C, D) if three are looked at (e.g. A, B, C), it would provide a measure of ¾ or 75%. When considering the following sample sequence of requesting visualisations (A, B, A, B, C, A), this would result in *visualisation requests* = 6, but *visualisations explored* = 75%, since only 3 out of the 4 available visualisations were looked at (i.e. A, B, C).

In the present scenario, *visualisations explored* was chosen as a metric, since it was considered a good proxy for how extensively the participants examined and discussed the set of available visualisations. The reason for this is that as part of the task, participants were asked to make sense of and discuss each visualisation they requested, which participants also did in most cases. Both measures, therefore, were used to capture how much users interact with the system and the task material.

Quantitative Analysis: Interactions Between the Participants

To test H2 regarding the human-human interactions *(The voice condition will encourage (a) more turn-taking and (b) question-asking between participants.),* we measured (a) the number of speaker changes made during a conversation and (b) the number of questions participants asked each other. Turn-taking was approximated through quantifying the number of speaker changes (see [119, 175]). As these metrics intend to capture participants' behaviour to compare it between conditions, turns by Vizzy itself (when it provided a prompt) were not counted towards these metrics.

Qualitative Analysis: Patterns of Collaboration and Sensemaking

To investigate how the conversations after Vizzy's interventions unfolded, a randomly selected set of segments was transcribed. From these, we examined in more detail the content of the conversations and the extent of turn-taking, as well as how initial ideas and inferences about the data visualisations were followed up. Excerpts are provided to illustrate the patterns of discussions and interactions that took place in Section 4.5.3. In addition, interviews were conducted at the end of the experiment to find out more about the participants' experience of using the system. An analysis of the interviews is provided in Section 4.5.4.

Data Analysis Methods

To achieve sufficiently accurate identification of the active speaker with the given set-up, we developed our own speaker diarisation model, which took the three audio streams (left speaker microphone, centre microphone, right speaker microphone), and compared the intensity values using *Parselmouth* [198]. Based on manually diarised recordings, threshold values for each microphone were defined using *Sequential Model-based Algorithm Configuration (SMAC)* [193]. Based on the thresholds, absolute differences between the microphones were defined to identify the active speaker, which was used to identify *speaker changes*. Each audio segment of the speaker diarisation model was then transcribed using *Azure Cognitive Speech Services*. While the accuracy was not perfect, it was sufficient to then also quantify the *number of words spoken* for each participant.

Vizzy's utterances, along with the visualisations that were requested by the participants, were automatically tracked using a log file, from which the timestamps were extracted. Timestamps were manually recorded after each prompt by Vizzy from the beginning to the end of the participants' discussion about that prompt to record the discussion lengths (if a prompt was ignored by participants, the discussion duration was set to 0).

The number of *questions* asked by each participant pair as part of their discussion was not based on the automatically generated transcription files but manually tagged and coded, since their questions were not always identified with the required accuracy by *Azure Cognitive Speech Services*. To reduce confounds, the questions that were not related to the data analysis task were excluded in the quantitative analyses, for example, questions about Vizzy and its capabilities (e.g. *"Do you think Vizzy can do this?")*. We did this as we hypothesised that the voice modality could lead to more questions about the interface capabilities compared to the screen modality (post-hoc analysis showed that this was indeed the case). Furthermore, in those cases where a participant requested a visualisation by asking a question, which occasionally happened (e.g. *"Vizzy, could you show developed and developing countries?"* instead of *"Vizzy, please show developed and developing countries.")*, this was also not considered for the *participants' questions* metric to avoid confounds.

The duration of the individual sessions varied across participant pairs when exploring the dataset. To control for this, averages per minute were calculated instead of totals for the metrics *requests* (for visualisations), *speaker turns,* and *questions* by participants.

4.5 Findings

This section describes the quantitative findings, the qualitative findings of participants' discussions following a prompt from Vizzy, and the findings from the interview. First, the main quantitative findings will be presented.

4.5.1 Main Quantitative Results

Overall, participants in both conditions looked at most of the available 22 visualisations. A main finding was that the participants in the voice condition interacted more with the system, explored and discussed more of the available visualisations, and asked more questions about the visualisations. Statistical significance was assessed using an alpha level of 0.05 for t-tests and U-tests[10].

We conducted (i) a U-test on the number of visualisations that were looked at and (ii) a t-test on the number of requests made per minute. Both null hypotheses were able to be rejected as the results were found to be significant. In support of H1 (U_{18} = 18, *p* = .017), the percentage of *visualisations explored* was higher in the voice condition (*M* = 96.97%, *SD* = 6.43%) compared with the screen condition (*M* = 89.39%, *SD* = 7.54%) and the difference in *requests* (per minute) to change the visualisations was also found to be significant (*t*(16) = 2.75, *p* = .007), where more requests were made in the voice condition (*M* = 1.08, *SD* = .15) than in the screen condition (*M* = .80, *SD* = .26). These two significant findings therefore support the first hypothesis (H1) that participants in the voice condition would interact more with the system and look at more of the available visualisations. Figure 4.5 shows the number of requests per minute for both conditions. See Table 4.1 below for an overview of all results.

The t-test was also found to be significant for H2 (*t*(16) = 3.51, *p* = .002); participants asked more questions per minute in the voice condition (*M* = .98, *SD* = .22) than in the screen condition (*M* = .55, *SD* = .29) as can also be seen in Figure 4.6. Specifically, participants in the voice condition asked 78% (0.98/0.55) more questions than in the screen condition. A

[10] Based on an examination of box plots, no significant outliers were identified. The *Shapiro-Wilk* statistic indicated that the data was distributed normally (*p* > .05), which was confirmed by examination of histograms as well as skewness and kurtosis values; only for the metric *exploration* this was not the case, which is why a Mann-Whitney U test was conducted here instead of a t-test.

significant difference was found between the two conditions for the number of changes of who spoke at any given time ($t(16) = 2.10$, $p = .026$) approximating the *turns taken*; this happened more often in the voice condition ($M = 6.39$, $SD = 1.11$) than in the screen condition ($M = 5.52$, $SD = .57$) as can also be seen in Figure 4.7.

Post-hoc analyses revealed that there were no relevant differences between conditions regarding how many of the 19 available NLUI prompts were triggered per minute[11].

Table 4.1: Summary of findings: Inferential statistics are *t-tests*, and effect sizes (ES) are *Cohen's d* except for *exploration*, which is based on a *Mann-Whitney U test* and *eta squared*; all *p*-values are significant.

Hypothesis/Analysis Category	Metric	Cond.	Mean	*SD*	Statistic	*p*	ES
Human-Computer Interactions (Hypothesis 1)	***exploration***	voice	96.97	6.43	18.00	.017	0.27
		screen	89.39	7.54			
	requests	voice	1.08	0.15	2.75	.007	-1.30
	(per minute)	screen	0.80	0.26			
Human-Human Interactions (Hypothesis 2)	***questions***	voice	0.98	0.22	3.51	.002	-1.65
	(per minute)	screen	0.55	0.29			
	turns taken	voice	6.39	1.11	2.10	.026	-0.99
	(per minute)	screen	5.52	0.57			

[11] Although the metrics were defined so that they would not be affected if there are differences in how often Vizzy asked questions, we aimed to provide agent questions in a similar way in both conditions. To ensure this was the case, we did a manipulation check, which showed that the number of questions asked per minute was indeed very similar in the voice condition ($M = .44$, $SD = .15$) and text condition ($M = .45$, $SD = .16$).

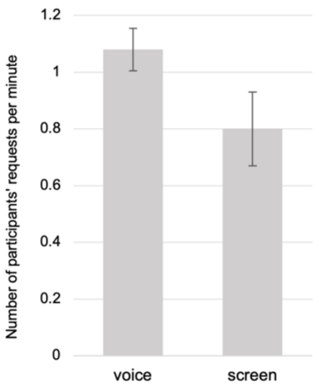


Figure 4.5: Requests made by participants per minute for voice and screen conditions.

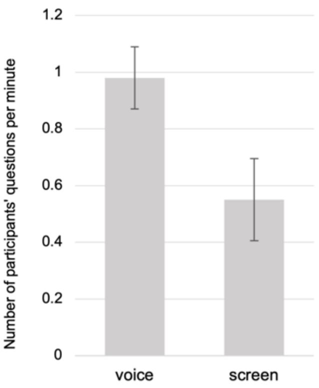


Figure 4.6: Questions asked by participants per minute for voice and screen conditions.

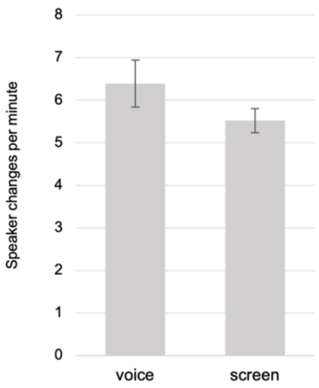


Figure 4.7: Turns taken per minute for voice and screen conditions.

The differences between *voice* and *screen* are significant for all three Figures ($p < 0.05$).

4.5.2 Quantitative Analysis of Types and Patterns of Responses

In addition to testing the two hypotheses, the patterns of responses across the two conditions were examined. In particular, the levels of participation between the pairs in the two conditions were analysed in terms of the activities and contributions of each participant. Overall, there was a tendency towards more equal participation in the voice condition. However, the differences were found to be not statistically significant. For this, we used two measures:

(1) Interactions between each participant and Vizzy Analytics: We calculated the average deviation from *equal contribution* (i.e. that both participants would make 50% of requests). The deviation in percentage points was found to be smaller in the voice condition (*M* = 18.07, *SD* = 13.41) than in the screen condition (*M* = 22.28, *SD* = 15.42), suggesting that the pairs interacted with the system 'more equally' in the voice condition.

(2) Interactions between participants: For *total words spoken*, the deviation in percentage points from *equal contribution* in the voice condition (*M* = 4.38, *SD* = 2.78) was also found to be smaller compared to the screen condition (*M* = 7.52, *SD* = 8.51). Similarly, for *total duration of speech*, the deviation from *equal contribution* in voice was (*M* = 5.36, *SD* = 5.37) slightly smaller than in the screen condition (*M* = 6.63, *SD* = 3.49). It is worth noting that there may be multiple factors

related to the tendency towards more balanced interactions in the voice condition. For example, in the screen condition, it often was the same participant who read Vizzy's prompts out loud, which may have somewhat affected the above metrics. However, these factors will not be considered in more detail here, since both (1) and (2) were not the primary focus of this study but rather an adjunct to our main analyses.

The duration of the discussions in response to Vizzy's prompts was found to be somewhat shorter in the voice condition (*M* = 35.10s, *SD* = 17.39) than in the screen condition (*M* = 47.67s, *SD* = 24.85). However, there was no significant difference. To examine further the patterns of conversation, we subsequently analysed how the pairs responded to Vizzy's prompts in terms of (i) the percentage of prompts that were responded to versus ignored by the pairs in the two conditions, (ii) how long they took before responding, and (iii) the length of their conversations. These analyses were conducted to better understand the general process/pattern of how participants interacted with the system and responded to its prompts regardless of the condition.

(i) Prompts responded to versus those ignored across the two conditions. Nearly all pairs responded to Vizzy's prompts in both conditions. Only 6.90% of its prompts were ignored. Pairs in the voice condition ignored fewer prompts (*M* = 4.56%, *SD* = 9.91) than in the screen condition (*M* = 9.30%, *SD* = 11.39).

(ii) Delay in responding to Vizzy. The average time taken by a pair to react/respond to a Vizzy prompt was roughly 4 seconds for both conditions (*M* = 3.85, *SD* = 1.70). As the screen-based prompts did not have to be attended to and considered by pairs immediately, we were interested in whether the time to respond was longer. It was found to be slightly longer in the screen condition (*M* = 4.19, *SD* = 1.69) than in the voice condition (*M* = 3.52, *SD* = 1.73).

(iii) Duration of responses after Vizzy prompts. We classified the discussions in response to Vizzy's prompts into what we defined as 'short' or 'long' responses. Short responses generally lasted up to 20 seconds (they were generally at least 5s long) and made up 20.90% of the total conversations – in these cases, pairs usually just agreed on a possible answer without discussing it further. However, there were far more longer responses across both conditions, comprising 72.20% (*SD* = 19.01). They generally lasted between 21-90 seconds, but in some

cases even more. The long responses usually consisted of pairs talking about the patterns they saw being depicted in the graph data, followed by hypothesising about the possible reasons for this. There was only a small difference in the number of long responses across the two conditions, with an average of 74.46% (*SD* = 18.99%) in the voice condition and of 69.95% (*SD* = 19.90%) in the screen condition.

4.5.3 Qualitative Analysis of the Participants' Discussions Following Vizzy's Interventions

The quantitative findings showed significant differences between the two conditions in terms of the number and kinds of interactions among participants themselves and with the NLUI. In both conditions, Vizzy's prompts acted as facilitators for participants' conversations, triggering them to talk about the possible reasons behind the changes in the obesity data for the different demographics. Here, we are interested in examining further the reflective thinking that occurred following Vizzy's prompts and to see if there were any differences in terms of what the pair said and did next. To do this, we carried out an in-depth conversation analysis of 4 conversation segments before and after one of Vizzy's prompts. We present here two randomly selected prompts by Vizzy for which we transcribed the participants' conversations before and after Vizzy asked each prompt in both conditions (voice and screen). Each of the transcribed segments was then analysed with respect to the patterns of how the conversation was structured and what was spoken about.

To examine the interactions and sensemaking that took place, we used an adapted form of conversational analysis that focused on the turn-taking between the participants following an intervention from Vizzy. We draw from Porcheron et al.'s method [326], who used it to describe in detail the various methods families use to organise their talk with and around their smart speaker in their everyday conversations. This has become an accepted method in HCI, where a number of segments are chosen to illustrate the types of conversations that unfold with and around voice assistants and aspects of social conduct. When transcribing the segments, we follow some of the standard transcription conventions [16, 180, 327] in conversation analysis used in HCI. For reference, we indicate where pauses take place (e.g. (1.7) for 1.7 seconds), where an utterance is <faster> than usual, or where it is elonga:::ted, where talk is LOUD or °quiet°. Empty parentheses (   ) are used where spoken words could not be recognised. Where speech overlaps indentation and [square brackets] are used and ((unspoken actions)) are given in double parentheses, which can be either actions of speakers or the system. Speakers are indicated by P1 and P2; the synthesised speech produced by Vizzy is identified by the label "VZ".

The segments start slightly before Vizzy's prompt and end after the participants either request another visualisation, start discussing another topic, or agree/conclude on their answer. After transcribing and analysing four segments, the recordings of all 18 pairs were listened to several times in full length by myself and another researcher and analysed for patterns in participants' behaviours – such as how Vizzy's prompts were responded to in both conditions. After that, the identified patterns were discussed between us. The insights from this preliminary analysis were used to describe if the behaviours that were observed in the four segments below were found to be typical for the respective condition. Overall, for both conditions, the pairs seemed to be at ease with each other, taking turns and engaging in a level of banter. Furthermore, participants seemed to quickly get used to the system not following up on their questions; they took on board the NLUI prompts and included them in their conversations, usually discussing them until they came to a conclusion or felt they had sufficiently discussed the prompt and then moved on to another visualisation.

With respect to the two conditions, it was found that the same prompt in voice or screen condition elicited similar levels and types of reflective thinking. However, there were also specific patterns in each condition regarding how the conversations and interactions took place. In particular, the segments reveal that a typical conversational pattern for the voice condition was for a pair to start or continue to discuss relatively quickly after an NLUI prompt by 'bouncing off' ideas of each other, asking each other questions, generating hypotheses about possible reasons behind the patterns in the graphs and then moving on to another topic and/or visualisation. The typical pattern in the screen condition was that one or both participants initially read the prompt out loud when it appeared on the screen, which was then followed by a discussion similar to those which took place in the voice condition. However, the conversation often resumed and progressed with a 'slower pace', and there seemed to be less 'thinking aloud' or bouncing off ideas as part of their reflective process.

Responses to the Prompt "What Might Have Caused the Sudden Spike?"

First, we present examples of the conversations in the voice and screen condition following the prompt *"What might have caused the sudden spike?"* asked by Vizzy when the visualisation "girls and boys, developed and developing countries" was being displayed (see also Figure 4.4). As the visualisation shows, there is a marked increase ("spike") in developed countries between 1996 and 2002 which the prompt refers to. This open-ended prompt was intended to trigger the participants to look at the data and generate their own hypotheses when responding.

The first segment (1-VOI) illustrates that there is a high number of speaker changes comprising a quite rapid back-and-forth of suggestions between the participants in the voice condition. They seem to iteratively speculate on and construct hypotheses about possible reasons, as to what may have caused the sharp increase in the time series graphs they are looking at by building upon (or contrasting) what the other person says. Often, they seem to just 'think out loud' while generating ideas (e.g. lines 1-2, 16, 19-23).

```
01   P2    It's the same pattern ((points at the visualisation and traces the line in
02         the air)) as the global time series.
03   P1    Ah, you mean that ((points at visualisation)) [the orange line] is the
04         same as…
05   P2                                                  [     Yeah.     ]
06   P1    Yeah. The global one?
07   P2    … The global increase in uhh…
08   P1    Yeah.
09   P2    … in the overweight.
10   P1    Yeah.
11   P2    But here it's an...
12         (3.6)
13   VZ    What might have caused the sudden spike?
14   P2    °Sudden spike°, ah…
15   P1    It is around the 2000s, shortly before 2000.
16   P2    I don't know, like (0.3) globalisation?
17         (2.2)
18   P1    I have no idea.
19   P2    And possibility of getting a lot of different foods (2.4) or could also
20         be...
21   P1    Oh, I think, obviously, electro::nics – compu::ters, PlayStations.
22   P2    Ah, yeah!
23   P1    So, children play less outside and get fat.
24   P2    Yeah (0.6) yeah, true.
25         (2.2)
26   P1    Uhm, but it's actually quite surprising. That's an (0.8) 5% increase
27         (0.4) °almost°.
28   P2    Yeah, even a bit more.
29   P1    Yeah, °or a little bit less°. (0.4) CRAZY!
30   P2    People stay less outside and play inside.
31   P1    And look, the other one is just linear.
32   P2    Yeah, this is developing, yeah.
```

```
33  P1    This is a cra:zy increase. We should maybe look if it’s the same, (1.6)
34        uhmm…
35  P2    Men, like, uhhm ((points at one of the cardboard cards on the table))
36  P1    Yeah, Vizzy, show (1.3) men and boys.
37        (2.3)
38        ((Vizzy shows the requested visualisation, and participants continue
39        discussing the newly opened visualisation.))
```

Segment 1-VOI: Pair 3, Visualisation on display: “girls and boys, developed and developing countries”.

In Segment 1-VOI presented above, P1 begins by describing the part of the graph which Vizzy’s prompt is related to, and then the two participants alternate between hypothesising about the reasons as to why this happened, and which historical events could be related to it. Before Vizzy triggered the prompt, the pair were involved in a discussion about how the pattern they are seeing compares to patterns in other visualisations they have previously explored. After a silence of more than three seconds, Vizzy then triggers its prompt. When it plays via the speaker, it appears to scaffold participants’ thinking around when and why there was a spike in the developed countries (see orange line in Figure 4.4). Vizzy does not contribute any further to the conversation. Instead, the participants engage in a discussion about possible reasons for the increase. The conversational interactions between both participants show how they speculate and reflect on different reasons following Vizzy’s prompt. After Vizzy’s prompt, P2 speculates on the possible reasons behind the spike, making two suggestions for possible answers (on line 16 and 19), which is then followed by P1 also making a suggestion, namely that new technologies were the main reason (line 21), which P2 then agrees to (line 22 and 24). The segment finishes with participants taking a closer look at the increase and an attempt to quantify it in a percentage increase (lines 26-29). After that, on line 30, P2 provides a ‘conclusion’. This leads to a new train of thought; to check if this increase can also be found in the graphs for *men and boys*. At which point they then request Vizzy to take a look at the visualisation with the corresponding time series shown side by side.

As mentioned previously, it can also be seen here that participants quickly got used to Vizzy not following up after asking a question, and thus usually just continued their conversation. Hence, the role the pairs see Vizzy playing is essentially ‘someone’ who follows their requests for showing new visualisations on the display and who will occasionally prompt them with a question to get them to reflect.

The participants sometimes ask speculative questions (e.g. see line 3), which could be directed to Vizzy or each other. As they have learnt to understand when Vizzy will intervene and what Vizzy will say, their question-asking is more of a way of thinking aloud to clarify what they are looking at in the data or to ascertain what the other participant meant when they said something.

In this segment, it is worth noting that P2 may have continued their thought after having paused mid-sentence *("But here it's an...")*. It is possible that the silence considered by Vizzy/the Wizard to be an opportune time to trigger a prompt, was in fact P2 (and P1) thinking about a possible argument/reasoning and not because they were 'lost' or stuck. Interestingly, though, from the analysis of the transcripts across both conditions, pairs often did not seem to mind when Vizzy's prompts were not perfectly 'aligned' with their ongoing conversation, and they just 'reoriented' their conversation towards the prompt, as it was also the case in this segment. In other cases, the pairs occasionally just ignored an imperfectly triggered prompt or answered it briefly to then carry on with another topic or with what they discussed prior to the prompt.

The second segment contains the same prompt but was asked in the screen condition (1-SCR). In this segment, participants also had an extensive discussion in response to Vizzy's prompt. However, in contrast to the first segment (1-VOI), the discussion unfolds at a somewhat 'slower pace'; there are several pauses in the conversation (e.g. lines 13, 15, 35), the turns are longer, and there is less of the rapid and iterative generation of ideas and hypotheses seen in the first segment, where participants' reflective process unfolded in shorter turns.

```
01  P2   Maybe there is not that much of a difference in terms of lifestyle for
02       children.
03  P1   Yeah.
04  P2   Than for adults.
05  P1   Yeah.
06  P2   (              ) I don't know, run around and playing.
07  P1   So basically, this can't show that there is no differences in (1.9)
08       hormonal stuff...
09       (3.2)
10  VZ   ((Vizzy displays prompt: What might have caused the sudden spike?))
11  P1   ((Reads prompt and mumbles part of what she is reading.))
12       Around 96…
13       (7.9)
14       °I don't know°
15       (6.3)
16  P2   Mmmmh, (1.7) there might be a lot of (0.3) things. I mean, maybe there was
17       some (0.8) ehmm (1.8) maybe around 2000 there were lots of companies like
18       (      ) companies
```

```
19  P1    Yeah, that's, that's possible or maybe more ehm women, more MOMS started
20        working and not cooking so much or something like that. Or, (1.6) uhmm.
21  P2    Yeah (2.7), it's very hard to tell to be honest.
22  P1    Yeah.
23  P2    Uhmm.
24  P1    So, there is a sudden spike. (1.3) Yeah, but I think this shows that it's
25        mostly something abou::t (0.4) the lifestyle, right?
26  P2    Yeah. And also, it's something that people realise is wrong, otherwise
27        (0.6) it wouldn't slow down.
28  P1    Yeah, and you need to take (0.8) care of it. (0.6) Yeah, that's what I
29        mean that (0.4) how much more can it go from 60%? (0.8) At some point you
30        will start (1.7) someho::w taking (     ) from the government or from I
31        don't know.
32  P2    Mmhm.
33  P1    They will start taking (1.7) initiatives to slow it down.
34  P2    Mhm. (1.4) OK, so.
35        (3.2)
36  P1    So, OK, let's recap. (0.8) Uhmm, (switches to another visualisation: Women
37        and Men, Global) So, basically (0.4) there are more women (2.2) but this
38        because it depends on (1.2) developed and (2.7) developing classification
39        (switches to another visualisation: Women and Men, Developed and
40        Developing Countries) this may be because of a lifestyle change or because
41        of lifestyle differences.
43        ((Subsequently P1 continues to summarise further findings which they have
44        made and they both discuss and conclude what the main patterns were.))
```

Segment 1-SCR: Pair 12, Visualisation on display: "girls and boys, developed and developing countries".

Segment 1-SCR shows how the participants begin with a discussion about how the differences in children between developed and developing countries are less pronounced than for adults, which the pair explains is most likely due to their different lifestyles, in the sense that children may have more similar levels of physical activity between developed and developing countries than adults have. After a silence (line 9), Vizzy displays the prompt, which both participants read while P1 mumbles it aloud to the other. This behaviour was found to be typical in the screen condition; either one or both of the participants often read Vizzy's prompt aloud or mumbled while reading it from the screen. In doing so, they let the other person know that they are currently reading it while drawing their joint attention to it.

It seems as if the pair struggles to know what to think or say at the beginning, which is also reflected by the long silence after Vizzy's intervention (lines 13/15). However, then Vizzy's prompt triggers a long discussion between the two, where there is some back and forth in the way they come up with different hypotheses as to why the trend they are looking at occurred. Sometimes the speculative hypotheses are formulated as questions for the other to consider, for example, P1 proposes: "*Yeah, but I think this shows that it's mostly something about the lifestyle, right?*" followed later by: "*Yeah, that's what I mean that how much more can it go from 60%?*".

Both reveal the level of reflection involved in their sensemaking of what the visualisations show and what the reasons for certain trends and patterns could be.

In this segment, both P1 and P2 make suggestions although P2 tends to agree with what P1 is suggesting following Vizzy's prompt with one-word answers or questions. This difference in who contributes the most was generally more marked in the screen condition, corroborating the quantitative findings. This may be due to one person taking the baton in steering the discussion as a result of having implicitly decided to be the one who reads out Vizzy's prompt, or it could also be as a result of more vocal partners taking the lead. However, as previously described in the quantitative analyses, this difference in how balanced the contributions was not found to be significant.

Responses to the Prompt "Is the Increase of One More Significant Than the Other?"

Next, we look at two examples of conversations where Vizzy triggered the prompt "*Is the increase of one more significant than the other*?" in the voice condition (segment 2-VOI) and the screen condition (segment 2-SCR) for the visualisation "women and girls, developed and developing countries" (see Figure 4.8). To literally answer this prompt, participants would just need to compare the increase in the different time series graphs. However, the question could be understood as implicitly asking participants to consider why this might be the case. Indeed, in both conditions, the participant pairs had extensive discussions following this prompt rather than simply answering "yes" or "no" before moving on to the next visualisation.

Similar to the previous conversations, segment 2-VOI illustrates how P1 and P2 take turns to bounce their ideas off each other in an exploratory way (e.g. lines 1-10) and how they generate hypotheses and what types of questions they ask each other while doing so (e.g. lines 8-10, 29). The conversation was again relatively 'fast-paced', consisting of shorter turns, and the pair often appeared to be thinking 'on their feet' and out loud (e.g. lines 3-6, 8-10, 15-18, 23-26).

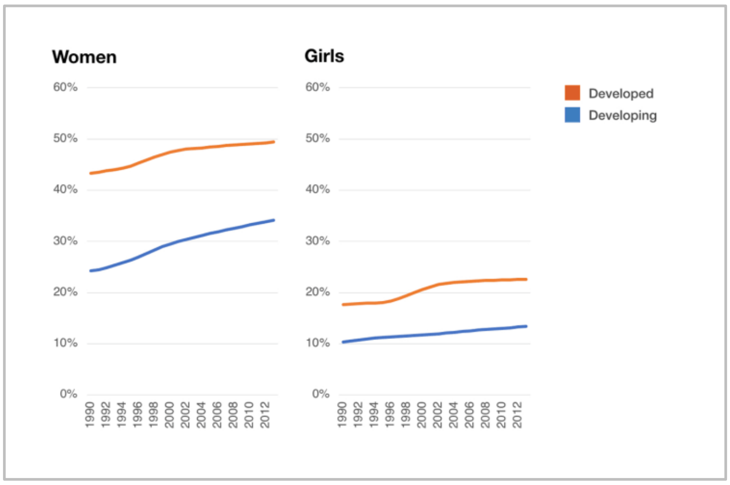


Figure 4.8: Visualisation "women and girls, developed and developing countries".

P2 I like the "developed girls". (0.8) It is FLAT. (0.4) It stopped (0.3) rising.

P1 Yeah, it's (0.4) <it's really weird> it's like, we:: went from what, 18% to 22% <then ( ) it>.((laughs))

P2 ((laughs)) (1.5) Ha. It hasn't changed at all.

P1 °Yeah°, but in developing countries it's steady:::y something.

P2 Hmm.

P1 We don't really know how it will evolve. (1.2) Like will it keep rising slowly::y ((shows with hands)? Will it ((shows with hands)) jump u::p? Will it stabili::ze?

(2.6)

VZ **Is the increase of one more significant than the other?**

(1.4)

P2 No.

P1 Ehm, yeah, but it...

P2 Which one?

P1 Adults are, the...

P2 OK, ehm...Yeah.

P1 The increase is more significant.

P2 If you look between those two ((points at the graphs)), yeah...

P1 Yeah, the children ARE SOMEWHAT protected, I guess. ((laughs))

P2 ((laughs))

P1 It's like it's influencing them less, <but>, (0.3) <in the meantime> I mean they are more ehhm (0.4) checked up by doctors, by everybody ( ) (0.4) which is their height their weight ( ). (0.5) Yeah, even in schools their food is checked and basically...

P2 Mhm.

P1 And adults they can do basically whatever they want, so...

P2 Vizzy, can you show global? (0.7) °What happens if we put them together?°

(3.5)

((new visualisation appears))

(8.6)

P1 Yeah, it doesn't add much I guess.

P2 All four of them were almost steady right so if you add them you get something almost steady. ((laughs))

P1 ((laughs)) Yeah, it makes sense ((laughs))

P2 ((laughs))

((They subsequently continue discussing the new visualisation on screen and comparing it with the previous one.))

Segment 2-VOI: Pair 8, Visualisation on display: "women and girls, developed and developing countries".

In this segment, after Vizzy triggered the prompt, *"Is the increase of one more significant than the other?"* P2 answers immediately, *"No."* However, P1 does not seem to agree with P2. Instead, P1 provides a reason why there is a more significant rise for adults compared with children: because they *"can do whatever they want"*, whereas children are monitored much more when growing up. P2 listens to his explanation, occasionally interjecting with disfluencies, suggesting she is considering P1's explanation but still does not seem to fully agree. Then, after they made sense of and reflected together on the pattern for a certain amount of time, they move on to another visualisation (line 29).

This segment shows how the participants generate their hypotheses by asking questions of the time series graph, connecting what they are seeing with possible reasons for the different trends while also reflecting on and hypothesising what might happen. For example, P1 asks three questions in succession when looking at a line graph: *"Like will it keep rising slowly? Will it jump up? Will it stabilise?"*. Furthermore, it can be seen how the pair seems to use humour when formulating certain tentative or more tentative or 'daring' hypotheses (e.g. line 21 where P1 says, *"Yeah, the children ARE SOMEWHAT protected, I guess."* which then both laugh at). Taken together, the segment shows how Vizzy's interjection led the pair to having a relatively fast-paced discussion, bouncing ideas off each other and asking each other questions while speculating on and reflecting on the differences.

The fourth segment below (2-SCR) illustrates the patterns of conversation before and after Vizzy's prompt, "*Is the increase of one more significant than the other*?" in the screen condition. In this segment, Vizzy triggered the prompt after a pause when the participants were deciding which visualisation to look at next (lines 6-10). This may have given the impression that Vizzy was not helping them choose but instead providing a prompt about the current visualisation they were looking at. The pair appeared not to mind Vizzy's intervention as they changed tack to think about what the answer might be.

As can be seen from the segment below, there is a long discussion between the participants which evolves somewhat more slowly over time, consisting of relatively long speaker turns. Following Vizzy's prompt *"Is the increase of one more significant than the other?"* the pair work out an analysis (including calculations). Similar to the previous voice segment (1-VOI), participants discussed their answer quite extensively despite the prompt/question being a “yes/no” one. The participants also appeared to engage in a process of thinking aloud (lines 1-3, 22-25, 28-32) while speculating and reflecting on the possible reasons for the increase.

In contrast to the previous segment (2-VOI), the pair spent considerable time looking at the visualisation without speaking, before discussing it in more detail (line 19). There were several other instances of long silences when they were reading and appearing to figure out what the change in the slope of lines in the graph meant. However, they spent more time examining the graphs; reading off and inferring the percentages that helped them work through why there might be a significant difference – which did not happen in the previous segment.

```
01  P1    But I think maybe it just looks like a straight line because the increase
02        is too, [      (2.6)      it's too slow maybe we can't,      (2.2)       ]
03        yeah, maybe we cannot really identify this from this figure.
04  P2            [(0.6) Mmmh (1.3) The difference or the percentage is too narrow.]
05  P2    Mhm. (4.2) OK.
06  P1    OK. So, what to do next.
07        ((Both start to look at the cardboard cards.))
08  P2    Up to you. (3.1) One we haven't done.
09  P1    Yeah.
10        ((Both continue to look at the cardboard cards.))
11  VZ    ((Vizzy displays prompt: Is the increase of one more significant than the
12        other?))
13        ((Both read the prompt and mumble part of what they read.))
14  P2    (4.1) What does it mean “one than the other”? It means the developed and
15        developing or women and girls?
16        (1.6)
17  P1    Mmh, (1.5) maybe (            ) the question.
18  P2    Hmm, I’m a bit confused.
19  P1    ((Both look at different parts of the visualisations for 21 seconds))
20  P2    More signi::fi::cant.
21        ((Both look at the visualisations for another 14 seconds))
22  P1    Yeah, for me personally I would say the increase of the, of the women
23        [are more significant], because... But, we don't, we just know the
24        absolute increase is more significant, we don't know the corresponding
25        difference. So...
26  P2    [ Yeah, I think so. ]
27  P2    Mmh.
28  P1    So, () I cannot really answer this question (0.8) ‘cause [ (0.9) ]...
29        Maybe to a simple conclusion, from about 25% to about 35% so it's, it's 10
30        over 25 is about 40% (1.5) and for the GIRLS (2.7) it seems a little bit
31        less than 40% so I, yeah, so I think the increase of the women is more
32        significant than the girls, yeah. So, do you see what I discovered?
33  P2                                                                [  Mhm. ]
34  P2    Mmh.
35  P1    Yeah, (1.4) yeah, under the condition that my calculation is correct.
36        ((laughs))
37  P1    ((Both mumble something and look at the cardboard cards and the
38  P2    touchscreen with the filter menu.))
39        (7.0)
40  P1    ((Starts making a selection on the touchscreen interface.)) Would we get
41        the same result for “male”?
42        ((Vizzy shows requested visualisation.))
43        ((Both look at the visualisation for about 8 seconds and then continue to
44        discuss the differences.))
```

Segment 2-SCR: Pair 17, Visualisation on display: “women and girls, developed and developing countries”.

The discussion in this segment was initially more focused on making sense of the prompt (lines 17-21), followed by reading the data from the visualisations (lines 22-35). Here, even though they examine and speculate on the patterns and ultimately provide an answer to Vizzy's prompt they do not reflect more extensively as to *why* this might be the case. Instead, they elaborate on Vizzy's prompt, focusing on the visualisation itself and the details in the change of the curves in the graphs. In doing so, they attempt to quantify the increase in percentage points. After exploring the data in this way, they conclude by conferring that there is a more significant increase in women's obesity as compared to girls.

Overall, the above four segments illustrate the varied types of conversations that followed after Vizzy's prompts, guiding the participants to speculate and reflect on possible reasons for the trends in the data while orienting them towards paying attention to particular aspects of the data. The main behaviours and characteristics identified in the qualitative analysis are summarised in Table 4.2. Some of the findings from the qualitative analysis corroborate those found in the quantitative analysis. For example, the 'pace' of the discussion in regard to qualitative aspects ('thinking out loud' and quickly starting to explore possible answers/ideas) corroborates the quantitative aspects described in Section 4.5.1 (more turn-taking). The kinds of ensuing conversations for both the voice and screen condition had a similar pattern: After receiving a prompt from Vizzy, participants formulated speculative questions towards the data, reflected on possible reasons, and suggested different hypotheses (sometimes also in the form of questions). In terms of whether there were any marked differences between the screen and voice condition, we observed how participants in the voice condition tended to recommence the conversation relatively quickly after Vizzy provided the prompt and did so at a fast pace by bouncing ideas/hypotheses off each other, as if they were brainstorming. In contrast, in the screen condition, the participants often took their time to 'start up' the conversation again after having read the prompt on the screen. Stopping to read the prompt, therefore, had the effect of slowing down the conversation and got participants to think more about the prompt and how to respond to it best; when their conversation resumed, there appeared to be less of the rapid and exploratory idea generation found in the voice condition as part of their reflection process, and instead, it was characterised more by careful examination of the trends and patterns. This finding corroborates the differences in turn-taking found in the quantitative analysis in Section 4.5.1.

Table 4.2: Summarised findings from the qualitative analyses of the discussions following NLUI prompts showing key similarities and differences between both conditions.

Voice Condition		Screen Condition
Extended discussions speculating on and exploring various hypotheses and possible explanations	=	Extended discussions speculating on and exploring various hypotheses and possible explanations
After a prompt the discussion recommences quickly	≠	After a prompt the discussion slows down/pauses
After a prompt quickly starting to speculate and explore possible or tentative ideas/answers (by 'thinking out loud')	≠	After a prompt discussing more about what the prompt means and *how* to answer it before discussing possible answers

4.5.4 Participants' Reflections on Vizzy

In the semi-structured interviews following the study, participants reflected upon the role Vizzy played in the task. However, as the experiment was designed to be between subjects, each pair only experienced one setting, so their reflections only refer to the experience they had. Thus, the interviews provided insights into how participants perceived and experienced Vizzy, its prompts, and the role it plays – on a general level, independent of the condition/modality. Overall, the interviews showed that most of the participants found Vizzy and its prompts useful. The section is split into two subsections reflecting two general aspects of how Vizzy's role was understood by participants – how Vizzy's prompts (i) scaffolded their thinking and (ii) fostered 'slow' and reflective thinking in both conditions.

Scaffolding Participants' Thinking

About half of the pairs mentioned that it felt like Vizzy is *part of the discussion*, like a facilitator or even a collaborator, that helps them when they needed some input. For example, participant 1 in pair 4 in the voice condition (hereafter, these identifiers will be abbreviated, in this case, VOI-4-1) mentioned: *"It's like talking to a colleague, like Vizzy, could you please check…"* Furthermore, it also seemed as if most participants understood and appreciated

Vizzy's behaviour of prompting them to consider when there was a silence, without subsequently following up on it, for example, participant 2 in pair 5 in the screen condition (SCR-5-2): *"The questions were interesting, they were all pointing to something that we have missed, for example, the steadiness, I wouldn't have analysed the steadiness myself."* Similarly, VOI-9-2 said: *"The suggestions were useful when we were having a break; it would help us see what else was there. It was waiting for us."* The majority of the participants pointed out that the assistant helped them to not get lost or stuck on a particular data visualisation. It also allowed them to find additional differences or trends in the data when they thought that they had already discovered everything or couldn't find any other patterns, for example, VOI-8-1: *"I think one thing that helped was that when we were kind of stuck and we were not saying anything, it would just generate a suggestion. I found that useful."*

About one third of the pairs mentioned that they would rather not have Vizzy interject when performing a specific, well-defined task/analysis or when they knew what to look for. VOI-8-1 also thought Vizzy could help to generate hypotheses (i.e. exploratory analysis) instead of testing existing hypotheses (i.e. confirmatory analysis): *"If you have a lot of variables and you are not really sure what you are looking for or if you are training someone it might be a good thing to use. If I know what I am looking for, I probably won't use it. (…) I would use it to GENERATE hypotheses instead of TESTING my hypotheses."* And similarly, SCR-1-1: *"It depends on if the data that I am working on is something that I am familiar with. If it is something that I have been working on for the past five years probably not [use such a system] – if I am using new data, sure."* These comments also suggest that participants consider the system to be more suitable for working with new, unfamiliar topics (or datasets in this case) rather than with familiar ones – for which it might not be necessary to engage in reflection and speculation to the same extent.

A couple of the pairs also reflected on how Vizzy helped them remember the data by making them talk about it. For instance, SCR-9-1 said: *"I am really impressed by what we all remember from that, so maybe it is also a good thing for remembering data by talking about it and having some kind of facilitator."* And similarly, VOI-5-2: *"Because it is so interactive, I think it stays in my memory as well."*

However, a few of the participants did not like the way Vizzy prompted them. SCR-3-1 commented that they were not always helpful: *"Sometimes it asked something that we already* *discussed or that we were in the middle of discussing."* (Even if the experimenter tried to provide prompts about aspects that participants have not previously discussed, it was not always possible to find prompts without any overlap.) This points to one of the key challenges of proactive prompts already introduced earlier, namely that they cannot always be delivered in a way that users will find them meaningful and useful. Furthermore, SCR-9-2 commented that *"It was more like an examiner as we need to find an answer to the question it asks. While the guidance is quite minimal."* SCR-7-2 also reflected: *"It is like they [Vizzy] are joining the conversation and immediately leaving it".* One participant, SCR-7-1, mentioned it would be better to *"have it help only when we want help"* rather than it being proactive. However, this was rather an exception – most of the other groups said Vizzy's interjections were helpful, guiding them to know what to look for in the data, probing and scaffolding their thinking, and helping them to reflect.

More than half of the pairs in the voice condition mentioned positive aspects related to the shareability of a voice interface and its suitability for collaborative situations, which was not the case in the screen condition. This corroborates the tendency towards more balanced interactions in the voice condition that was found in the quantitative analysis (Section 4.5.2). For example, VOI-7-2 mentioned: *"One big advantage is that we are both in control, whereas in a typical laptop or tablet scenario, it would either be my computer or his computer, and he says let's look at 'women and girls' and then I would have to change it. It is a nice interaction when we are both in control. We are exploring more actively."*

Fostering Slow and Reflective Thinking

About a third of the pairs mentioned that they had the impression Vizzy made them do the data analysis task more slowly than if they were just doing it by themselves or with common analytics tools. However, most of them acknowledged that this 'slowing down' of their thought process can also have benefits, for example, in situations where they are exploring a new dataset/topic or getting a new perspective on one they may already be familiar with, such as SCR-9-1: *"I think it is good [to use this system] if you have time and you are trying to figure out things."* Similarly, SCR-2-2: *"I mean, it was more time-consuming than traditional tools, but that also has benefits if you are not in a rush."* Related to this, SCR-5-1 also described how they understood the concept of the NLUI that does not follow up: *"Instead of an assistant, I would say it's like a tutor, so he or she has the answer already, and he or she is trying to guide me."* This illustrates the different purposes an NLUI can have, one that helps to get things done (i.e. metaphor of an assistant) versus one that guides the users in doing the thinking themselves (i.e. metaphor of a tutor), with the latter being what we aimed for in the present scenario/study. Both could be considered different types of NLUIs taking on more (i.e. assistant) or less (i.e. tutor) of the task.

Two pairs mentioned that this kind of slower interaction (i.e. the 'tutor role') would be most suitable for users who are getting familiar with a dataset (or data analysis more generally). Once they have become familiarised, they may then prefer for the system to become 'faster' (i.e. 'assistant role'), by making suggestions about what visualisations to look at first and in what sequence or by having additional commands/controls for switching between visualisations more quickly – in other words taking on more of the task and enabling faster exploration. For example, VOI-6-1 mentioned: *"If you are looking at the same data for an extended period of time, you mostly want it to be very fast to get data out. This isn't exactly fast. I guess this is more suitable if you are introducing a new topic or you are trying to get a new perspective on the same data."*

This comment summarises well the main purpose of Vizzy, which is to support users in exploring and reflecting on a (new) dataset/topic from different angles to gain new insights. Most participants understood that the purpose of Vizzy is less to enable users to quickly and efficiently conduct specific (confirmatory) analyses. They understood that it is mainly designed for users who are not (yet) experts in the given dataset and/or data analytics – or it can allow those who are experts to approach familiar datasets from a different angle (in the words of VOI-6-1: *"to get a new perspective").*

Furthermore, the participants' comments above illustrate the trade-off between providing an essentially proactive NLUI that probes (i.e. 'tutor role') versus one that allows to easily and quickly perform certain analyses and that takes on more of the work (i.e. 'assistant role'). The former can help users to think more for themselves to see connections or trends and use their common-sense knowledge, whereas wanting something easier or 'faster' is based on a desire for an NLUI that can 'do some of the thinking' on their behalf, generating certain graphs and corresponding hypotheses they could then concur with and accept. Here, the aim was to determine how NLUIs could act more as facilitators or tutors, probing the users so that they get a better understanding and are engaged in the sensemaking activity; in other contexts, it may be more desirable if the NLUI takes more the role of an assistant, doing more of the computation, making suggestions and drawing conclusions, rather than prompting the user to do the reasoning, learning, or decision-making themselves. Which kind of NLUI to model/design for will depend on the role desired of an NLUI in a given setting. Depending on the setting, it may be more important to support users in reflecting, sensemaking, and acquiring new knowledge (and in enabling them to transfer that knowledge) or to help them become more effective at solving a specific problem. In other settings, the goal may be to 'just' make users complete a specific task as quickly as possible.

4.6 Discussion

Our study has shown how voice versus screen-based human-NLUI interaction can affect users' conversation and collaborative reflection as well as users' interactions with the system incorporating an NLUI. Supporting our first hypothesis, we found participant pairs in the voice condition made significantly more requests to Vizzy and explored more of the available data visualisations. Supporting our second hypothesis, we found participant pairs took more turns and asked each other more questions when interacting with the system in the voice condition compared with the screen condition. When analysing the conversations to determine how this might be related to the way participants reflected on and made sense of the data visualisations, we observed the interactions in the voice condition to be at a 'faster pace' with more bouncing off ideas between the participants.

As proposed in the introduction of this chapter, one reason for these differences is that voice prompts were more seamless and better aligned with the human-human interactions and the conversations that took place, in the sense that participants could 'embed' their voice requests into the ongoing conversation while keeping their eyes on the screen depicting the visualisations (or on each other). The finding that participants took more turns and asked more questions in the voice condition can be further explained by voice-based interaction being more immediate and direct compared to screen-based interaction. This may have stimulated more discussion and encouraged more turn-taking as participants explored and reflected on different ideas and hypotheses. Furthermore, in the voice condition, Vizzy 'takes turns' itself by intervening in participants' conversations, which was not the case in the screen condition – where the participants instead had to stop their conversation to start reading from the screen. Vizzy actively intervening in the conversation by speaking aloud may have motivated participants to proactively take turns and generate ideas or questions like Vizzy did (as can also be seen in segment 1-VOI). This may have also been further reinforced because by using voice, Vizzy not only had a more direct impact on the social interaction but even became a temporary 'participant' in it, which might have led to increased 'social behaviour' in participants (also see Section 2.1.2) compared to screen-based interactions. One such social behaviour related to social response theory [337] could be imitation and mirroring, which resonates with Ceha et al.'s [76] study (where participants mirrored the NLUI's question-asking), leading to participants adopting the interface's behaviour more in the voice condition due to being more of a social and conversational 'actor' or agent in that sense compared to an NLUI that just displays a question rather than taking an active turn in the ongoing conversation. However, in the follow-up interviews, some of the participants also mentioned challenges of the system and the modality of voice, namely that it was sometimes awkward if the system just chimed in, in particular, if it did so in unexpected ways. This could be, for example, when Vizzy provided a prompt which overlapped with what they had previously discussed.

The findings resonate with those found by Gonzalez and Gordon [158] on the effects of different modalities in interactive narratives, where the voice modality resulted in participants behaving in a different way compared to text-based interaction. Also, it concurs with findings from previous studies, where participants had more interactions with each other after the system spoke aloud following certain actions they made when using a tangible interface (e.g. [132]). Furthermore, voice requires immediate joint attention when it occurs, whereas reading text from a screen together requires paying attention in a different way. For the latter, there may be a slight delay as one waits for the other to finish reading and knowing when it is appropriate to start the conversation again.

The qualitative analyses also indicated that there were nuanced differences between the conditions in how engaged and interactive the discussions were and the way in which ideas and hypotheses were generated. In particular, participants in the screen condition often needed a bit more time to start the discussion. In this condition, it seemed as if the participants were reflecting more about what the prompts mean and how they should answer before discussing possible answers, while in the voice condition, they often immediately started discussing possible answers as if they were 'thinking out loud'. It appeared as if they were willing to be more exploratory in their discussion and less focused on providing a 'correct' answer to Vizzy's prompt.

There could be a number of reasons for the differences mentioned above. Firstly, if a prompt is asked via voice, people may feel more compelled to answer it, as it feels more similar to interacting with another human being. Thus, the users may adhere more to the rules of human-human conversations where it would be awkward or inappropriate to wait more than a few seconds before responding, which would also be in line with social response theory [337]. Text messages, on the other hand, even if represented as chat bubbles, may lead to VoiceViz being perceived less like a 'social' entity or actor, and people may feel less compelled to answer them immediately. This suggests that the NLUI's modality impacts on how users conceptualise, think of, respond to, and 'treat' the NLUI. Thus, the difference between conditions in how participants responded to an NLUI prompt, is most likely not just due to how the human mind processes one or the other modality, but also due to how people conceptualise the NLUI *depending on* the modality. Here again, some of the social behaviours in response to having more qualities of a 'social actor' may have come into play. In other words, when an NLUI feels more like a 'participant' in social interaction (i.e. because it speaks), people may not want to 'let it wait' for too long when it asks them something. This may partly explain the more immediate responses to NLUI prompts observed in the voice condition, where participants tended to just start thinking out loud and reflecting on possible answers. Secondly, since in the voice condition, Vizzy actively intervened in the conversation, participants may have also been more proactive and just saying what they think. Text messages, on the other hand, are less of an active intervention in the ongoing conversation, which possibly resulted in participants being less proactive themselves. Thirdly, another reason may be that when participants were in the middle of a conversation and they were prompted through voice, it may have been more natural for them to integrate Vizzy's prompt into the flow of their conversation. In other words, as there is no change in modality (from reading text to speaking and vice versa, as it is the case in the screen condition), they could just continue with their discussion.

However, if the NLUI is perceived to be 'butting in' too often it could become annoying (in particular, in the voice condition). Given that Vizzy was designed to have a minimal level of interaction (i.e. one prompt approximately every two minutes), the pairs were usually forgiving on the occasion when being interrupted in their ongoing conversation. Moreover, Vizzy was also designed to only occasionally prompt the participant pairs at opportune times rather than 'joining in' their ongoing conversation. The pairs quickly understood this underlying 'interaction model', and it seemed from the interviews they were happy with it in general – not wanting or expecting Vizzy to be an equal partner in the conversation but just a facilitator. This limited form of 'proactive agency' in the form of a facilitator, therefore, may in the long run, be more effective than trying to design the NLUI to be a human-like conversationalist, at least for group settings where it may be undesirable to have a system that intervenes too often in an ongoing conversation between humans – and by that potentially limiting their autonomy and agency in undesirable ways.

Taken together, the quantitative and qualitative findings suggest that voice interfaces can enable a faster 'pace' in collaborative reflection and sensemaking in terms of its structure (turn-taking) but also in terms of its content (responding more immediately and 'bouncing off' ideas). Furthermore, the findings revealed that there is a tendency towards more balanced human-human and human-NLUI interactions in the voice conditions. In addition, the voice modality seems to lead to more exploratory behaviour and curiosity in terms of how many questions are being asked but also in how Vizzy's prompts are responded to (coming up with different ideas/hypotheses and 'thinking out loud'). Finally, the voice modality also showed a higher engagement in terms of interactions with the system (number of visualisations *requested* and *explored*). However, both modalities enabled different types of sensemaking and reflective thinking – while in the voice condition, it tended to be more exploratory and speculative, it seemed to be 'slower' and more 'deliberate' or 'calculated' in the screen condition. Which type of reflective thinking is more desirable might depend on the task.

4.6.1 Limitations

The study investigated how an NLUI could support users in an exploratory task via prompts, which it provided proactively at opportune times. It is worth noting that (i) these prompts were prepared prior to the study by the research team, and (ii) that they were triggered by a human experimenter/the 'wizard' (approximately every two minutes) based on simple rules, which were (a) the topic of the prompt was not previously discussed by participants and (b) there was a silence of at least three seconds. Hence, there is the aspect of (i) the quality of the prompt itself, (ii) the appropriateness of the prompt's content given the context (i.e. the ongoing discussion between participants) and (iii) the appropriateness of its specific timing. If the creation and delivery of the prompts were implemented/automated by a system, they may not reach the same accuracy and quality as in the present study – both in terms of their content[12] as well as their context-specificity and timing[13]. However, it is worth noting that in the present exploratory, open-ended task, the effect of potentially inaccurate timing may be smaller compared to others (e.g. well-defined ones and when people know what they need to do), and our findings showed that the majority of participants did not mind occasional imperfect timing. Furthermore, the NLUI prompts were designed to be rather infrequent so that the NLUI would not intervene too often. Therefore, even if a system should deliver the prompts with a somewhat worse timing, it might not have significant negative effects on users if it is in a similar scenario and task. Nevertheless, it is important to acknowledge that the range of tasks people perform in everyday life where they might be interested in and receptive to such proactive behaviours may be somewhat limited, since they may often not be in such an exploratory and open 'state of mind' as participants in the present study were.

The study compared two conditions: voice-based NLUI input and output versus screen-based NLUI input and output (along with the screen-based presentation of visualisations in both

[12] However, several prompts in this study were, in fact, not particularly complex and usually focused on a distinct pattern in the data (e.g. *"Is the increase of one more significant than the other?")*. Therefore, it is likely that a system that uses a suitable AI model would be able to scan for such patterns in the (time series) data and 'come up' with prompts comparable to those in this study.

[13] Although it is important to mention that this is also changing rapidly with ongoing technological developments/progress. Of particular relevance here are more recent AI models that offer real-time speech interaction, which could enable a system like VoiceViz to monitor the ongoing conversation and find appropriate moments to deliver a prompt.

conditions). It is possible that some of the effects observed were more strongly related to the output than to the input in the specific modality and vice versa. With the present study design, it is not possible to adequately disambiguate the effects of input and output in the respective modalities, given the inclusion of only two modality combinations (for the NLUI interactions). In the present study, the focus was intentionally on voice-based versus screen-based NLUIs – reflecting the two combinations we considered most natural and appropriate for a variety settings and applications.

There are also limitations with the chosen metrics. As such, the number of visualisations participants *requested* and *explored* need to be understood only as proxies of participants' engagement in the data exploration task. It is thus important to interpret these metrics in conjunction with the other quantitative and qualitative analyses to get a better understanding of participants' engagement and reflection process. However, since in the present scenario participants were asked to discuss the visualisations they chose to look at together, there was always some engagement and sensemaking for each requested visualisation, and both metrics can thus be considered adequate proxies for task engagement in the given scenario/task.

In summary, the findings from our study suggest the use of a proactive *voice* interface could be preferable compared with a *screen-based* one for (multi-user) settings where a high level of interaction with the tool/system is desired, along with a fast-paced conversation, more speculation and questioning, as well as 'think aloud' and rapid/more immediate responses and idea generation. However, this does not mean that voice is 'better' *in general*; in some cases, it may be desirable to 'slow down' or stop the current discussion by the use of screen-based prompts to trigger different types of reflective thinking or to get participants discuss the prompt, build a shared understanding before answering, and enable a more 'deliberate' or 'measured' reflection.

4.7 Conclusion

The first study reported here has shown how interacting with an interface that incorporates an NLUI either through a voice or a graphical/screen-based interface with chatbot-type messages was able to scaffold and prompt users when completing an open-ended sensemaking task, in this case, exploring a dataset using a data visualisation tool. The NLUI's role of being a facilitator that occasionally provides a prompt for things to consider and patterns to look at in the data was accepted and acted upon by participants. A conclusion that can be drawn from the findings is that NLUIs can be designed to become 'part of' and facilitate human-human conversation and collaboration in the future, without taking away control from the users and limiting their agency and autonomy. Participants using the voice interface were found to be more engaged in the exploratory data analysis task compared with the screen condition, and interactions tended to be more balanced. Having an NLUI that speaks directly to them led to them asking more questions and taking more turns, resulting in many of their discussions resuming and progressing more rapidly than when reading the same prompt in a chat window on a screen. However, presenting prompts on a screen also has potential benefits; users can decide when to read the prompts leading to fewer interruptions of their ongoing conversations. Furthermore, having to pause the conversation to read a prompt can have the effect of slowing down users' conversation and thinking, which in some contexts may result in users spending more time thinking about what the prompt means and how to answer it. In sum, interacting with an NLUI-enabled interface in either modality can have advantages: Voice-based interactions may encourage more fast-flowing talk, while screen-based interactions may slow down people's conversation and collaborative reflection more. Which is preferable depends on what the activity or task is about.

Given the promise of an NLUI to support reflection ('in action' [373]) in such an exploratory sensemaking task, the prototype and study described in the next chapter focused on how an NLUI (using a text/screen-based interaction) can be designed to support reflection when people not only make sense of data – as was the focus in this study – but also when making complex decisions based on data.

5. ProberBot: Fostering Reflective Thinking in a Decision-Making Task

The aim of the study reported in this chapter was to explore how cognitive prompting by a text/screen-based NLUI can support reflective thinking when making complex decisions. It builds upon the findings of the previous chapter, where it was found that an NLUI ('Vizzy') enabled people to explore new perspectives and to reflect on the possible reasons behind certain patterns in a dataset, by providing prompts that were targeted at supporting them in their sensemaking. Here, the focus is on helping people reflect when making complex decisions. On the one hand, the goal of the reflection here is to support the decision-making process itself (e.g. helping people evaluate the respective decision-making criteria), which could also be called critical thinking (see Section 2.2.1). On the other hand, the goal is to support *metacognitive* processes (e.g. becoming more aware of how one has considered different criteria in their decision-making and how this might align with one's intentions). The reason for taking this direction was to explore how reflective thinking can be supported through NLUIs in more complex cognitive tasks that involve different kinds of reflection. Thus, similar to the first study, the focus here was mainly on *reflection-in-action*, in the sense that participants receive reflection prompts during their decision-making process.

The task chosen for the second study was investing in the stock market, in which reflecting on and trying to make sense of one's decisions and decision-making process can be beneficial – for example, to avoid rash decisions. However, it is well known that it can be difficult when investing in the stock market to 'keep calm' and act in a way that is not driven by emotion but in line with one's strategy and one's long-term goals. Thus, the question addressed here is how to design an NLUI to support reflective thinking in a complex decision-making task. Or, more generally, how can NLUIs be embedded into software tools used for complex decision-making and designed to scaffold and probe human cognition? The chapter describes the research that was conducted to explore possible uses of such 'probing NLUIs'. A study was conducted in which investors interacted with a prototype called 'ProberBot'[14]. The goal was to determine how effective it was at supporting their thinking and reflection when carrying out investing tasks using a simulated stock trading platform.

[14] The prototype was developed by Warren Park who was an intern at UCLIC at that time under my supervision. He also contributed to the data preparation/clean-up and the qualitative analysis.

5.1 Introduction

The focus of this study was on how to design an NLUI that can support people when making sense of a range of different types of data and act out a series of decisions at the interface. How can the NLUI encourage a person to reflect on what they are doing when this might be deemed advantageous? To address this question, an NLUI was designed that was intended to probe and scaffold human decision-making in the moment it occurs. The chosen task of making stock investment decisions is complex and entails uncertainty. The idea here was to develop an NLUI embedded in the software platform that could slow investor's decision-making down at key moments so that they could reflect more on why they were choosing to sell or buy a stock at a given moment – thereby encouraging reflection-in-action [373] (also see Section 2.2.1). Hence, when an investor is about to make a decision at the interface, the NLUI would ask them specific questions that are meant to trigger reflective thinking and help them externalise their thought process. The NLUI in this decision-making context thus intends to *probe* the user's thinking, which is why it was called *'ProberBot'*.

Part of the rationale behind this approach of using an NLUI in this way is it could help people externalise and explain their thoughts akin to how they might explain their thinking to another person. It has been shown that making decisions with others can lead to better decisions when effective decision-making approaches are applied [254]. **The idea here is that 'externalising' one's thoughts (such as a rationale for a decision) and thereby constructing a written representation of them can foster reflection and lead to new insights.** The goal is that by explaining to the NLUI their ideas, reasoning, or understanding of something people might improve, clarify, and further develop their own thoughts.

For stock investing, joint decision-making has been found to help reduce overconfidence [321], which is one of the key challenges in investment decision-making, as it can lead to not adequately balancing and considering the breadth of criteria and aspects that should 'play into' one's decision (e.g. considering specific valuation metrics beyond a stock's momentum, popularity, etc.). The question posed here is, can this kind of shared decision-making process be 'transferred to' or at least approximated by an NLUI? While ProberBot cannot – and does not attempt to – replicate the same social interactions involved in human collaborative sensemaking and decision-making, it aims to emulate certain aspects of it, in particular, the process of explaining the reasoning behind an idea or a decision to some ‘entity’, which a person can take turns with through conversational interactions.

The design of ProberBot’s dialogues was informed by some of the cognitive and emotional biases that can occur during investment decision-making (e.g. [27, 466]), which have been found to lead to poor decision-making in investing, negatively affecting the investing performance/returns [33]. Some of the most well-known ones that people are susceptible to are (i) the *hindsight bias*, (ii) the *availability bias*, and (iii) the *disposition effect* (the latter can also represent a bias even if it does not carry the word ‘bias’ in its name). The *hindsight bias*, also known as the ‘knew-it-all-along’ effect [140, 350], refers to overestimating the ability of oneself to have predicted an outcome after it has already happened. As a result, the hindsight bias can lead (together with other factors) to people being overconfident, since they believe that they have accurately predicted past events. *Availability* or *recency bias* refers to overestimating the importance of recent events (e.g. news) and underestimating other information [14, 98]. The *disposition effect* refers to a common tendency of investors of holding ‘losing stocks’ too long (‘losing’ in the sense that the stocks have been losing value since when they were originally purchased, meaning that the current price is lower than the initial purchase price) while selling ‘winning stocks’ too soon (‘winning’ in the sense that the current price has risen from the initial purchase price) [438]. The idea of ProberBot was to prompt the investors at key times with relevant questions that could trigger reflective thinking so that they might then become aware of any potential biases in their decision-making and, as a result, try to avoid them.

Therefore, the main question addressed in the study reported here was whether having an NLUI probe the investors at certain times helps them better think through certain decisions which could otherwise be impulsive, driven by emotion, and potentially irrational? If so, how?

The NLUI that was designed was embedded into a simulated stock trading platform developed to resemble existing platforms. It simulated real-world conditions that stock investors face when having to make a series of decisions based on what is happening in the market, for example, if stocks start falling unexpectedly. To evaluate the efficacy of ProberBot in fostering reflective thinking when using the simulated stock trading platform, we asked experienced stock investors to conduct several scenario-based investment decision-making tasks. Following this, we conducted in-depth interviews with them, asking them about their perceptions of and thoughts on using the trading interface plus ProberBot and, in particular, how it impacted their thinking while using it, as well as afterwards. Next, the relevant literature underpinning the idea of an NLUI that aims to scaffold reflective thinking in investment decision-making is presented.

5.2 Background and Related Work

Explaining one's thoughts can help oneself build a better understanding of something [426] – e.g. a topic to be learned, an issue to be understood, or a problem to be solved. The next section further 'unpacks' this idea in terms of its theoretical and empirical grounding, which is followed by a section that gives a brief overview of how existing NLUIs aim to support investment decision-making. Following that, Section 5.2.2 will give an overview of investing NLUIs that have been built and studied in past research.

5.2.1 The Roles of Explanation in Sensemaking and Decision-Making

The fast, *'intuitive system'* in Kahneman's Dual-Process Theory [205] introduced in Chapter 2 has many advantages because it generally relies on heuristics, which can make decisions more efficient. However, it can sometimes also lead to rash decisions (e.g. "I will quickly buy this product/stock, as it looks cheap right now!"), in which certain important criteria may not have been considered (e.g. "How would buying this product/stock actually align with my goals or strategy?"). It is proposed that the process of explaining a decision can engage the *'reasoning system'* more, which can, in turn, be conducive to better decision-making. It could also trigger metacognitive processes leading to supporting metacognitive awareness, which can help people make better decisions [142].

The *Self-Explanation Effect* suggests that the process of explaining concepts to others prompts deeper information processing, identification of knowledge gaps, and clarification of complex ideas [78]. Self-explanation has been deployed in online learning environments where it was found to be successful at improving learners' abilities to apply decision rules [196] or led students to solve a code debugging task with better performance when prompted to use self-explanation [241]. Self-explanation can thus encourage reflection on one's knowledge and cognitive processes (which also involves metacognitive processes – see for example [276]), thereby having the potential to improve comprehension of the subject matter – for example, a specific stock and one's rationale to invest in it.

Based on this theoretical framing of explanations, it is proposed here that an NLUI that asks questions can trigger reflective processes when someone needs to make a decision. However, it is also important to note that explaining a decision itself is, of course, not a guarantee for improving it and, in some cases, can even lead to the opposite. This can happen when existing biases in a decision might, in fact, be reinforced by explaining or rationalising them (post hoc), for example, due to people's tendency to focus on information that supports their beliefs (see for example [135, 136, 239]). Next, an overview of existing NLUIs that support investment decision-making is provided – highlighting the ways in which most of them are different to the idea of a 'probing' NLUI.

5.2.2 NLUIs for Decision-Making in Financial Contexts

One popular 'use case' in investing contexts has been to use conversational investment advisors, which are often also referred to as 'robo-advisors' (e.g. [69, 103, 183, 293, 444]). Such advisors usually aim to (i) capture a person's interests and preferences (e.g. for specific industries), their values (for example, concerning environmental, social, and corporate governance aspects), investment time horizon, and risk aversion, to then (ii) provide suggestions for suitable financial products, investment strategies, as well as portfolio allocations. Given this focus on designing NLUIs to provide advice, research has thus also investigated how such advice could be delivered to increase people's inclination to follow it. A study by Milana, Costanza, and Fischer [287] found that the NLUI's variability in the responses (there were different ways in which the NLUI's messages were phrased while their meaning was the same) and reply suggestion buttons (providing the user with different options to choose from rather than having to type the message themselves) significantly increased the inclination to follow the investment-related advice of the NLUI.

Similar to some of the NLUIs outlined earlier in Section 2.1.4, which aimed to help people more easily analyse complex data by requesting certain analyses or graphs, a number of tools have been developed to support stock investing that provide specific data and information related to a stock through conversational interactions. For example, Lauren and Watta's [246] NLUI provides users with real-time stock price retrieval, the latest financial news, historical graphs of stock prices, stock sentiment based on tweets, and forecasting of stock prices. The latter was also the aim of Halder et al.'s NLUI [174]. In a similar vein, Sharma et al. [383] built an NLUI giving users the ability to query information on companies to get an overview of a stock, including accessing the latest news and trading recommendations, which participants in a small user study found helpful, although they also noted that sometimes they were not sure why the NLUI recommended them something.

To conclude, this research suggests that some of these existing investing NLUIs show promise in providing users with relevant information and recommendations for their decision-making. However, most of these tools have not (yet) been designed to *'scaffold'* investors' decision-making process. Beyond relevant data and recommendations, they are thus not providing investors with any other/further cognitive support, even though research in other domains/contexts has shown that they are able to do so – such as NLUIs that were found to successfully facilitate reflective thinking in sensemaking and learning tasks (as introduced in Chapter 2). Hence, the approach taken here was to design an NLUI that probes and scaffolds the investor's decision-making at specific moments to help them *stay on course* with their strategy and avoid emotional (re)actions. The motivation for doing this is to enable more strategic thinking and to improve decision-making. The idea is to have the NLUI embedded into the interface used for the task (i.e. the stock investing platform interface) akin to other NLUIs and AI co-pilots embedded into software tools (e.g. Microsoft Office or Analytics tools) as described in Section 2.1.4.

5.3 Method and Research Questions

The method adopted in this study was inspired by the *technology probe* approach developed by Hilary Hutchinson [192] over two decades ago, which has since become a mainstream prototyping technique in HCI. Essentially, a partially functioning prototype is built with the aim of collecting data about its use in/through a real-world scenario or context. Here, we use the method to discover more about how probing conversational agent interventions will be perceived and experienced by users, in particular, how they think these kinds of probes will affect their decision-making and reflections surrounding this. Using the technology probe prototype, we ask the following related research questions:

The Tool: How should a probing NLUI be designed to be embedded into a software tool that will encourage reflective thinking during decision-making?

The Domain: How do expert investors – familiar with online investing platforms – think about having a probing NLUI added/integrated into a trading platform, which asks them reflective questions regarding their investment decisions? Does it support their thinking or distract them when considering what to do with each investment?

5.4 ProberBot Design

The rationale behind the concept of ProberBot is that it can proactively intervene when a user is about to make an important and potentially risky decision while interacting with a software tool (in this case, a stock trading platform). The design process began by creating ProberBot dialogues, with the aim to address the three cognitive biases mentioned earlier; we then evaluated these dialogues with experts to see how they could be improved and to decide which of them to use in the subsequent study. The steps of the design process are described in more detail in the following sections.

5.4.1 Dialogue Design for the ProberBot

The three biases mentioned earlier that have previously been found to affect investment decision-making were considered as a basis for designing the initial ProberBot dialogues. The idea was to design the ProberBot's conversational prompts so that they could be triggered in situations where there may be a bias in the investor's decision-making, helping them realise that their decisions could be biased (such as keeping a losing stock in their portfolio without a clear reason to do so).

There have been various strategies proposed to help investors reduce the effect of biases 'creeping into' their thinking. These include following a clearly defined analytical process that can be tested and retested and adjusted throughout time [28, 311, 466] while keeping track of personal decisions, mistakes, and successes by keeping an 'investing journal/diary' and other accountability mechanisms (see also [181]). However, all of these are time-consuming and difficult to maintain. They may help with record keeping and reflecting, but they are not as suitable for and effective at supporting in-the-moment decision-making. One of the motivations of the proposed NLUI that is embedded into the trading platform interface is that it can potentially reduce the risk of certain biases 'head-on' by encouraging the user to engage in meta-level thinking when it matters most, notably during the ongoing decision-making process. Of interest here is whether it can help make parts of the decision-making clearer by answering the NLUI's questions. The idea is that this process of externalisation should make vague thoughts more explicit and bring to light inconsistencies in argumentations and decisions [78, 112].

Designing the Initial Dialogues

The dialogues were designed to '*indirectly*' address the three biases introduced earlier (see Section 5.1), rather than trying to avert them specifically. The reason for this approach was that although *general* patterns in people's behaviour can indicate the presence of a bias (such as systematically holding losing stocks longer than winning stocks), it can be difficult to determine if a bias is indeed present in a *specific/isolated* decision, as the decision-making tendencies that the chosen biases can lead to are often only visible *over a series of observations*. Thus, in contrast to some of the decision-making tasks in studies that have investigated cognitive biases, it is *not* possible for the chosen investment-related biases to see directly from a single, isolated decision that a bias is present. The intention is, therefore, to encourage the investor to think more about their decision in the moment and help them become aware themselves when they might succumb to one of the biases. The dialogues were designed to be related to the three biases:

Dialogue 1. *Part A: Asking the user to formulate an investment thesis/motivation when making a major buy decision. Part B: Reminding the user of their initial/previous investment thesis and asking them to what extent it still holds if they intend to sell the same stock soon after initial purchase (i.e. Part A).*

This dialogue asks the user to indicate the extent to which they think their previously defined investment thesis for a stock still holds when they intend to sell it soon after having purchased it. It also asks for the factors which made them change their mind from the initial investment thesis.

Dialogue 2. *Asking the user for the reasons for still holding a stock which has been 'falling' for an extended period of time.*

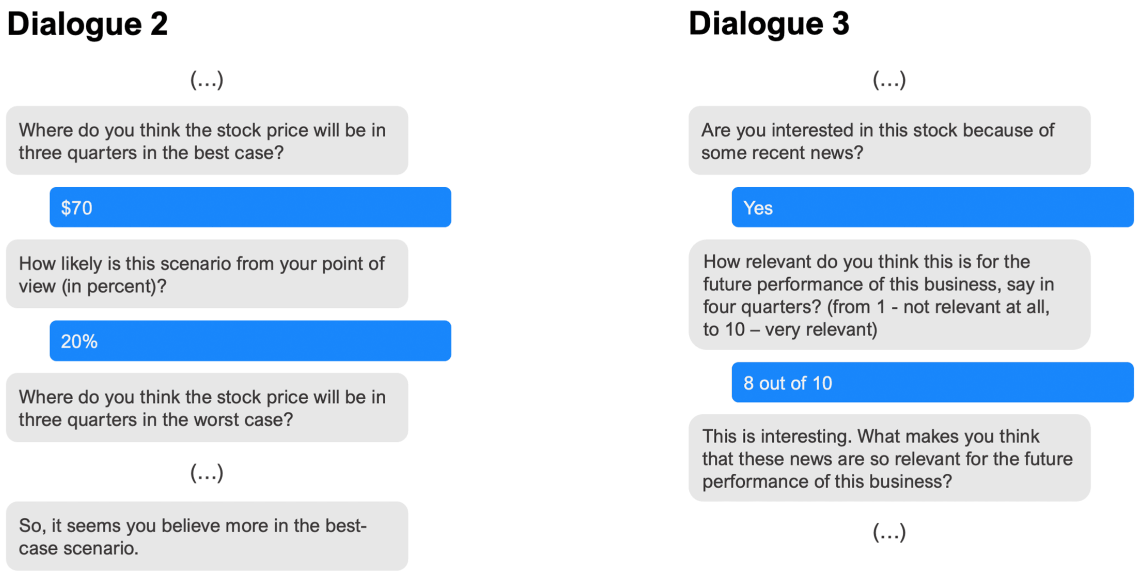
The dialogue starts off by asking the user how likely they think a recovery of the stock price is (i.e. to the price level at which they initially purchased it) and then asks them to provide the main reasons for their estimation/evaluation. It then gets the user to estimate the potential risks and gains in a structured way by coming up with a best- and worst-case scenario and rating the probabilities.

Through this, it is intended to help the user reflect on and become aware of the potential opportunity cost (i.e. the cost of not selling that loosing stock/replacing it with another). In the end, it presents back a summary of the user's evaluation.

Dialogue 3. *Asking the user how relevant they consider a recent news item to be for the future performance of a stock.*

On the one hand, company-related news can be selective and exaggerated, and on the other, they are usually quickly 'picked up' by the market and reflected in the stock price [129, 130]. Furthermore, they are often not relevant for long-term strategies and investing horizons, so *not* taking any direct actions in response to them is often the best approach [32, 231]. Therefore, to reduce the risk of potentially impulsive decisions based on an attention-grabbing news item, this dialogue asks users to reflect on its relevance, given their strategy and investment thesis for the specific stock.

Dialogue 1 (part A and B) is based on/informed by the hindsight bias; *Dialogue 2* is based on the disposition effect, and *Dialogue 3* is based on the recency/availability effect. As can be seen in Figure 5.1, the questions ProberBot asks do *not* give any specific directives (again, because identifying a flaw or bias in an isolated decision is difficult) but are intended to get the user to think about certain aspects, such as how likely they consider certain scenarios (an excerpt of Dialogue 1 can be found in Figure 5.2).



Dialogue 2

(...)

Where do you think the stock price will be in three quarters in the best case?

$70

How likely is this scenario from your point of view (in percent)?

20%

Where do you think the stock price will be in three quarters in the worst case?

(...)

So, it seems you believe more in the best-case scenario.

Dialogue 3

(...)

Are you interested in this stock because of some recent news?

Yes

How relevant do you think this is for the future performance of this business, say in four quarters? (from 1 - not relevant at all, to 10 – very relevant)

8 out of 10

This is interesting. What makes you think that these news are so relevant for the future performance of this business?

(...)

Figure 5.1: Excerpts of Dialogue 2 and Dialogue 3, which probe a user's thinking with context-dependent prompts/questions.

The dialogues were designed to promote short conversational turns to provide a scaffold for getting the user to make specific estimations and evaluations (such as how likely they consider a certain future event). This was also inspired by other NLUIs that provide certain choices for people to choose from when reflecting, such as the mental-health app *Woebot* [329], where simple options/inputs which people need to select from to express what they think or feel are often sufficient to trigger reflection (see also Section 2.2.2).

5.4.2 Expert Feedback

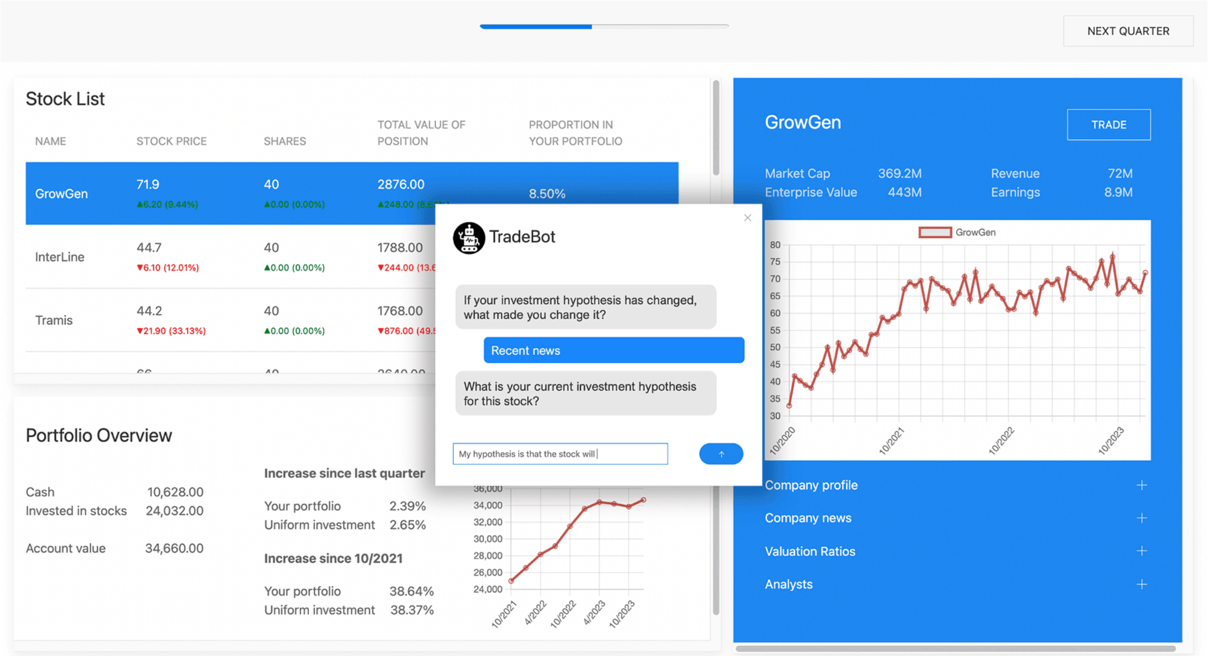
To evaluate the efficacy of ProberBot in increasing awareness of the three biases when trading using the simulated stock trading platform we built, two experts were contacted, both having extensive professional stock investing experience. The dialogue 'prototypes' for the three biases were presented to the two experts (using a digital whiteboard) for their feedback, showing the 'dialogue tree' of ProberBot with the different probing questions. A key comment they both made was that the ProberBot was too 'pushy'. It was argued instead that the ProberBot should just scaffold investors' thinking rather than be set up to ask 'leading' questions or to make suggestions (mainly due to the challenge of determining if a bias is in fact present, as they also pointed out). For example, they critiqued a part of the disposition effect dialogue, which initially was phrased as *"That's interesting.* ***Maybe you want to reconsider*** *why you keep holding this stock?"* and they suggested changing it to *"That's interesting.* ***So, you are currently holding a stock which you don't believe will recover."*** This shift in expression makes the ProberBot seem less pushy, appearing as a more neutral statement, leaving it up to the user what they conclude from it and how they want to act upon it. Based on their feedback, similar changes were made to the phrasing of the other dialogues to make them more neutral as well so that they would only provide 'cognitive scaffolds' instead of hinting at or suggesting what the user should do.

One aspect that was particularly appreciated by both experts was that the data captured in ProberBot's dialogues could provide a useful history of decisions that users could refer back to, akin to keeping an investing diary [86] (just in conversational form), allowing them to see and revisit what they decided previously and why – similar to scaffolds in educational settings, which also intend to help the student come to a conclusion or an insight by themselves.

Another comment they made was that it can be very useful to have 'someone' (i.e. this NLUI 'entity') to talk to when thinking about their decisions, since it would get them to express their thoughts (e.g. their thesis/rationale that motivates an intended investment). Based on the experts' feedback, the dialogues were revised so that they would enable people to externalise their thought process and, in so doing, enable them to reflect on and identify potential risks and biases by themselves rather than having the NLUI too explicitly referring to or hinting at certain biases – or at least less so than some of the initial dialogues did.

5.4.3 Design of the Technology Probe

Trading Platform. Figure 5.2 shows how the ProberBot dialogue was embedded in the simulated stock trading platform. On the left at the top is a list of all the stocks, underneath is an overview of a person's overall portfolio (including the performance throughout time) and on the right is a graph that shows the stock price over time. The information on the right would update when a stock is selected from the list. The ProberBot appears in a pop-up window in the middle at key moments when the user is about to make a trade.



Stock List

NAME	STOCK PRICE	SHARES	TOTAL VALUE OF POSITION	PROPORTION IN YOUR PORTFOLIO
GrowGen	71.9 ▲6.20 (9.44%)	40 ▲0.00 (0.00%)	2876.00 ▲248.00 (8.6	8.50%
InterLine	44.7 ▼6.10 (12.01%)	40 ▲0.00 (0.00%)	1788.00 ▼244.00 (13.6	
Tramis	44.2 ▼21.90 (33.13%)	40 ▲0.00 (0.00%)	1768.00 ▼876.00 (49.5	

Portfolio Overview

Cash	10,628.00
Invested in stocks	24,032.00
Account value	34,660.00

Increase since last quarter

Your portfolio	2.39%
Uniform investment	2.65%

Increase since 10/2021

Your portfolio	38.64%
Uniform investment	38.37%

TradeBot

If your investment hypothesis has changed, what made you change it?

Recent news

What is your current investment hypothesis for this stock?

My hypothesis is that the stock will

GrowGen

Market Cap	369.2M	Revenue	72M
Enterprise Value	443M	Earnings	8.9M

Figure 5.2: Simulated trading platform interface and ProberBot chat window.

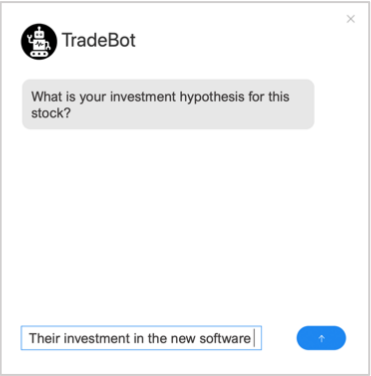
Both the simulated stock trading platform and the ProberBot were designed to look authentic by providing a realistic level of functionality and interactivity. Various types of data that are common in trading interfaces were provided, including stock price time series graphs, price/earnings ratio, price/book ratio, price/sales ratio, analyst ratings, consensus/average target price, market cap, revenue, company information/profile, and company news items. The simulated stock trading platform was designed to provide five years of synthetic stock data, with a 'Next Quarter' button that could be clicked on. Clicking the button would update the stock prices and market situation/context. Due to our focus on longer-term investing strategies, (financial) quarters were chosen as a common and meaningful time interval in which a stock investor with a long-term strategy may consider and assess new investments, as well as their existing investments and portfolio performance. Furthermore, long-term investors would only rarely buy and sell the same stock in time intervals significantly shorter than three months.

The interface shows the stock price data and relevant metrics for three companies, which were intended for our study scenarios. We also designed the combined trading tool and ProberBot so that rules could be set for each dialogue for when it should appear (e.g. when there would be volatility or certain trends in the stock price, duration of holding the stock, etc.).

ProberBot Interface. The NLUI was designed to support three kinds of responses: (i) text inputs (Figure 5.3), (ii) discrete scale inputs, such as 1-10 or 0-100% (Figure 5.4), and (iii) multiple-choice inputs (Figure 5.5) depending on the question asked. The choice of the response/input type depended on the question being asked, and usually, there was a mix of them in each dialogue. This enabled variability for the participants when considering how to react and reflect upon them. The NLUI dialogues were designed to have several branches using a determination logic based on predefined, context-dependent criteria. This allowed the NLUI to output context-aware responses based on previous user inputs, which were mostly follow-up questions or summaries of previous inputs. For the present study the dialogues were 'hard-coded' to appear at specific moments (as part of our scenarios) at which the user would intend to trade a specific stock. This was to assure increased control and comparability of the collected data. After interacting with the ProberBot, the user has the option to move on to confirm or cancel their intended trade.

5.5 Study Design

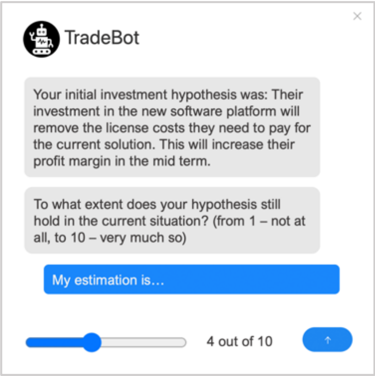
To evaluate the effectiveness of embedding various ProberBot dialogues in the trading software tool, we designed an interactive scenario-based study. Asking participants to interact with the tool themselves would enable them to experience it first-hand and get a sense of how such an embedded NLUI 'works'. The aim of this study was to explore (a) whether experienced investors would think it could help mitigate certain biases based on specific, realistic situations and decisions, (b) understand the reflective processes it could potentially trigger (including metacognitive ones) and (c) how disruptive or intrusive the different ProberBot dialogues may be perceived. Another objective was to obtain insights into people's understanding and conceptualisation of ProberBot and elicit their ideas on other cognitive tasks they would like a (future) probing NLUI to help them with (and which *not*) based on their experience of interacting with it.



What is your investment hypothesis for this stock?

Their investment in the new software

Figure 5.3: Text input example (from Dialogue 1 – part 1).



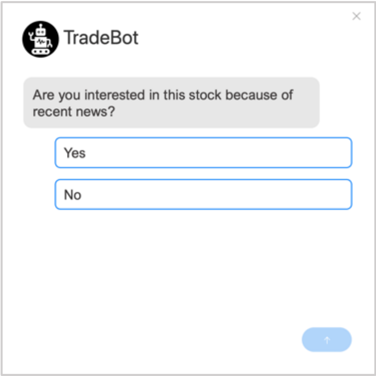
Your initial investment hypothesis was: Their investment in the new software platform will remove the license costs they need to pay for the current solution. This will increase their profit margin in the mid term.

To what extent does your hypothesis still hold in the current situation? (from 1 – not at all, to 10 – very much so)

My estimation is…

4 out of 10

Figure 5.4: Scale input example (from Dialogue 1 – part 2).



Are you interested in this stock because of recent news?

Figure 5.5: Multiple choice/yes-no example (from Dialogue 3).

To provide participants with a realistic experience of a possible use of ProberBot, we designed a series of scenarios for which they had to work out what to do under the guidance of the researcher. The scenarios were: (i) making an investment in a stock which would, in a second step, have to be revisited/re-evaluated after performing badly over several quarters, (ii) evaluating their portfolio after a certain amount of time and deciding which stocks to continue to hold, and (iii) deciding if an investment in a stock which is currently being hyped is adequate by considering the different news items and other information and metrics regarding that stock. For all these scenarios, participants interacted with the NLUI and the respective dialogues it provided. The researcher provided participants with certain hints and suggestions for their actions at specific points, since (a) we could not expect them to go through all the data and make a decision in the available time frame; (b) we wanted each participant to experience the same ProberBot dialogues to be able to better compare their reactions and thoughts. A key difference to VoiceViz thus was that ProberBot delivered the prompts at specific points, for example, when specific actions are performed in the simulated stock trading platform which were previously defined. The tool developed for the study had the capability to deliver the prompts based on different events in the trading simulation (e.g. changes in a stock price or one's portfolio value) or based on the user's activity/interactions within the tool (however, in this study the prompts were constrained to the moment when a participant was clicking on the trade button before confirming a stock trade). Ethics approval was obtained from UCL prior to the study (UCLIC/1819/008/RogersProgrammeEthics).

5.5.1 Participants

Six participants who had different levels of experience in investing in the stock market were recruited for the study. Four participants had professional trading experience in different contexts (e.g. working for stock markets or investment banks), and two participants had long-term private/retail investment experience. We intentionally only recruited experienced stock investors, since understanding the simulated stock trading platform (including the different types of data it provided) and the ProberBot dialogues (and why they appeared) required sufficient investing experience.

5.5.2 Procedure

Participants were informed of the purpose of the study and asked to fill out a consent form agreeing to being audio and video recorded during the study for subsequent analysis. After a walkthrough of the simulated stock trading platform and a familiarisation phase, the researcher provided the first scenario. The interaction with the trading interface was guided by the researcher but was operated by the participant on their computer/browser while they were sharing their browser window through the video conferencing software with the researcher.

The scenarios provided in the study involve significant changes in companies' financial results (e.g. their earnings), forecasts, and their stock price after several financial quarters to imply situations of increased uncertainty in which the ProberBot could be triggered. The ProberBot appeared in each scenario when the participant was about to make a trade (clicking the trade button), which resulted in the bot appearing 3-5 times across the three scenarios. After each interaction with the ProberBot, we asked the participants to guess why it was triggered, as well as whether they thought the questions/interactions were appropriate for the given situation. After their interactions with the ProberBot for each scenario were completed, the participants were interviewed about what they thought of having a probing NLUI embedded into a trading interface, their views on the dialogues it provided, and the extent of its intrusiveness and potential disruptiveness. Overall, the think-aloud interactions with the prototype together with the interview took M = 45 minutes (min. 35, max. 58).

5.5.3 Data Analysis

The recordings were transcribed verbatim. They covered statements made by participants while interacting with the tool and ProberBot and while thinking aloud, as well as the in-depth interviews following the investing task. I then reviewed the transcriptions using an iterative and open form of thematic analysis. I went through all the interviews and inductively coded them. A second researcher went over all the interviews and suggested changes to the codebook and the coding. Subsequently, all disagreements were discussed and resolved between the researchers. The identified themes were organised into (i) perceptions about the value of a ProberBot, (ii) challenges of having a ProberBot, (iii) expected use of a ProberBot and individual needs.

5.6 Findings

Taken together, all six participants had a similar understanding of what the ProberBot was trying to achieve; they thought that its probing interactions could help them reflect during their own decision-making by preventing certain impulsive actions or inconsistencies in their decisions. However, some of the participants also had concerns about whether they would find certain dialogues useful when they are making investment decisions. These and other findings are presented in more detail in the following sections.

5.6.1 Perceptions About the Value of a Probing NLUI

When the participants were asked what they thought the purpose of the ProberBot was, all of them mentioned its potential value of trying to support their decision-making and encourage them to reflect on their intended trades when there is a risk of (re)acting inadequately (e.g. not acting in line with their previous investing decisions and strategy). Participants could see for most questions "where the ProberBot is coming from" or why it asks these questions. For example, Participant 6 (P6) mentioned:

"It made you re-evaluate your thought rather than having a knee-jerk reaction or emotional response (...) so it gives you a little bit time to think (...) it gives you time to reflect upon why you're making these decisions really." (P6)

Participants also pointed out that having to articulate and justify their decision-making process enabled them to reflect more on what they were doing. For example, P1 commented, *"I think it's good to make you actually justify (...) why you are making the decision."* and similarly, P5 said:

"Sometimes I don't articulate to myself why I'm doing this. I suppose I have an idea of what my decision-making process is, but [the ProberBot] made me stop and think about it (…), it makes you articulate what's going on." (P5)

Besides these aspects of articulating and 'explaining' one's thought process, participants also mentioned that they could see how it could help investors keep track of their own past decisions to enable them to revisit or – if needed – revise their investment theses and/or strategy (which was a point particularly related to Dialogue 1):

"It's nice to be reminded of your earlier decision-making process, because sometimes you forget you've actually done quite a lot of due diligence and thought about something very carefully (…) it might make you think, well actually, that still holds now." (P1)

When asked about what they think could be the triggers for ProberBot to appear, their explanations were very similar. For example:

"You could interpret the situation for these companies as being, they were particularly, uhm, not risky decisions necessarily, but where I needed to stop and think about it. It wasn't just a day-to-day kind of trade." (P5)

5.6.2 Challenges of Having a Probing NLUI

Although all participants understood the value of having the ProberBot intervene while they were trading, and were generally in favour of it, they also made comments about certain challenges. For example, two participants were concerned it could interrupt them and be intrusive:

"It's a little bit intrusive, but it's something I think you could get used to if you could make it almost part of the [personal investment decision-making] process." (P1)

Another participant found the ProberBot to be ‘quite curt’. Two participants mentioned that some dialogues may not be that helpful for them, considering their experience and their investing approach. A more general problem was pointed out by P2 regarding Dialogue 1, which was about capturing the initial investment thesis and then re-evaluating it. They mentioned that people would usually find a way to explain their previous actions afterwards (‘post hoc’) and then proceed with their decision even if it contradicts their initial investment thesis or their general strategy:

“You know my challenge with a bot like this is it's forcing me to provide post hoc explanations, right. I've already decided what I'm going to do, now it's asking me to justify it.” (P2)

Dialogue 3, which asked the user about the relevance of a news item for a specific stock/company, was perceived to be less useful by three participants as they said they would *generally not* trade based on the news (but rather based on the fundamental data of the business and its valuation metrics). One participant had even stronger feelings about it, but for a different reason, for them Dialogue 3 was going beyond just scaffolding their decision-making, leading them into a certain way of thinking:

“Is the bot prompting me in a particular way to influence that train of thought somehow? (…) This feels more like it's no longer being a neutral bot, but it's actually leading me down a particular [direction]. So, previously [i.e. in previous dialogues], it was a reflection, but this one really felt like it's pushing me in this direction of making particular assumptions or decisions.” (P5)

Participants also made several comments about the additional time that the interaction with the ProberBot added to the decision-making process. However, this comment was usually made when they were considering other stock traders with short-term strategies (for example, so-called ‘day trading’), acknowledging that this would not apply to long-term investing strategies. One participant describes a dilemma of sometimes “not wanting but at the same time needing” the ProberBot when being in a rush and thus being at risk of acting against one’s long-term strategy or goals:

“So now you need to click through it, and some time passes, and if there's something which is really time sensitive so, for example, if I want to trade it now and I don't want to waste time telling the bot,

why I wanted to buy or sell now, then I think I should have the ability to skip it. So maybe this is when the market is moving really quickly, but at the same time, maybe someone else might say, this is exactly when you need it, so you don't overreact." (P4)

In other words, what P4 points out here is that some ProberBot interactions can feel disruptive and slow the user down to reflect, but that this may be exactly what is appropriate in certain situations where an investor may get emotional, impatient, or rushed.

5.6.3 Expected Use and Individual Needs of a Probing NLUI

Participants had several ideas for further situations when they would use probing NLUI like ProberBot and how it could prompt them with certain reflective questions, including what could trigger the questions to appear (e.g. certain user traits, states, or behaviours). For example:

"I would ask the bot to ask me whether I'm really sure I want to buy or sell. After 8:00 o'clock at night, I might have had too much to drink (...) I think that when the market dropped with COVID, it would have been very useful to have a bot saying. 'Are you sure [you want to sell]? Oh, go away and think about that.'" (P6)

Although this was a somewhat humorous suggestion, the underlying idea of the ProberBot probing their thinking and asking them about their emotional state and current context was also made by two other participants.

Participants also mentioned that a probing NLUI could help them better keep track of and make sense of their own performance and decision-making processes, including how a decision was made, for example:

"Was the reason [for a previous 'inappropriate' decision] I did not do my due diligence properly, or did I do something different this time, or was it 2:00 am in the morning? If you could reflect that back that would be very useful." (P1)

Somewhat related to these aspects of the context or the way in which a decision is being made, four participants mentioned that a probing NLUI should have certain options to control it. For example, P3 said:

"Sometimes you go against your own nature [knowingly] (...) and a bot kindly reminding me, then, maybe I don't care what the bot says. Like, I know what he's going to say because I've used him[15] *for months now, and I know his nature, so it's almost like, you know, sometimes you just don't want to hear it. I'm going to trade this regardless (…). You could have like three levels of intrusiveness right like 'standard', 'super helpful', or 'on the sidelines'."*

Finally, it was noted that investing decisions often depend on a variety of factors in addition to those currently reflected and considered in the ProberBot dialogues, such as desired portfolio proportions/allocations (e.g. due to a specific strategy) that should be achieved or maintained. For example, P1 and P3 said:

"Sometimes, that means I have to sell something else [first, before buying], so I have to justify not only the buying process, but actually is the thing I'm buying going to do better than me just hanging on to what I've got already." (P1)

"Was this – even if it's just trading one stock – a response to a market condition or is it in response to analysis of the portfolio, overall, or is it based on this instrument in, you know, in a silo." (P3)

The fact that (investment) decisions are often interdependent, as pointed out in this quote, is a key consideration for the design of probing/scaffolding NLUIs for decision support. As such, the dialogue design needs to take into account that decisions can build upon each other (i.e. depend on past decisions and affect future decisions), or they can be sub-decisions of a higher-level decision (e.g. the decision to sell a specific stock due to the higher-level decision to rebalance the portfolio in a specific way, as P3 alluded to in the above quote). Translating these interdependencies into equally (or at least comparably) interdependent dialogues, which are then triggered at appropriate points in the decision-making process, is a key design challenge of building such probing NLUIs – among various other challenges and considerations, which we will discuss in the next section.

[15] Interestingly, P3 and P5 referred to the ProberBot with "he/him" pronouns, although we introduced it without a specific gender. The other participants referred to it in a gender-neutral way.

5.7 Discussion

The findings from the user study with experts suggest that having a probing NLUI embedded into a trading platform can help a stock investor's decision-making by scaffolding their reflective thinking in the moment they make decisions (i.e. *reflection-in-action*). In line with the *self-explanation effect* [78] it seemed that externalising their decision rationales by further explaining and motivating them in response to the NLUI's questions helped participants better think through their decisions. Furthermore, our qualitative analysis suggests that the NLUI helped raise awareness of the potential biases the investors may not have been aware of at key moments.

In contrast to VoiceViz, the delivery of prompts was here *not* provided based on certain time intervals or decided in the moment when there seemed to be an opportunity for the NLUI to intervene but rather at specific points based on certain rules – in this study constrained to when participants were about to make a trade[16]. Given that participants interacted with a functional trading simulation, the moment of commencing a trade was one of the key user actions where encouraging reflection seemed most promising.

Encouraging this 'taking a step back' to reflect at this point may be most preferable for trading tasks that are not overly time-sensitive or following short-term strategies (as for example in 'day trading'), where rapid decision-making is critical. For long-term investing strategies, ProberBot was appreciated for its ability to question their motivations, helping them to critically reflect on and explain their decisions – which they may not do when acting 'in the moment'. The participants did not raise specific concerns about the time the interaction with the ProberBot may add to their decision-making process, but rather that this added thinking time is often even desirable for investors with long-term strategies. It seemed that the process of externalising their decision rationales enabled participants to question and reflect on their own decision-making process and to get an understanding and awareness of what some of the intricacies and potential pitfalls of a decision are as well as of their own decision-making process (i.e. metacognitive awareness). Thus, this 'NLUI-moderated self-explanation' seems

[16] This generally resulted in participants receiving a prompt approximately every 4-6 minutes (depending on how quickly they progressed with the task).

to be a promising approach to get people to reflect on and improve their decision-making and their strategies.

However, the study also revealed challenges of using a probing NLUI in this manner. For example, some participants thought that an NLUI like ProberBot could be intrusive or that its dialogues may sometimes seem not relevant or effective (corroborating some of the challenges of proactivity discussed in Section 2.1.5 and Section 2.1.6). This implies that one of the difficulties of integrating a probing NLUI into existing software tools, in the way we envisioned, is the risk of disrupting ongoing decision-making. The question was also raised as to how a probing NLUI could be designed, given it would not always have the relevant information for why the user is making a certain decision (e.g. someone selling stocks of a company due to having to pay back a student loan and not because they do not believe in that company anymore). This resonates with previous research, which highlighted the importance and relevance of proactive interventions for people to accept them [72, 122, 279, 280, 285, 286], which requires the NLUI to have sufficient 'awareness' of what they are doing and what their goals are. There were a few instances where participants were not sure of the benefits of specific ProberBot dialogues. For example, three participants mentioned Dialogue 3 appeared not relevant to how they usually make their investment decisions. As the findings also suggest that it may not be straightforward to develop dialogue scripts that users find helpful for all biases and that only certain biases are suitable to be addressed by a probing NLUI. Furthermore, people may require some time to get used to this type of probing dialogue to overcome the perception it is getting in their way.

Hence, it is not straightforward as to how to design a probing NLUI's dialogues so that they are 'delivered' at opportune times. They need to be sensitive to different user needs and when best to intervene (see Section 2.1.6). The user could be involved in helping here, by suggesting where and when they would like their NLUI to intervene. This could include 'telling' the probing NLUI, in the set-up phase, their personal level of experience, investing goals, and investing strategy and then, when in use, having specific settings for how much the NLUI should intervene (e.g. 'standard', 'super helpful', or 'on the sidelines' as suggested by one participant).

Having such controls could not only make the probing NLUI more tailored to the users' needs but also make them feel comfortable to receive its prompts/probing dialogues from an early stage of usage, as they would be able to anticipate to some extent when and how certain dialogues would be triggered. This could potentially also result in a higher willingness to interact and engage with the probing NLUI and consider its probing questions in their decision-making.

The findings from the study also suggest there may be several 'tensions' involved in using probing NLUIs in the context of decision-making, which need to be considered. These include:

Tension 1. The first tension is related to the dilemma of 'not wanting but at the same time needing' the ProberBot. This was alluded to by P4, where a person might be acting emotionally, responding to a recent market event, without having reflected sufficiently on their decision-making. Due to being driven by their emotions, they may not be receptive to the probing NLUI's cognitive scaffolding. One possible way to address this tension is to remind them of their strategy and goals before providing the prompts to them.

Tension 2. As was mentioned in the interviews, there is also the risk – even if the probing NLUI's questions are designed to be as neutral and non-leading as possible – that people will speculate why a certain question might be asked and wonder what aspect in their decision-making the probing NLUI might have 'picked up on' and consider as inadequate or flawed. This may get people to not just see the question as a *reflection prompt* but as a *suggestion* for them to (not) do something so that they might try to read 'between its lines' – as was the case for Dialogue 3, for example. This underlines that there is a fine line between an NLUI that just helps people externalise their thinking and reasoning effectively, versus an NLUI that is perceived as trying to influence their decisions and nudge them in a certain direction – in other words, when dialogues are not carefully designed, they may quickly be perceived as some form of ('hidden') advice, which is not what a probing NLUI that intends to support reflective thinking is trying to achieve. The phrasing of probing NLUI's dialogues thus needs to be carefully crafted.

Tension 3. Somewhat related to the second tension, even if the idea of a question might be to support reflective thinking and perhaps even debias a decision, a challenge is that a change in a decision might not always lead to a better outcome (e.g. of a specific stock's performance over time) – even if it might seem objectively better (e.g. better grounded on a range of different criteria). This could lead to frustration, in particular, if people changed their decision because they thought there to be an implicit suggestion in the NLUI's probing question as mentioned in the previous tension. One way to mitigate this challenge is to make clear to people that the purpose of the probing NLUI is only to help them reflect and not there to provide any suggestions.

Tension 4. There is the 'tension' that a question may not always lead to a more systematic or thorough consideration of different criteria but could even lead to a reinforcement of an intention already created in a person's mind – due to people's tendency to look for information that confirms their existing beliefs (confirmation bias). This is also related to the challenge P2 referred to, that a question might just get them to rationalise their decision *post hoc*. This could be partially mitigated by designing the NLUI's questions so that they generally try to get people to expand their explanation of and reasons for an investment thesis.

Tension 5. Some participants mentioned that giving the user control over when and how the probing NLUI intervenes in their decision-making could improve the experience and help them better integrate the NLUI and its prompts into their decision-making. However, participants at the same time also pointed out that it may be difficult to know how best to configure when the probing NLUI should or should *not* intervene – which is also related to the first tension, that sometimes when someone may not want to engage with the probing NLUI, they may in fact need it most. A possible solution is that the user could skip more extensive dialogues, but they would have to 'explain' to the probing NLUI why they would like to turn it off for certain transactions or for a certain period of time before being able to turn it off (e.g. for the example mentioned above where the reason for selling a stock might be that they just need some cash to pay back a student loan). However, no matter how this is addressed, there remains a certain 'tension', since allowing the user too much control could undermine the main purpose and the effectiveness of them as neutral 'probes' that are intended to get users to reflect on their decisions in the moment and to reduce the risk of being impulsive and regretting it later.

Addressing and trying to meaningfully balance these ‘tensions’ could increase the likelihood that cognitive tools like probing NLUIs can successfully enable people to engage in meaningful reflection processes as part of their decision-making while also making sure that the NLUI’s probing questions are not too intrusive, burdensome, and distracting for them.

However, there might be situations where such probing NLUIs are not the right approach because some of the tensions may just be too challenging to be meaningfully addressed and balanced. Yet, for certain tasks and contexts probing, NLUIs could be a promising approach for extending NLUIs to help human decision-making ‘within’ the respective interface used for decision-making. They might help people externalise their thought process and reasoning through a scaffolding conversational interaction, helping them make sense of, structure, clarify, and refine their decision-making.

5.8 Conclusion

The findings of this study have revealed the potential benefits of designing an NLUI embedded into a software tool used for decision-making that can support reflection. The reflective thinking ‘targeted’ the decision-making process itself (e.g. evaluating which information a person considers to be relevant for their rationale) as well as the *metacognitive* processes (e.g. becoming more aware of how they make decisions). The study also showed how it can be possible to augment human cognition by considering a probing NLUI that enables people to reflect on their thought process by expressing it in their own words and by using a range of interface elements. Hence the approach that is being advocated here is how designing an NLUI to be more of a *probing* NLUI can enable the user to reflect on their decision-making in the moment – especially in situations when there is a risk of impulsive and potentially biased decisions. However, the comments made by the participants suggest that such probing NLUIs should prompt users sparingly, so as not to annoy or distract them too much. The findings build upon the previous chapter, where the NLUI ‘VoiceViz’ was found to enable people to make sense of the trends and patterns in a dataset – here, the NLUI also enabled people to make decisions based on data (i.e. stock data) as also shown in Table 5.1 below.

Table 5.1: Outcomes of reflection targeted by the NLUIs in the first two studies.

What the Reflection Targets	
VoiceViz:	• Understanding the patterns in the data and their potential causes
ProberBot:	• Understanding the relevance of different data for a decision • Understanding one's own decision-making behaviours

The findings show how a probing NLUI can help people reflect on and make sense of the relevance of different types of (stock-related) data as well as their own decision-making behaviour. The question this led to was how could this be taken further – more specifically, how could NLUIs also support people in having other forms of self-insight when reflecting on their behaviour? The study described in the next chapter addresses this question by exploring how an NLUI can be designed to support people in reflecting on their past experiences and behaviours through creative expression to improve upon them in the future. Thus, instead of focusing on reflecting on the ways in which one makes decisions, the focus in the next chapter is on the ways in which someone can gain new understandings of and insights into challenging situations and experiences in their life through creative self-expression.

6. SelVReflect: Fostering Reflective Thinking in a Creative/Expressive Task

The study reported in this chapter explores how cognitive prompting by an NLUI can support self-reflection as part of/through an expressive task. In the previous study, reflective thinking was encouraged by the system to enable the participants to obtain new perspectives on their own decision-making. The study presented here focuses on (a) *self*-reflection and (b) externalisation of one's thoughts and emotions through visual expression. The question asked here is if it is possible for people to (a) engage in deeper forms of reflective thinking enabled by an NLUI, and (b) how can the externalisation of thoughts supported by the NLUI be performed through a different medium/modality than through verbalisation – such as a virtual 'canvas' for free expression.

The goal was to ask participants to visually express a past personal challenge in a VR-based 3D 'canvas' to understand how they dealt with it and ultimately overcame it. The task chosen for the study was inspired by art therapy, which can enable people to have new insights on themselves or give experiences a new meaning by expressing them. The idea was to enable people to discover ways how to deal with similar challenges in the future. The reason for choosing VR as a medium in/through which to do this is that it offers an immersive space for exploration and creation, providing opportunities for self-expression and reflection. SelVReflect was based on a VR tool that has been used in previous research for expressive activities, which was then combined with an NLUI built as part of this study. As in previous chapters, the aim was to provide scaffolding questions at opportune times as a form of guidance and encouragement to users while they engage in a task, which in this case was creative/expressive.

The specific research question addressed here was whether embedding a set of voice-based prompts into an expressive VR tool would be able to help people structure their thinking. The reason for selecting voice rather than text prompting is that they would augment the visual experience of expression and exploration within the VR space without directly interfering with it (i.e. instead of prompts being displayed in the virtual environment in VR), which was also confirmed by a user-centred design process conducted as part of this project that focused on the NLUI's characteristics. The question this raises is whether being immersed in a VR experience while being verbally prompted to think enables people to reflect in similar ways as in the two previous studies? If so, how does it achieve this?

6.1 Introduction

Challenging experiences in our everyday life – such as in professional contexts or in personal relationships – can be linked to anxiety, stress, or difficulties with planning, prioritising, or decision-making, and negatively affect our wellbeing [1]. Reflecting on such experiences can facilitate understanding, provide new meaning, offer more self-insight [36], and support life changes [399] and personal growth [59, 265, 390]. Here, we chose to enable people to reflect on a *successfully mastered* past challenge – which is related to the concept of 'mastery experiences' from self-efficacy theory (see Bandura [29]). Through that, people would be encouraged to specifically focus their reflection on the aspects that helped them overcome the challenge.

NLUIs have been shown to be effective for guiding and facilitating self-reflection in wellbeing and mental health contexts [10, 21, 43, 178, 220, 226, 228–230, 251, 253, 314]. They have been explored for their potential to provide listening services, where their design is less focused on the NLUI's ability to provide answers but rather on their ability to evoke compassion, examining the effects of vocalising difficult thoughts to a computer [314]. For example, Lee et al. [253] presented an NLUI which was able to encourage deep self-disclosure in people, as well as a chatbot which asks for help, with the goal of fostering the ability in users to take a more objective, compassionate look at their own challenges in life [251]. Kocielnik et al. [230] designed *Robota*, an NLUI that aims to stimulate reflection and self-learning in the workplace by asking questions and chatting with the user of the system. Perhaps the most commercially successful of these self-reflective NLUIs has been *Woebot*, a counselling chatbot which has been shown to reduce anxiety, depression, and unwanted behaviours through the use of Cognitive Behaviour Therapy (CBT) [141, 329]. This research shows how NLUIs can ask people questions about their (challenging) experiences in their lives and the ways in which this can enable people to express themselves and reflect on these experiences.

In contrast to this research on NLUIs for expression and self-reflection, the approach taken here is that people express themselves *visually* instead of through words in response to the NLUI's questions. The present approach was inspired by art therapy where (guided) visual expression can – among other things – help people 'access' and express emotions that may be difficult to articulate, explore complex issues in a non-threatening manner, gain personal insights, and give experiences a new meaning by representing them in new ways [267]. More specifically, this was enabled here through creating *three-dimensional visual representations* in VR. The main reason for choosing VR are its unique affordances that allow people to *construct* and *experience* personal creations and environments in immersive and engaging ways [262] that go beyond what other interfaces and modalities can offer. Furthermore, VR has been found to provide meaningful interventions for a number of domains, including self-reflection [25, 240, 400] and wellbeing [47, 128, 146, 428]. For example, research suggests that it can alleviate stress through guided imagery (e.g. [381]), support emotion regulation [290], enhance creative expression in art therapy [57], and elicit positive change in mood, meaning-making, and interpersonal connectedness [50, 225]. Based on this research, it is proposed here that VR offers a unique 'canvas' for self-expression and self-reflection, allowing the use of dynamic elements in a 3D virtual environment which can be used to represent different aspects, such as temporal sequences or thematic relationships. The idea is that the resulting 3D creations and their structures can then be further explored by physically *walking through* them and approaching them from different perspectives. This spatial exploration can not only allow people to 'immerse themselves' in their creations but also give them the opportunity to take on different perspectives, which can enable them to look at the events or experiences that they have represented from a bird's eye view. The latter can also be referred to as taking on a 'self-distanced perspective', which can be helpful when making sense of past challenging experiences and interpreting them in new ways (see [235]). Based on these characteristics, affordances, and opportunities of VR, the goal here is to use a VR 'canvas' to offer people a space to express personal experiences in response to questions they receive from an NLUI, which aim to enable in-depth reflection and provide inspiration for their creative process.

Like VoiceViz and ProberBot presented in the previous chapters, the NLUI's questions are thus also designed to encourage users to explore and take on new perspectives and help them express their thoughts. However, here the questions are designed to enable people to visually express certain thoughts and ideas – which is quite different to expressing oneself through words. The focus was, firstly, on how to design the NLUI's questions and then deliver them as an 'NLUI+VR' tool, given the different form of interaction; secondly, how people express and externalise their thoughts (in the VR 'canvas'), and thirdly, how the questions support self-reflection through visual expression. Specifically, the following two research questions were addressed:

- How can we design an NLUI-based VR experience that fosters self-reflection through guided creative expression?
- How does our design affect the overall experience and reflection?

6.2 Background and Related Work

Several kinds of reflection, as outlined in Section 2.2.1, have been covered so far: VoiceViz (Chapter 4) mainly focused on supporting reflection on 'external material' (the dataset with its trends and patterns), ProberBot covered both reflection on 'external material' (one's portfolio and the stocks and their characteristics) as well as on oneself (one's past decisions and one's decision-making process). The focus of SelVReflect, as the name suggests, is mainly to reflect on oneself in/through a creative activity.

As introduced in Section 2.2.1, there are different types of (self-)reflection and ways to conceptualise it. Of particular relevance here is Schön's [42] framing of reflection with its differentiation between *reflection-in-action* and *reflection-on-action*, which has been widely used [36, 373, 390]. While *reflection-in-action* was most relevant for the tasks studied in the previous two chapters, *reflection-on-action* is of most relevant here, as people use their memories of an event to reconstruct an experience. This effort of 'stepping back into' the experience and retrieving and organising the aspects we remember is done to understand what has happened and to draw out lessons for the future [373].

While self-reflection is generally considered to be helpful [36, 236, 289], it can also be challenging without support [390], and there is a risk of getting stuck in negative thought cycles [414], which needs to be carefully considered when designing systems to support reflection [123]. One way to mitigate these challenges and risks is through providing guidance. The research suggests that moderately directed guidance helps with reflection [80, 96, 141, 390]. There is a range of existing (human-human) practices of guiding someone through reflective and expressive activities from creative, educational, or therapeutic contexts (e.g. [121, 267, 339]). Such techniques can help someone more successfully navigate through different kinds of self-expression and reflection.

Psychological and psychotherapeutic approaches that are of particular relevance here are goal-oriented psychotherapy [358] and Positive Psychology [376, 446]. The main goal of these approaches is to help people understand how they can cope with challenges by focusing on which of their existing resources they could use to do so. The so-called Agentic Positive Psychology [31] specifically focuses on 'mastery experiences', which refer to personal (past) successful behaviours to deal with and adapt to certain situations (e.g. [30]). Such mastery experiences have been proposed as one of the most effective ways to instil the belief in one's own ability to succeed (e.g. in performing certain desired behaviours), and they can lead to positive behaviour change [29]. This is referred to as *self-efficacy* [29]. Self-efficacy can be measured with questionnaires, such as the General Self-Efficacy Scale [375]. An intervention procedure that builds upon self-efficacy and mastery experiences is *self-modelling* [116], in which people observe some representation of themselves (e.g. through a recording) of how they successfully complete/master a challenge.

Here, mastery experiences and self-modelling are combined with self-expression and reflection. Expressing and reflecting on oneself can help people understand challenging events better and give them a new meaning [59, 265, 390]. The use of self-expression to reconstruct, take on different perspectives on, and explore the meaning of an experience is also used in certain therapeutic approaches, such as art therapy [267]. A foundational theory in art therapy called the Expressive Therapies Continuum [204] forms the relationships between drawing conceptual meaning using reflective activities and how feelings of positive affect occur through expressive activities. As an immersive medium and multi-dimensional 'canvas' for creative expression, VR has also been recently explored as a tool to be administered to patients during advanced stages of art therapy [172, 173]. However, the authors point out that it is important to have (verbal) guidance in the process of creation as it can be difficult or overwhelming (see also in Section 2.2.1).

This can particularly apply to people who may have more difficulties with opening up and expressing their experiences and emotions [340]. This needs to be considered when designing systems for such purposes. Here, the Dimensions of Emotional Openness model (DOE) by

Reicherts et al. [340] can be a useful 'predictor' for how difficult a person may find it to perform such a task: People engaging in creative self-expression and reflection (in VR) need to become aware of what emotions they experience and how to internally and externally represent, express, or communicate them. Therefore, it can be assumed that people with lower DOE scores (i.e. who are 'less' emotionally open) will benefit the most from receiving guidance from an NLUI during the expressive process. The next section will describe how this NLUI guidance was designed and combined with the expressive 'VR canvas'.

6.3 Design

An iterative user-centred design process (UCD) was followed to design the prompts and the VR tool. The main reason for this was that the existing approaches for guiding reflection discussed in the previous section could not be directly translated to the given VR scenario due to its unique (expressive) affordances, immersion, and the different interactions it involves (e.g. compared to art therapy). The following questions were addressed: (i) what role the NLUI should play in the reflection process, (ii) how and when it should appear (in VR), and (iii) the types of guidance and specific prompts it should provide. The UCD was divided into four stages, which are shown in Figure 6.1. Five participants took part in each stage. To reduce the novelty effect, participants with prior knowledge of VR were invited to take part, comprising 2 women and 3 men with an average age of $M = 27.6$ years (range: 21 – 31). Two were students, two were PhD students, and one was a research assistant. While all were familiar with HCI research, none of them had previous experience with 3D drawing nor with comparable voice-based guiding NLUIs.

6.3.1 Individual Design of the Virtual Environment

The NLUI+VR system we developed was called *SelVReflect*. It comprised a VR tool that was embedded with a set of voice-based prompts. It uses an adapted version of a tool called *'Mood Worlds'* (which is an adaptation of *'Tilt Brush'* [17]), originally developed by Nadine Wagener [429]. Mood Worlds offers a palette consisting of various (animated) brushes, animated and static 3D objects, pre-set environments (e.g. mountain, beach, forest) and a

[17] https://www.tiltbrush.com

colour panel, allowing the user to create their own 3D creation in this VR canvas. Wagener et al. [429] found that this approach of letting people build their own representations of personal experiences in VR can enable not only re-experiencing them and re-engaging with them in a new way, but it also supports positive feelings. However, Mood Worlds was not specifically developed to support reflection, but rather as a way to express and 'engage with' past experiences involving positive emotions, and it did thus not provide any guidance as support for users to reflect.

6.3.2 Designing the NLUI and its Guiding Questions

Figure 6.1 shows the four stages of the UCD process. In *Stage 1: Creating a Personal Experience*, participants were asked to visualise a personal challenge using Mood Worlds (hence, without any guidance). This experience was developed to familiarise them with creating representations of personal experiences in VR and to help them with identifying possible hurdles that users might face without an NLUI supporting them. The drawing was followed by a short interview (approx. 5 min), in which the researcher inquired about thoughts and ideas behind the creation to stimulate users' reflection regarding their visual representation.

The second stage, *Stage 2: Rewatching to Identify Personal Needs*, focuses on identifying opportunities for prompts through watching a screen recording from Stage 1. Whenever participants remembered themselves struggling to express aspects of the challenging experience, they described difficulties and reasons in a template that provided a structure they could follow when taking their notes. They then added ideas on how guidance could have helped at this point, such as an affirmation, inspiration for a new idea or an approach to visually represent an aspect of the experience, or a question for reflection. They also included suggestions for specific wording for those prompts.

Stage 3: Discussing & Agreeing on Design Solutions brought all five participants together in a focus group to compare and discuss difficulties and opportunities for guidance. This UCD stage lasted one hour. The focus was placed on *when* and in *which form* (modality, phrasing, voice, etc.) the guidance should be delivered. Apart from the general desire to feel comfortable listening to the voice (rather than having text prompts), there was no agreement on its specific characteristics among participants, such as its gender and degree of human-likeness.

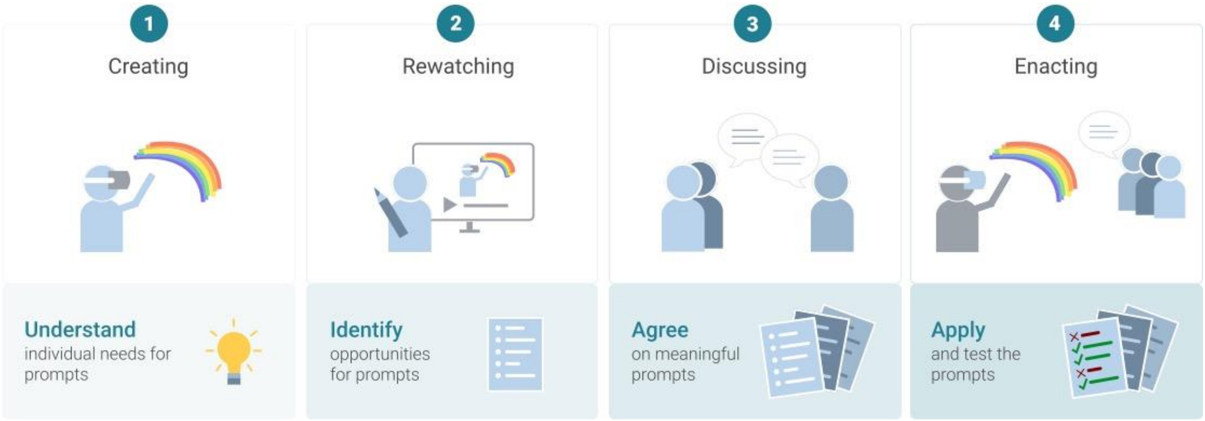


Figure 6.1: The four stages of UCD to design the NLUI guidance for SelVReflect.

When the flow of the open discussion seemed to hesitate, a moderator provided new questions to the discussion, e.g. getting them to think about the role and goals they imagined the NLUI to play and possess, or how it would compare to having human-driven guidance.

Finally, *Stage 4, Enacting & Evaluating Feasibility,* aimed to explore which of the previously identified aspects work well in practice, especially with a focus on the types of guiding questions and their timing. Another researcher, who had not experienced the VR tool before, took on the role of the 'SelVReflect user' and represented a personal challenge in VR. The participants observed and played the role of the guiding NLUI by intervening with reflective questions (approx. 20min). Afterwards, the participants and the 'user' discussed their insights (approx. 20min).

Each stage further informed and refined the design of the NLUI and its guiding prompts for the given usage scenario. Participants wished for a non-embodied NLUI. As a key reason, they mentioned that feeling watched could negatively affect creativity and the experience, while a non-embodied NLUI would act as a facilitator in such a creative self-reflection process. Its role in encouraging and scaffolding self-support through prompts and questions is different from a counsellor providing support, who in everyone's eyes, should always be a human. The NLUI should talk in a reassuring, non-judgmental voice and be available whenever the user requests inspiration. As there was no agreement on the voice characteristics, it was decided that users should be provided with different voice options.

Identified Phases of the Experience. The four-staged design process revealed that user needs differ and can be categorised into three phases. The NLUI's prompts, therefore, need to be designed accordingly. In the beginning, users may feel insecure about how to start with visually representing challenges within the VR canvas. This led to an insight that the prompts should facilitate decisions and help create an initial visual representation of the challenge *in their heads*. This would pave the way for getting an understanding of the experience and its different aspects, components, and emotions they might involve so that one could then more seamlessly build the external representation *in VR*. One participant of the UCD (UCD P2) wanted to *"be prompted to then think more abstract and just draw something that works for me"*. This comment refers to participants' desire to receive inspiration for how they could express certain things (in an abstract way) and what could be approaches to do so – in particular at the beginning when they might not know where and how to start. As soon as a user seems confident with the VR system, the NLUI should not be as prominent anymore. *"Less is more"* (UCD P5) for this phase, in which a user should get into the flow. This corroborates the findings of the studies in the previous two chapters. However, when progress stagnates, it should remind users to think in abstract ways, to take a new perspective into account or to choose another tool. Participants recommended avoiding questions starting with "why", because, as UCD P3 pointed out: *"It is an open creative space where people should not feel bad about their creation"*. Once the creation has been completed, the NLUI could ask if they were happy with their creation and nudge them to physically move around to 'walk through' and look at their creation from different perspectives. Reflective and guiding questions were suggested, such as *"Think about the situation again. Does your drawing reflect it sufficiently well?"*.

6.4 Final Prototype

Based on the design process, the NLUI-based VR experience *SelVReflect* was created. A schematic representation is shown in Figure 6.2 (based on an excerpt from the representation created by P8). As our findings and the previous literature do not provide a clear picture of the preferred gender and degree of human-likeness, participants were able to choose between female/male and human/synthetic voice. By allowing participants to choose their preferred voice, it was intended to reduce the risk of feeling discomfort when listening to the NLUI, which could lead to undesirable effects (confounding factors) on the expressive/reflective task.

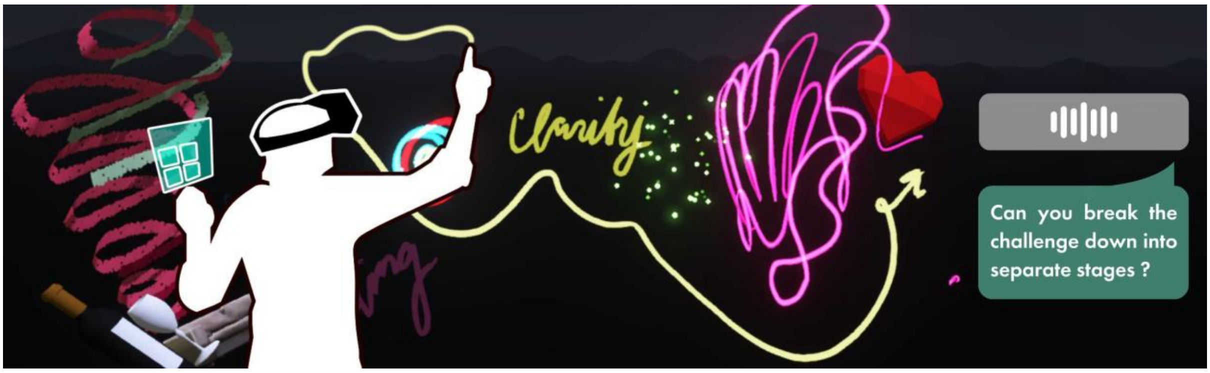


Figure 6.2: Schematic representation of a user creatively reproducing and reflecting on a past challenge using SelVReflect while being guided through the NLUI's prompts.

In line with the findings, we defined three phases for the guided VR experience: (i) *'Warm-up'*, (ii) *'Free-flow'*, and (iii) *'Re-walk'*, in which the NLUI takes on different roles and addresses specific needs. In *Warm-up*, users actively request prompts to get started that point out some tools and features of the palette. In *Free-flow*, NLUI prompts are either triggered by the user pressing a specific button on the VR controller or through a longer period of inactivity, indicating that the user is unsure how to proceed. This approach of both request-based and proactive delivery of prompts can be referred to as a so-called *mixed-initiative* interaction [7]. In a pilot test, four participants used an initial version of the VR tool without prompts; they were asked at different points in the process when they were inactive and appeared to be stuck what they were thinking about at the specific moment and what the reasons for their inactivity were. The findings from this testing suggested that an adequate threshold for triggering a prompt would be after about 30 seconds of the user either (a) looking at their creation without drawing or placing objects or (b) scrolling through the menu without making a choice (indicating that they might be stuck or unsure how to proceed with their creation). Finally, in the *Re-walk* phase deeper reflection questions are asked, which the user can request and proceed through at their own pace.

The NLUI addresses key needs identified in the UCD. For the first phase *(Warm-up)*, this was to receive concrete actions to get started (i.e. overcoming the "blank page syndrome") and creating a basic structure. For the main phase *(Free flow)* this was to receive (i) *inspiration* for new ideas about aspects to express and for using the tools, as well as (ii) *encouragement* to motivate users to continue drawing and to be expressive. For the last phase *(Re-walk)*, this was to receive thought-provoking prompts to enable in-depth reflection. Table 6.1 shows example prompts for each phase of the SelVReflect experience. Each prompt contains an inspiration and an encouragement part, for example, *"[Inspiration:] Have you considered how the stages are connected with each other? How could you design these connections? [Encouragement:] Remember, there is no right or wrong here – as long as you express it in a way that feels right to you, that's all that matters."* The inspiration part of the prompt is usually in the form of a guiding question. Another researcher and I reviewed the list of all prompts, merged similar ones, and improved the phrasing to make them clear, non-imposing, playful, and easy to understand.

The creation of the final set of prompts was informed by principles used to foster self-reflection, self-expression, and creative flow, drawing from the following areas:

1. Counselling [339]: Questions that are asked here by the counsellor often intend to prompt people to identify emotions and thought patterns, encouraging self-guided emotional insight and problem-solving.
2. Art therapy [267]: Here, the questions (which can be similar to those in Counselling) are combined with specific expressive tasks, such as creating an image that illustrates how a challenge or other experience can be 'decomposed' into different aspects and components.
3. (Self-)Reflection research [143]: Guidance here should explicitly structure and offer encouraging prompts, for example, to review and reflect on any produced material.
4. Education [121]: Here, prompts are often designed to make learners feel secure, supported and encouraged to explore and take risks.

Considering Fleck and Fitzpatrick [143], who describe a spectrum of five consecutive levels of (self-)reflective thought, ranging from "No Reflection" (R0) to "Critical Reflection" (R4), SelVReflect is designed to facilitate reflection for users to reach at least the "Dialogic Reflection" (R2), which refers to: *"Looking for relationships between pieces of experience or knowledge, evidence of cycles of interpreting and questioning, consideration of different explanations, hypothesis and other points of view."*

Table 6.1: Example prompts from the three phases of the SelVReflect experience.

Phase	Example Prompt	Purpose
Warm-up	Now think about the main stages of the challenge from the start until the end, when you ultimately overcame it. How many different stages or steps were there?	Suggestions for concrete actions
Free flow	Have you considered how the separate stages might be connected to each other? How could you design these connections? Could they differ? Remember, there is no right or wrong here – as long as you express it in a way that feels right to you, that's all that matters!	Receive inspiration and encouragement for expression and reflection-in-action
Re-walk	Now, focus again on the actions and ideas that helped you overcome the challenge. How did you represent these and how do they tie into the whole process?	Receive thought-provoking questions for reflection-on-action

6.5 Evaluation

We conducted an exploratory user study to see how people would use SelVReflect and whether it could help them reflect in the way envisioned. The overall aim was to evaluate how SelVReflect affects users, with a particular focus on the dependent variables affect (positive and negative), self-efficacy, and reflection. We further investigated how differences in emotional openness within our participant sample affect the above dependent variables. Ethics approval was obtained from UCL (UCLIC/1819/008/RogersProgrammeEthics) prior to the study.

6.5.1 Participants

We used our extended social network and snowball sampling to recruit participants. In total, 20 participants took part (7 females, 12 males, 1 non-binary). The age of the participants who took part in the study was *M* = 29 years (*min:* 29, *max:* 53). Table 6.2 shows more details for each. Participants were recruited from four research labs within different domains and from industry and received a remuneration of £10. Participants self-indicated that they felt mentally stable and healthy at the moment of participation. Most participants had some experience with VR (14 have used it a couple of times or less and 3 people on a regular basis).

Table 6.2: Overview of the participants.

	Age	Gender	Profession	VR Exper.	Chosen Voice	Duration	Challenge Context
P1	31	male	Student	minimal	Synth. Male	23 min	studies
P2	23	male	Student	minimal	Human Male	19 min	studies
P3	30	male	Project manager	minimal	Human Female	17 min	work
P4	25	male	Student	minimal	Human Female	21 min	studies
P5	28	female	Scient. assist.	minimal	Synth. Male	34 min	relationship
P6	31	male	Scient. assist.	extensive	Synth. Female	21 min	studies
P7	31	female	Scient. assist.	occasional	Human Male	18 min	work
P8	32	female	PhD student	minimal	Human Female	22 min	relationship
P9	32	male	PhD student	extensive	Human Male	10 min	work
P10	25	female	PhD student	occasional	Human Male	22 min	studies
P11	27	female	PhD student	minimal	Human Female	34 min	studies
P12	27	male	PhD student	minimal	Synth. Male	18 min	work
P13	26	male	PhD student	minimal	Synth. Female	28 min	work
P14	28	non-binary	PhD student	occasional	Human Female	36 min	studies
P15	27	male	IT specialist	minimal	Human Female	35 min	friends & family
P16	24	male	PhD student	minimal	Synth. Male	28 min	friends & family
P17	29	male	PhD student	occasional	Synth. Female	33 min	university
P18	26	female	PhD student	minimal	Synth. Female	14 min	friends & family
P19	53	male	Scient. assist.	minimal	Human Female	31 min	work
P20	27	female	PhD student	minimal	Human Female	20 min	studies

We used the DOE-20 questionnaire [340] to assess participants' affect processing. The questionnaire encompasses five components, including cognitive-conceptual representation of emotions (REPCOG) and communication and expression of emotions (COMEMO). Both traits are relevant for the task of expressing and representing emotions, which participants subsequently carried out in SelVReflect. The participant sample had scores similar to the reference values for DOE-20, indicating 'normal' affect processing: for REPCOG ($M = 2.26$, $SD = 1.11$) and COMEMO ($M = 1.89$, $SD = 1.10$) versus the reference values [56, 340] of REPCOG ($M = 2.24$, $SD = 0.77$) and COMEMO ($M = 2.01$, $SD = 0.82$). Participants' combined REPCOG-COMEMO score was used to form two groups, one with an elevated (upper half) and one with a lower (lower half) capability of representing and expressing emotions, which we will refer to as HI-EMO and LO-EMO. The two groups will be used to investigate how emotional openness affects participants' affect, self-efficacy, and 'levels' of reflection.

6.5.2 Study Set-up

The VR tool participants used for the study was developed in Unity. In the study, it was running in Unity on a computer which was connected to an *Oculus Quest 2* using *AirLink*. The visual representation of participants' chosen challenge was created by themselves, with the same toolkit previously used in Mood Worlds [429] (see Section 6.3.1 for a description). An additional component was added for the voice-based NLUI, delivering the guiding prompts when *requested* by the user (through the VR controller), or *proactively* based on the user's behaviours (i.e. 30-second inactivity thresholds) as described in Section 6.4.

6.5.3 Procedure

Similar to Prpa et al. [330], we chose an exploratory study design. After giving consent and sharing demographic data, participants completed the first set of questionnaires. They then started a tutorial phase, in which they familiarised themselves with SelVReflect's functionality. Afterwards, they were asked to choose their preferred voice (human/synthetic and female/male) for the NLUI. While the experimenter set up SelVReflect with the chosen voice, participants were instructed to recall an emotionally loaded challenge they had successfully overcome (i.e. a 'mastery experience' as described in Section 6.2) for which they did not (or no longer) experience any discomfort when thinking about it. Participants were then asked to write a paragraph describing the challenging experience – which was based on the Autobiographical Emotional Memory Task (AEMT) [199]. Then, they created a visual representation of the challenge, its stages and the emotions attached to each stage using SelVReflect. During that, they either requested guiding prompts or they would be proactively provided in case of inactivity. Towards the end, the NLUI would ask reflection prompts, inviting participants to 'walk through' their creation again and approach it from different perspectives. They could choose to think out loud at that moment and say what they are reflecting on, but they did not have to. When finished, participants filled out the second set of questionnaires. Additionally, they answered a set of questions specifically designed to assess their experience with SelVReflect as well as the guidance they received through the NLUI while using it. The study ended with a semi-structured interview. Each session lasted in total about 1h 26min on average (*range:* 55 min. – 1h 50 min).

6.5.4 Data Collection

For each participant, we measured the time spent on the drawing phase and the reflection phase separately. Quantitative data was collected from three validated questionnaires. Further, we collected qualitative data through interviews.

Measures

We used the PANAS questionnaire [435] to measure participants' affective states before and after using SelVReflect. Participants indicated on a 5-point Likert scale to what extent they felt a specific emotion (ten positive items, ten negative items) at that moment. Scores can range from 10 to 50. By using this measure, we can assess if SelVReflect creates positive affect, as can be assumed based on prior research [267, 429]. Increased positive emotions are further relevant as they can co-occur with a sense of achievement and mastery [30]. Moreover, with the PANAS we can measure if SelVReflect increases negative affect. Experiencing negative emotions could be a sign of rumination, one of the potential risks identified for self-care tools [123], which we aimed to avoid in *SelVReflect*. The GSE by Schwarzer and Jerusalem [375] was used to capture the perceived self-efficacy before and after the NLUI-based VR experience (one general factor, good psychometric properties). Participants indicated on a 4-point Likert scale to what extent they agree with ten items. Scores can range from 10 to 40. Although GSE is designed to capture traits rather than states, it has been successfully used for pre-post evaluation of short-term interventions (e.g. [35, 436]). Reflecting on mastery experiences, as is the case in SelVReflect, can increase self-efficacy.

After using SelVReflect, we used the Technology-Supported Reflection Inventory (TSRI) [41]. This scale specifically addresses how well a system supports reflection. Items 1-3 were reused with the same wording, 4-5 were slightly adjusted to fit the present tool and reflection task (see Figure 6.4 for more information), and items 6-9 were excluded, since they were not applicable (as they were related to long-term usage of a system and exchange with other people, which was not part of the present reflective activity). See Table 6.3 below for when each of the questionnaires was filled out by participants (i.e. before or after using SelVReflect).

Table 6.3: Showing the four questionnaires and when they were filled out.

	Before	After
DOE-20	✓	
PANAS	✓	✓
GSE	✓	✓
TSRI		✓

Interview Protocol

We conducted semi-structured interviews that lasted on average 15 min (*min:* 08:04 min, *max:* 24:40 min). In the interview, we asked participants to elaborate on the differences between creating in VR compared to in 2D (e.g. sketching or drawing on paper), how the NLUI's guiding prompts made them feel, if and how the prompts changed how they visualised and thought about the challenge, and what they took away from using SelVReflect. The full interview protocol can be found in the supplementary material of the underlying publication [FP4].

Data Analysis

A two-way repeated measures ANOVA was performed to examine the effect of time (i.e. *pre* and *post*) and DOE group on emotions (PANAS) and on self-efficacy (GSE). We further checked for interaction effects of Time and DOE Group.

All audio recordings were transcribed verbatim and imported into *MAXQDA* software. Nadine Wagener and I coded four interviews using open coding. Next, a coding tree was established through iterative discussion with all researchers involved in the project. The remaining transcripts were coded by several researchers using the coding tree. A discussion session between Nadine and I was conducted to identify themes using thematic analysis [55]. Those were discussed and agreed upon in a final discussion round between Nadine and I as well as two additional researchers who were experienced in psychology and reflection research.

6.6 Findings

Based on the evaluation, we gathered quantitative results from the questionnaires, as well as qualitative insights from the interviews. We first report on the quantitative findings.

6.6.1 Quantitative Findings

The data of the *pre* and *post* questionnaires were analysed using a two-way repeated measures ANOVA to investigate how the experience affected participants' self-efficacy and their affective state before and after using SelVReflect. Due to a misunderstanding in the task instructions, one participant was excluded in the analysis. Based on visual inspection of our data and the *Shapiro–Wilk* statistic we could not assume normally distributed data. Therefore, we applied the Aligned Rank Transformation (ART) [450].

When examining participants' choices of the available voices using descriptive statistics, we found that the *human female* voice was chosen 8 times, while the other voices – *human male, synthetic female, synthetic male* – were all chosen 4 times. Participants received a guiding prompt approximately every 92s (*SD* = 33s).

Emotions (PANAS)

A two-way repeated measures ANOVA showed a significant effect of TIME on POSITIVE EMOTIONS $F(1, 17) = 10.720$, $p = .004$ (see Table 6.4 and Figure 6.3). We found no significant interaction effects of TIME and DOE GROUP $F(1, 17) = 0.335$, $p = .570$. We did the same analysis for NEGATIVE EMOTIONS (PANAS). The test showed neither a significant effect of TIME on the NEGATIVE EMOTIONS $F(1, 17) = 0.446$, $p = .513$ nor an interaction effect of TIME and DOE Group $F(1, 17) = 1.998$, $p = .176$.

This shows that there was a significant increase in participants' positive affect from before to after using SelVReflect. Furthermore, it shows that this increase does not seem to be affected by how emotionally open participants are.

Table 6.4: ANOVA statistics for PANAS and GSE scores for factors TIME, DOE_GROUP, TIME DOE_GOUP. Significant results are marked with asterisks (* *p* < .05, ** *p* < .001).

Factor	PANAS Pos.	PANAS Neg.	GSE
TIME	$F = 10.720$	$F = 0.446$	$F = 6.189$
	$p = 0.004$**	$p = 0.513$	$p = 0.024$*
	$\eta^2 = 0.387$	$\eta^2 = 0.026$	$\eta^2 = 0.267$
DOE_GROUP	$F = 0.998$	$F = 4.781$	$F = 2.041$
	$p = 0.332$	$p = 0.043$*	$p = 0.171$
	$\eta^2 = 0.055$	$\eta^2 = 0.220$	$\eta^2 = 0.107$
TIME:DOE_GROUP	$F = 0.335$	$F = 1.998$	$F = 1.224$
	$p = 0.570$	$p = 0.176$	$p = 0.284$
	$\eta^2 = 0.019$	$\eta^2 = 0.105$	$\eta^2 = 0.067$

Self-Efficacy (GSE)

Another two-way repeated measures ANOVA was conducted with TIME on *GSE* $F(1, 17) = 6.189$, $p = .024$, showing a significant effect (see Table 6.4 and Figure 6.3). Again, we found no significant interaction effects of TIME and DOE GROUP $F(1, 17) = 1.224$, $p = .284$.

This shows that there was a significant increase in participants' self-efficacy from before to after using SelVReflect. Furthermore, it shows that this increase does not seem to be affected by how emotionally open participants are.

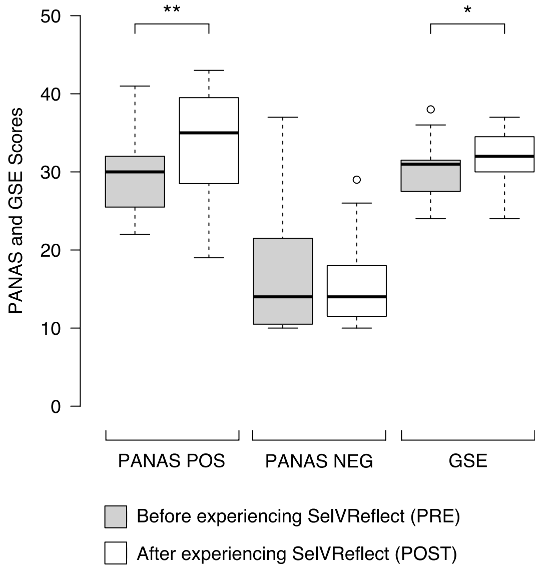
Reflection (TSRI)

When considering how participants rated the reflection they engaged in while using SelVReflect, neutral ratings were given to the first two items related to (1) making changes in one's life (*Md* = 3, *SD* = 1.305) or to (2) the ways in which one approaches things (*Md* = 2, *SD* = 1.170), as can also be seen in Figure 6.4. High ratings were given for item (3) the extent to which the system gives ideas to overcome challenges (*Md* = 4, *SD* = 0.994), (4) the enjoyment of exploring the challenge (*Md* = 4, *SD* = 0.911), and (5) the ease of getting an overview of the challenge (*Md* = 4, *SD* = 0.946). This shows that overall SelVReflect enabled participants to get new ideas for how they can overcome challenges; it gave them an 'overview' of their challenging experience, which they enjoyed exploring using the tool. However, SelVReflect received neutral ratings for how it might lead to changes in participants' lives and how they might approach things differently through using the system. The ratings can be seen in Figure 6.4, which also divides the results further into participants with higher (HI-EMO) and lower capability (LO-EMO) of representing and expressing emotions. As can be seen in Figure 6.4, the ratings are very similar for both groups, suggesting the system worked similarly 'well' for both groups.

Experience Ratings

When asked about the experience with SelVReflect, participants gave it positive ratings along various dimensions (on a Likert scale from 1 to 5), including how engaging (*Md* = 4.5), creative (*Md* = 5), and insightful (*Md* = 4) they found the experience. When considering these ratings from the HI-EMO and LO-EMO groups separately, they are generally very similar, as can be seen in Figure 6.5. However, LO-EMO ratings for difficulty were lower (*Md* = 3) than for HI-EMO (*Md* = 1).

When participants rated the received guidance, it was also highly rated for how engaging (*Md* = 4) and useful (*Md* = 4.5) it was. The guidance from the NLUI was given a rating of *Md* = 3 for how insightful it was. Again, the ratings from the HI-EMO and LO-EMO are very similar (see Figure 6.6). However, there was a more noticeable difference in the ratings for how challenged they felt by the NLUI in their process of expression and reflection, with LO-EMO participants *Md* = 2.5 versus HI-EMO *Md* = 1.



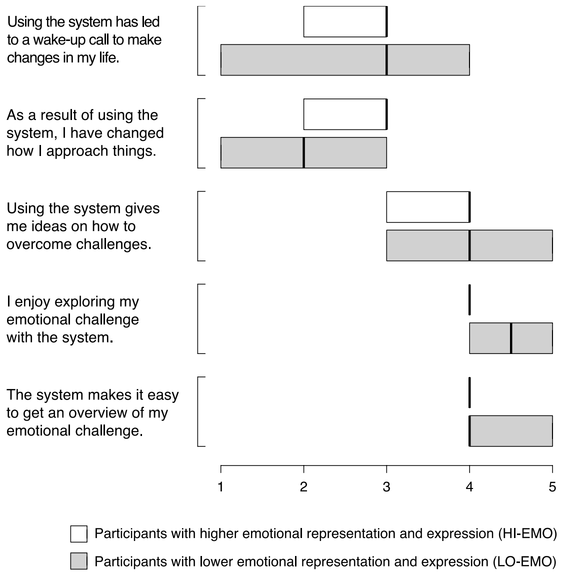
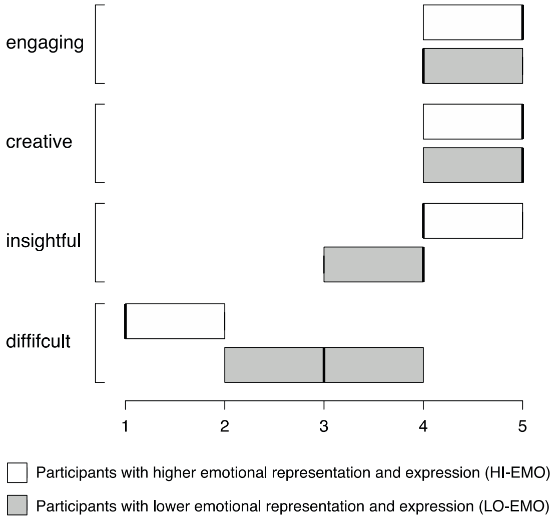


Figure 6.3: Mean scores for pre and post measurements of PANAS (positive and negative) and GSE scales. Significant results are indicated with * p < .05 and ** p < .01.

Figure 6.4: Median ratings and interquartile range for the first five items of the TSRI scale. They are split up for HI-EMO and LO-EMO groups.



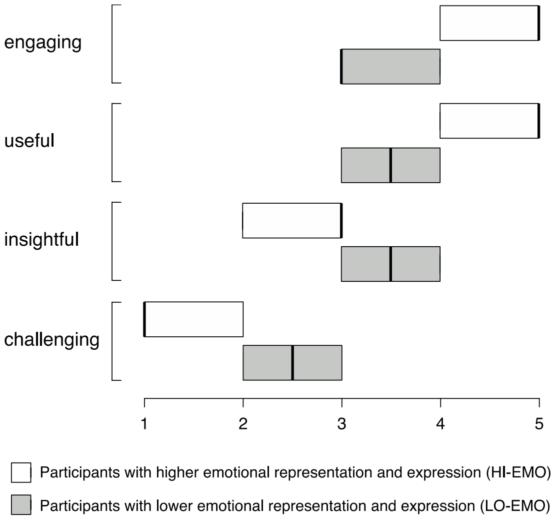


Figure 6.5: Median ratings and interquartile range for the ratings for the experience of using SelVReflect, split up for HI-EMO and LO-EMO groups.

Figure 6.6: Median ratings and interquartile range for the ratings for the guidance received, split up for HI-EMO and LO-EMO groups.

Summary

Overall, the quantitative results suggest that SelVReflect was not only successful in supporting people's self-reflection and exploration of their past challenges but that *the overall experience* was also perceived as engaging, creative, and insightful. Participants also found *the NLUI* engaging and useful. This might have enabled the significant increase in participants' self-efficacy and positive affect that was observed from before to after using SelVReflect.

There generally did not seem to be noticeable differences for all the questionnaires and ratings depending on participants' capability of representing and expressing emotions. The only rating that showed a slightly larger difference between the groups was the difficulty rating (see Figure 6.5). Although the experience did not seem to be difficult overall, participants with lower capability (LO-EMO) of representing and expressing emotions seemed to find it somewhat more difficult than those with higher capability HI-EMO).

6.6.2 Qualitative Findings

To complement the quantitative analysis, a qualitative analysis of the interview data was conducted through a thematic analysis. Four core themes were derived from this data: *VR Providing the Space*, *Palette Providing the Creative Tools*, *Guidance Providing the Scaffolds*, and *Experience of Transformative Reflection* (see Figure 6.7). The findings are described below and illustrated with excerpts from the interviews.

Most of the chosen challenges were related to participants' studies in university (9) or work settings (6), while the remaining (5) were related to friends and family or romantic partners (see also Table 6.2). Five out of the latter were about deciding on how to allocate time between different friend groups, family, or work. Six challenges were related to interpersonal difficulties as part of studies (e.g. group projects, theses) or professional settings. Four challenges dealt with approaching a major decision (e.g. switching jobs or moving homes), and three were related to adjusting to a new situation or setting (e.g. a new home or job).

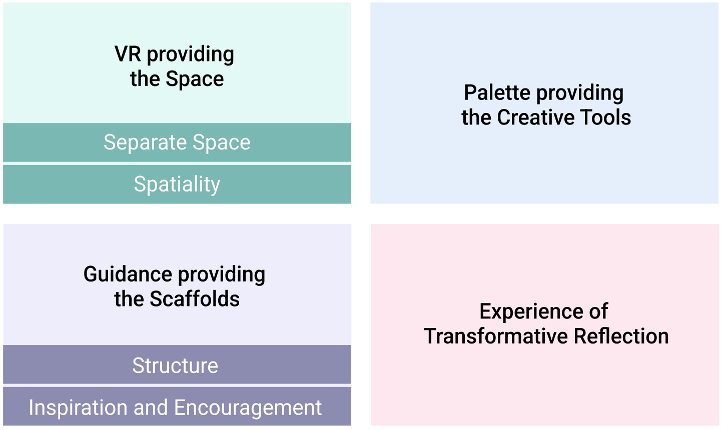

VR providing the Space
Separate Space
Spatiality
Palette providing the Creative Tools
Guidance providing the Scaffolds
Structure
Inspiration and Encouragement
Experience of Transformative Reflection

Figure 6.7: The four themes and codes identified in the semi-structured interviews.

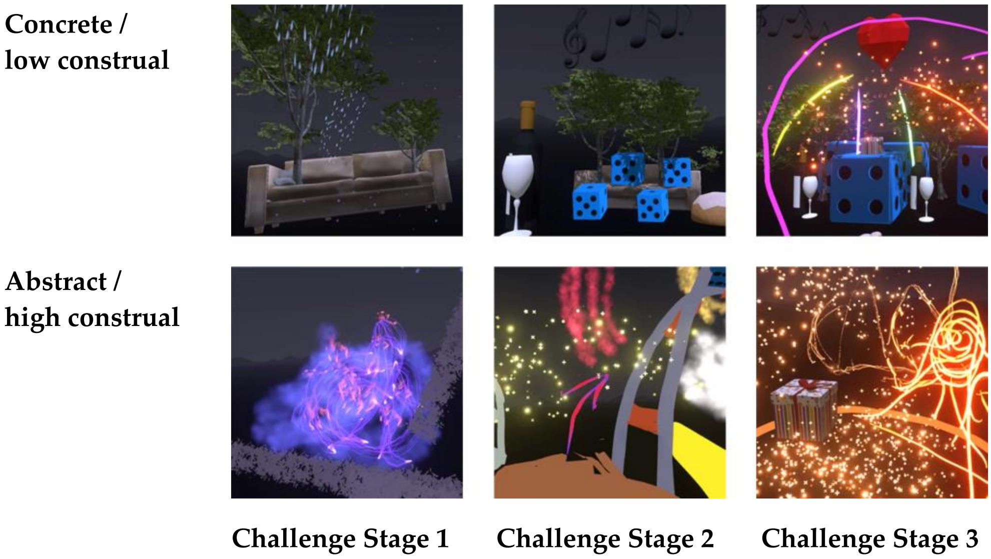


Figure 6.8: Two examples of concrete and abstract SelVReflect creations made by two participants.

Figure 6.8 shows two examples of the 3D representations created using the VR tool. The first row depicts a concrete representation of a challenge (P4). Trees and dice stand for different friend groups (stage 1), sharing an evening together (stage 2), and becoming friends (stage 3). The second row depicts an abstract representation (P11). Uncertainty and anger are represented with blue smoke, plasma, and fire (stage 1), transitioning from mud through smoke by going up a ladder (stage 2), and joy of overcoming the challenge in bright colours (stage 3). As can be seen, they are imaginative, using a variety of representations. Regardless of the levels of abstraction, the figures show the high degree of creativity and expressivity in participants' representations of their challenging experiences. The next section will describe the findings of the thematic analysis.

Theme 1: VR Providing the Space

The first theme focuses on the specific benefits of VR for facilitating creativity and reflection. It encompasses the codes *Separate Space* and *Spatiality*.

Participants reported that being immersed in VR enabled them to enter a different 'mindset'. A key reason was that VR offers a *separate space* without external disturbing factors, which allowed them to dive deeper into their thoughts. One participant elaborated on the benefits of VR:

> *"You're like cut off from everything. You're like in this empty void. It helps a lot of people to be with their thoughts and explore them more because they're cut off and for themselves."* (P9)

The VR canvas also allowed participants to utilise the virtual 3D space to express components of their challenge beyond what would be possible in physical reality. For example, they used the third dimension as a representation of time or to link the relationships between components of their challenge, as visually representing their thoughts in 3D "*makes it easier to show correlations between several things"* (P7).

Further, they utilised the *spatiality* of VR to immerse themselves in, as well as physically, mentally, and emotionally distance themselves from different elements (of their challenge). On the one hand, they enjoyed being surrounded by their creation, exploring their 'challenge environment' through a first-person perspective. This allowed participants to enclose themselves within their creations, 'break through' them, and become more physically involved in the depiction of their challenge.

> *"You can actually 'paint yourself in', completely all around you, if you like, and take up different positions."* (P13)

Although 2D figures cannot convey the feeling of being 'cornered' by one's own creation, or physically moving through a wall when overcoming a challenge, excerpts are provided in Figure 6.8. On the other hand, participants also enjoyed being able to take a literal 'step back' from their visual representations (i.e. to perform 'self-distancing' [235]). This change of perspective led to another experience and enabled them to see the whole picture, which sometimes made the problem seem smaller than before.

> *"I just recognise kind of the third person perspective on your decisions, so putting yourself not in your shoes, but just having a bird's eye view. There might be something that is interesting and is now more tangible with having done it yourself [in 3D]."* (P16)

However, for some, the blank space surrounding them created a feeling of being lost, and not knowing where to start. On a similar note, some participants found it challenging to "think in 3D" when drawing and to utilise the complete space available for their creation.

Theme 2: Palette Providing the Creative Tools

As a second theme, it was found that the variety of tools provided the means for creative self-expression and reflection. Most participants emphasised that using the palette increased their motivation, and made them think in a more abstract way, so that *"visual elements [are used] as a sort of analogy or metaphor"* (P19). Objects were mainly used as placeholders for people (see Figure 6.8) or abstract constructs, such as loss of agency (e.g. dice), personal growth (e.g. tree) or emotions (e.g. fire for anger). Animated brushes were often used to represent emotions and relationships between components or different people involved in the personal experience. For an example of abstract representation, see Figure 6.8. One participant stated:

> *"I feel like the choice of tools was surprisingly wide enough to try out different things and also to, yeah, express more complex emotions."* (P5)

However, this availability of choice was also a limitation for some participants. Roughly half of the participants reported difficulties in choosing a tool, especially at the beginning. As well as some (technical) problems with actually drawing or resizing objects, four participants reported difficulties in identifying emotions, and eight participants were unsure how to visualise them with the provided tools.

Theme 3: NLUI's Guiding Questions Providing the Scaffolds

The third theme encompasses the reactions to and effects of the NLUI's guidance throughout the study. It included the codes *Structure* and *Inspiration and Encouragement*. Overall, participants readily understood that the prompts which they received every few minutes were there to support their expression and reflection and that they did not have to verbally respond to them (although some were thinking aloud – in particular in the final *Re-walk* phase).

Encouraging participants to decompose their challenge into smaller stages structured their thinking and representation of the challenge. Overall, participants reported feeling reassured by this guidance. Participants emphasised the importance of feeling inspired and encouraged by the spoken prompts. This was due to both their content, which led them to approach components differently, and their tone: *"it's not just about what they say, but how they say it. It helps you relax and ease into it"* (P9). This led them to think in greater detail about the challenge they were depicting. Furthermore, their self-confidence was strengthened by both the tone of the voice and the affirmations it offered. P8 further elaborated:

> *"I really loved the guidance. [...] In taking the time to dissect the situation where I was in, I think that really helps also because it made me feel more confident about what I did. It [the guidance] supported me in looking back on it [the situation] and seeing it from multiple perspectives, probably more than what I thought about so far."* (P8)

Another aspect was that participants generally thought that when they got lost in some thought or how to express something that the guiding questions were *"helpful to stay on track or get back on track"* (P15). Similarly, participants sometimes felt like they got stuck in a certain way of looking at something and that the questions would help them change their perspective, as expressed by P14:

> *"Yes, they totally got me to think more, because sometimes you're stuck within the viewpoint you have and [...] you're just caught within one phase [of the challenging experience], and then if the suggestion is how does this interconnect with others, you suddenly realise that it interconnects [with another aspect of the experience]."* (P14)

Furthermore, participants also acknowledged that the structuring from the NLUI was helpful for their reflection – even if it might occasionally have felt as if it 'imposes' something they might not have done otherwise, as pointed out by P19:

> *"I would just have looked at it, sort of, "yeah, this was the challenge, and then the end" but now you* ***have to*** *break it down […] but for reflecting on it, I think it* ***helped me*** *[…] think about phases and actions, maybe decisions, options, alternatives." (P19)*

Related to this feeling that they occasionally "have to" do something that might not have been in line with their ideas, more than half of the participants also expressed a desire for a form of verbal guidance that is more context-aware in the sense that the guidance would be more aligned with what they are doing and the approach that they take. Examples of what they suggested were having more individual timing and content-specific questioning towards aspects of their creation, such as intervening at a specific moment when an important choice was made. They also desired more specific help and inspiration regarding how they could visually represent certain elements of their emotional experience (e.g. in the form of symbols, visual metaphors, etc.). This suggests that a customised kind of verbal guidance throughout the VR experience might help individuals progress through and express their personal challenges even more. Furthermore, some participants commented on prompts occasionally being triggered prematurely when they took more time in choosing and adjusting the settings for a tool (e.g. selecting colour hues). However, they usually did not find it disruptive to the general flow, even if it may not have been helpful for reflection in those cases.

Theme 4: Experience of Transformative Reflection

Participants reported that they reflected *"along the way" (P15)* by decomposing the challenge, abstractly visualising their emotions, and adding details. This was especially prominent during the *Free-flow* phase. In the *Re-walk* phase, while (re-)experiencing their creation, some participants reported that prompts enabled them to engage in deeper reflection, while others felt less affected by them. However, all the participants agreed that they were successfully reflecting at some point during the SelVReflect experience.

By making use of the opportunities for reflection, the participants approached their challenge and its (emotional) components from different perspectives. This led to a change in their conceptualisations of the following:

1. *The challenge itself.* They gained new (or deeper) knowledge about the reasons behind the challenge, the emotions involved, and their role in the process. For example:

    - *"It's an even deeper engagement with the situation [in contrast to writing about it]."* (P19)
    - *"I thought more about what kicked it off."* (P9)
    - *"It's a sense of accomplishment, a sense of resolution, also a sense of closure."* (P5)

2. *Themselves as a person.* This encompasses self-awareness about one's character, one's beliefs, and one's role within the social ecosystem. They also felt proud of themselves. For example:

    > *"That was actually a realisation that I never had before: That talking to people and the opening up and not always trying to solve things by myself, which is what I do now, [is healthy]."* (P8)

3. *The bigger picture.* Most of the participants appreciated their relationships with their friends more than before, discovering alternative approaches or solutions to their challenges, such as thinking in stages (which the guiding questions encouraged them to do). This made them feel better equipped to act more effectively in similar situations in the future. For example:

    > *"It [SelVReflect] is a possibility to reflect on certain problems, especially also from an emotional point of view, and to find other approaches and therefore to be able to adapt one's behaviour better."* (P13)

6.7 Discussion

Both the quantitative and qualitative findings showed how SelVReflect provided a 'canvas' for the participants to visually represent their challenging experiences. The process of expressing and externalising their thoughts and ideas was structured and guided by the questions. Participants could 'respond' to the NLUI questions by adding to and modifying the visual representations of their experience. They readily understood that they did not have to verbally respond to the prompt (although some participants chose to think aloud). The participants enjoyed creating their own visual representation and appreciated being assisted by the voice-based NLUI. The findings suggest that they encouraged and supported their expressivity and externalisation of their ideas and motivated them to reflect. Overall, SelVReflect thus gave participants a feeling of accomplishment – regarding their view of the challenge itself, how they managed to visually represent it, and what they could learn from this process of creation, exploration, and reflection.

This corroborates the quantitative findings, which showed that SelVReflect had a significant effect on positive affect and self-efficacy of the participants. However, the findings also indicate that the experience was somewhat more difficult for participants with lower affect processing scores compared to those with higher scores – yet, despite differences in difficulty, both reached similar outcomes and seemed to have a similar experience.

All participants mentioned aspects that refer to existing conceptualisations of reflection in literature. Some reported discovering new constructive approaches for challenges [143], gaining (self-)awareness [270] (such as general self-knowledge of how to deal with problems such as their relationship with friends), developing new understandings and appreciation (such as about reasons for the challenge and appreciation of relationships to friends), and feeling empowered or better equipped for the future [291].

The 'levels' of reflection [143] that participants reached through SelVReflect differed – some reached level 1 (Reflective Description), while others progressed to level 2 (Dialogic Reflection). While some felt confirmed in their previous perspective on the challenge, others discovered new ways of how they could approach them more effectively in the future. This suggests that even some ‘transformative reflection’ took place. Discovering new ways of dealing more effectively with challenging situations is closely related to self-efficacy.

As established in previous research, effective self-reflection benefits from guidance and encouragement [143, 390]. The findings suggest that the NLUI achieved this and that it enabled participants to express and reflect on the challenge in different ways, helping them discover new aspects about it and themselves. However, as Agapie et al. [3] emphasise, deep reflection also requires effort, which can decrease the enjoyment of the task itself and lead to a loss of motivation to persevere. It seems the NLUI could successfully mitigate this, as participants pointed out the voice guidance not only made them feel comfortable and confident to be exploratory and creative in both their expression and reflection, but it also brought structure, helped them break down the process of creation into smaller parts, and supported them in ‘dissecting’ the experience and its different components. Furthermore, the guidance enabled them to take on new perspectives and reflect on how different components might be interconnected.

The approach of having different forms of interaction with the NLUI – request-based in the beginning, followed by mixed-initiative (both request-based and proactive), seemed to work for the present study. This way, participants could start off and do the first parts of their creation at their own pace. Then, once a basic structure was in place and participants had ‘warmed up’, the NLUI would occasionally also intervene proactively to give them some inspiration for what to express and/or reflect on.

Furthermore, the approach of having a certain structure or ‘evolution’ of the guiding prompts over time also seemed to be effective with participants finding them to help them “stay on track”: In the beginning, the prompts were mainly designed to facilitate expression and externalisation and, in the end, to encourage participants to explore the representation and reflect on it more deeply.

Similar to VoiceViz and ProberBot the present prompts were not designed to provide participants with any instructions but rather to inspire them to consider and explore different aspects while performing their task. However, what was given particular attention in SelVReflect was to design the prompts in a way that they encourage the user. The reason for doing so was that many people find creative/expressive tasks challenging, in particular if they involve personal topics/issues as was the case with SelVReflect. The findings suggest that this was generally achieved (e.g. *"It's not just about what they [the prompts] say, but how they say it. It helps you relax and ease into it"* – P9). It seems that the approach of having prompts consisting of an *inspiration* and an *encouragement* part is effective for expressive activities where it is particularly important that people feel comfortable and confident to be able to achieve a positive experience and outcome (e.g. activities that involve more personal topics, experiences, feelings). Furthermore, it helped participants to get used to and comfortable with the opportunities of expression that the VR canvas provides, which go beyond other means and modalities they may be familiar with (e.g. drawing and writing).

Participants' descriptions of how they understood the visualised relationships between parts/aspects of the challenge in new ways through representing and exploring them (i.e. by literally walking through their creation) further suggest that the experience enabled effective forms of external cognition. The findings also underline the crucial role that the prompts from the NLUI guidance played specifically with regard to external cognition, enabling them (a) to think the representation through (e.g. its stages and components) and visualise its structure in their head, to then (b) *externalise it to* and *represent it in* the VR canvas by asking them to focus on specific aspects, to finally (c) explore the representation they created and encouraging them to reflect on what they discover in the process and what new connections and dependencies they might see, with the aim to help them build a new understanding. This further suggests that apart from the *reflection-on-action* (i.e. the reflection on the past challenging experience), which SelVReflect enabled, an important role of the prompts was also to facilitate forms of *reflection-in-action* (where 'in-action' represents the process of creation and going through the past challenging experience again while using SelVReflect) which enabled participants to explore different ways in which they could best express and represent certain aspects of their experience.

6.8 Conclusion

SelVReflect, with its three-dimensional 'canvas' for expression and its 'embedded' NLUI guiding the process of expression and reflection, was able to help participants make sense of the challenging experience, discover new aspects in it, and get new understandings of how they could approach comparable situations in the future. The findings suggest that reflecting on ourselves – and how we act in our respective environments – through scaffolded self-expression has the potential to not only foster *situational* awareness but also *self*-awareness, emotional intelligence, and even the formation of new knowledge about ourselves. However, as in the previous studies, the findings also revealed that the proactive reflection prompts also involve certain challenges – they were sometimes not relevant for what the participants were doing at a given moment or pointed out things they had already reflected on.

Similar to the previous two studies, SelVReflect was designed to support people in reflective thinking – in particular, *reflection-on-action* – to make sense of a topic or of oneself. While in the case of VoiceViz, most of the reflective thinking (in response to the NLUI's questions) was expressed and externalised in a conversation between pairs of people, in ProberBot, this was done by the user through written answers and explanations in response to its questions. Here, another *modality* or *way* to express one's thoughts, namely *visual expression*, was explored. As the study showed, this enabled effective forms of external cognition which supported reflective thinking: Reflecting on different aspects of the past experience helped participants figure out ways to visually express/externalise it – at the same time, exploring what they expressed enabled them to reflect on different ways to look at the challenge and what helped them overcome it (e.g. reflecting on the connections between stages and components). This further suggests that in the case of SelVReflect, there seemed to be a particularly strong interconnection between both 'cognitive externalisation' and reflection. This not only gave them insights into their own behaviour and how they might be able to improve it in the future in similar situations (which is similar to what ProberBot intended to achieve with respect to investors' decision-making – see Table 6.5 below) but also it enabled a number of participants to give the challenging experience a new meaning in the sense that they might look at it in a new way. Thus, a particular strength of SelVReflect seems to be that the rich representations it allows people to create provide a 'fertile ground' for reflection.

Table 6.5: Outcomes of reflection targeted by the NLUIs in each study.

	Outcomes of Reflection for User
VoiceViz:	• Understanding the patterns in the data and their potential causes
ProberBot:	• Understanding the relevance of different data for a decision • Understanding one's own decision-making behaviours
SelVReflect:	• Understanding one's behaviour in a challenging situation • Understanding the meaning of this challenging situation (in new ways)

Taken together, the three studies show how proactive question-asking NLUIs that are 'embedded' into interfaces used to perform a range of open-ended tasks with the aim of triggering users' reflective thinking can lead to different forms of sensemaking and enable a range of different insights. The idea of NLUIs proactively asking users questions specific to their ongoing task to get them to reflect on different approaches, perspectives, and aspects to consider seems to augment their thinking in various ways. All studies showed that the proactive questions can provide opportunities for *reflection-in-action*, and in particular, SelVReflect showed how they can enable *reflection-on-action* [373].

Having examined the potential of these types of task-embedded NLUIs, a question was what other forms of proactive NLUIs could there be – in terms of the types of support that they provide and how they intervene – and how they could be embedded into everyday situations. This is explored in the next chapter. Instead of focusing on specific tasks that are performed using certain interfaces and how the NLUI can be embedded into them, as was the case for the three studies reported so far, the next chapter examined which possibilities and opportunities there are for proactive NLUIs to become part of people's everyday lives. The goal is to provide a counter-perspective to the types of 'cognitive co-pilots' looked at so far and explore their possible 'design space' more broadly.

7. Cognitive Co-pilots in Our Everyday Lives

This chapter focuses on how NLUIs that proactively prompt people could be embedded into their everyday lives. The aim of the final set of studies reported in this chapter is to explore 'what might happen' if cognitive co-pilots could be designed to 'venture out' from task-specific interfaces and instead weave themselves into people's everyday activities.

The aim of the research was to consider how to deliver information that might be relevant for an ongoing everyday activity. This research was conducted in collaboration with colleagues from the University of Bremen as part of the Excellence Chair project, which was also the basis of the previous chapter (SelVReflect)[18].

The NLUIs here were designed as proactive voice assistants **(VAs)** that would appear in the form of a smart speaker. To investigate people's perspectives on proactive VAs, a set of storyboards were designed depicting a variety of proactive actions by the VA in everyday situations and social settings. The question that was addressed was how people might react to having a proactive VA which makes suggestions in different everyday settings. Would they find the advice helpful, or conversely, too intrusive for the ongoing activity? How would they feel about a VA observing different (social) activities? And how and when would they want – and not want – a VA to intervene?

[18] While the first publication had shared first authorship, I was the second author on the second publication underlying this chapter (with Nima Zargham being the first author). However, I was actively involved in the research throughout the process, from initial study ideas to the data analysis and the final write-up of publication arising from this work.

7.1 Introduction

Voice assistants (VAs) are accessible across a spectrum of devices, including smartphones, tablets, PCs, vehicles, smart home gadgets, and smart speakers. As more and more VAs are finding their way into our homes, in particular in the form of smart speakers, they play a greater role as digital everyday helpers. They are being widely used for a range of activities such as smart home control, information retrieval, entertainment, online shopping, and managing schedules [332] and are further evolving to handle more complex tasks and dialogues. Advances in AI, natural language processing, and sensing techniques are expected to give these systems a better sense of people's behaviours, preferences, intentions, and surroundings, enabling them to become increasingly more proactive (e.g. [105, 118, 286, 367]) – so that they can suggest a person things or ask them questions that might be relevant for their ongoing activity. In the near future, the increase in VA's (proactive) capabilities is likely to increase even further with the proliferation of LLMs, which are also used to extend the abilities of these VAs.

Section 2.1.5 reviewed the literature on proactive voice assistants, suggesting that despite some proactive behaviours being perceived as uncomfortable, disruptive, and invasive, people also recognise the numerous benefits of such interactions. Other research has examined the appropriate timing and delivery of proactive interventions, for example to reduce interference with ongoing tasks (see Section 2.1.6). While proactive services can provide useful information for assisting, inspiring, and engaging users, the timing and relevance of interventions are critical to the user experience (e.g. [8]) but are also very challenging to get right (e.g. [286]). The importance of timing and appropriateness of proactive interventions is even more pronounced for voice user interfaces (VUIs). Attending to GUI-based notifications can more easily be delayed until the user is ready, which is not possible with VUIs as speech demands immediate attention and can thus interfere with ongoing user activities or social interactions (as it was also found in the VoiceViz study in Chapter 4). While there seems to be a demand for proactivity, there is limited evidence about what makes a proactive voice assistant desirable (e.g. [8]). However, proactive interactions in such devices could open up new opportunities and potentially empower a broad range of applications [441]. Yet, certain proactive behaviours can cause discomfort and be perceived as disruptive and invasive (e.g. [12]). For successful proactive interventions, not only users' current mood but also cultural and social context need to be considered. Proactive features also present challenges regarding privacy (see Section 2.1.5), as they constantly need to monitor and process their environment and the users' behaviour.

When compared with the NLUIs presented in previous chapters, these concerns are – unsurprisingly – significantly more pronounced for proactive NLUIs that are embedded in everyday (domestic) settings. Proactivity in the former 'only' requires monitoring a person's activity within the software tool used to perform the specific task – there is generally no need to observe the physical world – while in the latter, the smart speakers' physical environment might need to be continuously monitored in order to provide meaningful interventions based on what is currently happening, which might also include intimate activities and interactions between people. As highlighted by Tabassum et al. [407], privacy is one of the main concerns people have and most likely a decisive factor as to why they might not want to use a proactive VA.

Here, we begin to address the desirability and usefulness of proactivity, given these previous concerns, by exploring people's attitudes towards various everyday (social) scenarios in which a VA proactively addresses the user(s) in different ways based on their current activities and conversations. For this purpose, we designed a set of storyboards illustrating a range of possible proactive interventions in a home environment, which were used for three separate studies (two of which are reported in this chapter) to investigate people's perceptions of proactive VAs in everyday situations.

7.2 Study Design

To investigate circumstances for a desirable proactive VA in everyday situations, we used an approach inspired by scenario-based design methods [70] and vignette studies [4], which allows us to investigate (future) technologies despite current technological limitations – and has also been used by other researchers in similar studies [264]. This comprised questionnaires and online interviews in which participants were asked about their perceptions of a range of storyboards that were shown to them, representing various scenarios within home environments. The storyboards show different people, what they are doing, including their ongoing interactions and conversations with each other, as well as a smart speaker proactively saying something.

7.2.1 Storyboards

Two of the research team and I held multiple brainstorming sessions in which we came up with 30 scenarios. The creation of the scenarios was based on what was imagined could possibly be useful proactive interventions in everyday situations. The scenarios were all situated in a home environment, including a single person or multiple people – which also reflected one of the main ways the scenarios were classified (single versus multi-user). The scenarios were further classified according to the interruption of a conversation among people, whether the action was 'imposed' on the user or rather suggestive, and the potential to be perceived positively or negatively by the user(s). Several iterations of narrowing down the set of scenarios were performed, mainly focusing on how well the chosen scenarios covered the different classifications. This resulted in a final selection of eight scenarios for which graphical storyboards were created. A pilot study was conducted on the final set of storyboards with three participants, which showed that they were successful in getting participants to contemplate the VA's intervention and what they believe the outcomes of the intervention might be, including how they expect the people depicted in the scenario to react to the VA's intervention.

All storyboards were in a comic style with two or three separate panels. Several different styles were explored with the aim to convey the situation without any ethnic or cultural cues so that all participants should be able to put themselves in the shoes of the characters. The fictional agent was given the gender-ambiguous name 'Jay' to reduce gender bias. To avoid an influence from the reactions of the depicted characters on the participants' opinion, no facial expressions or responses to Jay's behaviour were included. The cylinder-shaped appearance of the voice assistant was similar to a conventional smart speaker (see Figure 7.1 for example, a full list of storyboards can be found in [FP3]). The set of storyboards was generally similar across the studies despite two replacements and small adjustments to phrasing (for example, based on further feedback received from other researchers or in pilot testing conducted before the studies).

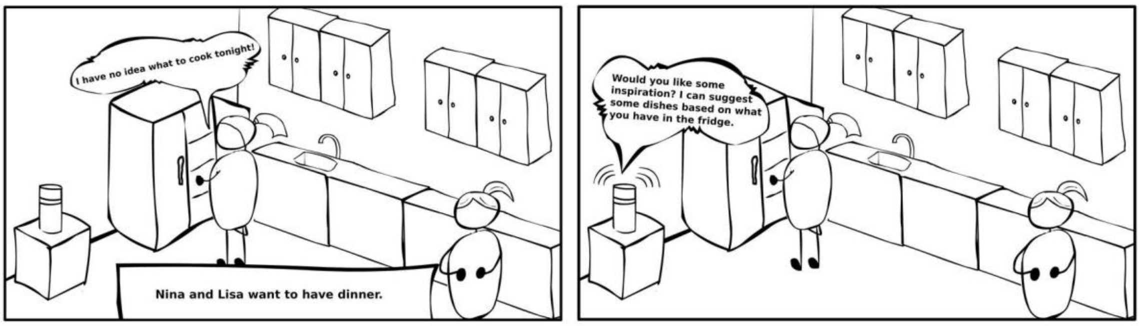

I have no idea what to cook tonight!
Nina and Lisa want to have dinner.
Would you like some inspiration? I can suggest some dishes based on what you have in the fridge.

Figure 7.1: Scenario 1 – Cooking Inspiration: Two friends are contemplating what to cook for dinner when Jay offers to suggest recipes based on what is in the fridge.

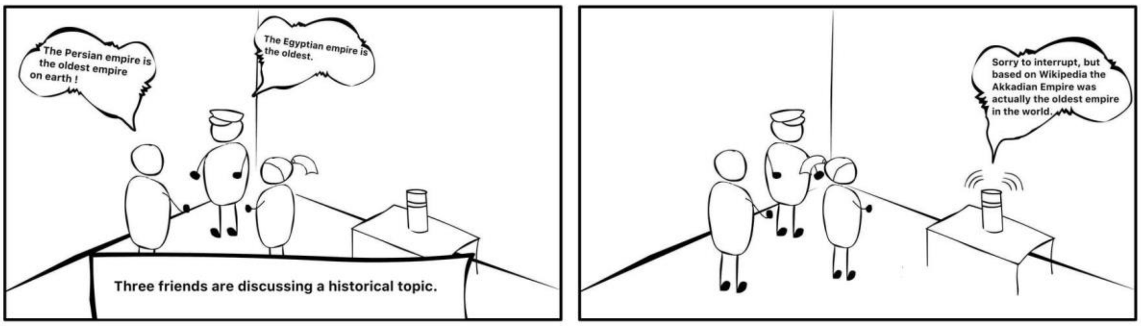

The Persian empire is the oldest empire on earth !
The Egyptian empire is the oldest.
Three friends are discussing a historical topic.
Sorry to interrupt, but based on Wikipedia the Akkadian Empire was actually the oldest empire in the world.

Figure 7.2: Scenario 2 – Fact Checking: Three friends discuss a historical topic when Jay interrupts them to get a fact right.

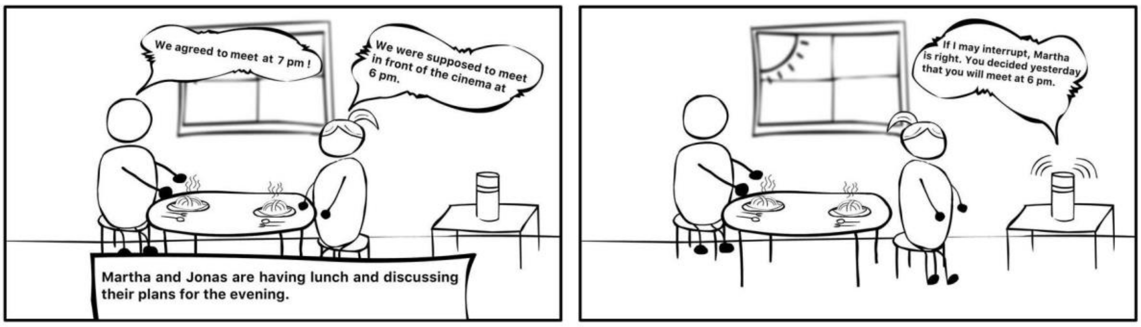

We agreed to meet at 7 pm !
We were supposed to meet in front of the cinema at 6 pm.
Martha and Jonas are having lunch and discussing their plans for the evening.
If I may interrupt, Martha is right. You decided yesterday that you will meet at 6 pm.

Figure 7.3: Scenario 3 – Disagreement Clarification: Two people remember differently what they agreed on when Jay settles the disagreement by quoting what they said.

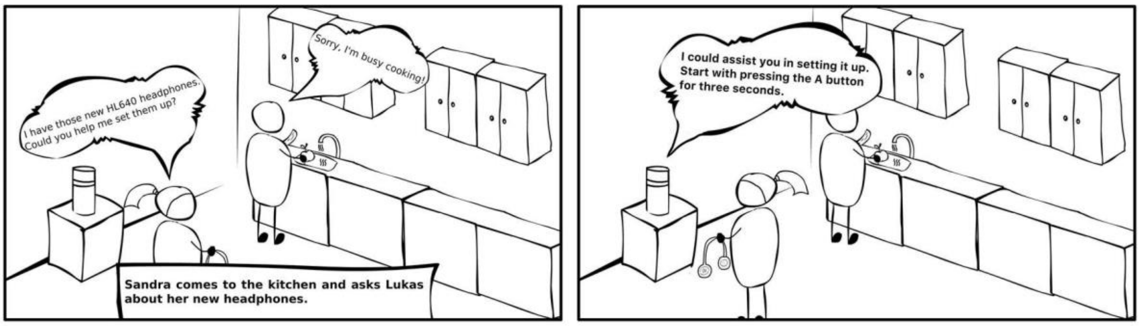

I have those new HL640 headphones. Could you help me set them up?
Sorry, I'm busy cooking!
Sandra comes to the kitchen and asks Lukas about her new headphones.
I could assist you in setting it up. Start with pressing the A button for three seconds.

Figure 7.4: Scenario 4 – Technical Support: A person asks their friend for help with setting up new headphones. As the friend is busy, Jay offers to assist.

7.3 Study 1: Survey

An online questionnaire was designed so that the scenarios could be evaluated. It was designed so that participants would rate Jay's proactive interactions in terms of *usefulness*, *appropriateness*, *pleasantness*, and how positive or negative their *overall impression* is, using a five-point Likert scale for each scenario individually. Ethics approval was obtained from UCL (UCLIC/1819/008/RogersProgrammeEthics) prior to the study. The survey concluded with a set of questions on demographics. Since current smart speakers are used by a wide range of users of different age groups, we did not have any inclusion criteria apart from being fluent in English.

7.3.1 Study Design

After a welcome text and a short introduction, participants gave informed consent. They were then introduced to the concept of a proactive VA and the fictional agent 'Jay'. They were asked about their typical usage of VAs and if they own a smart speaker. They were then presented with the eight scenarios, one by one, in randomised order. In the end, participants were asked to share what they liked or disliked regarding Jay's proactive behaviour.

Participants

A quota sampling approach was used to recruit participants. The acquisition was based on mailing lists, social networks, and word-of-mouth. Participation was voluntary and uncompensated. Of the $N = 47$ participants 25 self-identified as female, 18 as male, 1 as non-binary, and 3 preferred not to say. 34 participants were 18 to 34 years old, 7 were between 35 to 54, and 6 were older than that. 26 of the participants have previously used VAs (10 rarely, 16 often). 12 participants owned a smart speaker.

7.3.2 Findings

The following findings give an impression of the participants' diverse opinions on the proactive abilities of the VA in the scenarios. First, quantitative and then qualitative results are presented. Due to the highly exploratory nature of this research, we refrained from inference testing and only use mean *(M)* and standard deviation *(SD)* as descriptive statistics[19].

Usefulness of the interactions received high ratings with a mean of *M* = 3.73 out of 5 (*SD* = 1.33) across all scenarios compared to how *pleasant* (*M* = 2.95, *SD* = 1.32) and *appropriate* (*M* = 2.94, *SD* = 1.43) the participants found the scenarios. However, there was considerable variation in participants' ratings: All scenarios received the highest and the lowest possible ratings on all tested dimensions by at least one participant. Overall, this suggests that participants could generally see the use of many of the scenarios overall giving them rather elevated usefulness ratings; however, pleasantness and appropriateness are generally perceived to be less elevated and around a neutral rating of 3 (on the scale from 1 to 5).

The voice assistant was designed to intervene when there is either one or more than one person present. This was found to have an effect on the attitudes expressed – for example, concerning *appropriateness*, which can be seen in Figure 7.5. However, the pattern was similar for how *useful*, *pleasant*, or *positive* the interaction was perceived. For example, the scenarios in which Jay addressed the user in reaction to an ongoing conversation were rated worse than when the user was not engaged in a conversation. Similarly, the interactions in which the user was alone when being addressed by Jay received better ratings than when being with others.

Whether the action was classified as being imposed or was suggestive also had an influence on participants' ratings (see Figure 7.5). The scenarios in which Jay framed the assistance as a suggestion, instead of imposing the help on the user, were judged more positively by the participants.

[19] Given the custom questions, it would have been more appropriate to only use the *median* and *interquartile range*, as custom Likert-scale questions should generally be considered ordinal. However, this methodological issue was only identified after the publication of the paper which this section is based on (see the short paper in the list of publications at the beginning of this thesis) and it was decided to keep the descriptive statistics presented here consistent with the publication.

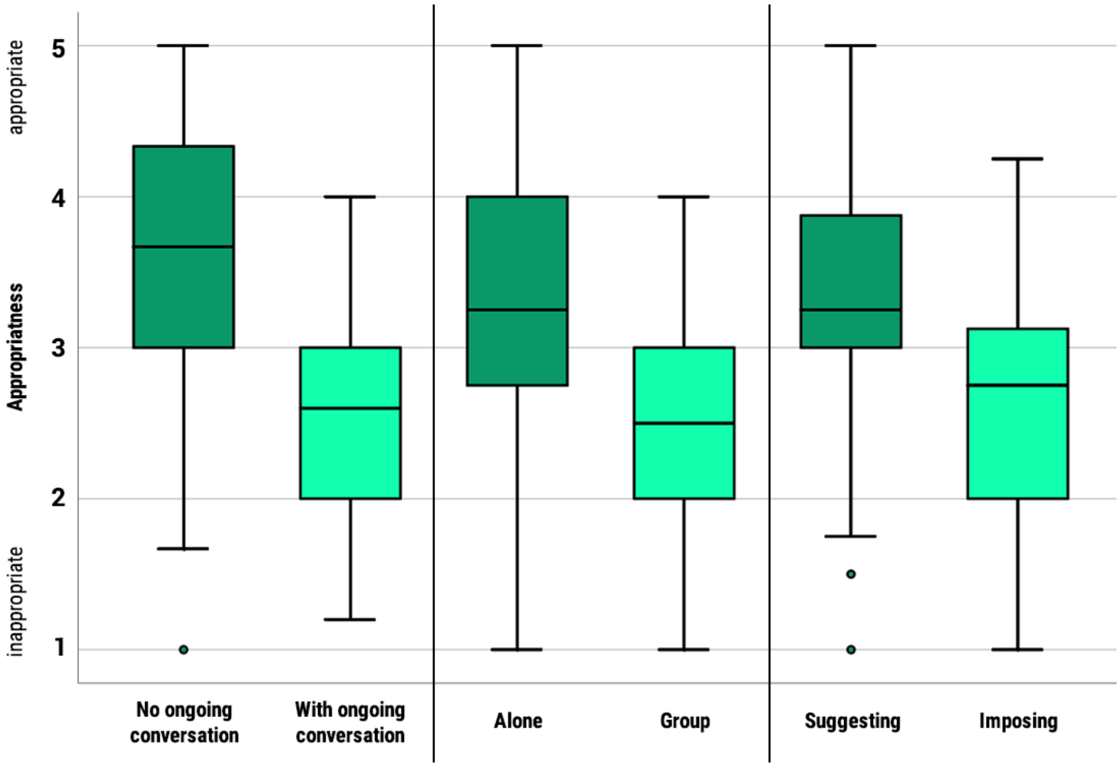


Figure 7.5: Box plot of *appropriateness* ratings comparing the three scenario classifications.

Participants' Reflections on Proactivity

To evaluate the answers to the open-ended questions, three researchers agreed on a codebook that was generated from a random selection of ten participants' responses. Subsequently, all responses were coded along this categorisation and summarised.

Overall, participants found the proactive behaviour of Jay helpful. The most favoured aspect of Jay were the proactive reminders. 20 people mentioned that they would benefit from such a feature. On the other hand, one participant (P4) had concerns about whether this would become a habit: *"I think it will make me lazy and will have a bad effect on my memory overall"*.

15 participants pointed out that the timing for initiating a proactive action is crucial. P9 mentioned: *"When Jay is proactive, it should basically behave like a person. Jumping in every discussion or argument is going to be annoying."* Four people stated that Jay's proactive behaviour is fine only when being alone. When more people are present, they would not like to be interrupted by the VA: *"If I am in the middle of an interaction with one or more persons, I do not want Jay to interrupt."*

Five participants were sceptical about the social sensitivity of a proactive smart speaker. They raised concerns about an AI's understanding of the conversational context, which can sometimes even be difficult for humans. P27 mentioned: *"It would be great if Jay could learn some basic good manners and develop a certain level of social sensitivity by interacting with humans like children do. I could easily imagine a young kid interrupting a social interaction and being told off by his parents."* Seven participants pointed out that certain proactive behaviours could damage human-human interaction. P32 speculated*: "If the relationships in the household are suffering from a lack of time spent together, it may exacerbate the circumstances by taking time away from the families."*

7.3.3 Discussion and Conclusion

The survey results suggest that most of the participants tend to see proactive interventions as *useful*. However, the ratings for *appropriateness* were much lower, suggesting that appropriateness given (social) context will affect the overall acceptability of the interactions. Furthermore, there were various concerns regarding the timing of interventions and appropriateness in certain contexts, which resonates with previous studies (e.g. [8, 72, 245]).

The quantitative analysis revealed that in settings where users were alone with the VA, the interventions were generally rated more positively than when other people were present. This suggests people do not mind that much if they are interrupted when alone but feel that it is more intrusive when they are with others.

There were many comments about the social awareness of the VA, questioning its understanding of context and people's intentions. Related to that, participants mentioned that not all questions in an ongoing human-human conversation are meant to be responded to (by a VA). Social skills, such as when to speak or when to approach others, are complex abilities that are difficult for computer systems to master. A possible approach that was suggested, which could reduce inappropriateness in social situations, is that the assistant would ask more politely if it should suggest or remind about something, such as, "Would you like me to help you with that?" or "May I suggest something concerning ... ?" which resonates with the findings by Edwards et al. [122] on how proactive interventions could be initiated.

Overall, the scenario-based questionnaire study shows that people generally found proactive VAs useful, but many raised concerns about the timing of interventions and loss of control. Furthermore, the diverging opinions suggest that proactive VAs may be desirable only in certain situations and for some users. One approach that was suggested by participants is to be able to decide when to allow a VA to observe the environment and to be proactive. To better understand the perceptions beyond some of the aspects measured in this survey, an interview study was conducted, which is presented in the next section.

7.4 Study 2: Interview Study

This study was designed to gain a more comprehensive understanding of the factors that make a proactive intervention to be perceived as desirable (or not) for which an 'interactive interview' was developed that involved a range of tasks to elicit further people's reflections on the same storyboards used in the above study.

7.4.1 Study Design

A sequence of different interview parts combined with specific tasks were used to find out how participants perceive the depicted (social) situation and how they think Jay's intervention affects it, as well as to understand how proactive interventions need to be designed to mitigate any negative effects on people's (social) activities. An aim was to enable participants to think about the scenarios from different perspectives. For example, they were asked to order the 8 scenarios in terms of their usefulness, appropriateness, and invasiveness, and how they think people might react to the VA's interventions. They were also asked to provide ideas for how the VA should best intervene, including how it should initiate its interventions.

Participants

N = 15 people participated in the study, of which seven self-identified as female and eight as male. They were between 22 and 35 years of age (*M* = 27.86, *SD* = 4.47). Five participants had a bachelor's degree, nine had a master's degree, and one had a PhD. Participants were recruited using convenience sampling. The participation was voluntary and uncompensated. The recruitment continued until data saturation was reached, satisfying the recommended sample sizes of theoretical saturation from the literature [167, 423]. Two thirds of the participants had previously used VAs (four rarely, six often). Seven participants owned a smart speaker. All participants were proficient in English.

Procedure

The participants were asked to give informed consent and fill in the demographics questionnaire prior to the session. Before starting the interview, the interviewer explained that participants should assume the data is processed locally on the device[20]. The interviews were all held online. At the beginning of each one, participants were informed about the study procedure and the concept of a proactive VA. The tasks set as part of the interview were performed through a virtual whiteboard tool *Miro*[21]. All participants were given a short familiarisation phase with Miro and the virtual board. During each session, the participants shared their screens with the interviewer to be guided through the tasks. All sessions were audio-recorded for later analysis. The sessions took 51.3 minutes on average (*SD* = 10.6).

Data Analysis

The data was analysed in terms of: (i) content from the virtual whiteboards and spoken statements from the interviews and (ii) the information from the completed tasks on each participant's board. The findings reported in what follows focus on participants' perceptions of the scenarios and are thus mainly based on the interview data. More details on some of the findings of the separate tasks can be found in the underlying publication [FP3].

The analysis was reviewed and discussed by two other researchers and me. The transcripts of the interviews were independently coded by me and another researcher using inductive coding and subsequently merged and consolidated. Together with two of the researchers I discussed the codes, resolved disagreements, and derived themes. The themes can be categorised into (I) perceived helpfulness, (II) privacy and mistrust, (III) consideration of social context, (IV) configuration and control, (V) and initiating and phrasing of interventions.

[20] While some of Jay's features may not yet be feasible today with offline/on-device processing, we wanted to avoid participants solely worrying about data privacy, as this aspect is already well-researched.

[21] https://miro.com

7.4.2 Findings

Overall, participants had diverse opinions about the proactive behaviours in VAs. Some were favourable of the different kinds of proactive interventions Jay presented in the scenarios and valued the additional features, while others disliked them: *"I would rather ask [for help] than getting help without asking"* (P6). Some had mixed feelings: *"It's like a double-edged sword: both helps and can intrude"* (P5). Next, we consider the five themes. In each subsection, the findings are presented followed by an interpretation of them.

Theme 1: Perceived Helpfulness

The proactive assistance for the *Technical Support* scenario was positively perceived by most participants: *"[Jay] was smart enough to understand the initial question was aimed at another person. After seeing that no solution can be found, it jumps in and helps"* (P6). Reacting to indirect calls for assistance was also highlighted for the *Cooking Inspiration* scenario, for example, P5 said, *"The character is mentioning that she has no clue, and she needs help"* without addressing the VA. P6 also mentioned *"It's not just answering a question, but rather trying to solve a problem it has detected"* This suggests the participants considered this situation to be a meaningful 'entry point' for the agent to proactively intervene.

An observation was the participants' indecisiveness on whether proactivity is desirable or not, when they found interactions intrusive but at the same time useful. About the *Disagreement Clarification*, P8, said: *"Very useful but very scary. It can destroy you, but it will also cut the discussion short."* Similarly, for the *Fact Checking* scenario, P15 said, *"I think in this case, none of them are right, so the speaker was being helpful. If one of them was right, then they would feel bad about it, no one wants to be corrected."*

Interpretation. These results show that there are several situations in which participants find the proactivity both useful and appropriate. However, a common pattern in their comments was the dilemma of proactive interventions being perceived as helpful but at the same time intrusive – a *proactivity dilemma*. For several scenarios, participants were ambivalent about whether the intervention was overall desirable or not, hence a 'double-edged sword'. It also underlines how helpful an intervention is perceived often depends on the appropriateness of the intervention for the social situation. In other words, helpfulness is often not just dependent on the content of the intervention and what problem it can solve (e.g. helping a couple remember the time they agreed to meet) but also what it could mean for the social situation. This interdependence is captured by the following quote from P5: "*This can be helpful, but it can hurt people's feelings – that makes it not really helpful.*"

Theme 2: Privacy and Mistrust

The second theme of privacy and mistrust is key. Even though we asked our participants *not to focus* on privacy and data protection concerns (as this would have most likely biased the results largely toward this theme), they were identified as the biggest concern among participants. All interviewees wanted transparency and control in data processing, for example: "*If I know where my information is being processed and used, I can decide better to use such systems or not*" (P12). Some participants were concerned about the misuse of personal data for hidden agendas or providing proactive advertisements, for example, P10 said*: "[the agent] might give me suggestions that are influenced by political reasons or advertisements and try to control my behaviour based on that."* Another concern was about an entity intruding into the private environment: "*It's like another person is always at your home*" (P12).

One participant found it "*really scary that everything could be monitored*" (P8). The participants also pointed out that people might constantly feel 'observed' or 'judged'. This was especially prominent for scenarios where the agent interrupted a conversation. For example, for the *Fact Checking* scenario, P7 said, "*This would be intruding my privacy. It's an intrusive move. I see the assistant as a tool rather than an equal conversation partner.*" Mistrust was further expressed about 'false alarms' and 'misinterpretations' of certain situations and user states or behaviours by the agent, which might create confusion, frustration, or even conflict. It was also stated that for certain conversations which are not about a private or intimate topic, it might be acceptable for the VA to intervene, for example in the *Cooking Inspiration* scenario, "*The person herself is mentioning that she needs help ('she has no clue'). This is good timing [of the proactive intervention], and the topic is not really private.*" (P6)

Interpretation. In order for VAs to be proactive, they require more information about users' environment and behaviour, meaning more personal data needs to be processed to provide such services. Not surprisingly, interviews revealed how participants' main hurdle for adopting proactive VAs were privacy concerns as has been identified by others [264, 407]. Participants were worried about the misuse of their personal data by companies providing such VAs and third parties. Another concern was related to having an additional 'entity' in the home that is not just a passive servant – like current smart speakers – but rather some form of (social) 'actor' that takes an active role in their private space and family life. The participants associated this worry with paternalism and a lack of control over the device, fearing negative social repercussions, especially when there was more than one person in a space.

Theme 3: Consideration of Social Context

Generally, participants were sceptical about the agent's social awareness. Seven participants found Jay's interventions disruptive and intrusive when they interfered with ongoing conversations, for example, P1 said, *"[Jay] should not stop the thinking process and break conversations. It damages the human-human interaction"*. The proactive intervention was then considered by P16 as *"ruining the magic of the discussio*n." Two participants even perceived these interruptions as "creepy". Jay's interjections were considered unwelcome because the agent was seen *"as a tool rather than an equal conversation partner"* (P7). One participant considered it to be *"like a contract: everything is noted down. That's very stressful"* (P15). The content of the conversation was described as an important factor for proactivity by seven participants: *"If it is an intimate conversation, [Jay] should not really intervene"* (P10). Two participants were concerned about the missed opportunity of socialising and bonding with another person due to the imposed help by the agent: *"This is not received as an act of helping, but rather programmed"* (P15). Further, the presence of people in the room was a common theme: *"Emotional connection between me and my visitors is the key factor"* (P3). In the presence of other people, 12 of the participants preferred the agent to be proactive only if it was an urgent matter.

Moreover, most participants found it frustrating or unpleasant when the agent corrected users: *"People would feel bad about it. No one wants to be corrected"* (P14). One participant was torn as *"this can be helpful, but it can hurt people's feelings"* (P3). When the agent was contradicting one user while supporting another, participants found it even more insensitive. Regarding the *Disagreement Clarification* scenario, verifying what was previously agreed was seen as the assistant taking sides. Such well-intended interventions were thought to *"potentially cause users to argue"* (P13), and they *"could add more oil to the flame"* (P1). For the *Fact Checking* scenario, however, one participant assumed: *"I think in this case, none of [the users] is correct, so the speaker was being helpful."* (P14). Four participants speculated that the users in this scenario might feel offended, and three presumed that the proactive intervention would cause social awkwardness. In contrast, a small number of participants were in favour of these interventions, because *"it's nice to be corrected"* (P7) or *"it's factual and cuts the discussion short"* (P8). Similarly, two people appreciated the *Disagreement Clarification* scenario: *"I love this example. I think these arguments come up quite often, and everyone thinks they are right. Personally, in this situation, I would like to have that. I always dreamed about having such a system to check for the truth"* (P6).

Interpretation. In multi-user scenarios, the interventions in which the agent would help people resolve an issue and save time were perceived positively. However, other than time-critical or urgent situations, these were only perceived to be appropriate when the people had a chance to first try to resolve the matter by themselves. Participants generally thought that when the agent detected a question that was aimed at other people, responding to such questions before the intended person got a chance to respond was perceived as annoying and interfering. However, if the intended person could not properly respond to these questions or inquiries, the agent's intervention was considered useful and appropriate. For example, in the *Technical Support* scenario, the agent intervenes based on a request for help but only does so after the addressed person says they are not able to help at that point. Participants assumed that the agent was aware of the context and could appropriately detect an opportune moment to engage in the ongoing conversation. However, participants raised a concern about the agent taking away an opportunity of bonding, even if it is being helpful. They frequently mentioned that the agent's intervention in social situations is disruptive and could potentially damage human-human interaction. In accordance with previous research [286, 440], understanding the relationship between the people who are co-located, as well as the seriousness and intimacy of the conversation, were pointed out as important factors for the appropriateness of the agent's intervention in these situations.

Moreover, when the agent corrected people, some participants found it inappropriate, annoying, patronising or even insulting. The *Disagreement Clarification* scenario was rated most invasive and ranked second to last in terms of appropriateness. One reason for this was that in this scenario, the conversation was perceived as private. Additionally, the agent's intervention contradicts one of the people present and approves of the other, which resolves the disagreement but could further 'fuel' the conflict. Nevertheless, some participants still found this highly useful and wished for such systems in their households, e.g., to cut discussions short. This example illustrates well that there seem to be major individual differences in how the proactive interventions are perceived.

Theme 4: Configuration and Control

Most participants mentioned the importance of being in control and being able to configure the system's proactive actions, in particular concerning the timing and topics. Three participants suggested the possibility to switch proactivity off temporarily. Four wanted to regulate interventions based on who is present in the room. Limiting proactive interventions at specific times of the day was suggested by three participants. One proposed to set the agent's 'proactivity extent' using a slider in the settings. Hence, the users' agency was raised as a concern among participants (similar to the suggestion made for ProberBot in Chapter 5, see Section 5.6.3). They found certain proactive interventions of Jay patronising and imperious. Participants did not like the assistant playing the role of someone who is controlling certain aspects of their lives: *"I'm a person and I decide for my life. AI should not decide for me"* (P4) or *"If I have activated this [type of proactive intervention] in the settings, I would be more open to it. But if it is unasked for, I would be really annoyed"* (P10). For example, for the *Fact checking* scenario, one participant also pointed out, *"This is positive if I have previously activated it. I don't want it to do that in every gathering. Sometimes I may want to lie and it's none of Jay's business."* (P2). Beyond customisation, participants also hoped for the system to automatically adjust over time. Whether manual or automatic, for one participant *"it needs to be adapted enough to the user's needs in order to understand when it's really needed – and when not"* (P9).

Interpretation. Participants were concerned about their possible loss of agency. The feeling of being controlled and patronised by an agent was expressed as a worry. Based on our observations, the factors that would increase the chance of appropriateness for such interventions were the phrasing and the predictability of the interaction based on pre-configuration by the users. Participants wanted to be able to configure times and topics so that they could anticipate interactions to some extent and have more authority. For example, it was suggested that the 'feature' where Jay might intervene when it detects statements that are not in line with other existing evidence on a topic (e.g. based on encyclopaedias or other information the VA can access), may only be desirable in certain contexts, as sometimes social interactions do not need to be aligned with other evidence available to the VA – for example if the interactions are playful, not that serious, or when inaccuracies might not cause any harm. One way to overcome the concerns is to enable proactive VAs to be customisable, such as letting the user decide on how short they want their VA's responses to be [170].

Theme 5: Initiating and Phrasing Interventions

How to introduce proactive interventions was a recurring theme during the interviews. For most of the interactions, participants suggested – similar to the findings of the survey study – that the agent should ask for permission or give some kind of cue before speaking: "*Maybe it is more acceptable if [Jay] says 'sorry to interrupt'*" (P14). Some thought it is a good compromise to first announce that the VA is able to help or has a suggestion without being too specific yet, such as in the *Cooking Inspiration* scenario, P8 said: "*I also like that Jay asks before giving a direct answer. It's not intrusive but a possible solution for a problem they have.*" (P8) Based participants suggestions for how the VA should intervene we identified three different kinds of '*initiations*' for proactive interventions:

1. **Non-verbal cues** where the agent indicates an intervention with a visual or auditory signal but then waits for the user's prompt to proceed.
2. **Verbal cues** where the agent announces the subject but waits for the user's permission to proceed.
3. **Direct interventions** where the agent brings up the subject directly.

Direct interventions were mainly suggested for urgent or health-related scenarios but also more generally when there might be a need/benefit to act quickly. When interrupting conversations between people, non-verbal cues were preferred as they were considered to be the least distracting. Otherwise, the VA might be perceived as "*the annoying kid in the class that screams the answer*" (P10) instead of raising their hand, which might be the equivalent of a non-verbal cue. If a situation is not urgent or if the VA has previously been specifically set by the user to intervene without asking (e.g. for a specific task like technical support or to remind the person of their appointments, etc.), it was generally suggested that the VA should first provide certain cues/signs to let the people know that the VA has a suggestion or idea. For example, for the *Disagreement Clarification* scenario, one participant said that *"It shouldn't correct information in such situations. At least not proactively without asking me first.* ***Maybe a sign before*** *and then the information."* (P2) Similarly, it was also suggested that *"It is better if [the agent] gathers more information before making a conclusion and providing suggestions"*. Generally, participants suggested initiating the intervention in a polite and calming manner, gently 'building up' potentially distressing topics while keeping them goal-oriented and succinct.

Interpretation. The findings show participants expected the agent to ask for permission before conversing. This supports Arias et al. [12], who suggested that the agent should make sure the users are willing to interact at the specific moment. This permission request could be communicated in various forms. Verbal cues would have high conversational 'fidelity' in relation to human conversations, such as addressing the user by name ("Excuse me, Alex?") or polite phrases ("Sorry to interrupt?" – P14). A more subtle approach could be non-verbal cues of different modalities, such as abstract audio or light indicators. Depending on the ongoing activity, the preferences of our participants differed. The cue should not distract people from their activity unless it is an urgent matter requiring a striking cue. Verbal cues were described as the most distracting, followed by audible cues. Visual cues were described as the least distracting.

7.4.3 Discussion

This interview study showed in more detail when proactive interventions by VAs in everyday scenarios are perceived to be desirable. The findings demonstrate that the participants saw benefits in proactivity, specifically in cases of providing timely support or relevant information for an ongoing task. However, concerns such as privacy implications, potential loss of agency, and interference with ongoing (social) activities may negatively affect people's experience of such systems. To address the identified 'proactivity dilemma', it is important to consider how useful and appropriate an intervention might be for the ongoing (social) activity. Below is a set of six considerations for designing proactive VAs for everyday life. On a general level, these considerations also apply to the NLUIs covered in the previous three chapters; however, they are likely to be less ‘pronounced’ there. The reason for this is that the NLUIs in previous chapters ‘operated in’ more constrained task-specific contexts, which did not involve equally complex (social) settings that needed to be ‘navigated’ for proactive interventions. The *Fact Checking* (Figure 7.2) scenario is used for most considerations below to illustrate what they refer to.

1. **Permission to observe the environment in specific (social) contexts or at specific times.** Which permissions was the VA given to observe its environment (e.g. time of day, people present, etc.)?

2. **Permission to intervene in a specific way in a given context.** Which settings or preferences were made regarding the VA’s proactive interventions? What types of conversations and topics is the VA allowed to intervene in?

   Scenario Example: Was the VA configured by the users to fact-check the content of ongoing conversations based on information it can access (such as encyclopaedias) at certain times or for certain topics?

3. **Alignment with the goal of people’s ongoing activity.** (1) Is the intervention aligned with the overall goal of the ongoing (social) activity? (2) Does this activity or conversation involve any private, intimate, or (inter)personal aspects which could be negatively affected or disrupted by an intervention of the VA?

   Scenario Example: Regarding question (1), once the VA detects inaccurate facts in the ongoing conversation (statements on which empire is the oldest), it needs to determine if delivering the correct information is aligned with the goal of the ongoing activity. For example, is the goal of the ongoing conversation to agree on or learn historical facts (e.g. as part of a history assignment), or is it more playful and speculative (e.g. perhaps even part of a historical fact ‘guessing game’ and thus giving the correct answer would ‘break’ the game)?

   Despite its importance, question (2) might be less critical for the *Fact Checking* scenario, since the historic topic is less private/personal. In other words, the potential harm that a proactive intervention could cause for the social interaction may be smaller than in other, more private/personal conversations – as it might be the case in the *Disagreement Clarification* scenario, for example.

4. **Alignment with goals, interests, and abilities of individuals.** Is the intervention aligned with the goals, interests, and abilities of the individuals present?

Scenario Example: The VA needs to determine if the information of its intended proactive intervention is relevant for and contributing to the goals of the individuals present – do these individuals have an interest in knowing a fact (e.g. are some or perhaps even all of them going to have a history exam on the next day)?

Here, the provided information contradicts both parties and can thus be considered 'neutral'. However, in other cases where the interests compete, and the intervention may be in favour of one of the parties but not the other, the VA would have to more carefully decide if it should 'get involved' or not (e.g. as it is the case in the *Disagreement Clarification* scenario, for example).

In some cases, the VA might also need to consider if users would be able to understand (and execute) what it suggests given their mental and physical abilities and skill at a certain task (e.g. considering the *Cooking Inspiration* scenario, it might suggest a recipe that is too complicated given a person's cooking skills, or in the *Technical Support* scenario, it might suggest a solution that the person may not be able to follow given their technical skills).

5. **Urgency and importance.** How urgently is the information needed to effectively contribute to people's goals and interests?

   Scenario Example: If all the above factors have been considered and the VA 'concludes' that an intervention is adequate (i.e. the people present are likely to have an interest in knowing which Empire is the oldest), the next question is how important and urgent it is for them to receive this information. And does the information need to be delivered at this specific moment, i.e. while they are having their conversation, or could it also be delayed and provided at a later point and perhaps through a different channel (e.g. a notification on their phone)? For example, assuming these are indeed students preparing for a history exam, is the exam just about to take place or is it only going to be in a few days?

6. **Appropriately tailored initiation and delivery.** How can the intervention be delivered so that it reduces any negative side effects, in particular regarding the previously assessed goal of the ongoing activity and of the individual people?

   Scenario Example: Is it most effective to just 'insert' the information into the ongoing conversation, or should the VA first check with the people present if they want to receive the information? For example, are the people already familiar with the VA and its interjections? Could it cause unease to be corrected and realising that they were wrong? The goal is to intervene in a way that does not create a feeling of unease, while at the same time being goal-oriented and concise to avoid causing confusion.

Taken together, these factors show the nuanced design considerations that are involved in designing VAs to proactively intervene in people's everyday lives and the activities they engage in, in particular, when these activities are social (i.e. involving multiple people). Many of them will only become feasible with further advances in real-time (on-device) AI capabilities, as a VA needs to have an 'understanding' of what is happening at specific moments, what the goals of the ongoing activities are, as well as what people's personal goals might be.

As indicated by the findings and the positive attitudes of some participants, there might be certain groups of people who are more open and forgiving to proactive interventions from VAs – even to interactions that most others did not find acceptable. For example, there were some people who seemed to "love" the idea of a VA getting certain facts straight in conversations and even disagreements among individuals. For those individuals, it may not be a major issue if some of the above factors cannot always be accurately assessed (e.g. how well an intervention might align with people's interests, with the ongoing activity, or how well it might be timed). They might also perceive a VA's proactive interventions as less disruptive, unpleasant, or inadequate. In short, depending on who is using a system, there might be certain differences in how 'well' the above considerations need to be addressed by a VA for an intervention to be seen as desirable or not. Nevertheless, it is unlikely that a proactive VA will be able to offer meaningful and desirable interactions for a wider population if these considerations are not adequately incorporated when implementing and deploying them more widely. But even if they are carefully considered, there are more fundamental questions that will still need further research and discourse as to whether and how such proactive features can be designed so that they do not undermine human agency and autonomy or interfere with or disturb interactions and relationships between humans.

7.5 Conclusions

Since proactive VAs that are comparable to those illustrated in the set of storyboards are not yet available in the market, a *speculative design/design fiction* approach [17] was employed for the questionnaire and interview study. This method enables evaluating aspects of the system that would be difficult to explore from an ethical point of view in a study in which functional prototypes would be deployed, such as intimate private settings and conversations. However, since participants did not experience the situations and proactive behaviours themselves in the studies, their perceptions are likely to only reflect in a limited way the experiences they would have with such a VA in the real world. While some of the proactive behaviours may turn out to be less problematic than people might have assumed based on the scenarios – many of them may also turn out to be more disruptive, annoying, or intrusive than people expected them to be based on the present scenarios.

The two studies showed that proactive VAs in everyday situations can support people in their ongoing tasks in new ways and that participants generally consider many of the hypothetical proactive interventions to be useful. However, there are various challenges concerning the desirability (e.g. appropriateness and invasiveness) of their proactive interventions – in particular when multiple people are present. As the studies showed, some of these challenges could be addressed by adapting when and how the VAs intervene depending on people's preferences and by adjusting the way prompts are delivered and phrased.

The findings from the studies demonstrate that the perception of the desirability of proactive interventions is highly contextual. It depends on the type of ongoing activity, the urgency of the topic, the user's current emotional state, the agent's initiation and phrasing of the intervention, as well as how aligned the intervention is with the people's interests. Developments in AI models that can process information (e.g. natural language) in real time will likely help address some of these challenges. However, there are many other open questions that are mainly related to the (interaction) design of these VAs which will also need to be addressed. For example, how the VA could be configured for people's preferences and how it would adapt its behaviour during specific (social) activities including the phrasing and delivery of its interventions.

It is also important to note that although a majority of the participants thought that many of the VA's interventions were undesirable for the given ongoing (social) activities, some of them gave much more positive ratings. For example, even for those interventions that were seen as being undesirable by most – such as when a VA intervenes in a disagreement to get a fact right – some mentioned that they would be happy for a VA to intervene in that way.

In contrast to the NLUIs described in previous studies (VoiceViz, ProberBot, and SelVReflect) proactive VAs 'operate in' a significantly more complex environment – while the NLUIs in the former three chapters were all embedded within a specific software tool used for a task and thus the environment they 'operated in' was confined by what the task involved, the NLUIs presented in this chapter are embedded in the real world and thus need to make sense of a much wider range of activities, tasks and interactions between people. While the previous three studies showed that people generally found the proactive interventions not only useful but, in most cases, also meaningful and appropriate for what they were doing and thinking at a given point, there are various challenges involved in achieving this for proactive VAs in everyday contexts. NLUIs that point out something to a user concerning their current activity may generally just work better when this happens in a way that is 'embedded' into cognitive tasks performed at a (computer) interface where one tries to work on a specific task for an extended period of time. This is in contrast to the messy, nuanced, and interwoven activities and social interactions in everyday life, where it can be much harder for an NLUI to intervene in meaningful and desirable ways and at the right time.

In the next chapter, the final discussion is presented, which examines the body of research covered in the previous chapters, beginning with answering the three research questions.

8. Discussion

This chapter discusses the findings of all the studies reported in the previous four chapters, which investigated different aspects of how NLUIs, which take the role of 'cognitive co-pilots' (in this chapter abbreviated as CCs), can be designed to support different types of tasks. The chapter begins by summarising the main findings and then addresses the three main research questions outlined in Chapter 1. Building upon this, Section 8.2 then provides a set of design principles and considerations. This is followed by Section 8.3, which summarises the main contributions to knowledge of this PhD research. Following this, Section 8.4 discusses some of the main limitations of the conducted research, leading to Section 8.5, which outlines some of the possibilities for future research on CCs. The chapter then concludes with an overview of some of the ethical considerations of CCs and possible ways forward in Section 8.6.

8.1 What was Found?

The main contribution of this PhD thesis was to demonstrate how NLUIs could be designed to be proactive (i.e. initiating a dialogue with a person or triggering inner thoughts and actions) and, in doing so, support human cognition in a range of different tasks both when using software tools to perform specific activities or in everyday settings. The vision that motivated the programme of research was to design various interfaces that would make people stop and think about what they are doing. This was done through scaffolding, probing, or guiding questions to enable people to reflect and support their sensemaking.

The research examined both the advantages and challenges of NLUIs (e.g. to support reflection) and proactive (natural language) interfaces and how they can support ongoing tasks as well as task/interface-embedded NLUIs. The novelty of the research contribution lies in the way in which these existing ideas were brought together and how they were applied to different tasks and contexts:

1. With the exception of learning tools, NLUIs have not yet been more widely embedded into software tools to scaffold and support cognitive tasks (see Section 2.1.4). Here, we showed how they can also scaffold people's cognition for a range of tasks.

2. Proactive interventions (by NLUIs) have mostly been applied to providing information or recommendations (see Section 2.1.5). Here, we demonstrated how proactive CCs could also be designed to help people in discovering new things, expressing something, or identifying ways to improve their decision-making.
3. The benefits of facilitating reflective thinking through NLUIs have so far focused largely on wellbeing and educational domains (see Section 2.2.2). Here, we show how reflective thinking facilitated by an NLUI can be useful for a wide range of tasks that benefit from or require reflective thinking.

The NLUIs that were developed were able to support people in performing tasks such as decision-making, by proactively asking them questions that facilitated and scaffolded reflective thinking. Three different prototypes were designed and evaluated: VoiceViz (Chapter 4), ProberBot (Chapter 5), and SelVReflect (Chapter 6). They demonstrated how it is possible to extend cognition in a variety of ways, through essentially getting people to stop and think and approach a task in a different (and potentially more systematic) way, give them an idea for how to proceed, or what else they could consider. At the same time, the findings also showed that besides the opportunities of proactivity for certain tasks and contexts, there are also major challenges, such as privacy, autonomy, and the agency of people interacting with such proactive NLUIs. Some of these challenges were also examined in more depth in the final set of studies, which explored how proactive NLUIs support people in their everyday activities by providing them with relevant information for their ongoing activities.

This section discusses these and other key findings and addresses each of the three research questions outlined in Chapter 1. Below, a brief summary of the findings for each research question is provided before they are addressed in more detail in the following sub-sections.

RQ1: How can ‘cognitive co-pilots’ be designed to proactively support people in tasks they engage in? (Section 8.1.1)

The set of studies conducted in this PhD have shown how CCs can be designed to intervene at opportune times to help users perform a variety of tasks, for example, analytical, decision-making, and creative ones. Study 1 (VoiceViz) showed that designing a CC to have a set of open and closed questions either provided through text or voice triggered further trains of thought and discussion in a collaborative setting. It did this by getting the participants to ‘change tack’ and think about things they had not considered before. The use of voice was particularly effective at making them stop in the moment and enable them to reflect. In contrast, when the prompts were presented as text messages on the screen requiring participants to read them, they chose *when* to do this, enabling them to be more in control of the conversation. Study 2 (ProberBot) showed how scaffolding prompts can also be provided at key points in decision-making processes to help people reflect and elaborate on their rationale before proceeding with a decision. Here, the prompts were text-based and designed to get the user to evaluate and make their intended decision more explicit through a set of different interactive UI elements. The study showed how this can help consider multiple perspectives and criteria in one’s decision-making. Study 3 (SelVReflect) showed how voice prompts can be designed to intervene in a creative task by providing both encouragement and inspiration and by evolving over time – from initial ‘hands-on’ guidance, which helped participants to put a basic structure in place, to more exploratory prompts, and finally ending with higher-level reflection prompts, which enabled participants to make sense of their creation. The prompts in this and the other two studies were triggered when users seemed stuck or when they seemed to miss an important aspect in their thought process. This approach worked well in most cases for the chosen tasks, but there were also cases when the prompts interfered with participants’ ongoing thinking.

RQ2: How can cognitive co-pilots support reflective thinking? (Section 8.1.2)

The findings of the studies showed that there are many ways in which reflective thinking can be supported. The tasks investigated in this thesis were all open-ended, and thus, they generally benefitted from or even required reflective thinking in order to progress effectively with them. The CCs were designed to encourage people in their reflective thinking when it might be difficult to do so without any scaffolding or guidance and when they might not know how to proceed with the task. Some of the main ways this was achieved in the present tasks were to encourage the user(s) to explore a set of data from different perspectives; to consider different possibilities before making a decision, and in doing so, reduce the risk of making a rash decision; or to be systematic when making sense of an experience and expressing it. The different kinds of reflective thinking involved in these tasks and how they are supported are discussed in Section 8.1.2.

RQ3: How can the findings of the studies be conceptualised and lead to a model of how scaffolding NLUIs, like cognitive co-pilots, extend people's minds? (Section 8.1.3)

As discussed in the literature review, there are many different models describing human cognition and the different processes and types of cognitive activities it involves (see Section 2.2.1 for example). The model developed in this thesis does not intend to provide another 'description' of human cognition, but it rather intends to conceptualise how technologies can scaffold and encourage people's reflective thinking by supporting an iterative process of external cognition. This model is presented in the second part of Section 8.1.3.

In addition, the present research extends the conceptualisation of proactivity and the role it could play in augmenting human cognition, describing its (potential) impacts in different scenarios in more nuanced ways than has previously been done in the literature. Previously, much of the concern in HCI about proactivity has been negative, worrying about its disruptiveness and intrusiveness. Even though most of these concerns remain, the research conducted here has demonstrated how it can also have positive effects for certain types of tasks, for example, encouraging the user to think in different ways and be more reflective in their sensemaking, decision-making, or self-expression.

The following three sub-sections address these findings in more detail for each RQ, starting off with RQ1. In these three sub-sections and the following parts of the discussion terms, ***scaffolding CCs*** are used to refer to CCs in the former three chapters (Chapters 4-6), and ***informational CCs*** to refer to the CCs presented in Chapter 7[22]. The reason for this is that the CCs presented in Chapter 7 do not aim to scaffold human cognition by asking guiding questions like the other three CCs but rather provide information and/or suggestions for what people should do.

[22] Note that RQ2 (Section 8.1.2) and RQ3 (Section 8.1.3) mainly apply to the ***scaffolding CCs*** but less to the ***informational CCs.*** The findings of Chapter 7 on the ***informational CCs*** will thus be discussed in Section 8.1.1 – in Section 8.1.2 and Section 8.1.3 they are mainly used to 'contrast' the findings on the ***scaffolding CCs*** rather than to address the RQs themselves.

8.1.1 Designing Cognitive Co-pilots to Proactively Support People (RQ1)

To answer RQ1 on how CCs should (best) be designed for the types of tasks that they intend to support, we considered how the CC should help users progress with their ongoing activity by intervening at adequate moments in a task with the aim to reduce interferences. By designing the CC in ways that would be integrated into the interface of a software tool, it could be shown how it is possible to provide effective prompting, guidance, and scaffolding that would facilitate people's thinking rather than distract them from their task. For VoiceViz, even though the embedding varied slightly depending on the experimental condition (i.e. the modality being voice or screen), it was similar to ProberBot and SelVReflect in the sense that the prompts were delivered within the interface used for the ongoing task. For the *informational CC*, the CC was not embedded in an interface but 'embedded' in the environment in which the everyday social interactions and activities depicted in the storyboards took place.

The idea of embedding CCs in task-specific software tools was that CCs could be designed specifically for a given task. As such, the ways in which the CC intervenes could be informed by how people tend to perform the given task, what they might have difficulties with, and how this is reflected in their behaviours while using the interface. For many software tools, the proactive interventions can thus be triggered by specific user interactions with specific interface elements (e.g. pressing a specific button), as these interactions can be tied to what the user might be doing or thinking (i.e. 'proxies' for cognitive processes). This was the case for ProberBot, where when a user was looking at certain stock-related information, this could be used to trigger questions that would be specific to what information the user considered in their decision-making (e.g. if someone looked at the news items on a specific stock). However, in some cases, it can be more difficult to make inferences about a person's thought process only based on their interactions with a software tool. This was the case for SelVReflect, for example, where the process of visually representing an experience was highly abstract and individual, and it was thus more challenging to infer what a person might be working on and thinking about at a given moment during the activity. Nevertheless, even there it was possible through the user-centred design process to identify prompts that seemed to work sufficiently well for how the task might be performed by most people.

The types of tasks and activities that were investigated in the studies were deliberately chosen to be diverse – covering analytical/sensemaking (C4), decision-making (C5), and creative activities (C6), as well as a range of everyday social activities (C7). This was with the aim to 'test' and investigate the concept of a CC in different contexts to examine what role it could play, how the CC could be designed to best support the ongoing task, and how people respond to it and perceive it depending on the task.

A user-centred approach was followed to design the various CCs and their prompts in order to determine when and how the CC should best intervene in specific activities. Given the differences in the tasks which the CCs were designed to support, there were thus also various differences in the CCs' characteristics. In the following subsections, the differences in the design of the CCs will be discussed following the structure outlined in Table 8.1[23].

Table 8.1: Design characteristics of the studies and NLUIs.

Design characteristics	VoiceViz Chapter 4	ProberBot Chapter 5	SelVReflect Chapter 6	Scenarios Chapter 7
(1) Single- / multi-user	multi	single	single	multi/single
(2) Modality of interaction	voice/screen	screen	voice	voice
(3) Main trigger for proactivity	conversation	specific action	inactivity	conversation

Characteristic 1: Single or Multi-user

Starting with the first row of Table 8.1, one way in which the studies differed was in terms of focusing on a **multi-user** versus **single-user** context. While VoiceViz (C4) focused on multi-user interactions, ProberBot (C5) and SelVReflect (C6) focused on single-user scenarios, and the storyboards (C7) mainly on multi-user but also some single-user interactions. There are clearly differences in how the CC can and should (not) interact if there are multiple people present who may engage in an interaction with each other. For example, when an activity

[23] It is worth reiterating that the studies in Chapter 7 did not involve the design of a system that people used; however, it is still included in the table for the sake of completeness.

involves social interaction and conversation with more than one person present, a CC can get a more accurate 'picture' of what is happening and how people might be progressing with an activity based on what is being discussed (e.g. [11, 13]). Beyond knowing how people are progressing, there is also the advantage that interventions can be triggered based on statements that indicate that there is a clear need for information or inspiration (for example, when people make statements like "I wonder what we could do to…" as it was the case when people were not sure what they should cook in Chapter 7).

This is in contrast to single-user contexts, where a CC cannot rely on people's speech to get a sense of what is 'going on' (unless a person is talking to themselves). The proactive interventions in this setting need to be triggered in different ways (e.g. based on specific task-related activities). However, single-user settings also have advantages, since proactive interventions may not interfere with ongoing activities and interactions in the same way. More specifically, when a proactive intervention interrupts a single user – even if it can disrupt their train of thought – they may be able to decide relatively quickly if and how they should consider the CC's prompt when progressing with a task. In a multi-user scenario, the response to the intervention of the CC needs to be coordinated; first, people will usually need to build a shared understanding of the prompt, and then they need to agree on how to respond to it and how to carry on with their conversation. This is because the CC becomes – at least temporarily – a 'participant' in the ongoing collaborative interaction, which requires coordination among the collaborators. This resonates with existing research on how multiple people interact with commercially available voice assistants [39, 326, 327]. Even though these voice assistants have (so far) generally not been proactive, interacting with them in multi-party settings requires similar forms of coordination in order to decide which requests should be made to the voice assistant, who should make them, and to then decide how to proceed with the conversation following the voice assistant's response.

An example for the required coordination in VoiceViz was that the pairs had to decide if and how to respond to a prompt[24]. This was in contrast to ProberBot or SelVReflect, where both steps (making sense of and deciding what to do with the CC's prompt) were only done by one

[24] While this could be observed in both modalities, it was particularly pronounced in the text condition, where the pairs spent time reading through and discussing the prompt together.

user and, arguably, happened more 'fluidly', as a single person does not need to coordinate their activities with the other(s). Beyond that, there is also the aspect that a collaborative/conversation activity may sometimes be more difficult to re-commence at the point where it had been 'left off'. One reason for this is that certain dynamics of social interactions may sometimes need time to 'build up'. For example, considering the *informational CCs*, it could be possible that the people in the storyboards who were debating about a topic might enjoy doing so, and it might not be an issue for them to not have the correct information/data (e.g. on a historical fact). The CC's intervention during such a discussion may change and potentially disrupt its 'momentum', as participants speculated in the interview.

Hence, the studies reported here showed how collaborative and single-user activities might benefit from certain interruptions while others might be negatively affected. Which one should be chosen for or promoted by the interface depends on the activity to be supported – for example, how straightforward it is to identify opportunities/needs for proactive interventions based on the ongoing activity (e.g. based on existing literature), what can the CC contribute to the activity, and in which ways can the CC become integrated into it.

Characteristic 2: Modality of Interaction

Considering row 2 in Table 8.1, the modality which the CC used for its prompts was mainly chosen based on the task and the interface that was used to perform it. In the case of VoiceViz, two modalities were compared to shed light on how this would impact the way in which people would perform an exploratory task and interact with each other when trying to make sense of a set of data visualisations. In ProberBot, the interface built was a GUI based on those used in typical trading platforms (albeit simplified for the purpose of the study). Here, the embedding was via a chat-based interaction (i.e. text/screen modality), as this modality was considered most suitable for this single-user task that involves working with numbers, various graphs, metrics, and other visual information. Furthermore, the questions provided by the NLUI, as well as the user's responses, were relatively complex and required users to consider and synthesise different pieces of information from the trading interface (e.g. stock-related metrics), which was thought to be more effective in written/visual form. In SelVReflect, the voice interaction seemed to be most suitable for this type of creative task in VR, where having to read the questions may interfere with someone's visual expression. This was also supported by the user-centred design process in which participants expressed their preference for receiving guidance through voice-based prompts, as they considered this to be less distracting. Finally, for the *informational CC* scenarios, voice was used as the proactive interjections were mostly targeted at social interactions, most of which did not involve the use of any other technologies or displays where the prompts could be provided. Furthermore, voice made the most sense here, as the aim of the CC here was to contribute to and 'weave itself into' ongoing conversations between people. Having considered the design rationales behind the modalities used for the different CCs the remainder of this section will discuss some of the impacts which the different modalities had and/or how they were perceived.

The VoiceViz study showed that there are differences in how pairs carried out their task depending on whether they are speaking and listening to an NLUI versus selecting commands and reading the NLUI's prompts off the screen. Participants in the voice condition interacted more with the system, explored more of the available visualisations, and asked more questions. This resonates with previous research, which showed that voice interactions can be more interactive and engaging, as found by Kocielnik et al. [227]. However, differences in the conversation patterns between the conditions were more nuanced. Participants in the screen condition often needed a bit more time to commence their discussion after a prompt and had more silent pauses. In contrast, in the voice condition, the pairs spoke more as if they were 'thinking out loud' and more willing to brainstorm, as well as more exploratory and speculative in their collaborative sensemaking and reflection. Furthermore, voice prompts could directly be responded to by participants in the voice condition, whereas in the screen condition, participants had to first decide when to direct their attention to the text prompt and read it. In addition, in the screen condition, participants seemed to spend more time thinking and discussing the prompt and building a shared understanding of it – taking a somewhat more 'analytical approach'. In the voice condition, there seemed to be a tendency to avoid longer silences after a prompt and keep the conversation going. Participants might have spent less time thinking about the prompt and instead more directly proceeded to exploring what possible answers could be by generating hypotheses.

In the ProberBot study, the participants appreciated that they could type their answers and at the same time interact with interface elements (e.g. sliders, multiple choices), while in the SelVReflect study, participants commented on how the use of voice for prompts made them feel more comfortable exploring and expressing themselves as well as reflecting on certain aspects of their past experience. In the case of the scenario-based studies on the *informational CCs*, it became apparent how designing voice interactions involves various challenges – in particular, when saying something during an ongoing social interaction between the people in the setting, where it often was perceived to be too intrusive.

More generally, having a voice interface may elicit or facilitate more creative and exploratory (collaborative) behaviours, whereas text/screen interactions may be suitable for supporting activities involving more analytical aspects or involving data, which might be easier to convey and interact with visually. Furthermore, voice prompts require more immediate attention, while considering/reading text or screen prompts can be more easily delayed – which of both is more effective and desirable depends on the task and how disruptive prompts might be.

Characteristic 3: Behavioural Triggers for Proactive Interventions

Row 3 in Table 8.1 is concerned with how to design the trigger for a proactive intervention. In short, all the CCs provided their proactive questions based on people's ongoing activity. However, there were differences in which aspects of the ongoing activity were most relevant for triggering a question. Since VoiceViz and the *informational CC*[25] storyboards focused on situations where multiple people interacted, the conversation provided the basis for the CC's intervention – in the sense that it would 'follow' the conversation and determine when it could 'contribute' to it. In the case of ProberBot and SelVReflect, which were both single-user activities, the decision of when to intervene was made based on what the person was doing (within the respective interface) while performing the activity. In the case of ProberBot, the

[25] At least the four scenarios covered in Chapter 7 all involve social interactions. There are a few other scenarios in the publications underlying this chapter where only one person is present, and the intervention is thus based on the person's ongoing activity and other information that might be available and relevant for the situation (such as calendar appointments, etc.).

interventions were triggered every time the person would make a buy or sell trade[26], as this was considered one of the key points where reflective thinking might be beneficial, as the aim was to get users to consider their rationale and motivation when investing. In the case of SelVReflect, the triggers were based on the person's activities involved in expressing themselves using the VR canvas. In particular, when there was extended inactivity within the canvas and/or extended activity within the 'palette' (the toolbox/menu with the expressive tools), which was used as an indication that the person might not be sure how to represent one aspect of their experience.

One key difference in how SelVReflect was conceived with regard to proactivity was that the person could also request questions by themselves (also referred to as *mixed-initiative* interfaces [189]). This was done as the design process showed that although there are specific behaviours that might indicate that a person is 'stuck' or not sure how to proceed (e.g. inactivity while choosing a tool) there are other situations where it might not be directly clear from the behaviour (e.g. when a person might be actively drawing but feeling unable to represent something in the way they want). This was in contrast to VoiceViz, for example, where it was generally more straightforward to determine if a prompt might contribute to the pair's sensemaking process, as this process was more 'visible' (or rather *audible*) given that participants were continuously engaged in a conversation as mentioned in the previous section (see also [11, 13]). Hence, there are different ways to determine the points in a task when people might benefit from a prompt.

The prompts that were delivered by the CCs were designed to be triggered when users might need some support and scaffolding for what they were doing or discussing. Sometimes, this can be 'hit or miss' depending on whether it can be inferred what they might need in a situation. Furthermore, when there is a longer silence or inactivity, it might be because users are stuck but they could also just be thinking about something. For example, in VoiceViz and SelVReflect, there might have occasionally been long silences or inactivity due to participants thinking about a certain aspect. In such situations, the NLUI might sometimes have

[26] In the study reported in this thesis, the triggers for ProberBot to intervene were constrained to buying and selling a stock, even though other triggers based on past investment decisions and other behaviours within the interface were also implemented (such as looking at specific stock-related information).

interrupted a thought (for both voice and screen-based interfaces). However, this generally did not seem to annoy or bother participants that much, and they just continued with what they were doing, suggesting that it was usually not that critical when to intervene in these more open-ended tasks. In many ways, this is similar to a human-human conversation, when sometimes two speakers overlap, or another interrupts a conversation. This also resonates with research that showed that some interruptions by proactive interventions can be accepted by people as long as they are perceived to be helpful, as found by Peng et al. [318], for example.

In SelVReflect this was similar, since the creative task could be done in various ways, and thus the CC's interactions would not be overly disruptive even if a person was working on or thinking about something different at the time when a prompt was delivered. For ProberBot, this was somewhat different in the sense that the proactive interventions were always delivered at specific points that were relevant in the decision-making process (i.e. when people were about to make a trade). Although there might have sometimes been certain forms of interruption, ProberBot's interventions were triggered based on specific interactions or activities within the trading interface (rather than silence or inactivity like in VoiceViz and SelVReflect) and were thus more directly tied to the user's activities. Here, participants generally made no comments about the ProberBot disrupting their thoughts but rather that it was occasionally just intrusive, asking them questions that they did not feel the need to consider. In the *informational CC* scenarios, the CC's hypothetical interventions were triggered based on ongoing conversations. Although participants found the interventions useful – in the sense that the CC seemed to have adequately inferred a need for information in the given conversation – they had more concerns regarding how desirable they would find the interventions in the specific social situation. This suggests that a wide range of parameters – beyond the need for information – might need to be considered to make proactive interventions in everyday social settings acceptable.

To conclude, the studies showed that the task and context generally determine how challenging it is to identify opportune moments for proactive interventions. However, by adopting a range of approaches to designing and delivering the CCs' prompts, they were generally successful in supporting participants in what they were doing – to help them reflect on and consider something that they might not have noticed, discovered, or understood. As described above, there were different rules for how prompts were triggered; for VoiceViz and SelVReflect, this was when the participant(s) seemed to be stuck or unsure how to proceed with the task and their reflective process (with both providing a prompt every couple of minutes) whereas for ProberBot this was when people performed specific actions in the interface (i.e. making a trade). As long as there is limited interaction from the NLUI (e.g. only every few minutes and/or only at critical points in a task), the findings of the studies suggest that interruptions generally do not seem to matter that much for open-ended tasks like those explored in the present studies. Yet, it is important to note that there are many tasks and contexts in which people might be less interested in, receptive to, and forgiving to such proactive interventions. This could be the case, for example, in more well-defined tasks, and where people might know (or think that they know) what they need to do. Thus, it needs to be carefully evaluated for a given context and activity if proactive interventions might be desirable and effective. Next, we examine how the prompts triggered reflective thinking.

8.1.2 How Cognitive Co-pilots Support Reflective Thinking (RQ2)

While the previous section mainly looked at the design of CCs and how and when they could deliver their prompts, the focus of this section is on how CCs' questions trigger reflective thinking to enable them to make sense of themselves or the materials they engage/interact with as part of a task (RQ2: How can cognitive co-pilots support reflective thinking?).

VoiceViz (Chapter 4), ProberBot (Chapter 5), and SelVReflect (Chapter 6) all covered open-ended tasks involving different things people had to make sense of. Furthermore, they all had a similar goal in terms of what each CC's prompts intended to achieve, namely, to provide questions that could support reflective thinking when progressing with the task. However, all the CCs differed in how they provided these prompts, which was mainly determined by the following questions: (1) In which ways can the task benefit from reflective thinking (e.g. to support the sensemaking)? (2) How should the questions be phrased and delivered to support forms of reflective thinking that are beneficial for the task?

Many open-ended tasks require reflective thinking in order to perform them – for example, to figure out *how* to perform a task, what to start with, what conclusions can be drawn from a set of data, etc. Therefore, when performing such tasks, people will generally have to engage in *some* reflective thinking – even without being prompted. It is thus likely that in the present tasks, participants (would) have also engaged to some extent in reflective thinking without any question prompts from the NLUI. However, since reflective thinking is often challenging, it can usually benefit from being scaffolded and supported (e.g. [291, 390, 393]) – such as by giving people ideas for what to look for or to consider. For example, a person might find it hard to decide which aspects to consider and what questions they should ask themselves when looking back at a past decision/experience without any guidance. Thus, in the same way a student can learn more effectively through scaffolding questions from a teacher or tutor [81, 312, 342, 393, 448], this thesis started out with the idea that people can in many open-ended tasks benefit from scaffolds that support their reflective thinking. But how do these scaffolding prompts trigger reflective thinking?

An essential part is to understand what the different tasks involved – in terms of what people are making sense of, what insights they might (want to) gain, and which forms of reflection might be required for this. Next, we consider how this might have occurred in the different studies.

VoiceViz

In the case of VoiceViz, the thinking required by the participants mainly involved determining some of the key trends and patterns in the data, including different rates of increase, similarities, differences, and potential relationships between the graphs, and speculating on possible reasons that could have caused them. The reflection that VoiceViz intended to facilitate to help participants make sense of the different data visualisations was thus to get them to think critically about some of the patterns and what they could mean as well as relating them to their own experiences and perceptions of the world (e.g. triggering reflections like “Are really six out of ten men overweight? In my family it surely isn’t more than one out of five.”). Participants were able to use the question prompts that were occasionally presented to them to explore and discover the questions by themselves, look at the data visualisations from different perspectives, and/or speculate on the reasons behind the trends and patterns they showed. In particular, some prompts were able to help participants to come up with ideas for what to look *at* or look *for* without pointing at a possible answer. In this sense, Vizzy was able to take a ‘back seat’ in the pairs’ ongoing conversations. They were able to provide ‘discussion hooks’ which the pairs could then pick up and reflect on or also ignore if they did not consider the questions to be relevant. The pairs in both conditions readily understood that the NLUI was not designed to have a conversation with them or to be an equal partner in the conversation but that it would occasionally prompt them when it appeared that they were stuck or in need of help.

Although the question prompts were designed so that they would ideally work ‘on their own’ without further scaffolding or follow-up questions, not all participants liked that Vizzy just asked the questions without engaging further in the conversation, for example: *“It was more like an examiner as we need to find an answer to the question it asks. While the guidance is quite minimal.”* This shows that there is a challenge in striking the right balance between reducing interventions to not disrupt the ongoing conversation too often versus also providing additional guidance (to avoid the NLUI being perceived to be an examiner). However, this perception was rather an exception, as participants generally knew what to expect from the beginning based on how Vizzy was introduced to them, and so they understood that Vizzy was only there to provide ideas rather than to tell them what to do. Nevertheless, the questions did not always get people to reflect, for example, when participants already thought about certain aspects themselves and/or when they were more experienced with analysing data. Yet, in most cases, the questions were able to get participants to reflect as their speculative conversations following VoiceViz questions showed.

ProberBot

The way ProberBot worked was somewhat different as it entailed a different kind of cognitive task. Given the complexity of making stock investment decisions, the questions were informed by existing literature on the common challenges, pitfalls, and biases involved in investment decision-making. These were then translated into a set of questions. The questions were put together into short dialogues which intended to probe the person’s thinking on various aspects and criteria that might be relevant for their decision-making.

Overall, the scaffolding questions and dialogues seemed to ‘work’ as they got participants to consider things that they did not think of but which they perceived as relevant. Participants appreciated the way in which the probing questions got them to more explicitly express and externalise their rationale behind a decision and how doing so got them to reflect on the way in which they made their decision. The reflective thinking that the CC intended to support here was thus concerned both with critical thinking about stock-related information (e.g. news items, valuation metrics, their price, and the trends and patterns in them) as well as metacognitive reflection concerned with how participants made their decisions. The findings showed that ProberBot generally seemed to be successful in triggering these kinds of reflection. However, there were also certain challenges which were captured in five '*tensions*' described in the findings of the ProberBot chapter and summarised in Table 8.2 below.

Table 8.2: Possible tensions involved in designing cognitive co-pilots that probe decision-making.

	→ Tension ←	
	Aim of an NLUI that intends to probe human thinking/decision-making	**User's potentially misinterpreted aim of the NLUI and their resulting behaviour**
1.	Slowing down the user's thinking for systematic decision-making	In situations where there is a need to 'think slowly' the user may, in fact, often be acting emotionally, thus feel a need to act quickly and ultimately ignore the prompts
2.	Scaffolding and probing the user's thinking with the aim to enable reflective thinking	Having a feeling of being nudged (by a hidden/implicit nudge in a prompt)
3.	Aiming to debias a user's decision by highlighting certain aspects of the specific decision and its context	There is a risk that changing a user's decision may sometimes not lead to a better outcome
4.	Getting a user to reason and explain to better understand and formulate a decision and their rationale for it	Post-hoc rationalisation of a (potentially flawed) decision without any improvement (or perhaps even a reinforcement of an existing bias due to confirmation bias)
5.	Giving the user control to tailor the probing NLUI to their decision-making process, strategy, and goals	Risk of configuring the NLUI so that it does not provide its probing dialogues when the user might need it most

The first tension points towards the 'dilemma' of needing but not necessarily wanting the ProberBot, in particular when someone might actually benefit the most from it (i.e. acting emotionally when they should ideally slow down, take a step back, and think systematically).

The second one refers to the challenge that despite trying to formulate scaffolding questions, they can still be perceived to have 'hidden agendas', which can negatively affect how people respond to them. The third one refers to the challenge that although there are certain best practices and criteria in investing (and in many other contexts), there is generally not a specific decision that can be considered to be ideal (as nobody can predict the future performance of a stock) – thus changing a decision can also be 'for the worse'. The fourth refers to a general tendency of people that when having to reconsider a decision, they might just engage in post-hoc rationalisation rather than more deeply reflecting on it, which might only reinforce the existing beliefs (and potential biases). The fifth tension refers to the challenge that although participants wished for controls for when and how a CC should appear, they also believe that there need to be certain constraints to what can be controlled. The reason is that without certain constraints, the purpose of the ProberBot could be undermined, for example, when the user might not be keen to reflect on a decision and disable the ProberBot to do a quick (and possibly rash) trade.

The tensions discussed above reveal how interventions are sometimes not aligned with people's short-term intentions (even if they might be aligned with their longer-term goals). This can make it more challenging to design tools that get people to engage in reflective thinking while they perform a (decision-making) task. However, if users can configure the CC for their own strategies, goals, and interests, as suggested by the findings, some of the tensions may be 'softened' to some extent, and they might more readily use the questions to reflect.

SelVReflect

Considering SelVReflect, the reflective thinking that took place involved people making sense of and representing a past challenging experience. The questions seemed to encourage reflection on (a) the challenging experience itself, (b) how to visually represent its different components/aspects, (c) what can be discovered in the representation, and (d) what can be learned from it. The putative reflection that took place was when a participant understood and approached their personal challenge in certain and possibly new ways (if they managed to break it down, order it, identify relationships, etc.), which could enable them to express it and visually explore it better. Participants reported discovering new aspects of/related to the challenging experience they were not fully aware of before, gaining (self-)awareness, gaining new insights, discovering new constructive approaches for challenges, and feeling empowered or better equipped for the future. While some felt confirmed in their previous perspective on the challenge, others discovered new ways of approaching them more effectively in the future. This suggests that for a number of participants, 'transformative reflection' [143] might have taken place, which also corroborates the increase in participants' self-efficacy scores after using SelVReflect. However, not all participants reached such 'deeper reflection', which can be for various reasons – this might depend on how experienced they are in self-reflection, or it could also be contingent on how well the CC's questions worked for them and their chosen challenge.

To conclude, CCs can be designed to support a range of tasks by triggering different kinds of reflection, which can lead to different outcomes, as shown in Table 8.3.

Table 8.3: Outcomes of reflection for each study.

Outcomes of Reflection for User	
VoiceViz:	• Understanding the patterns in the data and their potential causes
ProberBot:	• Understanding the relevance of different data for a decision • Understanding one's own decision-making behaviours
SelVReflect:	• Understanding one's behaviour in a challenging situation • Understanding the meaning of this challenging situation (in new ways)

The last question that needs to be answered is in which ways did the CCs 'extend' people's minds and how this can be conceptualised in a model (RQ3), which is covered in the next section.

8.1.3 A Model for How Cognitive Co-pilots Extend People's Minds (RQ3)

Building on RQ2, addressed in the previous section on how CCs can support reflective thinking, this section addresses RQ3: How can the findings of the studies be conceptualised and lead to a model of how scaffolding NLUIs, like cognitive co-pilots, extend people's minds? To answer this question, the focus is on the ways in which CCs can be understood as extending the mind. As was discussed in the introduction (Chapter 1), there are various ways in which technology can extend the mind, such as by augmenting perceptive capabilities, problem-solving, reasoning and so on. Here, the emphasis was on extending how people carry out open-ended tasks, where reflective thinking is meaningful or even needed to proceed with the task, explore different ideas, approaches, alternatives, and perspectives, and to gain new insights. Hence, to address the question of how the CCs augmented the mind, not only the types of cognitive tasks need to be considered but also *if* and *how* the reflective thinking that the CC enabled might normally occur without it. It is thus worth revisiting some of the findings from the studies with a particular focus on the reflective thinking people engaged in and how this supported the participants in proceeding with the task.

In the case of **VoiceViz**, the qualitative analysis revealed that – independent of the condition – following a question by the CC (Vizzy), participants speculated on and explored various hypotheses and possible explanations. When participants were then asked to reflect on what role Vizzy played in their sensemaking and exploration of the dataset, they pointed out that the questions got them to consider things that they had not thought of before. Participants said that the questions were *"pointing to something that we have missed"* or helped them *"see what else was there."* This suggests that VoiceViz supported their thought process by extending the ways in which they made sense of the data visualisations. In the case of **ProberBot**, participants similarly pointed out that it got them to think about things that they might otherwise not have considered in their decision-making process. Compared to VoiceViz, the support from the CC here was less about discovering something but rather to reflect on an aspect which might otherwise have been 'overridden' by a rash/emotional response – i.e. to think '*fast*' when there might be a need to think *'slow'* (considering Kahneman's framing [205]). As the findings from the user study suggest, ProberBot seemed to have been able to trigger reflective thinking when this was adequate or even desirable. Furthermore, ProberBot was found to get people to reconsider and re-evaluate their previous decisions and rationale, which participants said they tend to forget about, if not actively prompted for – this reflects another way in which a CC can extend someone’s mind. In the case of **SelVReflect**, the CC was there to give the user ideas for how to express or reflect on something, which might be difficult without guidance (as also suggested by the user-centred design process). Considering our findings, it seems the CC was successful in getting participants to approach, conceptualise, and express the past challenge in ways they did not previously think of. This is further corroborated by the ratings participants gave the guide on how *useful* as well as *insightful* it was for representing and reflecting on the past challenge. The way thinking might hypothetically be extended in the ***informational CC*** scenarios, was towards providing procedures to solve a problem or perform a task, receiving factual information, or rectifying inaccurate memories. Despite being often viewed as invasive, participants generally thought of the interventions as being useful for the ongoing (cognitive) tasks – thereby extending what a person might have thought of (e.g. a solution for a problem they had not yet considered) or have readily available in their mind. However, as this final study was concerned with understanding people’s perceptions of the proactive interventions rather than letting them experience the proactive questions themselves, it is not possible to draw further conclusions on the CCs’ effects on cognition.

Taken together, the studies showed that CCs can support cognitive activities (mainly reflective thinking), which can be useful for performing and progressing with different types of cognitive tasks – and thus, in some ways, extending what people would have thought of without the CC. They do so by asking questions that can enable reflective thinking, as described in Section 8.1.2. However, there is another way in which CCs can ‘extend the mind’. This is explored at a specific ‘instantiation’ of the Extended Mind Thesis, which re-appeared several times throughout the previous chapters: external cognition (e.g. [362]). Therefore, the question of how the CCs extend people’s minds (RQ3) is also approached here by looking at how CCs can support special forms of external cognition. As such, an important aspect of how reflective thinking was enabled is by externalising one’s thoughts, as can be seen in the second row of Table 8.4. The idea here is that technology can take on the role of encouraging, supporting, and scaffolding this externalisation to help the reflective thought process further ‘progress’. Table 8.4 also shows what people made sense of and what they reflected on.

Table 8.4: Comparison of the different subjects/'materials' of sensemaking, modalities of externalisation, and foci of reflection.

	VoiceViz	ProberBot	SelVReflect
Making sense of:	Data visualisations	Stock-related data & information (change in price, metrics, news)	A past personal experience and its different aspects/components (e.g. people it involved)
Externalisation of:	Observations or hypotheses **by verbalising them**	One's reasoning and rationale behind an investment thesis **by writing it down**	Perceptions and understanding of the experience **by expressing it visually as a 3D structure**
Reflective thinking on:	What specific trends or patterns in the data show and the possible reasons behind them	One's decision-making process and the factors that were involved in it, how one could adjust their decision-making	What the experience means for oneself, which further aspects the 3D structure reveals, and how one could approach similar situations in the future

ProberBot and SelVReflect – in particular – involved different forms of external cognition mainly by constructing external representations which would *shape* and *be shaped by* the person's ongoing thought process. Building representations – in visual and/or written form – can often play an important role in reflection, sensemaking, and learning [124, 291]. In the case of ProberBot, participants were prompted to make various evaluations regarding their decisions by interacting with different interface elements. By doing so, they created and 'worked with' external representations of their thoughts in the chat window/interface panel. As part of this process, participants refined and sometimes reformulated their hypotheses and rationale while interacting with ProberBot. In SelVReflect, the external cognition particularly manifested in participants literally 'walking through' their creation, as part of that changing their perspective on different aspects, and refining and re-expressing parts of the visual representation. However, even in VoiceViz, where no external representations in visual or written form were constructed by participants, similar cycles of interaction and 'external cognition' can be found: While interacting with VoiceViz the participant pairs also (i) *externalised* their ideas following a prompt, (ii) built a shared understanding (or *'representation'*) of what the visualisations showed them through their conversation, and (iii) *iteratively developed* their shared understanding of the visualisations and their hypotheses about them. Therefore, in all three studies, reflection and the externalisation of thoughts through representations and/or conversation played a role in similar ways: When the CC asked a question, reflection was usually needed to externalise one's thought/idea, and then the externalised thoughts generally enabled further reflection.

The SECC Model

In this section, a conceptual model is presented that describes the assumed 'cycles of external cognition' which are scaffolded by the CCs. All the tasks in the three studies involved externalising people's thoughts – through *verbalisation* in the case of VoiceViz, through *writing/typing* in ProberBot, or through *visual expression* or 3D sketching in SelVReflect. These representations would then be revised and iterated over, through discussion[27] (VoiceViz), through a chat interaction in which participants had to motivate and evaluate their rationale in multiple ways (ProberBot), or through changing the visual representation (SelVReflect). To conceptualise these iterative cycles of (external) cognition and how they are scaffolded by tools like CCs, I propose the SECC (Scaffolded External Cognition Cycle) model, which is shown in Figure 8.1. The model mainly intends to highlight the relationships and the transitions between (i) people's understanding or 'representation' of what they are working on (which is referred to in the model as people's *'internal representation'* of the task and its materials) and (ii) how people express this internal representation using external means/resources (e.g. UI elements, text, drawings, speech) referred to as the *'external representation'* (see Figure 8.1). The model intends to illustrate the role the (iterative) construction of representations plays in (reflective) thought processes (see also Dix [112, 113]). In many ways, the model is similar to the idea of the cycles of aligning and updating internal and external representations described in existing accounts of external cognition (such as Kirsh [224]). However, the main goal of the model is to provide a framework that highlights the different ways in which scaffolding prompts provided by cognitive tools like CCs can *support* cognition in distinct ways.

[27] Although VoiceViz did not involve the creation of *external representations* as they are commonly understood, such as texts, diagrams, sketches, etc. (see also [224]), it involved the construction of a *shared understanding* among pairs of their observations and hypotheses through conversation, which followed similar cycles as the construction of external representations. See, for example, excerpt Segment 1-VOI: Pair 3 in Section 4.5.3, which shows how pairs started off from an idea that was initially verbalised, which was then revised and adjusted through discussion.

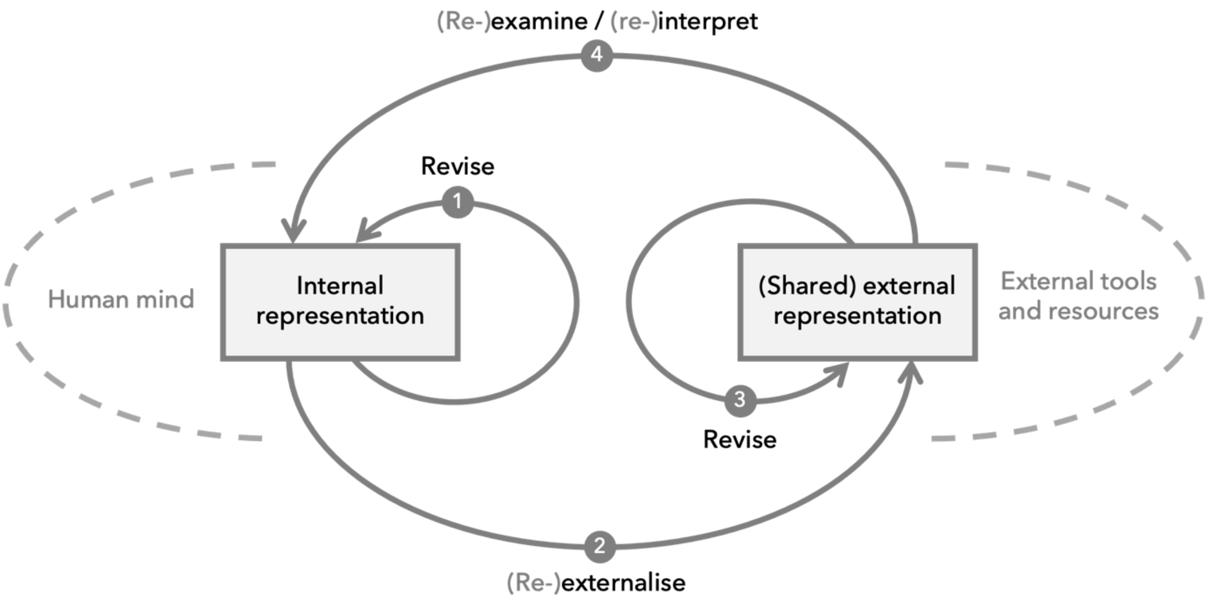

(Re-)examine / (re-)interpret
4
Revise
1
Human mind
Internal representation
(Shared) external representation
External tools and resources
3
Revise
2
(Re-)externalise

Figure 8.1: The SECC model showing the transitions between internal and external representations in cognitive tasks that involve/rely on external resources.

The model 'starts off' with the formation of a thought, which is developed and refined in a person's mind (which refers to **arrow 1** in the model, the internal **'revise loop'**). This formation of an idea can remain for some time in this loop of internal 'revision', but in the case of the tasks considered in this thesis, it would generally be externalised at some point to produce an external representation of it – which can be through verbalisation, writing, sketching, etc. (referring to **arrow 2** or the **'externalise transition'**). This created representation might be further refined (referring to **arrow 3** or the external **'revise loop'**) by oneself (e.g. ProberBot or SelVReflect) or collaboratively – for example, in the form of a conversation (e.g. VoiceViz). Once the external representation has been refined to some extent, it would usually be examined by the person to see if it reflects the internal representation and/or if it leads to new insights or interpretations (referring to **arrow 4** or the **'examine transition'**). The internal representation might then be updated or revised based on new insights (referring again to **arrow 1**, the internal **'revise loop'**). Following this, the person might externalise another (aspect of their) thought/idea and modify or extend the external representation (referring once again to the **'(re-)externalise' transition**) and thereby starting another cycle. If at some point there is a sufficient match/overlap between a person's internal and the external representation, and/or they think that they have reached a point where they have a sufficient understanding (of what they are working on – a data visualisation, a decision to be made, an experience they tried to break down in its key components) the current thought process might conclude. This could refer to a new insight, a decision being made or revised (e.g. ProberBot), or an aspect of the past experience successfully expressed (e.g. SelVReflect), following which a new 'cycle' through the model to perform another (sub-)task might start.

To illustrate the entire cycle with a concrete example, a scenario based on ProberBot is used: An investor might have an internal representation in the form of a (partially thought through) intention to invest in a stock, including the relevant metrics, predictions, etc. The ProberBot's question then gets the person to write down their intention and their rationale for it **(externalise transition)**. Following that, the ProberBot asks them various questions about their rationale, which stock-related information or news it is based on, etc., which might help the person to add various further aspects to the external representation of their rationale **(external revise loop)**. As part of this process, the person might already start to examine their external representation (i.e. their responses to ProberBot's questions) and reflect on how well it actually represents their intention of investing in the stock as well as if it adequately motivates and supports their intention **(examine transition)**. The **examine transition** might then be explicitly triggered when ProberBot prompts them if they still want to proceed with their intended stock investment based on what they have responded to its questions. In light of their responses, the person might then re-evaluate if their original investment rationale/thesis still 'holds' and if they want to proceed with their investment or if they should rather revise their intention and the rationale/thesis behind it **(internal revise loop)** which would conclude the decision-making process or 'cycle' in the SECC model.

It is worth noting that the aim of the model is not to accurately 'describe' when which loop **(arrow 1 and 3)** or transition **(arrow 2 and 4)** might take place at which point in a cognitive task that involves creating and working with external representations. In fact, humans might often rapidly alternate between these loops and transitions defined in the model, perform them in different orders, or even in parallel. For example, while someone is sketching out an idea on paper of a (***'externalise'*** transition, **arrow 2**), they might at the same time be thinking about how they can do so to best approximate their 'mental image' of their idea (***'examine'*** transition, **arrow 4**) and so on. Thus, rather than providing a tool to identify and describe which of these loops and transitions occur at a specific point, **the goal of the SECC is to**

highlight the different ways in which cognitive tools like CCs can support human cognition in tasks that involve or benefit from creating and working with external representations. As such, whenever someone needs to find a way to express something, to revise an internal or external representation, or to get new insights, some reflection is required, which can be facilitated and scaffolded by the CCs in order to raise and help address questions like: *How could I best express X? Does my expression match my thinking or understanding of X? What insights can I get from how I represented X?*

Considering the components of the SECC model, cognitive tools like CCs can support people with their scaffolding prompts in the following activities (note that the numbering corresponds to the four arrows in the model):

1. Revising and further developing a thought 'in one's mind' – without the need to use external resources (e.g. by providing scaffolds which help people consider a certain aspect in their thought process but without having to explicitly respond to it or act on it through external means). Some examples:

    a. *Are you interested in this stock because of some recent news?* (ProberBot)

    b. *Now think about the main stages of the challenge from the start until the end, when you ultimately overcame it. How many different stages or steps were there?* (SelVReflect)

2. Expressing thoughts through visual means, through writing, or verbalisation and thereby requiring people to make their thinking explicit and express it in a specific form, structure, or modality (e.g. verbalisation of observations in VoiceViz, ratings in ProberBot, abstract drawing in SelVReflect). Some examples:

    a. *What is your intuition for how this news item might affect the performance of the stock? Try to describe it in a few sentences.* (ProberBot)

    b. *Have you considered how the separate stages might be connected to each other? How could you design these connections? Could they differ?* (SelVReflect)

    c. *So, if you look at this, would you say that the increase is slowing down in the number of overweight people for all four groups?* (VoiceViz)

3. Revising and further developing external representations by pointing out possibilities of how the representation could be extended or enhanced (e.g. aspects that could also be introduced/added to a representation). Some examples:

   a. *Overall, do you feel all the important elements of the challenging experience are represented in your creation?* (SelVReflect)

   b. *Apart from the overall increase, is there anything else that you notice when looking at the shape of both line graphs?* (VoiceViz)

4. Using external representations to get new insights and interpretations of what people are working on/thinking about. This can be done by enabling them to examine the representation they have constructed (or are in the process of constructing) to get new insights and interpretations (e.g. of the decision or problem) from it. Some examples:

   a. *How might your intuition, which you've just described* [see example 2a above], *affect your original investment thesis?* (ProberBot)

   b. *Now, focus again on the actions and ideas that helped you overcome the challenge. How did you represent these and how do they tie into the whole process? Does the way you expressed them reveal something new to you?* (SelVReflect)

These are just a few examples of the prompts that were used in the present CCs to scaffold and facilitate the four different activities described by the SECC model, which are usually involved in tasks that rely on external representations.

In sum, the SECC model draws from and complements existing theoretical accounts on external cognition, putting a focus on the 'loops' and 'transitions' that are involved in constructing, working, and thinking 'with' external representations. Building upon the NLUIs presented in this thesis, the model highlights **four ways in which cognitive tools can support and scaffold this process**, as described in this section. The model thereby also illustrates how NLUIs can be *facilitators* or *'co-pilots'* of human thinking by supporting different aspects of (external) cognition. To revisit Douglas Engelbart's [125] vision introduced at the beginning of this thesis, one way to augment human cognition is by designing tools that support people in creating, rearranging, and manipulating various types of representations in real time as they progress with their thinking. Here, it is proposed that these forms of external cognition can be further augmented through CCs which scaffold that process of creating and 'thinking with' external representations. To draw from Douglas Engelbart's [125] definition of human augmentation, it could even be argued that the CC's **proactive cognitive scaffolds provide a form of *'superstructure'* to human cognition** and the processes of external cognition. In the next section, a set of design principles and considerations for designing CCs will be provided.

8.2 Considerations for Designing Effective Copilots

The findings of the three studies on the *scaffolding CCs* generally suggest that their proactive prompts can foster effective external cognition and reflection in a range of open-ended tasks. However, designing a CC's proactive behaviours in a way that they are meaningful and relevant – and at the same time acceptable and appropriate to the user(s) can be challenging. To achieve this in the best way possible a range of design considerations need to be taken into account, which are described in this section. First, two high-level design 'principles' are introduced, followed by seven design 'parameters' that are relevant for building tools like CCs.

Principle 1 – designing for human augmentation: CCs should enable people to find their own ways of performing and proceeding with a task. The CCs in this thesis intend to augment human cognition rather than replacing it (see also [364]). While this is a common aspiration for the design of (AI) systems [5, 60], it is proposed here that one way to achieve this is to design tools to scaffold human thinking. Where possible, a CC's prompts should thus not be in the form of instructions or recommendations for how to perform a task, as there can be a risk of undermining people's agency in performing the given task (some of the participants' comments in Chapter 5 and Chapter 7 underline this). Where this is feasible, CC should rather point out possibilities, which people can test and further develop to perform their task themselves, and which enable them in their *own* learning and thinking. This approach might also help overcome some of the challenges that have been identified for other (AI-based) systems, namely that people often over-rely on them and disengage with what they are doing, which can lead to flawed decision-making, lack of understanding, deskilling, etc. [26, 315, 388, 445]. As already mentioned earlier, there are, of course, also various exceptions to this principle, for example in situations where efficiently performing a task might be prioritised over scaffolding the human thought process.

Principle 2 – designing for external cognition: CCs should enable people to externalise their thoughts and 'think with' external resources. In other words, rather than prompting people to only imagine something in their mind (i.e. to build a mental representation), a CC should also provide ways to help them externalise their thoughts through external representations that can help them perform a specific task. If the CC supports people in expressing a thought through an external representation, this can (i) enable them to more effectively proceed with the task and their thought process, as they need to make their thoughts explicit (e.g. through verbal or visual expression) and (ii) this can further augment and empower their thought process by helping them get into a cycle of 'thinking with' and iterating on the representation they construct, as described in the previous section (see also [224]). The studies showed that such 'cycles of external cognition' can be enabled and facilitated by CCs for a range of tasks involving different modalities. The goal should be to help the user get into an interaction of iteratively refining their internal and external representations, which can be facilitated or 'moderated' by the CC. Furthermore, the studies suggest that giving users different ways to respond to and 'act on' prompts can further support and facilitate their thinking – ProberBot, for example, provided different types of UI elements (scales, multiple choice options, and text inputs) helping to express one's reasoning with respect to various criteria.

These two principles resonate with approaches and theories from educational and learning sciences, which posit that people need to solve problems and figure out solutions themselves to understand and learn something – as it is proposed by the Constructivist Learning Theory, for example [319, 427]. Related to this is the technique of asking scaffolding questions, which can enable learners to progress with their own thinking [144, 217, 341] or to reflect more deeply on something [51, 169, 232, 291, 373, 390]. Furthermore, research suggests that externalising one's thoughts can support cognitive processes [224] and building external representations can contribute to reflective thinking, and help people gain new understanding and insights [112, 113, 143, 291].

Taken together, the two *principles* aim to provide high-level guidelines that researchers, designers, and engineers working on or building AI tools (with proactive capabilities) might consider. However, it is worth noting that the principles might mainly apply to tasks for which at least some sensemaking and learning is desired or required from the human user when performing it. For example, familiar or everyday tasks in which the user knows what to do and how to proceed or other activities which do not need more extensive cognitive involvement from the user may not need (the same amount of) scaffolding from the system.

When a task or activity is considered to be suitable for being supported by a cognitive tool like a CC that intends to scaffold human thinking, a range of decisions need to be made regarding the CC's characteristics – that go beyond the two general *principles*. To guide some of these decisions, a set of design *'parameters'* based on the findings of this PhD research are described in the following – which are:

1) Proactive, reactive, or mixed-initiative interaction
2) Opportune moments for prompts
3) Frequency of prompts
4) Delivery of prompts
5) Phrasing of prompts
6) Modality of prompts
7) 'Evolution' of prompts over time

In general, many aspects of a CC, including these seven 'parameters', **should be configurable** to some extent by the user(s). For example, as seen in Chapter 7, participants' views and preferences differed significantly with respect to what characteristics they expect from a CC, suggesting that it would be best to give people some control to personalise their CC. However, it may often not be feasible or adequate to give users full control over a CC and its 'behaviours' (as could also be seen in the tensions discussed for ProberBot in Section 8.1.2), which means that some decisions will generally have to be made by the designers and engineers building the CC. Therefore, to guide these decisions, some of the main considerations are provided in the following sub-sections for each of the seven design parameters.

(1) Proactive, reactive, or mixed-initiative: CCs should only be proactive when there is a clear benefit for it.

First, the most obvious question that needs to be addressed is if a system even has to be proactive. Proactivity with the aim to augment human cognition may only be promising for (i) more complex non-routine tasks where people might benefit from help (without exactly knowing which form of help they might benefit from), (ii) for tasks that are open-ended and possibly ill-defined where it is meaningful to introduce different approaches, considerations, or perspectives, (iii) related to that, where deeper thinking and reflection may be beneficial, and finally, on a more practical level, (iv) tasks for which it is possible to determine with sufficient certainty when and where people might benefit from help/input. If none of these points apply, it might be more appropriate to choose a reactive interaction model instead (i.e.

as is the case with most technologies in the sense that users control them and make requests to them). The more of these four criteria apply to a task, the better it might be suited for a CC to proactively intervene. However, even if a task appears suitable for proactive interventions, it may generally still be advisable to also design the CC to *respond* to user requests (i.e. mixed-initiative, as it was also the case for SelVReflect), as a user might have specific and often highly individual/unique needs at certain points in a task, which might not be possible to predict and implement using any trigger rules for the CC's proactive interventions.

(2) Opportune Moments: CCs should only intervene based on clear signs that the user might benefit from a specific input.

As previously described, there is a range of 'indicators' for identifying opportune moments to provide a prompt. Tasks that involve external tools and resources and/or that involve human-human conversation can make a thought process and how it progresses more 'observable' for a CC. Indicators for opportune moments could be that based on what is discussed or created it can be seen that people are **(i) not making clear progress** for extended periods (e.g. SelVReflect or VoiceViz), **(ii) missing important aspects** in the materials they are working on (for the tasks where this can be determined, e.g. VoiceViz), or **(iii) expressing confusion or a clear need for information or inspiration** (e.g. SelVReflect and the Scenarios in Chapter 7). For other tasks, the opportune moments might be clearly tied to specific actions, such as **(iv) being in the process of making a major decision** that might benefit from critical thinking, which could be informed by or based on existing evidence on potentially suboptimal tendencies in human behaviour in the given type of task (e.g. ProberBot and the fallacies that can be involved in certain investment decisions). User studies can help determine when and where there might be opportune moments and how this can be identified based on the user's behaviours, which can then be translated into specific prompts and rules for when they are triggered.

(3) Frequency: CCs should generally prompt sparingly.

How often and when to be proactive is a core question for designing CCs. Obviously, it depends on the type of task, the length of the task, and the context. A 'heuristic' is that proactive interventions should only intervene in human thought processes and human-

human interactions in limited ways. For most tasks, the goal should be to provide only occasional 'input' to human thinking. What the appropriate frequency is will depend on the task, but as a rule of thumb, prompts might generally need to be a few minutes apart (unless the user specifically requests a prompt) to avoid disrupting people's train of thought too much, in particular when they are *spoken* to people. Although this will depend on the type of CC and the kinds of prompts they provide, there is also a risk of over-reliance, as too frequent interventions could habituate people. As a result, users might start to just wait until they receive the next prompt from the CC rather than trying to figure out how to proceed with the task themselves (for example, one participant in Chapter 7 pointed out that they are concerned that knowing that the CC will proactively intervene might make them lazy). Thus, one of the main challenges when designing proactive systems is to try to not be 'too helpful too often' (e.g. Microsoft Clippy). This also relates to some of the tensions introduced earlier in the context of ProberBot. In short, when the guidance from the CC gets too extensive, there is a risk that it might make people become *less* exploratory, engaged, and empowered.

(4) Delivery: CC should only intervene 'without warning' when this is thought to benefit ongoing thought processes and conversations.

It may often be adequate to provide some form of cue before a CC intervenes, in particular when it might be difficult to determine with sufficient confidence if people are receptive to the prompt (e.g. as was the case for many of the scenarios in Chapter 7). However, when there might be advantages in designing the CC to interject directly, for example, to provoke people's thinking or stimulate conversation (e.g. VoiceViz) the delivery might be more direct. Related to that, users should also be given the choice if they want to engage further with what the CC might have to say (e.g. after it has given them a cue that it has something to tell them) unless there is a clear indication that not doing so might be risky and potentially harmful (e.g. as it was the case for ProberBot which prompted participants to slow down and think before proceeding with a major and potentially risky trade). Finally, there should also be some flexibility in how the user is expected to respond to or act upon the prompt – as it was the case in SelVReflect, where some participants only expressed themselves visually while others also did so verbally when they were responding to a question, depending on their preference. In short, the user should generally be given some control over *if* and *how* a prompt is delivered to them unless it can be assumed that there are clear (objective) benefits of prompting them directly without being particularly configured to do so (this might be the case, for example, when the user forgets or overlooks something that is important for what they are doing or trying to achieve, or when there is an unexpected emergency, etc.).

(5) Phrasing: To support human cognition, CCs should inspire, encourage, and 'probe' human thinking.

To achieve this, questions that scaffold human thinking generally seem to work well [144, 217, 341]. Scaffolding questions aim to help someone progress with their thinking but without pointing out the/a solution (e.g. to a problem or task that is being performed). This can be achieved, for example, by enabling someone to look at an issue from a different angle. Questions can also enable cognitive engagement with a task as well as sensemaking and learning [40, 412, 467]. One consideration is also whether questions should be open or closed, which might depend on what the CC aims to achieve and if the task is to be more divergent or more convergent (e.g. in VoiceViz, the goal was to achieve an adequate balance between both).

Furthermore, formulating the prompts as questions can also have the advantage – even if they may implicitly contain a suggestion – that they might generally be perceived as less patronising or to be "*the annoying kid in the class that screams the answer*" (as also seen in Chapter 7). Related to this is the question of the extent to which the content of the prompt should be 'embellished' in a kind, encouraging, or motivating form/phrasing. For example, in some cases, the process of reflection and expression can be particularly challenging for people. In such cases, the prompts should be designed to be encouraging and motivating – as it was the case in SelVReflect. In other contexts, some friendly 'embellishments' might help to make the intervention less intrusive (for example, when a CC intervenes in an ongoing conversation to correct what a person said, as seen in Chapter 7 – e.g. "May I suggest something different ... ?"). However, there can, of course, also be cases where such embellishments might not be required or appropriate, for example, when important information needs to be delivered as quickly and efficiently as possible.

(6) Modality: Choose the modality that best matches the given task/interface and setting (and human-human interactions).

When deciding whether to use a voice or text-based CC, it is important to consider how it fits into the task. This includes (i) how much immediate attention its interventions should receive from users, (ii) how human users should attend to, engage with, and 'process' a prompt, and (iii) how well they can do so considering the modalities that the specific task involves (e.g. writing, drawing, or speaking). Regarding the first two points, using voice prompts can get people to more immediately pay attention to them, as was also the case in VoiceViz, where voice-based prompts were considered and responded to more quickly than screen-based ones. In VoiceViz the voice-based prompts also had the advantage that they 'activated' and 'accelerated' the ongoing conversation between the participants. On the other hand, text-based prompts are effective for scenarios where it is not desirable to interrupt people's train of thought and to allow them to decide when to engage with the prompt. Textual prompts also allow participants to re-read specific parts when needed. Furthermore, certain types of information generally tend to be more suitable to be provided/communicated in textual or visual form (such as numeric data, in particular if it is in structured/tabular form, etc.). Then there are tasks where being prompted through voice just appears to be more natural and meaningful, as was the case for SelVReflect – here, participants expressed a preference for auditory guidance as they thought it would better complement the creative visual task in VR. However, when choosing modalities it is also important to consider that certain people might generally find it easier to process spoken than written and/or visual information and vice-versa depending on their preferences and the way they process information (e.g. [345]). Therefore, some of the above points also have to be considered with respect to the needs of specific users/user groups for whom a CC should be designed.

If a CC is voice-based, an important consideration is also the type(s) of voice used (e.g. male, female, natural versus synthetic). Not only are there *individual* preferences for the type of voice, as seen in the SelVReflect study, but it also seems that there are differences *depending on the kind of task*. For example, some participants in SelVReflect pointed out that for the given creative and personal task, they prefer a more synthetic voice (while they might prefer a more natural voice for others). These individual needs and preferences regarding the type of voice might generally be best accounted for by giving users the option to choose the type of voice they would like to have for a specific CC.

(7) Evolution: Design the prompts in a way that they evolve with the human's knowledge, needs, and progress on a task.

The findings of the studies showed that it is important to consider how the prompts might evolve over time – based on a person's progress on a specific task, their developing understanding and knowledge, and related to that, their changing needs and preferences for the prompts they might require/desire. As the findings of the VoiceViz and ProberBot studies showed, participants generally pointed out that whether the prompts are useful will strongly depend on the person's expertise and familiarity with the specific task and their specific domain knowledge (which may also develop over time). For VoiceViz, in particular, it was mentioned that while prompts could be designed to support *'slower'* exploration and thinking *in the beginning*, they could be designed to support *faster* interactions/progress on the given task with *repeated use*. With *faster* they meant the NLUI could sometimes also provide them with specific insights rather than just asking them questions which they need to explore and answer themselves, for example. Similarly, one of the main outcomes of the user-centred design process of the SelVReflect study was that the prompts and the types of support they provide should adapt as the user progresses through the task. SelVReflect thus provided prompts that evolved from initial 'hands-on' guidance helping to put a basic structure in place, to more exploratory prompts, giving people ideas for things they could introduce to and consider in their creation, and ending with higher-level reflection prompts.

In fact, when considering how a CC's prompts might evolve over time, as part of a specific task and over longer periods of use, all the aforementioned 'parameters' should be considered. Or, to put it the other way round, **when addressing the previous seven design parameters, it is important to also consider for each of them how it might need to be adjusted over the course of using a CC**. For example, it is likely that over time, the CC's input might be needed for many tasks at different moments (i.e. parameter 1), potentially less frequently (i.e. parameter 2), and using a phrasing that evolves (i.e. parameter 5, such as by being more direct and ‘getting straight to the point’ when users already have some familiarity with the task and the CC).

To conclude, the design ‘principles’ and ‘parameters’ presented in this section provide some key design considerations for cognitive tools like CCs, which aim to proactively support cognitive processes. When designing such tools, however, the design decisions will mainly depend on the specific task/activity, the context in which it is performed, and the abilities and needs of the people who are performing it. Therefore, the decisions will generally have to mainly be informed by thorough user-centred design processes. It is hoped, however, that the considerations presented in this section might guide this design process as well as the user research and (prototype) evaluations that might be performed as part of it.

8.3 The Key Contributions to Knowledge

The present research contributes to the evolving body of research on how AI can be designed to augment and empower humans in their abilities [5, 8, 85, 138, 207, 365, 377, 386, 461]. In particular, it shows how AI systems that are designed as a CC[28] could augment human cognition in a range of tasks – covering analytical, decision-making, and creative ones. The idea of CCs that proactively scaffold cognition – as they are proposed here – can be particularly relevant for GenAI tools, where such scaffolding could help people figure out how they could use the tool to achieve their goals and which strategies could be most effective – similar to the scaffolding that Tankelevitch et al. [409] propose in their paper on the metacognitive opportunities and demands of GenAI. Thus, the idea of CCs could be seen as a way to conceptualise AI, which focuses less on *offloading* human cognition – by giving specific outputs – and instead more on providing *scaffolds* for cognition.

Related to that, as it was proposed in Section 8.1.3, probing and scaffolding NLUIs can extend existing theories of external cognition [224, 362] and, more broadly, the ways in which technology can be understood to extend the human mind [87] as they can be framed and designed as facilitators for reflection and external cognition. The main reason for this is that

[28] Although the CC prototypes in this PhD thesis did technically not involve any AI models, they would most likely do so if they were implemented for real-world uses.

NLUIs can prompt people through questions, enabling them to express their thoughts in different ways than other interfaces (e.g. 'regular' GUIs) are capable of. More specifically, NLUIs can provide conversational interactions that can elicit certain 'social' responses [155, 271, 300, 337] leading to people expressing themselves better [431, 459, 460], also because they might feel more compelled to answer a question from an NLUI. It is likely that in the present studies this also enabled participants to externalise their thoughts in more effective ways, as they generally reported that the CC's questions helped them express new aspects – such as speculations for patterns in the dataset (VoiceViz), rationales for their decisions (ProberBot), or new aspects and components of a past experience (SelVReflect).

The present work further contributes to the body of research on how (conversational) AI tools and – more broadly – NLUIs can be designed to intervene proactively in meaningful ways. Similar to previous work, the findings show how proactive interventions that are relevant to an ongoing task can support it effectively and are also perceived positively [280, 285, 318], even in collaborative interactions among humans [11]. While resonating with these previous studies on proactive interventions of NLUIs, the novel contribution here relates to the evidence on how such proactive interventions can facilitate reflection in a range of open-ended tasks. The present work thus also extends the research that has been done on NLUIs supporting reflection in learning contexts [71, 203, 448] by also investigating their use for open-ended sensemaking tasks that go beyond common learning activities.

The studies further corroborate previous findings on the importance of determining opportune moments for the NLUI to intervene to avoid undesirable disruptions [72, 139, 306, 441]. However, the findings also revealed that when designing the CCs to support specific open-ended tasks, relatively simple rules to trigger the interventions can be sufficient, as people are generally more receptive to prompts and to exploring different directions and approaches given the nature of such tasks. However, the research also showed that this still requires thorough design processes in which various considerations have to be taken into account in order to get the NLUI's interventions right (as described in the previous section).

Furthermore, this research also supports previous findings with regard to how the adequate framing of NLUIs can play an important role in how they are understood, used, and interacted with, and how this can lead to more effective use and a better experience for users [214, 458].

In all the studies, the NLUIs were introduced to participants as only being there to provide them with possible aspects to consider in the task they were doing. Through that, it is assumed that participants understood what to expect from the NLUI, which might have 'narrowed' the 'gulf of expectation and experience' [263].

More broadly, the research also showed how CCs can facilitate and scaffold different forms of reflective thinking across a range of tasks, covering both introspective reflection (also referred to as self-reflection) and reflection on external materials (also referred to as critical thinking [126], as introduced in Section 2.2.1). In the present studies, the reflection that the CCs facilitated was found to lead to new insights and improved understanding of what participants were working on – which is also in line with some of the general benefits of reflection highlighted by Kolb [232] or Moon [291]. The research thus also extends research within HCI on NLUIs and how they can be designed to support reflection (e.g. [43]). However, what distinguishes the present PhD research is its focus on designing NLUIs that provide proactive prompts that facilitate reflective thinking in different types of tasks than those considered by most existing research – covering data analysis, decision-making, and creative expression.

The last chapter also revealed how it can be more challenging to provide proactive interventions in desirable ways when they are not constrained and tailored to a specific task but provided in everyday situations, which is in line with previous research [264, 286]. The novelty of the present research lies in the exploration of a range of everyday situations, in particular social ones, which have not yet been more closely considered. As the findings show, there seems to be a wide range of views on the desirability of such NLUIs, suggesting that individual preferences must be accounted for by future CCs that aim to weave themselves into people's everyday lives.

Taken together, this research contributes to and extends existing knowledge in various ways, which are hoped to inform future research and design of (conversational) AI, NLUIs, and other tools that augment human cognition.

8.4 Limitations

Given this breadth in the researched prototypes and the tasks that they support, a wide range of methods were used throughout the thesis to explore how best to design NLUIs and how effective they were at supporting various forms of reflection. These included experiments, (scenario-based) interview studies, questionnaires, user-centred design, as well as 'interventional' designs with pre and post quantitative and qualitative assessments. The prototypes themselves covered purely scenario-based ones, to Wizard-of-Oz, as well as partly to mostly implemented ones. Taken together, the PhD has been following a *mixed methods* approach 'within' and 'across' studies that made it possible to approach and evaluate the idea of CCs from different angles and through different lenses.

However, the methods used in the studies have their limitations. First of all, given the breadth of methods used, most of the research remained exploratory, and the causal effects that such CCs might have on specific tasks have not been addressed. This would require further experiments with 'no co-pilot' baselines or control conditions to compare to. Furthermore, it would involve identifying adequate metrics that can measure the effects CCs have on cognition, which is a challenge for open-ended tasks like those explored in this thesis, as task 'performance' cannot be as easily defined for them. However, more importantly, the main reason this thesis did not employ a strictly experimental approach was that there seemed to be many exploratory questions that appeared to be important to consider before doing so – specifically on what the opportunities are of using CCs and how they could be designed for these different uses.

Another restriction of the studies is their limited generalisability. One reason is that most of the studies were conducted using a controlled set-up – in the case of VoiceViz and SelVReflect in a lab setting and in the case of ProberBot in a video call with screen sharing (during which participants were making investment decisions based on certain scenarios using the simulated stock trading platform). This had to be done because most prototypes were not functional to a level where they could be used in real-world settings without supervision – i.e. 'in the wild' [73]. This also meant that some of the tasks had to be performed within certain bounds of what the study prototypes were capable of – for example, VoiceViz could only provide a certain set of visualisations for the given dataset, and ProberBot could only be used in certain situations.

Related to that, it is important to point out that many participants might have generally had a rather positive attitude towards the proactive interventions within these constrained tasks and experiences – which also meant that many of the proactive interventions could be relatively well controlled/delivered. It is likely that participants' perceptions would be noticeably different – and often less positive – if they had to use the same or comparable tools to those investigated in the present studies in real-life settings. Some of these limitations were, however, counteracted/mitigated by choosing tasks that have some degree of realism and where the NLUI addresses actual challenges people may have experienced with analogous real-world tasks themselves (e.g. not knowing what to look for in a graph, making a rash decision, finding it difficult to express something) which supported the tasks' ecological validity.

Related to ecological validity, it is also important to note that the studies did not cover the use of any of these NLUIs over longer periods, which means, on the one hand, that many of the studies may have been affected by the novelty effect, and on the other hand, that some of the prototypes might have only been able to reveal their real effects if they were used for some time (e.g. the effect of a ProberBot on someone's decisions). To thoroughly understand how people would use such CCs in comparable types of sensemaking or other tasks, longitudinal studies would be required. Although a longitudinal design was planned for a second study on ProberBot, various technical challenges in developing ProberBot further to allow its longitudinal use kept delaying the study so that it was ultimately not part of this PhD thesis. If longitudinal studies were conducted, they might have either revealed that some of the interest and willingness to interact with them 'wears off' over time, or they could also find that there might be an uptake as people get used to and more comfortable with the proactive behaviours of the NLUI, as mentioned in the ProberBot study for example (Section 5.6.2).

Another limitation is that some of the study samples were rather small. However, they were generally chosen in function of the research questions and method used so that they would at least be large enough to address the research questions in an adequate way (e.g. smaller for purely qualitative inquiries, larger for purely quantitative ones, and somewhere between for the mixed method ones [66]). Furthermore, it is important to mention that despite attempts to make the samples more diverse, they mostly consisted of young, well-educated people from

Western countries. There might be differences depending on cultural background and demographics in terms of how NLUIs are responded to and used. For example, different cultures and societies may have different views on the ways in which it would be acceptable for a CC to proactively intervene in an ongoing conversation. Overall, the research has revealed that there are many opportunities for CCs for a range of tasks, which are worth exploring further. However, there are also still a range of challenges and open questions to be addressed – which will have to be considered before deploying such tools more widely.

8.5 Future Research

Beyond addressing some of the limitations covered in the previous section by extending sample sizes, running studies with longitudinal designs, and conducting them in real-world settings, there are many other opportunities for further research on CCs.

For example, there are various other tasks that might benefit from reflective thinking, but where doing so might also be challenging without guidance. For example, CC could be developed to facilitate other introspective activities, where personal decisions, strategies, or goals are being reflected on and refined. Apart from that, future research could explore the use of CCs for other open-ended sensemaking tasks where information/data needs to be analysed and synthesised to get new insights (e.g. discovering relationships in different health-related activities which a wearable device might track), to make complex decisions (e.g. career choices), or to produce different types of artefacts (e.g. a report on a specific topic) to name just a few.

Scaffolding the creation of artefacts is particularly relevant with the current developments in GenAI, as mentioned earlier. For example, there seems to be potential to adapt some of the AI co-pilot tools (like Microsoft Copilot) to employ some of the scaffolding approaches of the CCs proposed in this thesis. Besides helping with creating certain artefacts like current AI co-pilots do, they could support the user in structuring, reflecting on, and making sense of the task and its components by scaffolding the person's thought process.

Another line of research could explore different kinds of questions and how they are delivered, for example, regarding their modality, length, the extent to which they are formulated to be probing and encouraging, and how this should be adjusted to the type of task. As VoiceViz showed, the way the questions are delivered (i.e. modality) can affect the way in which people engage in a task and interact with each other and the system. There could be further differences like these that might be relevant for how well certain (cognitive) activities/tasks can be performed by users.

Finally, one line of research or rather a *lens* to approach some of the ideas suggested in the previous paragraphs would be to investigate further how CC can enable people to effectively build external representations for different types of cognitive tasks. Through that, a better understanding could be developed of the types of prompts that are most effective in getting people to create and use external representations for their thought processes. For example, which prompts help a person formulate a vague intuition, structure a 'nascent thought', or sketch out a rough mental image or model of something? And which prompts then enable people to structure and refine such internal representations by externalising them in certain ways? Some of these questions have been partially addressed in the studies in this thesis through the different design processes used to identify which types of support people might need for certain tasks. However, there is an opportunity for a more systematic inquiry into how certain prompts might enable and support different forms of external cognition.

8.6 Key Ethical Considerations

Despite the opportunities of CCs, there are various important ethical considerations involved in developing them and when deploying them in the 'real world'. If these considerations are not given adequate attention when designing CCs, they could have negative effects for the people using them. Many of these ethical considerations overlap with those for AI and autonomous systems (e.g. [200]). Although CCs may, in some cases, not even need AI capabilities to support certain tasks, they have similar characteristics to AI/autonomous systems in the sense that they have a certain agency, and their proactive interventions can impact human behaviour (e.g. affect their decision-making). For example, it is important to make *transparent* what a CC is designed for, when and why it intervenes, and in which ways it intends to support users. Furthermore, it is important to keep in mind that a CC may have its *biases* with regard to how it was designed (for example, to support people in adopting certain strategies to solve a task), which is also relevant if the CC includes AI models that might have been trained with biased data [308]. This could affect the user in undesirable ways and introduce (the AI's) bias to their behaviour. Beyond some of these more general concerns, this section puts the focus on *autonomy, agency, alignment,* and *privacy*, as these are of particular relevance for CCs.

Autonomy. Although one aim of CCs is to get people to be more 'cognitively engaged' in a task – by triggering reflective thinking and encouraging externalisation of their thinking – there is also a certain risk of over-reliance and dependence on a CC, which can inhibit people's autonomy. For example, people may get used to the CC assisting them when they are not progressing and might, as a result, just 'wait for' the next question from the CC to 'get them to think'. Furthermore, once people are used to having a CC for certain tasks, the question is how they can still perform them without a CC. To mitigate this risk, the CCs should, on the one hand, generally be designed so that it limits its proactive interventions only to specific moments or parts of a task. On the other hand, this issue should be mitigated by the design of the prompts – as such, only some of the questions should hint at a possible direction to pursue when progressing further with the task, most should ideally *probe* a person on their ongoing thought process – such as encouraging externalisation rather than steering their thinking in a very specific direction. This way, people may not develop the expectation that a CC will generally give them a possible 'next step' for a task, and the risk of over-reliance could be limited to some extent. However, to achieve this, the tasks to be supported need to be carefully examined first (for example, using similar user-centred design approaches as some of the studies used) to get an understanding of where the critical points are when people might benefit from scaffolding questions.

Agency. Even if a goal of CCs is to empower people in their own thinking by giving them certain 'hooks' or triggers to enable them to reflect and make sense of something, they could also be experienced as inhibiting people's agency in certain situations if they are not carefully designed. On the one hand, their scaffolding questions can be perceived as patronising, pushy, or too prescriptive if they are not sufficiently tested and refined (as seen in the design process of the ProberBot questions – see Section 5.4.2). Furthermore, another aspect related to disempowerment is that CCs will sometimes not be able to provide input that aligns with a person's ongoing thought process. Even if for the present tasks, this was generally not an issue due to their open-endedness and people being in an 'exploratory mindset', there can be situations where a person is keen to focus on a specific idea or to follow their own approach when performing a task. As mentioned by a participant in VoiceViz, often, someone may just want to quickly get a specific insight or result from the data (i.e. a quick confirmatory analysis) rather than doing a more extensive exploratory analysis of the dataset. Empowering people in what they are doing and want to achieve may thus often require a CC to give people certain controls, but it also needs to be 'aware of' what the goals of someone are – in general, but also for a specific task/interaction – which brings us to the next ethical challenge of *alignment* an issue that is also widely discussed in general in the context of AI [149].

Alignment. The main idea of CCs is that they scaffold cognitive processes through questions. The goal is generally not to recommend specific actions or to provide answers or solutions. However, even scaffolding questions can be 'leading' and directing a person towards a certain direction. For example, a teacher scaffolding their student's sensemaking or problem-solving also intends to help them to proceed with their thinking 'towards a direction' where they might be able to find the/a possible solution [40, 393]. Hence, the scaffolding itself can generally not be fully objective as it is often based on certain scaffolding approaches, best practices, useful heuristics, etc., which may contain certain assumptions on how a task could (or should) be performed. In ProberBot, for example, the scaffolding was mostly geared towards so-called value-based long-term investing strategies (which also overlapped with the investing strategies of the participants who were recruited for the study, but it might not overlap with the strategies of other people). As a result, it can happen that a CC affects people's behaviours in a way that may not be aligned with their goals. It is thus crucial to make it clear what the purpose of a CC's scaffolding questions is (before people start using the system) and, whenever possible, allow people to specify their goals so that the CC can adjust the way in which it scaffolds a person's thinking.

Privacy. CCs will generally need to have some form of 'awareness' of a specific context in which they operate and in which they intervene – which means that they need to sense what the people who it intends to support are doing. Furthermore, as discussed in the previous paragraph, the CC needs to have some representation of what the people's intentions or goals are. Both knowing people's goals and sensing specific contexts has important privacy implications. When designing CCs it is important to balance the amount and the sensitivity of the data to be collected versus the expected positive impact of the CC on people's (task-specific) behaviour (which is related to the 'data minimisation' principle – also outlined in Article 5(1)(c) in the GDPR). If there is not an adequate balance, the deployment of a CC should be reconsidered or, wherever possible, the data that it collects constrained. Furthermore, the machine learning models underlying the NLUIs that are required for processing the persons' behavioural data (e.g. what they do or say) should ideally run on the user's device (also referred to as 'edge computing') [190] so that the data does not need to be uploaded to the cloud. Although these considerations were not a challenge for the present studies, as the CCs were not deployed in real-world settings and did not need such sensing and computational capabilities, it can be seen from the studies what privacy implications there might be if the systems were deployed. For SelVReflect and ProberBot, the sensing could be largely constrained to people's behaviours within a specific software application; however, if designed to intervene in everyday situations (like in the scenario-based study), the NLUI would require collecting significantly more data, which would generally also be highly sensitive (e.g. private conversations). Tools like VoiceViz would fall somewhere in the middle of that spectrum, as the conversations that are monitored are specific to a given task (and thus most likely less sensitive than everyday conversations). When designing CCs, it is thus crucial to ask what data is really needed to provide meaningful cognitive scaffolds. If the scaffolds are designed in a way that they can work for many people (e.g. based on a user-centred design process), the CC may not even need to know exactly what people might be doing or thinking at a given moment and what their goals are (as it was also the case in the present studies).

Moving forward, many of these ethical considerations will also benefit from a continuing critical discourse within HCI and other fields on the roles that AI and tools like CCs should play in different areas of our lives – now and in the future.

9. Conclusion

This thesis explored how NLUIs can support people in cognitive tasks by proactively asking them questions. The questions were designed to be scaffolding, probing, and guiding with the aim of enabling reflective thinking. The focus was on open-ended tasks that benefit from reflective thinking, covering a collaborative exploratory data analysis task, a complex decision-making task, and a creative 3D drawing task. The findings revealed that the questions enabled people to explore and discover new ideas, approaches, or perspectives in ways that helped them progress with these tasks. In particular, by considering and responding to the NLUI's questions while performing these tasks, people were able to externalise their thoughts and ideas in ways that led to new understandings and insights. These findings show how NLUIs can extend our cognition in new ways, which is why they were given the framing/metaphor of being 'cognitive co-pilots'. However, there are also various challenges involved in such cognitive co-pilots. For example, as they can often only roughly 'guess' what a person might currently be doing or thinking about, their questions may sometimes not be adequately timed or relevant for what a person wants to do or achieve at a given moment in an ongoing task. As demonstrated in this thesis, one approach to mitigate this risk is to employ user-centred design approaches to define different types of scaffolding questions that can work for a range of people and for different ways of performing the given task. However, as misplaced questions cannot completely be avoided, it might be best to design the cognitive co-pilot with limited proactivity so that it only intervenes occasionally. The reason is that often a proactive question might trigger new thoughts, but sometimes it may also disrupt them – it is thus best to use them judiciously. The same applies to the question-asking NLUIs more generally – while there are many tasks where people can benefit from being prompted, there are others where this can be too disruptive and add too much 'friction'. However, when designed in sensible ways, proactive NLUIs hold promise to become 'co-pilots' to human cognition.

10. Bibliography

[1] Rochelle Ackerley, Jean-Marc Aimonetti, and Edith Ribot-Ciscar. 2017. Emotions Alter Muscle Proprioceptive Coding of Movements in Humans. *Scientific Reports* 7, 1 (2017), 1–9.

[2] Martin Adam, Michael Wessel, and Alexander Benlian. 2021. AI-Based Chatbots in Customer Service and Their Effects on User Compliance. *Electronic Markets* 31, 2 (June 2021), 427–445. https://doi.org/10.1007/s12525-020-00414-7

[3] Elena Agapie, Patricia A. Areán, Gary Hsieh, and Sean A. Munson. 2022. A Longitudinal Goal Setting Model for Addressing Complex Personal Problems in Mental Health. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (November 2022). https://doi.org/10.1145/3555160

[4] Herman Aguinis and Kyle J. Bradley. 2014. Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organizational Research Methods* 17, 4 (October 2014), 351–371. https://doi.org/10.1177/1094428114547952

[5] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* 53, 8 (August 2020), 18–28. https://doi.org/10.1109/MC.2020.2996587

[6] Mehdi Alaimi, Edith Law, Kevin Daniel Pantasdo, Pierre-Yves Oudeyer, and Hélène Sauzeon. 2020. Pedagogical Agents for Fostering Question-Asking Skills in Children. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (*CHI '20*), April 21, 2020. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376776

[7] James Allen, Curry Guinn, and Eric Horvitz. 1999. Mixed-Initiative Interaction. *IEEE Intelligent Systems and Their Applications* 14, 5 (September 1999), 14–23. https://doi.org/10.1109/5254.796083

[8] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*), 2019. ACM, New York, NY, USA, 3:1-3:13. https://doi.org/10.1145/3290605.3300233

[9] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction* 26, 3 (April 2019), 1–28. https://doi.org/10.1145/3311956

[10] Masayuki Ando, Kouyou Otsu, and Tomoko Izumi. 2023. The Impact of Parent-Like Chatbot Narratives on Daily Reflection. In *Human-Computer Interaction*, 2023. Springer Nature Switzerland, Cham, 279–293. https://doi.org/10.1007/978-3-031-35602-5_21

[11] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. Investigating Proactive Search Support in Conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference* (*DIS '18*), June 08, 2018. Association for Computing Machinery, New York, NY, USA, 1295–1307. https://doi.org/10.1145/3196709.3196734

[12] Kika Arias, Sooyeon Jeong, Hae Won Park, and Cynthia Breazeal. 2020. Toward Designing User-Centered Idle Behaviors for Social Robots in the Home. In *Proceedings of the 1st International Workshop on Designerly HRI Knowledge. Held in Conjunction with the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2020)*, September 2020. Naples, Italy. Retrieved from https://www.media.mit.edu/publications/toward-designing-user-centered-idle-behaviors-for-social-robots-in-the-home/

[13] Richard Arias-Hernandez, Linda T. Kaastra, Tera M. Green, and Brian Fisher. 2011. Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics. In *Proceedings of the 2011 44th Hawaii International Conference on System Sciences* (*HICSS '11*), January 04, 2011. IEEE Computer Society, USA, 1–10. https://doi.org/10.1109/HICSS.2011.339

[14] Vicky Arnold, Philip A. Collier, Stewart A. Leech, and Steve G. Sutton. 2000. The Effect of Experience and Complexity on Order and Recency Bias in Decision Making by Professional Accountants. *Accounting & Finance* 40, 2 (2000), 109–134. https://doi.org/10.1111/1467-629X.00039

[15] Daisuke Asai, Jarrod Orszulak, Richard Myrick, Chaiwoo Lee, Joseph F Coughlin, and Olivier L De Weck. 2011. Context-Aware Reminder System to Support Medication Compliance. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, 2011. IEEE, New York, USA, 3213–3218.

[16] J. Maxwell Atkinson and John Heritage. 1999. Transcript Notation - Structures of Social Action: Studies in Conversation Analysis. *Aphasiology* 13, 4–5 (April 1999), 243–249. https://doi.org/10.1080/026870399402073

[17] James Auger. 2013. Speculative Design: Crafting the Speculation. *Digital Creativity* 24, 1 (March 2013), 11–35. https://doi.org/10.1080/14626268.2013.767276

[18] Jillian Aurisano, Abhinav Kumar, Alberto Gonzalez, Jason Leigh, Barbara DiEugenio, and Andrew Johnson. 2016. Articulate2: Toward a Conversational Interface for Visual Data Exploration. In *IEEE Visualization 2016*, July 19, 2016. Baltimore, Maryland, USA.

[19] Vino Avanesi, Johanna Rockstroh, Thomas Mildner, Nima Zargham, Leon Reicherts, Maximilian A. Friehs, Dimosthenis Kontogiorgos, Nina Wenig, and Rainer Malaka. 2023. From C-3PO to HAL: Opening The Discourse About The Dark Side of Multi-Modal Social Agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (*CUI '23*), July 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3571884.3597441

[20] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. SearchBots: User Engagement with ChatBots during Collaborative Search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (*CHIIR '18*), March 01, 2018. Association for Computing Machinery, New York, NY, USA, 52–61. https://doi.org/10.1145/3176349.3176380

[21] Petter Bae Bae Brandtzæg, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. 2021. When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 06, 2021. ACM, Yokohama Japan, 1–13. https://doi.org/10.1145/3411764.3445318

[22] Marcos Baez, Florian Daniel, and Fabio Casati. 2020. Conversational Web Interaction: Proposal of a Dialog-Based Natural Language Interaction Paradigm for the Web. In *Chatbot Research and Design* (*Lecture Notes in Computer Science*), 2020. Springer International Publishing, Cham, 94–110. https://doi.org/10.1007/978-3-030-39540-7_7

[23] Marcos Baez, Florian Daniel, Fabio Casati, and Boualem Benatallah. 2021. Chatbot Integration in Few Patterns. *IEEE Internet Computing* 25, 3 (May 2021), 52–59. https://doi.org/10.1109/MIC.2020.3024605

[24] Bence Bago, David G. Rand, and Gordon Pennycook. 2020. Fake News, Fast and Slow: Deliberation Reduces Belief in False (but Not True) News Headlines. *Journal of Experimental Psychology. General* 149, 8 (August 2020), 1608–1613. https://doi.org/10.1037/xge0000729

[25] Sojung Bahng, Ryan M. Kelly, and Jon McCormack. 2020. Reflexive VR Storytelling Design Beyond Immersion: Facilitating Self-Reflection on Death and Loneliness. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (*CHI '20*), 2020. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376582

[26] Lisanne Bainbridge. 1983. Ironies of Automation. *Automatica* 19, 6 (November 1983), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

[27] H. Kent Baker and John R. Nofsinger. 2002. Psychological Biases of Investors. *Financial Services Review* 11, 2 (Summer 2002), 97–116.

[28] H. Kent Baker and Victor Ricciardi. 2014. How Biases Affect Investor Behaviour. *The European Financial Review* (2014), 7–10.

[29] Albert Bandura. 1977. Self-Efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review* 84, (1977), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

[30] Albert Bandura. 1997. *Self-Efficacy: The Exercise of Control*. W. H. Freeman, New York, NY, US.

[31] Albert Bandura. 2008. An Agentic Perspective on Positive Psychology. In *Positive Psychology: Exploring the Best in People, Vol 1: Discovering Human Strengths*. Praeger Publishers/Greenwood Publishing Group, Westport, CT, US, 167–196.

[32] Brad M. Barber and Terrance Odean. 2008. All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *The Review of Financial Studies* 21, 2 (April 2008), 785–818. https://doi.org/10.1093/rfs/hhm079

[33] Brad M. Barber and Terrance Odean. 2013. The Behavior of Individual Investors. In *Handbook of the Economics of Finance*. Elsevier, 1533–1570. https://doi.org/10.1016/B978-0-44-459406-8.00022-6

[34] Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and Gustavo Federizzi. 2020. Intelligent Personal Assistants: A Systematic Literature Review. *Expert Systems with Applications* 147, (June 2020), 113193. https://doi.org/10.1016/j.eswa.2020.113193

[35] Julie H. Barlow, Bethan Williams, and Chris Wright. 1996. The Generalized Self-Efficacy Scale in People with Arthritis. *Arthritis & Rheumatism* 9, 3 (1996), 189–196. https://doi.org/10.1002/1529-0131(199606)9:3<189::AID-ANR1790090307>3.0.CO;2-#

[36] Eric P.S. Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing Reflection: On the Use of Reflection in Interactive System Design. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (*DIS '14*), June 21, 2014. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2598510.2598598

[37] Clare Beatty, Tanya Malik, Saha Meheli, and Chaitali Sinha. 2022. Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. *Frontiers in Digital Health* 4, (2022). Retrieved January 5, 2024 from https://www.frontiersin.org/articles/10.3389/fdgth.2022.847991

[38] Grace M. Begany, Ning Sa, and Xiaojun Yuan. 2016. Factors Affecting User Perception of a Spoken Language vs. Textual Search Interface: A Content Analysis. *Interacting with Computers* 28, 2 (March 2016), 170–180. https://doi.org/10.1093/iwc/iwv029

[39] Diana Beirl, Nicola Yuill, and Yvonne Rogers. 2019. Using Voice Assistant Skills in Family Life. In *Proceedings of the 13th International Conference on Computer Supported Collaborative Learning*, June 2019. International Society of the Learning Sciences (ISLS), 96–103.

[40] Amanda Benedict-Chambers, Sylvie M. Kademian, Elizabeth A. Davis, and Annemarie Sullivan Palincsar. 2017. Guiding Students towards Sensemaking: Teacher Questions Focused on Integrating Scientific Practices with Science Content. *International Journal of Science Education* 39, 15 (2017), 1977–2001. https://doi.org/10.1080/09500693.2017.1366674

[41] Marit Bentvelzen, Jasmin Niess, Mikołaj P. Woźniak, and Paweł W. Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 06, 2021. ACM, Yokohama Japan, 1–8. https://doi.org/10.1145/3411764.3445673

[42] Marit Bentvelzen, Jasmin Niess, and Paweł W. Woźniak. 2021. The Technology-Mediated Reflection Model: Barriers and Assistance in Data-Driven Reflection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 06, 2021. ACM, Yokohama Japan, 1–12. https://doi.org/10.1145/3411764.3445505

[43] Marit Bentvelzen, Paweł W. Woźniak, Pia S.F. Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting Reflection in HCI: Four Design Resources for Technologies That Support Reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (March 2022), 1–27. https://doi.org/10.1145/3517233

[44] Daniel Ellis Berlyne, William Edgar Vinacke, and Robert Sternberg. 2023. Types of Thinking. *Encyclopedia Britannica*. Retrieved November 9, 2023 from https://www.britannica.com/topic/thought/Types-of-thinking

[45] Tao Bi, Raffaele Andrea Buono, Temitayo Olugbade, Aneesha Singh, Catherine Holloway, Enrico Costanza, Amanda C de C Williams, Nicolas E. Gold, and Nadia Berthouze. 2021. Towards Chatbot-Supported Self-Reporting for Increased Reliability and Richness of Ground Truth for Automatic Pain Recognition: Reflections on Long-Distance Runners and People with Chronic Pain. In *Companion Publication of the 2021 International Conference on Multimodal Interaction* (*ICMI '21 Companion*), Dezember 2021. Association for Computing Machinery, New York, NY, USA, 43–53. https://doi.org/10.1145/3461615.3485670

[46] Ludovic Le Bigot, Patrice Terrier, Eric Jamet, Valérie Botherel, and Jean-François Rouet. 2010. Does Textual Feedback Hinder Spoken Interaction in Natural Language? *Ergonomics* 53, 1 (January 2010), 43–55. https://doi.org/10.1080/00140130903306666

[47] Corey J. Bohil, Bradly Alicea, and Frank A. Biocca. 2011. Virtual Reality in Neuroscience Research and Therapy. *Nature Reviews Neuroscience* 12, 12 (December 2011), 752–762. https://doi.org/10.1038/nrn3122

[48] Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. 2021. An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (*CUI '21*), 2021. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3469595.3469611

[49] Antoine Bordes, Y.-Lan Boureau, and Jason Weston. 2017. Learning End-to-End Goal-Oriented Dialog. In *Proceedings of the International Conference on Learning Representations*, February 06, 2017. Toulon, France. Retrieved November 20, 2023 from https://openreview.net/forum?id=S1Bb3D5gg

[50] Cristina Botella, Giuseppe Riva, Andrea Gaggioli, Brenda Wiederhold, Mariano Alcañiz Raya, and Rosa Baños. 2011. The Present and Future of Positive Technologies. *Cyberpsychology, Behavior, and Social Networking* 15, (December 2011), 78–84. https://doi.org/10.1089/cyber.2011.0140

[51] David Boud, Rosemary Keogh, and David Walker (Eds.). 2013. *Reflection: Turning Experience into Learning*. Routledge, London. https://doi.org/10.4324/9781315059051

[52] Robert Bowman, Benjamin R. Cowan, Anja Thieme, and Gavin Doherty. 2022. Beyond Subservience: Using Joint Commitment to Enable Proactive CUIs for Mood Logging. In *Proceedings of the 4th Conference on Conversational User Interfaces* (*CUI '22*), July 26, 2022. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3543829.3544512

[53] Sheryl Brahnam and Antonella De Angeli. 2012. Gender Affordances of Conversational Agents. *Interacting with Computers* 24, 3 (2012), 139–153.

[54] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*), May 02, 2019. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300270

[55] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (January 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[56] Miriam Brintzinger, Wolfgang Tschacher, Katrin Endtner, Kurt Bachmann, Michael Reicherts, Hansjörg Znoj, and Mario Pfammatter. 2021. Patients' Style of Emotional Processing Moderates the Impact of Common Factors in Psychotherapy. *Psychotherapy* 58, 4 (2021), 472.

[57] Christian Brown and Rick Garner. 2017. Serious Gaming, Virtual, and Immersive Environments in Art Therapy. In *Digital Art Therapy: Material, Methods, and Applications*. Jessica Kingsley Publishers, London, 192–205.

[58] Duncan P. Brumby, Christian P. Janssen, and Gloria Mark. 2019. How Do Interruptions Affect Productivity? In *Rethinking Productivity in Software Engineering*, Caitlin Sadowski and Thomas Zimmermann (eds.). Apress, Berkeley, CA, 85–107. https://doi.org/10.1007/978-1-4842-4221-6_9

[59] Fred B. Bryant, Colette M. Smart, and Scott P. King. 2005. Using the Past to Enhance the Present: Boosting Happiness through Positive Reminiscence. *Journal of Happiness Studies* 6, 3 (2005), 227–260.

[60] Erik Brynjolfsson. 2022. The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. *Daedalus* 151, 2 (May 2022), 272–287. https://doi.org/10.1162/daed_a_01915

[61] Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. 2023. *Generative AI at Work*. National Bureau of Economic Research. https://doi.org/10.3386/w31161

[62] Vannevar Bush. 1945. As We May Think. *The Atlantic 176*. Retrieved August 25, 2023 from https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

[63] Martin Buss, Daniel Carton, Barbara Gonsior, Kolja Kuehnlenz, Christian Landsiedel, Nikos Mitsou, Roderick de Nijs, Jakub Zlotowski, Stefan Sosnowski, Ewald Strasser, Manfred Tscheligi, Astrid Weiss, and Dirk Wollherr. 2011. Towards Proactive Human-Robot Interaction in Human Environments. In *2011 2nd International Conference on Cognitive Infocommunications (CogInfoCom)*, July 2011. 1–6. Retrieved November 26, 2023 from https://ieeexplore.ieee.org/abstract/document/5999453/similar#similar

[64] Bill Buxton. 2007. Video Envisionment. In *Sketching User Experiences*, Bill Buxton (ed.). Morgan Kaufmann, Burlington, 349–370. https://doi.org/10.1016/B978-012374037-3/50079-1

[65] Victoria Cabales. 2019. Muse: Scaffolding Metacognitive Reflection in Design-Based Research. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI EA '19*), May 02, 2019. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3308450

[66] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (*CHI '16*), May 07, 2016. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[67] Alexia Cambon, Brent Hecht, Benjamin Edelman, Donald Ngwe, Sonia Jaffe, Amy Heger, Mihaela Vorvoreanu, Sida Peng, Jake Hofman, Alex Farach, Margarita Bermejo-Cano, Eric Knudsen, James Bono, Hardik Sanghavi, Sofia Spatharioti, David Rothschild, Daniel G. Goldstein, Eirini Kalliamvakou, Peter Cihon, Mert Demirer, Michael Schwarz, and Jaime Teevan. 2023. *Early LLM-Based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity*. Microsoft Research. Retrieved from https://www.microsoft.com/en-us/research/publication/early-llm-based-tools-for-enterprise-information-workers-likely-provide-meaningful-boosts-to-productivity/

[68] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (*CHI '20*), 2020. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376789

[69] Heloisa Candello, Claudio Pinhanez, David Millen, and Bruna Daniele Andrade. 2017. Shaping the Experience of a Cognitive Investment Adviser. In *Design, User Experience, and Usability: Understanding Users and Contexts* (*Lecture Notes in Computer Science*), 2017. Springer International Publishing, Cham, 594–613. https://doi.org/10.1007/978-3-319-58640-3_43

[70] John M. Carroll. 1999. Five Reasons for Scenario-Based Design. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 3 - Volume 3* (*HICSS '99*), January 05, 1999. IEEE Computer Society, USA, 3051.

[71] Jessy Ceha, Nalin Chhibber, Joslin Goh, Corina McDonald, Pierre-Yves Oudeyer, Dana Kulić, and Edith Law. 2019. Expression of Curiosity in Social Robots: Design, Perception, and Effects on Behaviour. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*), May 02, 2019. Association for Computing Machinery, Glasgow, Scotland UK, 1–12. https://doi.org/10.1145/3290605.3300636

[72] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello There! Is Now a Good Time to Talk? Opportune Moments for Proactive Interactions with Smart Speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (September 2020), 74:1-74:28. https://doi.org/10.1145/3411810

[73] Alan Chamberlain, Andy Crabtree, Tom Rodden, Matt Jones, and Yvonne Rogers. 2012. Research in the Wild: Understanding "in the Wild" Approaches to Design and Development. In *Proceedings of the Designing Interactive Systems Conference* (*DIS '12*), 2012. ACM, New York, NY, USA, 795–796. https://doi.org/10.1145/2317956.2318078

[74] Samantha W. T. Chan, Shardul Sapkota, Rebecca Mathews, Haimo Zhang, and Suranga Nanayakkara. 2020. Prompto: Investigating Receptivity to Prompts Based on Cognitive Load from Memory Training Conversational Agent. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (December 2020), 1–23. https://doi.org/10.1145/3432190

[75] Samantha W.T. Chan, Shardul Sapkota, Rebecca Mathews, Haimo Zhang, and Suranga Nanayakkara. 2022. Prompto: Cognition-Aware Prompts with Conversational Interfaces Using Physiological Signals. *GetMobile: Mobile Computing and Communications* 26, 2 (July 2022), 34–38. https://doi.org/10.1145/3551670.3551681

[76] Tanya L. Chartrand and John A. Bargh. 1999. The Chameleon Effect: The Perception–Behavior Link and Social Interaction. *Journal of Personality and Social Psychology* 76, 6 (1999), 893–910. https://doi.org/10.1037/0022-3514.76.6.893

[77] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (May 2021), 729–758. https://doi.org/10.1080/10447318.2020.1841438

[78] Michelene T. H. Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian Lavancher. 1994. Eliciting Self-Explanations Improves Understanding. *Cognitive Science* 18, 3 (July 1994), 439–477. https://doi.org/10.1016/0364-0213(94)90016-7

[79] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (October 2014), 219–243. https://doi.org/10.1080/00461520.2014.965823

[80] Gioia Chilton. 2013. Art Therapy and Flow: A Review of the Literature and Applications. *Art Therapy* 30, 2 (2013), 64–70. https://doi.org/10.1080/07421656.2013.787211

[81] Christine Chin. 2007. Teacher Questioning in Science Classrooms: Approaches That Stimulate Productive Thinking. *Journal of Research in Science Teaching* 44, 6 (2007), 815–843. https://doi.org/10.1002/tea.20171

[82] Christine Chin and Jonathan Osborne. 2008. Students' Questions: A Potential Resource for Teaching and Learning Science. *Studies in Science Education* 44, 1 (March 2008), 1–39. https://doi.org/10.1080/03057260701828101

[83] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. 2019. Once a Kind Friend Is Now a Thing: Understanding How Conversational Agents at Home Are Forgotten. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (*DIS '19*), June 18, 2019. Association for Computing Machinery, New York, NY, USA, 1557–1569. https://doi.org/10.1145/3322276.3322332

[84] Heeryung Choi, Jelena Jovanovic, Oleksandra Poquet, Christopher Brooks, Srećko Joksimović, and Joseph Jay Williams. 2023. The Benefit of Reflection Prompts for Encouraging Learning with Hints in an Online Programming Course. *The Internet and Higher Education* (March 2023), 100903. https://doi.org/10.1016/j.iheduc.2023.100903

[85] Elizabeth Churchill and Mikael Wiberg. 2024. From Humans to AI: A Timely Debate on Human-AI Relations. *Interactions* 31, 1 (January 2024), 5. https://doi.org/10.1145/3637223

[86] Michael C. Cipriano, Thomas S. Gruca, and Jennie Jiao. 2020. Can Investing Diaries Be Hazardous to Your Financial Health? *The Journal of Prediction Markets* 14, 1 (September 2020), 105–125. https://doi.org/10.5750/jpm.v14i1.1805

[87] Andy Clark and David Chalmers. 1998. The Extended Mind. *Analysis* 58, 1 (1998), 7–19.

[88] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and

Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (June 2019), 349–371. https://doi.org/10.1093/iwc/iwz016

[89] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*), 2019. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[90] Bud Colligan. 2011. *How the Knowledge Navigator Video Came About*. Dubberly Design Office. Retrieved November 15, 2023 from https://www.dubberly.com/wp-content/uploads/2011/11/knowledge_navigator_video.pdf

[91] Benjamin R. Cowan, Leigh Clark, Heloisa Candello, and Janice Tsai. 2023. Introduction to This Special Issue: Guiding the Conversation: New Theory and Design Perspectives for Conversational User Interfaces. *Human–Computer Interaction* 38, 3–4 (July 2023), 159–167. https://doi.org/10.1080/07370024.2022.2161905

[92] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can i Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, September 04, 2017. ACM, Vienna Austria, 1–12. https://doi.org/10.1145/3098279.3098539

[93] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (*CHI EA '16*), May 07, 2016. Association for Computing Machinery, New York, NY, USA, 1389–1397. https://doi.org/10.1145/2851581.2892410

[94] David R. Cross and Scott G. Paris. 1988. Developmental and Instructional Analyses of Children's Metacognition and Reading Comprehension. *Journal of Educational Psychology* 80, 2 (1988), 131–142. https://doi.org/10.1037/0022-0663.80.2.131

[95] Pietro Crovari, Fabio Catania, Pietro Pinoli, Philipp Roytburg, Asier Salzar, Franca Garzotto, and Stefano Ceri. 2020. OK, DNA! A Conversational Interface to Explore Genomic Data. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (*CUI '20*), July 22, 2020. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3405755.3406163

[96] Mihaly Csikszentmihalyi and Isabella SelegaEditors Csikszentmihalyi (Eds.). 1988. The Flow Experience and Its Significance for Human Psychology. In *Optimal*

Experience: Psychological Studies of Flow in Consciousness. Cambridge University Press, Cambridge, London, 15–35. https://doi.org/10.1017/CBO9780511621956.002

[97] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperAgent: A Customer Service Chatbot for E-Commerce Websites. In *Proceedings of ACL 2017, System Demonstrations*, July 2017. Association for Computational Linguistics, Vancouver, Canada, 97–102.

[98] Barry E. Cushing and Sunita S. Ahlawat. 1996. Mitigation of Recency Bias in Audit Judgment: The Effect of Documentation. *Auditing* 15, 2 (Fall 1996), 110.

[99] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz Studies: Why and How. In *Proceedings of the 1st International Conference on Intelligent User Interfaces* (*IUI '93*), February 01, 1993. Association for Computing Machinery, New York, NY, USA, 193–200. https://doi.org/10.1145/169891.169968

[100] Laurie Damianos, Dan Loehr, Carl Burke, Steve Hansen, and Michael Viszmeg. 2003. The MSIIA Experiment: Using Speech to Enhance Human Performance on a Cognitive Task. *International Journal of Speech Technology* 6, 2 (April 2003), 133–144. https://doi.org/10.1023/A:1022334530417

[101] Beth Davey and Susan McBride. 1986. Generating Self-Questions after Reading: A Comprehension Assist for Elementary Students. *The Journal of Educational Research* 80, 1 (September 1986), 43–46. https://doi.org/10.1080/00220671.1986.10885720

[102] Gerald Dawe, Rolf Jucker, and Stephen Martin. 2005. Sustainable Development in Higher Education: Current Practice and Future Developments. *A Report for The Higher Education Academy, York (UK)* (January 2005).

[103] Min-Yuh Day, Jian-Ting Lin, and Yuan-Chih Chen. 2018. Artificial Intelligence for Conversational Robo-Advisor. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 1057–1064.

[104] Fabrizio Dell'Acqua, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. Harvard Business School, Rochester, NY. Retrieved November 30, 2023 from https://papers.ssrn.com/abstract=4573321

[105] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A Survey on Proactive Dialogue Systems: Problems, Methods, and Prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, August 2023. International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China, 6583–6591. https://doi.org/10.24963/ijcai.2023/738

[106] Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and Evaluating Large Language Models for Proactive Dialogues:

Clarification, Target-Guided, and Non-Collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, December 2023. Association for Computational Linguistics, Singapore, 10602–10621. https://doi.org/10.18653/v1/2023.findings-emnlp.711

[107] Smit Desai and Jessie Chin. 2021. Hey Google, Can You Help Me Learn? In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, July 27, 2021. ACM, Bilbao (online) Spain, 1–4. https://doi.org/10.1145/3469595.3469601

[108] Smit Desai and Michael Twidale. 2023. Metaphors in Voice User Interfaces: A Slippery Fish. *ACM Transactions on Computer-Human Interaction* 30, 6 (September 2023), 89:1-89:37. https://doi.org/10.1145/3609326

[109] John Dewey. 1910. *How We Think*. D.C. Heath, Boston, MA, USA. https://doi.org/10.1037/10903-000

[110] John Dewey. 1933. *How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process*. D.C. Heath, Boston, MA, USA.

[111] Kedar Dhamdhere, Kevin S. McCurley, Ralfi Nahmias, Mukund Sundararajan, and Qiqi Yan. 2017. Analyza: Exploring Data with Conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces - IUI '17*, 2017. ACM Press, Limassol, Cyprus, 493–504. https://doi.org/10.1145/3025171.3025227

[112] Alan Dix. 2008. Externalisation–How Writing Changes Thinking. *Interfaces* 76, Autumn 2008 (2008), 18–19.

[113] Alan Dix and Layda Gongora. 2011. Externalisation and Design. In *Procedings of the Second Conference on Creativity and Innovation in Design* (*DESIRE '11*), October 19, 2011. Association for Computing Machinery, New York, NY, USA, 31–42. https://doi.org/10.1145/2079216.2079220

[114] Sidney K. D'Mello, Nia Dowell, and Arthur Graesser. 2011. Does It Really Matter Whether Students' Contributions Are Spoken versus Typed in an Intelligent Tutoring System with Natural Language? *Journal of Experimental Psychology: Applied* 17, 1 (2011), 1–17. https://doi.org/10.1037/a0022674

[115] Kohji Dohsaka, Ryota Asai, Ryuichiro Higashinaka, Yasuhiro Minami, and Eisaku Maeda. 2009. Effects of Conversational Agents on Human Communication in Thought-Evoking Multi-Party Dialogues. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (*SIGDIAL '09*), September 11, 2009. Association for Computational Linguistics, USA, 217–224.

[116] Peter W. Dowrick. 1999. A Review of Self Modeling and Related Interventions. *Applied and Preventive Psychology* 8, 1 (December 1999), 23–39. https://doi.org/10.1016/S0962-1849(99)80009-2

[117] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (*MobileHCI '19*), 2019. ACM, New York, NY, USA, 5:1-5:12. https://doi.org/10.1145/3338286.3340116

[118] Mateusz Dubiel, Kerstin Bongard-Blanchy, Luis A. Leiva, and Anastasia Sergeeva. 2023. Are You Sure You Want to Order That? On Appropriateness of Voice-Only Proactive Feedback Strategies. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (*CUI '23*), July 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3571884.3604312

[119] Sylvie Dubois, Martine Boutin, and David Sankoff. 1996. The Quantitative Analysis of Turntaking in Multiparticipant Conversations. *University of Pennsylvania Working Papers in Linguistics* 3, 1 (1996), 20.

[120] Maya Eden and Paul Gaggl. 2018. On the Welfare Implications of Automation. *Review of Economic Dynamics* 29, (July 2018), 15–43. https://doi.org/10.1016/j.red.2017.12.003

[121] Carolyn Edwards and Kay Springate. 1995. The Lion Comes out of the Stone: Helping Young Children Achieve Their Creative Potential. *Dimensions of Early Childhood* 23, 4 (1995), 24–29.

[122] Justin Edwards, Christian Janssen, Sandy Gould, and Benjamin R. Cowan. 2021. Eliciting Spoken Interruptions to Inform Proactive Speech Agent Design. *CUI 2021 - 3rd Conference on Conversational User Interfaces* (July 2021), 1–12. https://doi.org/10.1145/3469595.3469618.

[123] Elizabeth Victoria Eikey, Clara Marques Caldeira, Mayara Costa Figueiredo, Yunan Chen, Jessica L. Borelli, Melissa Mazmanian, and Kai Zheng. 2021. Beyond Self-Reflection: Introducing the Concept of Rumination in Personal Informatics. *Personal and Ubiquitous Computing* 25, 3 (June 2021), 601–616. https://doi.org/10.1007/s00779-021-01573-w

[124] Elliot W. Eisner. 1993. Forms of Understanding and the Future of Educational Research. *Educational Researcher* 22, 7 (October 1993), 5–11. https://doi.org/10.3102/0013189X022007005

[125] Douglas C. Engelbart. 2023. Augmenting Human Intellect: A Conceptual Framework. In *Augmented Education in the Global Age*. Routledge.

[126] Robert H. Ennis. 1987. A Taxonomy of Critical Thinking Dispositions and Abilities. In *Teaching Thinking Skills: Theory and Practice*. W. H. Freeman, New York, NY, US, 9–26.

[127] Rochelle E. Evans and Philip Kortum. 2010. The Impact of Voice Characteristics on User Response in an Interactive Voice Response System. *Interacting with Computers* 22, 6 (November 2010), 606–614. https://doi.org/10.1016/j.intcom.2010.07.001

[128] Caroline J Falconer, Aitor Rovira, John A King, Paul Gilbert, Angus Antley, Pasco Fearon, Neil Ralph, Mel Slater, and Chris R Brewin. 2016. Embodying Self-Compassion within Virtual Reality and Its Effects on Patients with Depression. *BJPsych Open* 2, 1 (2016), 74–80.

[129] Eugene F. Fama. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance* 25, 2 (1970), 383–417. https://doi.org/10.2307/2325486

[130] Eugene F. Fama. 1991. Efficient Capital Markets: II. *The Journal of Finance* 46, 5 (1991), 1575–1617. https://doi.org/10.1111/j.1540-6261.1991.tb04636.x

[131] Juan Carlos Farah, Basile Spaenlehauer, Vandit Sharma, María Jesús Rodríguez-Triana, Sandy Ingram, and Denis Gillet. 2022. Impersonating Chatbots in a Code Review Exercise to Teach Software Engineering Best Practices. In *2022 IEEE Global Engineering Education Conference (EDUCON)*, March 2022. 1634–1642. https://doi.org/10.1109/EDUCON52537.2022.9766793

[132] William Farr, Nicola Yuill, Eric Harris, and Steve Hinske. 2010. In My Own Words: Configuration of Tangibles, Object Interaction and Children with Autism. In *Proceedings of the 9th International Conference on Interaction Design and Children* (*IDC '10*), June 09, 2010. Association for Computing Machinery, New York, NY, USA, 30–38. https://doi.org/10.1145/1810543.1810548

[133] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S. Bernstein. 2018. Iris: A Conversational Agent for Complex Tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 2018. ACM Press, Montreal QC, Canada, 1–12. https://doi.org/10.1145/3173574.3174047

[134] Lisa Fazio. 2020. Pausing to Consider Why a Headline Is True or False Can Help Reduce the Sharing of False News. *Harvard Kennedy School Misinformation Review* 1, 2 (February 2020). https://doi.org/10.37016/mr-2020-009

[135] Philip M. Fernbach, Todd Rogers, Craig R. Fox, and Steven A. Sloman. 2013. Political Extremism Is Supported by an Illusion of Understanding. *Psychological Science* 24, 6 (June 2013), 939–946. https://doi.org/10.1177/0956797612464058

[136] Leon Festinger. 1957. *A Theory of Cognitive Dissonance*. Stanford University Press.

[137] Joel Fischer. 2011. *Understanding Receptivity to Interruptions in Mobile Human-Computer Interaction*. University of Nottingham. Retrieved from http://eprints.nottingham.ac.uk/12499/

[138] Joel E Fischer. 2023. Generative AI Considered Harmful. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (*CUI '23*), July 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3571884.3603756

[139] Joel E. Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. 2010. Effects of Content and Time of Delivery on Receptivity to Mobile

Interruptions. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services* (*MobileHCI '10*), September 07, 2010. Association for Computing Machinery, New York, NY, USA, 103–112. https://doi.org/10.1145/1851600.1851620

[140] Baruch Fischhoff and Ruth Beyth. 1975. I Knew It Would Happen: Remembered Probabilities of Once—Future Things. *Organizational Behavior and Human Performance* 13, 1 (February 1975), 1–16. https://doi.org/10.1016/0030-5073(75)90002-1

[141] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4, 2 (June 2017), e7785. https://doi.org/10.2196/mental.7785

[142] John H. Flavell. 1979. Metacognition and Cognitive Monitoring: A New Area of Cognitive–Developmental Inquiry. *American Psychologist* 34, 10 (1979), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

[143] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on Reflection: Framing a Design Landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (*OZCHI '10*), November 22, 2010. Association for Computing Machinery, Brisbane, Australia, 216–223. https://doi.org/10.1145/1952222.1952269

[144] Lawrence B. Flick. 2000. Cognitive Scaffolding That Fosters Scientific Inquiry in Middle Level Science. *Journal of Science Teacher Education* 11, 2 (2000), 109–129.

[145] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, Rebecca Wald, Fabio Catania, Raphael Meyer von Wolff, Sebastian Hobert, and Ewa Luger. 2021. Future Directions for Chatbot Research: An Interdisciplinary Research Agenda. *Computing* 103, 12 (December 2021), 2915–2942. https://doi.org/10.1007/s00607-021-01016-7

[146] Daniel Freeman, Sarah Reeve, A Robinson, Anke Ehlers, David Clark, Bernhard Spanlang, and Mel Slater. 2017. Virtual Reality in the Assessment, Understanding, and Treatment of Mental Health Disorders. *Psychological Medicine* 47, 14 (2017), 2393–2400.

[147] Brett Frischmann and Evan Selinger. 2018. *Re-Engineering Humanity*. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781316544846

[148] Luke Fryer and Rollo Carpenter. 2006. Bots as Language Learning Tools. *Language Learning & Technology* 10, 3 (2006), 8–14.

[149] Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30, 3 (September 2020), 411–437. https://doi.org/10.1007/s11023-020-09539-2

[150] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software Technology* (*UIST '15*), 2015. ACM, New York, NY, USA, 489–500. https://doi.org/10.1145/2807442.2807478

[151] Katy Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing Using Language Models. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, May 2022. Association for Computational Linguistics, Dublin, Ireland, 83–84. https://doi.org/10.18653/v1/2022.in2writing-1.12

[152] Gerd Gigerenzer. 2020. When All Is Just a Click Away: Is Critical Thinking Obsolete in the Digital Age? *Critical Thinking in Psychology* (2020), 197–223. https://doi.org/10.1017/9781108684354.010

[153] Gerd Gigerenzer and Wolfgang Gaissmaier. 2011. Heuristic Decision Making. *Annual Review of Psychology* 62, 1 (2011), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

[154] Sam J. Gilbert, Arabella Bird, Jason M. Carpenter, Stephen M. Fleming, Chhavi Sachdeva, and Pei-Chun Tsai. 2020. Optimal Use of Reminders: Metacognition, Effort, and Cognitive Offloading. *Journal of Experimental Psychology: General* 149, 3 (2020), 501–517. https://doi.org/10.1037/xge0000652

[155] Ulrich Gnewuch, Stefan Morana, and A. Maedche. 2017. Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *Proceedings of the 38th International Conference on Interaction Sciences*, October 17, 2017. Association for Information Systems, Seoul, South Korea. Retrieved November 1, 2023 from https://www.semanticscholar.org/paper/Towards-Designing-Cooperative-and-Social-Agents-for-Gnewuch-Morana/61ba3584885142e46673943142a4f2280ac14387

[156] Yoshiko Goda, Masanori Yamada, Hideya Matsukawa, Kojiro Hata, and Seisuke Yasunami. 2014. Conversation with a Chatbot before an Online EFL Group Discussion and the Effects on Critical Thinking. *The Journal of Information and Systems in Education* 13, 1 (2014), 1–7. https://doi.org/10.12937/ejsise.13.1

[157] Daniel Goleman. 1995. *Emotional Intelligence*. Bantam Books.

[158] Diego Gonzalez and Andrew S. Gordon. 2018. Comparing Speech and Text Input in Interactive Narratives. In *23rd International Conference on Intelligent User Interfaces* (*IUI '18*), March 05, 2018. Association for Computing Machinery, Tokyo, Japan, 141–145. https://doi.org/10.1145/3172944.3172999

[159] Arthur C. Graesser. 2011. Learning, Thinking, and Emoting with Discourse Technologies. *The American Psychologist* 66, 8 (November 2011), 746–757. https://doi.org/10.1037/a0024974

[160] Arthur C. Graesser. 2016. Conversations with AutoTutor Help Students Learn. *International Journal of Artificial Intelligence in Education* 26, 1 (March 2016), 124–132. https://doi.org/10.1007/s40593-015-0086-4

[161] Arthur C. Graesser, Haiying Li, and Carol Forsyth. 2014. Learning by Communicating in Natural Language With Conversational Agents. *Current Directions in Psychological Science* 23, 5 (October 2014), 374–380. https://doi.org/10.1177/0963721414540680

[162] Arthur C. Graesser, Kurt VanLehn, Carolyn P. Rosé, Pamela W. Jordan, and Derek Harter. 2001. Intelligent Tutoring Systems with Conversational Dialogue. *AI Magazine* 22, 4 (2001), 39–39.

[163] Lars Grammel, Melanie Tory, and Margaret-Anne Storey. 2010. How Information Visualization Novices Construct Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (December 2010), 943–952. https://doi.org/10.1109/TVCG.2010.164

[164] Jane Gravill, Deborah Compeau, and Barbara Marcolin. 2002. Metacognition and IT: The Influence of Self-Efficacy and Self-Awareness. *AMCIS 2002 Proceedings* (2002), 147.

[165] Jasmin Grosinger. 2022. On Proactive Human-AI Systems. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Cognition* (*CEUR Workshop Proceedings*), June 15, 2022. CEUR-WS.org, Örebro, Sweden, 140–146. Retrieved from https://ceur-ws.org/Vol-3400/paper12.pdf

[166] Dan Gruen, Candy Sidner, Carolyn Boettner, and Charles Rich. 1999. A Collaborative Assistant for Email. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems* (*CHI EA '99*), May 15, 1999. Association for Computing Machinery, New York, NY, USA, 196–197. https://doi.org/10.1145/632716.632839

[167] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How Many Interviews Are Enough? An Experiment with Data Saturation and Variability. *Field Methods* 18, 1 (2006), 59–82.

[168] Jingya Guo, Jiajing Guo, Changyuan Yang, Yanjing Wu, and Lingyun Sun. 2021. Shing: A Conversational Agent to Alert Customers of Suspected Online-Payment Fraud with Empathetical Communication Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 06, 2021. ACM, Yokohama Japan, 1–11. https://doi.org/10.1145/3411764.3445129

[169] Kent L. Gustafson, Jr Bennett, and Winston. 2002. *Promoting Learner Reflection: Issues and Difficulties Emerging from a Three-Year Study*. Defense Technical Information Center, Fort Belvoir, VA. https://doi.org/10.21236/ADA472616

[170] Gabriel Haas, Michael Rietzler, Matt Jones, and Enrico Rukzio. 2022. Keep It Short: A Comparison of Voice Assistants' Response Behavior. In *CHI Conference on Human*

Factors in Computing Systems (*CHI '22*), 2022. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3491102.3517684

[171] Florian Habler, Valentin Schwind, and Niels Henze. 2019. Effects of Smart Virtual Assistants' Gender and Language. In *Proceedings of Mensch und Computer 2019* (*MuC '19*), September 08, 2019. Association for Computing Machinery, New York, NY, USA, 469–473. https://doi.org/10.1145/3340764.3344441

[172] Irit Hacmun, Dafna Regev, and Roy Salomon. 2018. The Principles of Art Therapy in Virtual Reality. *Frontiers in Psychology* 9, (2018), 2082.

[173] Irit Hacmun, Dafna Regev, and Roy Salomon. 2021. Artistic Creation in Virtual Reality for Art Therapy: A Qualitative Study with Expert Art Therapists. *The Arts in Psychotherapy* 72, (February 2021), 101745. https://doi.org/10.1016/j.aip.2020.101745

[174] Anoushka Halder, Aayush Saxena, and S. Priya. 2022. Stock Market Prediction Through a Chatbot: A Human-Centered AI Approach. In *Ubiquitous Intelligent Systems* (*Smart Innovation, Systems and Technologies*), 2022. Springer Nature, Singapore, 435–446. https://doi.org/10.1007/978-981-19-2541-2_34

[175] Florian Hammer, Peter Reichl, and Alexander Raake. 2004. Elements of Interactivity in Telephone Conversations. In *Proceedings of the 8th International Conference on Spoken Language Processing*, October 04, 2004. Jeju Island, Korea. https://doi.org/10.21437/Interspeech.2004-592

[176] Jeffrey T. Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (March 2020), 89–100. https://doi.org/10.1093/jcmc/zmz022

[177] Harinder Hari, Radha Iyer, and Brinda Sampat. 2022. Customer Brand Engagement through Chatbots on Bank Websites– Examining the Antecedents and Consequences. *International Journal of Human–Computer Interaction* 38, 13 (August 2022), 1212–1227. https://doi.org/10.1080/10447318.2021.1988487

[178] Wieke Noa Harmsen, Jelte Van Waterschoot, Iris Hendrickx, and Mariët Theune. 2023. Eliciting User Self-Disclosure Using Reciprocity in Human-Voicebot Conversations. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (*CUI '23*), July 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3571884.3604301

[179] Alexander G. Hauptmann and Alexander I. Rudnicky. 1990. A Comparison of Speech and Typed Input. In *Proceedings of the Workshop on Speech and Natural Language - HLT '90*, 1990. Association for Computational Linguistics, Hidden Valley, Pennsylvania, 219–224. https://doi.org/10.3115/116580.116652

[180] Christian Heath, Jon Hindmarsh, and Paul Luff. 2010. *Video in Qualitative Research: Analysing Social Interaction in Everyday Life*. SAGE Publications, Inc., London. https://doi.org/10.4135/9781526435385

[181] Thorsten Hens and Anna Meier. 2015. *Behavioral Finance: The Psychology of Investing*. Credit Suisse.

[182] Khe Foon Hew, Weijiao Huang, Jiahui Du, and Chengyuan Jia. 2022. Using Chatbots to Support Student Goal Setting and Social Presence in Fully Online Activities: Learner Engagement and Perceptions. *Journal of Computing in Higher Education* (September 2022). https://doi.org/10.1007/s12528-022-09338-x

[183] Christian Hildebrand and Anouk Bergner. 2020. Conversational Robo Advisors as Surrogates of Trust: Onboarding Experience, Firm Perception, and Consumer Financial Decision Making. *Journal of the Academy of Marketing Science* (November 2020). https://doi.org/10.1007/s11747-020-00753-z

[184] David Hitchcock. 2022. Critical Thinking. In *The Stanford Encyclopedia of Philosophy* (Winter 2022), Edward N. Zalta and Uri Nodelman (eds.). Metaphysics Research Lab, Stanford University. Retrieved November 9, 2023 from https://plato.stanford.edu/archives/win2022/entries/critical-thinking/

[185] Joyce Ho and Stephen S. Intille. 2005. Using Context-Aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '05*), April 02, 2005. Association for Computing Machinery, New York, NY, USA, 909–918. https://doi.org/10.1145/1054972.1055100

[186] Carmen Holotescu. 2016. MOOCBuddy: A Chatbot for Personalized Learning with MOOCs. In *RoCHI Conference Proceedings*, September 08, 2016. Iași, Romania. Retrieved December 10, 2023 from https://www.semanticscholar.org/paper/MOOCBuddy%3A-a-Chatbot-for-personalized-learning-with-Holotescu/832c8de6424644765f98094c0127981120fc66e5

[187] Gan Keng Hoon, Loo Ji Yong, and Goh Kau Yang. 2020. Interfacing Chatbot with Data Retrieval and Analytics Queries for Decision Making. In *RITA 2018* (*Lecture Notes in Mechanical Engineering*), 2020. Springer, Singapore, 385–394. https://doi.org/10.1007/978-981-13-8323-6_32

[188] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2018. Applying Pragmatics Principles for Interaction with Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (January 2018), 309–318. https://doi.org/10.1109/TVCG.2017.2744684

[189] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '99*), May 01, 1999. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[190] Haochen Hua, Yutong Li, Tonghe Wang, Nanqing Dong, Wei Li, and Junwei Cao. 2023. Edge Computing with Artificial Intelligence: A Machine Learning Perspective. *ACM Computing Surveys* 55, 9 (January 2023), 184:1-184:35. https://doi.org/10.1145/3555802

[191] Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. 2003. Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '03*), April 05, 2003. Association for Computing Machinery, New York, NY, USA, 257–264. https://doi.org/10.1145/642611.642657

[192] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology Probes: Inspiring Design for and with Families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '03*), April 05, 2003. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[193] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. 2011. Sequential Model-Based Optimization for General Algorithm Configuration. In *Proceedings of the 5th International Conference on Learning and Intelligent Optimization* (*LION'05*), 2011. Springer-Verlag, Berlin, Heidelberg, 507–523. https://doi.org/10.1007/978-3-642-25566-3_40

[194] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2021. IdeaBot: Investigating Social Facilitation in Human-Machine Team Creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (*CHI '21*), May 06, 2021. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3411764.3445270

[195] Takamasa Iio, Satoru Satake, Takayuki Kanda, Kotaro Hayashi, Florent Ferreri, and Norihiro Hagita. 2020. Human-Like Guide Robot That Proactively Explains Exhibits. *International Journal of Social Robotics* 12, 2 (May 2020), 549–566. https://doi.org/10.1007/s12369-019-00587-y

[196] Yahya İltüzer and Yasemin Demiraslan Çevik. 2021. Effects of Self-Explanation on Applying Decision Rules in an Online Learning Environment. *Education and Information Technologies* 26, 4 (July 2021), 4771–4794. https://doi.org/10.1007/s10639-021-10499-y

[197] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth* 6, 11 (November 2018), e12106. https://doi.org/10.2196/12106

[198] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python Interface to Praat. *Journal of Phonetics* 71, (November 2018), 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

[199] Christophe Jallais and Anne-Laure Gilet. 2010. Inducing Changes in Arousal and Valence: Comparison of Two Mood Induction Procedures. *Behavior Research Methods* 42, 1 (2010), 318–325. https://doi.org/10.3758/BRM.42.1.318

[200] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1, 9 (September 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[201] Rogers Jeffrey Leo John, Navneet Potti, and Jignesh M. Patel. 2017. Ava: From Data to Insights Through Conversations. In *Proceedings of the 8th Biennial Conference on Innovative Data Systems Research*, January 08, 2017. Chaminade, CA, USA. Retrieved from https://api.semanticscholar.org/CorpusID:1519997

[202] Bernhard Jordan, Laura Koesten, and Torsten Möller. 2023. Chatting About Data - Interacting with Voice Interfaces to Engage with Election Panel Data. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (*CUI '23*), July 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3571884.3597126

[203] Malte F. Jung, Nik Martelaro, Halsey Hoster, and Clifford Nass. 2014. Participatory Materials: Having a Reflective Conversation with an Artifact in the Making. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, June 21, 2014. ACM, Vancouver BC Canada, 25–34. https://doi.org/10.1145/2598510.2598591

[204] Sandra L. Kagin and Vija B. Lusebrink. 1978. The Expressive Therapies Continuum. *Art Psychotherapy* 5, 4 (January 1978), 171–180. https://doi.org/10.1016/0090-9092(78)90031-5

[205] Daniel Kahneman. 2011. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York, NY, US.

[206] Soowon Kang, Heepyung Kim, Youngtae Noh, and Uichin Lee. 2021. Poster: Toward Context-Aware Proactive Conversation for Smart Speakers. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, September 21, 2021. ACM, Virtual USA, 38–40. https://doi.org/10.1145/3460418.3479306

[207] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education.

Learning and Individual Differences 103, (April 2023), 102274. https://doi.org/10.1016/j.lindif.2023.102274

[208] Jan-Frederik Kassel and Michael Rohs. 2018. Valletto: A Multimodal Interface for Ubiquitous Visual Analytics. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[209] Kim A. Kastens, Melissa Zrada, and Margie Turrin. 2020. What Kinds of Questions Do Students Ask While Exploring Data Visualizations? *Journal of Geoscience Education* 68, 3 (July 2020), 199–219. https://doi.org/10.1080/10899995.2019.1675447

[210] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. 2015. May I Help You?: Design of Human-like Polite Approaching Behavior. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*, 2015. ACM Press, Portland, Oregon, USA, 35–42. https://doi.org/10.1145/2696454.2696463

[211] Michael W. Eysenck Keane Mark T. 2020. *Cognitive Psychology: A Student's Handbook* (8th ed.). Psychology Press, London. https://doi.org/10.4324/9781351058513

[212] Samuel Kernan Freire, Mina Foosherian, Chaofan Wang, and Evangelos Niforatos. 2023. Harnessing Large Language Models for Cognitive Assistants in Factories. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (*CUI '23*), July 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3571884.3604313

[213] Samuel Kernan Freire, Evangelos Niforatos, Chaofan Wang, Santiago Ruiz-Arenas, Mina Foosherian, Stefan Wellsandt, and Alessandro Bozzon. 2023. Lessons Learned from Designing and Evaluating CLAICA: A Continuously Learning AI Cognitive Assistant. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (*IUI '23*), March 27, 2023. Association for Computing Machinery, New York, NY, USA, 553–568. https://doi.org/10.1145/3581641.3584042

[214] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oktober 2020), 163:1-163:26. https://doi.org/10.1145/3415234

[215] Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. 2018. Interrupting Drivers for Interactions: Predicting Opportune Moments for In-Vehicle Proactive Auditory-Verbal Tasks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (December 2018), 175:1-175:28. https://doi.org/10.1145/3287053

[216] Auk Kim, Jung-Mi Park, and Uichin Lee. 2020. Interruptibility for In-Vehicle Multitasking: Influence of Voice Task Demands and Adaptive Behaviors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (March 2020), 14:1-14:22. https://doi.org/10.1145/3381009

[217] Minchi C. Kim and Michael J. Hannafin. 2011. Scaffolding Problem Solving in Technology-Enhanced Learning Environments (TELEs): Bridging Research and Theory with Practice. *Computers & Education* 56, 2 (February 2011), 403–417. https://doi.org/10.1016/j.compedu.2010.08.024

[218] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. 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*, April 21, 2020. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376785

[219] Yeongdae Kim, Takane Ueno, Katie Seaborn, Hiroki Oura, Jacqueline Urakami, and Yuto Sawa. 2023. Exoskeleton for the Mind: Exploring Strategies Against Misinformation with a Metacognitive Agent. In *Proceedings of the Augmented Humans International Conference 2023* (*AHs '23*), March 14, 2023. Association for Computing Machinery, New York, NY, USA, 209–220. https://doi.org/10.1145/3582700.3582725

[220] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. 2019. A Conversational Agent in Support of Productivity and Wellbeing at Work. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, September 2019. 1–7. https://doi.org/10.1109/ACII.2019.8925488

[221] Alison King. 1989. Effects of Self-Questioning Training on College Students' Comprehension of Lectures. *Contemporary Educational Psychology* 14, 4 (October 1989), 366–381. https://doi.org/10.1016/0361-476X(89)90022-2

[222] Alison King. 1994. Autonomy and Question Asking: The Role of Personal Control in Guided Student-Generated Questioning. *Learning and Individual Differences* 6, 2 (January 1994), 163–185. https://doi.org/10.1016/1041-6080(94)90008-6

[223] Patricia M. King and Karen Strohm Kitchener. 1994. *Developing Reflective Judgment*. Jossey-Bass, San Francisco.

[224] David Kirsh. 2010. Thinking with External Representations. *AI & Society* 25, 4 (November 2010), 441–454. https://doi.org/10.1007/s00146-010-0272-8

[225] Alexandra Kitson, Mirjana Prpa, and Bernhard E. Riecke. 2018. Immersive Interactive Technologies for Positive Change: A Scoping Review and Design Considerations. *Frontiers in Psychology* 9, (2018), 1354. https://doi.org/10.3389/fpsyg.2018.01354

[226] A. Baki Kocaballi, Liliana Laranjo, Leigh Clark, Rafał Kocielnik, Robert J. Moore, Q. Vera Liao, and Timothy Bickmore. 2022. Special Issue on Conversational Agents for Healthcare and Wellbeing. *ACM Transactions on Interactive Intelligent Systems* 12, 2 (July 2022), 9:1-9:3. https://doi.org/10.1145/3532860

[227] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of the 2018 Designing Interactive Systems Conference* (*DIS '18*), 2018.

Association for Computing Machinery, New York, NY, USA, 881–894. https://doi.org/10.1145/3196709.3196784

[228] Rafal Kocielnik, Gary Hsieh, and Daniel Avrahami. 2018. Helping Users Reflect on Their Own Health-Related Behaviors. In *Studies in Conversational UX Design*, Robert J. Moore, Margaret H. Szymanski, Raphael Arar and Guang-Jie Ren (eds.). Springer International Publishing, Cham, 85–115. https://doi.org/10.1007/978-3-319-95579-7_5

[229] Rafal Kocielnik, Raina Langevin, James S. George, Shota Akenaga, Amelia Wang, Darwin P. Jones, Alexander Argyle, Callan Fockele, Layla Anderson, Dennis T. Hsieh, Kabir Yadav, Herbert Duber, Gary Hsieh, and Andrea L. Hartzler. 2021. Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (*CUI '21*), July 27, 2021. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3469595.3469599

[230] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (July 2018), 70:1-70:26. https://doi.org/10.1145/3214273

[231] Maximilian Koestner, Benjamin Loos, Steffen Meyer, and Andreas Hackethal. 2017. Do Individual Investors Learn from Their Mistakes? *Journal of Business Economics* 87, 5 (July 2017), 669–703. https://doi.org/10.1007/s11573-017-0855-7

[232] David A. Kolb. 2015. *Experiential Learning: Experience as the Source of Learning and Development* (2nd ed.). Pearson Education, Upper Saddle River, New Jersey, USA.

[233] Mitsuki Komori, Yuichiro Fujimoto, Jianfeng Xu, Kazuyuki Tasaka, Hiromasa Yanagihara, and Kinya Fujita. 2019. Experimental Study on Estimation of Opportune Moments for Proactive Voice Information Service Based on Activity Transition for People Living Alone. In *Human-Computer Interaction. Perspectives on Design*, 2019. Springer International Publishing, Cham, 527–539.

[234] Matthias Kraus, Fabian Fischbach, Pascal Jansen, and Wolfgang Minker. 2020. A Comparison of Explicit and Implicit Proactive Dialogue Strategies for Conversational Recommendation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, May 2020. European Language Resources Association, Marseille, France, 429–435. Retrieved October 24, 2023 from https://aclanthology.org/2020.lrec-1.54

[235] Ethan Kross and Ozlem Ayduk. 2011. Making Meaning out of Negative Experiences by Self-Distancing. *Current Directions in Psychological Science* 20, 3 (June 2011), 187–191. https://doi.org/10.1177/0963721411408883

[236] Ethan Kross, Madeline Ong, and Ozlem Ayduk. 2023. Self-Reflection at Work: Why It Matters and How to Harness Its Potential and Avoid Its Pitfalls. *Annual Review of Organizational Psychology and Organizational Behavior* 10, 1 (January 2023), 441–464. https://doi.org/10.1146/annurev-orgpsych-031921-024406

[237] Steve Krug. 2005. *Don't Make Me Think: A Common Sense Approach to the Web* (2nd ed.). New Riders Publishing, Berkeley, CA, USA.

[238] Harsh Kumar, Yiyi Wang, Jiakai Shi, Ilya Musabirov, Norman A. S. Farb, and Joseph Jay Williams. 2023. Exploring the Use of Large Language Models for Improving the Awareness of Mindfulness. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (*CHI EA '23*), April 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3544549.3585614

[239] Ziva Kunda. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* 108, 3 (November 1990), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

[240] Chongsan Kwon. 2019. Verification of the Possibility and Effectiveness of Experiential Learning Using HMD-Based Immersive VR Technologies. *Virtual Reality* 23, 1 (2019), 101–118.

[241] Kyungbin Kwon, Christiana Kumalasari, and Jane Howland. 2011. Self-Explanation Prompts on Problem-Solving Performance in an Interactive Learning Environment. *Journal of Interactive Online Learning* (2011). Retrieved December 3, 2023 from https://www.semanticscholar.org/paper/Self-Explanation-Prompts-on-Problem-Solving-in-an-Kwon-Kumalasari/29cd7b4f9e839441900b2af5dad7f7015a96a6d8

[242] Marcello L'Abbate, Ulrich Thiel, and Thomas Kamps. 2005. Can Proactive Behavior Turn Chatterbots into Conversational Agents? In *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology* (*IAT '05*), September 19, 2005. IEEE Computer Society, USA, 173–179. https://doi.org/10.1109/IAT.2005.49

[243] H. Chad Lane and Kurt VanLehn. 2003. Coached Program Planning: Dialogue-Based Support for Novice Program Design. *ACM SIGCSE Bulletin* 35, 1 (January 2003), 148–152. https://doi.org/10.1145/792548.611955

[244] Helmut Lang, Melina Klepsch, Florian Nothdurft, Tina Seufert, and Wolfgang Minker. 2013. The Influence of Proactivity on Interactive Help Agents. In *Human Factors in Computing and Informatics* (*Lecture Notes in Computer Science*), 2013. Springer, Berlin, Heidelberg, 748–767. https://doi.org/10.1007/978-3-642-39062-3_54

[245] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (November 2018), 102:1-102:31. https://doi.org/10.1145/3274371

[246] Paula Lauren and Paul Watta. 2019. A Conversational User Interface for Stock Analysis. In *2019 IEEE International Conference on Big Data (Big Data)*, December 2019. 5298–5305. https://doi.org/10.1109/BigData47090.2019.9005635

[247] Ludovic Le Bigot, Eric Jamet, Jean-François Rouet, and Virginie Amiel. 2006. Mode and Modal Transfer Effects on Performance and Discourse Organization with an

Information Retrieval Dialogue System in Natural Language. *Computers in Human Behavior* 22, 3 (May 2006), 467–500. https://doi.org/10.1016/j.chb.2004.10.006

[248] Ludovic Le Bigot, Patrice Terrier, Virginie Amiel, Gérard Poulain, Eric Jamet, and Jean-François Rouet. 2007. Effect of Modality on Collaboration with a Dialogue System. *International Journal of Human-Computer Studies* 65, 12 (December 2007), 983–991. https://doi.org/10.1016/j.ijhcs.2007.07.002

[249] Diana Lea and Jennifer Bradbery. 2020. *Oxford Advanced Learner's Dictionary*. Oxford University Press.

[250] Eun-Ju Lee. 2008. Flattery May Get Computers Somewhere, Sometimes: The Moderating Role of Output Modality, Computer Gender, and User Gender. *International Journal of Human-Computer Studies* 66, 11 (November 2008), 789–800. https://doi.org/10.1016/j.ijhcs.2008.07.009

[251] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: A Chatbot for Self-Compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, May 02, 2019. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300932

[252] Minha Lee, Lily Frank, and Wijnand IJsselsteijn. 2021. Brokerbot: A Cryptocurrency Chatbot in the Social-Technical Gap of Trust. *Computer Supported Cooperative Work (CSCW)* 30, 1 (February 2021), 79–117. https://doi.org/10.1007/s10606-021-09392-6

[253] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-Disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 21, 2020. ACM, Honolulu HI USA, 1–12. https://doi.org/10.1145/3313831.3376175

[254] Daniel Levi. 2001. *Group Dynamics for Teams*. Sage Publications, Thousand Oaks, CA, US.

[255] Zhuoyang Li, Minhui Liang, Hai Trung Le, Ray Lc, and Yuhan Luo. 2023. Exploring Design Opportunities for Reflective Conversational Agents to Reduce Compulsive Smartphone Use. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (*CUI '23*), July 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3571884.3604305

[256] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive Conversational Agents. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (*WSDM '23*), February 27, 2023. Association for Computing Machinery, New York, NY, USA, 1244–1247. https://doi.org/10.1145/3539597.3572724

[257] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive Conversational Agents in the Post-ChatGPT World. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '23*), July 18, 2023.

Association for Computing Machinery, New York, NY, USA, 3452–3455. https://doi.org/10.1145/3539618.3594250

[258] Joseph Carl Robnett Licklider. 1960. Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics* HFE-1, 1 (March 1960), 4–11. https://doi.org/10.1109/THFE2.1960.4503259

[259] Hannah Limerick, James W. Moore, and David Coyle. 2015. Empirical Evidence for a Diminished Sense of Agency in Speech Interfaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (*CHI '15*), 2015. ACM, New York, NY, USA, 3967–3970. https://doi.org/10.1145/2702123.2702379

[260] Lijia Lin, Paul Ginns, Tianhui Wang, and Peilin Zhang. 2020. Using a Pedagogical Agent to Deliver Conversational Style Instruction: What Benefits Can You Obtain? *Computers & Education* 143, (January 2020), 103658. https://doi.org/10.1016/j.compedu.2019.103658

[261] Diane J. Litman, Carolyn P. Rosé, Kate Forbes-Riley, Kurt VanLehn, Dumisizwe Bhembe, and Scott Silliman. 2004. Spoken Versus Typed Human and Computer Dialogue Tutoring. In *Intelligent Tutoring Systems*, James C. Lester, Rosa Maria Vicari and Fábio Paraguaçu (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 368–379. https://doi.org/10.1007/978-3-540-30139-4_35

[262] Jack M. Loomis, James J. Blascovich, and Andrew C. Beall. 1999. Immersive Virtual Environment Technology as a Basic Research Tool in Psychology. *Behavior Research Methods, Instruments, & Computers* 31, 4 (December 1999), 557–564. https://doi.org/10.3758/BF03200735

[263] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (*CHI '16*), 2016. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[264] Michal Luria, Rebecca Zheng, Bennett Huffman, Shuangni Huang, John Zimmerman, and Jodi Forlizzi. 2020. Social Boundaries for Personal Agents in the Interpersonal Space of the Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (*CHI '20*), April 21, 2020. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376311

[265] Sonja Lyubomirsky, Kennon M Sheldon, and David Schkade. 2005. Pursuing Happiness: The Architecture of Sustainable Change. *Review of General Psychology* 9, 2 (2005), 111–131.

[266] Raju Maharjan, Darius Adam Rohani, Per Bækgaard, Jakob Bardram, and Kevin Doherty. 2021. Can We Talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, July 27, 2021. ACM, Bilbao (online) Spain, 1–11. https://doi.org/10.1145/3469595.3469600

[267] Cathy A. Malchiodi (Ed.). 2012. *Handbook of Art Therapy* (2nd ed.). The Guilford Press, New York, NY, US.

[268] Tanya Malik, Adrian Jacques Ambrose, and Chaitali Sinha. 2022. Evaluating User Feedback for an Artificial Intelligence–Enabled, Cognitive Behavioral Therapy–Based Mental Health App (Wysa): Qualitative Thematic Analysis. *JMIR Human Factors* 9, 2 (April 2022), e35668. https://doi.org/10.2196/35668

[269] Nathan Malkin, David Wagner, and Serge Egelman. 2022. Runtime Permissions for Privacy in Proactive Intelligent Assistants. In *Proceedings of the Eighteenth USENIX Conference on Usable Privacy and Security* (*SOUPS'22*), August 08, 2022. USENIX Association, USA, 633–651.

[270] Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. 2008. MAHI: Investigation of Social Scaffolding for Reflective Thinking in Diabetes Management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '08*), 2008. Association for Computing Machinery, New York, NY, USA, 477–486. https://doi.org/10.1145/1357054.1357131

[271] Marcello M. Mariani, Novin Hashemi, and Jochen Wirtz. 2023. Artificial Intelligence Empowered Conversational Agents: A Systematic Literature Review and Research Agenda. *Journal of Business Research* 161, (June 2023), 113838. https://doi.org/10.1016/j.jbusres.2023.113838

[272] Ramon Marinez. 2015. Level and Trends of Overweight and Obesity. *Tableau Public*. Retrieved November 20, 2023 from https://public.tableau.com/profile/ramon.martinez#!/vizhome/LevelandTrendsofOverweightandObesity/Overweightandobesitylevel

[273] Paul Marshall, Eva Hornecker, Richard Morris, Nick Sheep Dalton, and Yvonne Rogers. 2008. When the Fingers Do the Talking: A Study of Group Participation with Varying Constraints to a Tabletop Interface. In *2008 3rd IEEE International Workshop on Horizontal Interactive Human Computer Systems*, October 2008. 33–40. https://doi.org/10.1109/TABLETOP.2008.4660181

[274] Manolis Mavrikis, Beate Grawemeyer, Alice Hansen, and Sergio Gutierrez-Santos. 2014. Exploring the Potential of Speech Recognition to Support Problem Solving and Reflection. In *Open Learning and Teaching in Educational Communities* (*Lecture Notes in Computer Science*), 2014. Springer International Publishing, Cham, 263–276. https://doi.org/10.1007/978-3-319-11200-8_20

[275] Moira McGregor and John C. Tang. 2017. More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (*CSCW '17*), February 25, 2017. Association for Computing Machinery, Portland, Oregon, USA, 2208–2220. https://doi.org/10.1145/2998181.2998335

[276] Danielle S. McNamara and Joseph P. Magliano. 2009. Self-Explanation and Metacognition: The Dynamics of Reading. In *Handbook of Metacognition in Education*. Routledge/Taylor & Francis Group, New York, NY, US, 60–81.

[277] Michael McTear. 2021. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-02176-3

[278] Michael McTear, Zoraida Callejas, and David Griol. 2016. Conversational Interfaces: Devices, Wearables, Virtual Agents, and Robots. In *The Conversational Interface: Talking to Smart Devices*, Michael McTear, Zoraida Callejas and David Griol (eds.). Springer International Publishing, Cham, 283–308. https://doi.org/10.1007/978-3-319-32967-3_13

[279] Anna-Maria Meck. 2023. Secure, Comfortable or Functional: Exploring Domain-Sensitive Prompt Design for In-Car Voice Assistants. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (*CUI '23*), July 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3571884.3604314

[280] Anna-Maria Meck, Christoph Draxler, and Thurid Vogt. 2023. How May I Interrupt? Linguistic-Driven Design Guidelines for Proactive in-Car Voice Assistants. *International Journal of Human–Computer Interaction* 40, 22 (October 2023), 7517–7531. https://doi.org/10.1080/10447318.2023.2266251

[281] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. PrefMiner: Mining User's Preferences for Intelligent Mobile Notification Management. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*UbiComp '16*), September 12, 2016. Association for Computing Machinery, Heidelberg, Germany, 1223–1234. https://doi.org/10.1145/2971648.2971747

[282] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2019. NotifyMeHere: Intelligent Notification Delivery in Multi-Device Environments. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval* (*CHIIR '19*), March 08, 2019. Association for Computing Machinery, Glasgow, Scotland UK, 103–111. https://doi.org/10.1145/3295750.3298932

[283] Abhinav Mehrotra and Mirco Musolesi. 2020. Intelligent Notification Systems. *Synthesis Lectures on Mobile and Pervasive Computing* 11, 1 (January 2020), 1–75. https://doi.org/10.2200/S00965ED1V01Y201911MPC014

[284] Muhsin Menekse and Michelene T. H. Chi. 2019. The Role of Collaborative Interactions versus Individual Construction on Students' Learning of Engineering Concepts. *European Journal of Engineering Education* 44, 5 (September 2019), 702–725. https://doi.org/10.1080/03043797.2018.1538324

[285] Christian Meurisch, Cristina A. Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring User Expectations of Proactive AI Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and*

Ubiquitous Technologies 4, 4 (Dezember 2020), 146:1-146:22. https://doi.org/10.1145/3432193

[286] O. Miksik, I. Munasinghe, J. Asensio-Cubero, S. Reddy Bethi, S.-T. Huang, S. Zylfo, X. Liu, T. Nica, A. Mitrocsak, S. Mezza, R. Beard, R. Shi, R. Ng, P. Mediano, Z. Fountas, S.-H. Lee, J. Medvesek, H. Zhuang, Y. Rogers, and P. Swietojanski. 2020. *Building Proactive Voice Assistants: When and How (Not) to Interact*. Retrieved May 5, 2022 from http://arxiv.org/abs/2005.01322

[287] Federico Milana, Enrico Costanza, and Joel E Fischer. 2023. Chatbots as Advisers: The Effects of Response Variability and Reply Suggestion Buttons. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (*CUI '23*), July 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3571884.3597132

[288] Inge Molenaar, Ming Ming Chiu, Peter Sleegers, and Carla van Boxtel. 2011. Scaffolding of Small Groups' Metacognitive Activities with an Avatar. *International Journal of Computer-Supported Collaborative Learning* 6, 4 (December 2011), 601–624. https://doi.org/10.1007/s11412-011-9130-z

[289] Ine Mols, Elise van de Hoven, and Barry Egen. 2020. Everyday Life Reflection: Exploring Media Interaction with Balance, Cogito & Dott. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction* (*TEI'20*), 2020. Association for Computing Machinery, New York, NY, United States. https://doi.org/10.1145/3374920.3374928

[290] Jessica Isbely Montana, Marta Matamala-Gomez, Marta Maisto, Petar Aleksandrov Mavrodiev, Cesare Massimo Cavalera, Barbara Diana, Fabrizia Mantovani, and Olivia Realdon. 2020. The Benefits of Emotion Regulation Interventions in Virtual Reality for the Improvement of Wellbeing in Adults and Older Adults: A Systematic Review. *Journal of Clinical Medicine* 9, 2 (2020), 500. https://doi.org/10.3390/jcm9020500

[291] Jennifer A. Moon. 2000. *Reflection in Learning and Professional Development: Theory and Practice*. Routledge, London. https://doi.org/10.4324/9780203822296

[292] Roger K. Moore. 2017. Appropriate Voices for Artefacts: Some Key Insights. In *1st International Workshop on Vocal Interactivity In-and-between Humans, Animals and Robots*, August 25, 2017. Skövde, Sweden. Retrieved November 22, 2023 from https://www.semanticscholar.org/paper/Appropriate-Voices-for-Artefacts%3A-Some-Key-Insights-Moore/724c080edbfefff16ce610e69614a3e4e9787d70

[293] Stefan Morana, Ulrich Gnewuch, Dominik Jung, and Carsten Granig. 2020. The Effect of Anthropomorphism on Investment Decision-Making with Robo-Advisor Chatbots. *ECIS 2020 Research Papers* (June 2020). Retrieved from https://aisel.aisnet.org/ecis2020_rp/63

[294] Patrick Moriarty and Damon Honnery. 2014. Reconnecting Technological Development with Human Welfare. *Futures* 55, (January 2014), 32–40. https://doi.org/10.1016/j.futures.2013.12.003

[295] Daniel G. Morrow, H. Chad Lane, and Wendy A. Rogers. 2021. A Framework for Design of Conversational Agents to Support Health Self-Care for Older Adults. *Human Factors* 63, 3 (2021), 369–378. https://doi.org/10.1177/0018720820964085

[296] Shabnam Mousavi and Gerd Gigerenzer. 2017. Heuristics Are Tools for Uncertainty. *Homo Oeconomicus* 34, 4 (December 2017), 361–379. https://doi.org/10.1007/s41412-017-0058-z

[297] Anwesha Mukherjee, Vagner Figueredo De Santana, and Alexis Baria. 2023. ImpactBot: Chatbot Leveraging Language Models to Automate Feedback and Promote Critical Thinking Around Impact Statements. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (*CHI EA '23*), April 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3544549.3573844

[298] Jaya Narain, Tina Quach, Monique Davey, Hae Won Park, Cynthia Breazeal, and Rosalind Picard. 2020. Promoting Wellbeing with Sunny, a Chatbot That Facilitates Positive Messages within Social Groups. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, April 25, 2020. ACM, Honolulu HI USA, 1–8. https://doi.org/10.1145/3334480.3383062

[299] Clifford Nass and Scott Brave. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. The MIT Press.

[300] Clifford Nass, Youngme Moon, and Paul Carney. 1999. Are People Polite to Computers? Responses to Computer-Based Interviewing Systems. *Journal of Applied Social Psychology* 29, 5 (1999), 1093–1110. https://doi.org/10.1111/j.1559-1816.1999.tb00142.x

[301] Clifford Nass, Erica Robles, Charles Heenan, Hilary Bienstock, and Marissa Treinen. 2003. Speech-Based Disclosure Systems: Effects of Modality, Gender of Prompt, and Gender of User. *International Journal of Speech Technology* 6, 2 (April 2003), 113–121. https://doi.org/10.1023/A:1022378312670

[302] Dennis C. Neale and John M. Carroll. 1997. Chapter 20 - The Role of Metaphors in User Interface Design. In *Handbook of Human-Computer Interaction* (2nd ed.), Marting G. Helander, Thomas K. Landauer and Prasad V. Prabhu (eds.). North-Holland, Amsterdam, 441–462. https://doi.org/10.1016/B978-044481862-1.50086-8

[303] Marie Ng, Tom Fleming, Margaret Robinson, Blake Thomson, Nicholas Graetz, Christopher Margono, Erin C Mullany, Stan Biryukov, Cristiana Abbafati, Semaw Ferede Abera, Jerry P Abraham, Niveen M E Abu-Rmeileh, Tom Achoki, Fadia S AlBuhairan, Zewdie A Alemu, Rafael Alfonso, Mohammed K Ali, Raghib Ali, Nelson Alvis Guzman, Walid Ammar, Palwasha Anwari, Amitava Banerjee, Simon Barquera,

Sanjay Basu, Derrick A Bennett, Zulfiqar Bhutta, Jed Blore, Norberto Cabral, Ismael Campos Nonato, Jung-Chen Chang, Rajiv Chowdhury, Karen J Courville, Michael H Criqui, David K Cundiff, Kaustubh C Dabhadkar, Lalit Dandona, Adrian Davis, Anand Dayama, Samath D Dharmaratne, Eric L Ding, Adnan M Durrani, Alireza Esteghamati, Farshad Farzadfar, Derek F J Fay, Valery L Feigin, Abraham Flaxman, Mohammad H Forouzanfar, Atsushi Goto, Mark A Green, Rajeev Gupta, Nima Hafezi-Nejad, Graeme J Hankey, Heather C Harewood, Rasmus Havmoeller, Simon Hay, Lucia Hernandez, Abdullatif Husseini, Bulat T Idrisov, Nayu Ikeda, Farhad Islami, Eiman Jahangir, Simerjot K Jassal, Sun Ha Jee, Mona Jeffreys, Jost B Jonas, Edmond K Kabagambe, Shams Eldin Ali Hassan Khalifa, Andre Pascal Kengne, Yousef Saleh Khader, Young-Ho Khang, Daniel Kim, Ruth W Kimokoti, Jonas M Kinge, Yoshihiro Kokubo, Soewarta Kosen, Gene Kwan, Taavi Lai, Mall Leinsalu, Yichong Li, Xiaofeng Liang, Shiwei Liu, Giancarlo Logroscino, Paulo A Lotufo, Yuan Lu, Jixiang Ma, Nana Kwaku Mainoo, George A Mensah, Tony R Merriman, Ali H Mokdad, Joanna Moschandreas, Mohsen Naghavi, Aliya Naheed, Devina Nand, K M Venkat Narayan, Erica Leigh Nelson, Marian L Neuhouser, Muhammad Imran Nisar, Takayoshi Ohkubo, Samuel O Oti, Andrea Pedroza, Dorairaj Prabhakaran, Nobhojit Roy, Uchechukwu Sampson, Hyeyoung Seo, Sadaf G Sepanlou, Kenji Shibuya, Rahman Shiri, Ivy Shiue, Gitanjali M Singh, Jasvinder A Singh, Vegard Skirbekk, Nicolas J C Stapelberg, Lela Sturua, Bryan L Sykes, Martin Tobias, Bach X Tran, Leonardo Trasande, Hideaki Toyoshima, Steven van de Vijver, Tommi J Vasankari, J Lennert Veerman, Gustavo Velasquez-Melendez, Vasiliy Victorovich Vlassov, Stein Emil Vollset, Theo Vos, Claire Wang, XiaoRong Wang, Elisabete Weiderpass, Andrea Werdecker, Jonathan L Wright, Y Claire Yang, Hiroshi Yatsuya, Jihyun Yoon, Seok-Jun Yoon, Yong Zhao, Maigeng Zhou, Shankuan Zhu, Alan D Lopez, Christopher J L Murray, and Emmanuela Gakidou. 2014. Global, Regional, and National Prevalence of Overweight and Obesity in Children and Adults during 1980-2013: A Systematic Analysis for the Global Burden of Disease Study 2013. *Lancet* 384, 9945 (August 2014), 766–781. https://doi.org/10.1016/S0140-6736(14)60460-8

[304] Donald A. Norman. 1993. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Addison-Wesley Longman Publishing Co., Inc., USA.

[305] Elisabeth Norman, Gerit Pfuhl, Rannveig Grøm Sæle, Frode Svartdal, Torstein Låg, and Tove Irene Dahl. 2019. Metacognition in Psychology. *Review of General Psychology* 23, 4 (December 2019), 403–424. https://doi.org/10.1177/1089268019883821

[306] Florian Nothdurft, Stefan Ultes, and Wolfgang Minker. 2014. Finding Appropriate Interaction Strategies for Proactive Dialogue Systems—an Open Quest. In *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication*, 2014. Citeseer, 73–80.

[307] Shakked Noy and Whitney Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* 381, 6654 (July 2023), 187–192. https://doi.org/10.1126/science.adh2586

[308] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in Data-Driven Artificial Intelligence Systems—An Introductory Survey. *WIREs Data Mining and Knowledge Discovery* 10, 3 (2020), e1356. https://doi.org/10.1002/widm.1356

[309] Andrew Olewnik, Randy Yerrick, Amanda Simmons, Yonghee Lee, and Brian Stuhlmiller. 2020. Defining Open-Ended Problem Solving Through Problem Typology Framework. *International Journal of Engineering Pedagogy (iJEP)* 10, 1 (January 2020), 7–30. https://doi.org/10.3991/ijep.v10i1.11033

[310] Andrew M. Olney, Sidney D'Mello, Natalie Person, Whitney Cade, Patrick Hays, Claire Williams, Blair Lehman, and Arthur Graesser. 2012. Guru: A Computer Tutor That Models Expert Human Tutors. In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (*ITS'12*), June 14, 2012. Springer-Verlag, Berlin, Heidelberg, 256–261. https://doi.org/10.1007/978-3-642-30950-2_32

[311] Eben Otuteye and Mohammad Siddiquee. 2015. Overcoming Cognitive Biases: A Heuristic for Making Value Investing Decisions. *Journal of Behavioral Finance* 16, 2 (April 2015), 140–149. https://doi.org/10.1080/15427560.2015.1034859

[312] Annemarie Sullivan Palincsar. 1986. The Role of Dialogue in Providing Scaffolded Instruction. *Educational Psychologist* 21, 1–2 (January 1986), 73–98. https://doi.org/10.1080/00461520.1986.9653025

[313] Hyanghee Park and Joonhwan Lee. 2021. Designing a Conversational Agent for Sexual Assault Survivors: Defining Burden of Self-Disclosure and Envisioning Survivor-Centered Solutions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 06, 2021. ACM, Yokohama Japan, 1–17. https://doi.org/10.1145/3411764.3445133

[314] SoHyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. 2021. "I Wrote as If I Were Telling a Story to Someone I Knew.": Designing Chatbot Interactions for Expressive Writing in Mental Health. In *Designing Interactive Systems Conference 2021* (*DIS '21*), June 28, 2021. Association for Computing Machinery, New York, NY, USA, 926–941. https://doi.org/10.1145/3461778.3462143

[315] Samir Passi and Mihaela Vorvoreanu. 2022. *Overreliance on AI: Literature Review*. Microsoft. Retrieved November 10, 2023 from https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

[316] Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing Human–AI Interaction by Priming Beliefs about AI Can Increase Perceived Trustworthiness, Empathy and Effectiveness. *Nature Machine Intelligence* 5, 10 (October 2023), 1076–1086. https://doi.org/10.1038/s42256-023-00720-7

[317] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (September 2014), 897–908. https://doi.org/10.1145/2632048.2632062

[318] Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. 2019. Design and Evaluation of Service Robot's Proactivity in Decision-Making Support Process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*), May 02, 2019. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300328

[319] Jean Piaget. 1952. *The Origins of Intelligence in Children*. International Universities Press. https://doi.org/10.1037/11494-000

[320] Lara S. G. Piccolo, Martino Mensio, and Harith Alani. 2019. Chasing the Chatbots: Directions for Interaction and Design Research. In *International Conference on Internet Science* (*Lecture Notes in Computer Science*), 2019. 157–169. https://doi.org/10.1007/978-3-030-17705-8_14

[321] Dominik M. Piehlmaier. 2023. The One-Man Show: The Effect of Joint Decision-Making on Investor Overconfidence. *Journal of Consumer Research* 50, 2 (August 2023), 426–446. https://doi.org/10.1093/jcr/ucac054

[322] Peter Pirolli. 2007. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195173321.001.0001

[323] Peter Pirolli and Daniel Russell. 2011. Introduction to This Special Issue on Sensemaking. *Human-Computer Interaction* 26, 1 (January 2011), 1–8. https://doi.org/10.1080/07370024.2011.556557

[324] Janneke van de Pol, Monique Volman, and Jos Beishuizen. 2010. Scaffolding in Teacher–Student Interaction: A Decade of Research. *Educational Psychology Review* 22, 3 (September 2010), 271–296. https://doi.org/10.1007/s10648-010-9127-6

[325] Martin Porcheron, Joel E. Fischer, and Stuart Reeves. 2021. Pulling Back the Curtain on the Wizards of Oz. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (January 2021), 243:1-243:22. https://doi.org/10.1145/3432942

[326] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (*CHI '18*), April 21, 2018. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[327] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (*CSCW '17*), February 25, 2017. Association for Computing Machinery, Portland, Oregon, USA, 207–219. https://doi.org/10.1145/2998181.2998298

[328] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. “Accessibility Came by Accident”: Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (*CHI '18*), April 21, 2018. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174033

[329] Judith J. Prochaska, Erin A. Vogel, Amy Chieng, Matthew Kendra, Michael Baiocchi, Sarah Pajarito, and Athena Robinson. 2021. A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study. *Journal of Medical Internet Research* 23, 3 (March 2021), e24850. https://doi.org/10.2196/24850

[330] Mirjana Prpa, Kıvanç Tatar, Jules Françoise, Bernhard Riecke, Thecla Schiphorst, and Philippe Pasquier. 2018. Attending to Breath: Exploring How the Cues in a Virtual Environment Guide the Attention to Breath and Shape the Quality of Experience to Support Mindfulness. In *Proceedings of the 2018 Designing Interactive Systems Conference* (*DIS '18*), 2018. Association for Computing Machinery, New York, NY, USA, 71–84. https://doi.org/10.1145/3196709.3196765

[331] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. “Alexa Is My New BFF”: Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (*CHI EA '17*), May 06, 2017. Association for Computing Machinery, Denver, Colorado, USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[332] Aung Pyae and Tapani N. Joelsson. 2018. Investigating the Usability and User Experiences of Voice User Interface: A Case of Google Home Smart Speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (*MobileHCI '18*), 2018. Association for Computing Machinery, New York, NY, USA, 127–131. https://doi.org/10.1145/3236112.3236130

[333] Lingyun Qiu and Izak Benbasat. 2009. Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems* 25, 4 (April 2009), 145–182. https://doi.org/10.2753/MIS0742-1222250405

[334] Pernilla Qvarfordt, Arne Jönsson, and Nils Dahlbäck. 2003. The Role of Spoken Feedback in Experiencing Multimodal Interfaces as Human-Like. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (*ICMI '03*), November 05, 2003. Association for Computing Machinery, Vancouver, British Columbia, Canada, 250–257. https://doi.org/10.1145/958432.958478

[335] Rifat Rahman, Md. Rishadur Rahman, Nafis Irtiza Tripto, Mohammed Eunus Ali, Sajid Hasan Apon, and Rifat Shahriyar. 2021. AdolescentBot: Understanding Opportunities for Chatbots in Combating Adolescent Sexual and Reproductive Health Problems in Bangladesh. In *Proceedings of the 2021 CHI Conference on Human Factors in*

Computing Systems, May 06, 2021. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445694

[336] Aditi Ramachandran, Chien-Ming Huang, Edward Gartland, and Brian Scassellati. 2018. Thinking Aloud with a Tutoring Robot to Enhance Learning. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI '18*), February 26, 2018. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3171221.3171250

[337] Byron Reeves and Clifford Ivar Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press, New York, NY, US.

[338] Stuart Reeves, Martin Porcheron, and Joel Fischer. 2018. "This Is Not What We Wanted": Designing for Conversation with Voice Interfaces. *Interactions* 26, 1 (Dezember 2018), 46–51. https://doi.org/10.1145/3296699

[339] Michael Reicherts. 2015. *L'entretien Psychologique et Le Counselling. De l'approche Centrée Sur La Personne Aux Interventions Ciblées*. Edition ZKS-Verlag, Coburg, Germany. Retrieved from https://www.researchgate.net/publication/242655672_L

[340] Michael Reicherts. 2022. *Dimensions of Openness to Emotions (DOE). A Model of Affect Processing. Manual with Instruments, Recent Studies and Reference Values (Technical Report 168-C, 2022)*. University of Fribourg, Fribourg, Switzerland. https://doi.org/10.13140/RG.2.2.28225.02401

[341] Brian J. Reiser. 2002. Why Scaffolding Should Sometimes Make Tasks More Difficult for Learners. In *Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community* (*CSCL '02*), 2002. International Society of the Learning Sciences, 255–264. Retrieved September 12, 2022 from http://dl.acm.org/citation.cfm?id=1658616.1658652

[342] Brian J. Reiser. 2004. Scaffolding Complex Learning: The Mechanisms of Structuring and Problematizing Student Work. *Journal of the Learning Sciences* 13, 3 (2004), 273–304. https://doi.org/10.1207/s15327809jls1303_2

[343] René Reitsma. 2019. The Future of Data: Too Much Visualization, Too Little Understanding? *Dialectic* 2, 2 (Summer 2019). https://doi.org/10.3998/dialectic.14932326.0002.207

[344] Maryam Rezaie, Melanie Tory, and Sheelagh Carpendale. 2024. Struggles and Strategies in Understanding Information Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 30, 6 (April 2024), 3035–3048. https://doi.org/10.1109/TVCG.2024.3388560

[345] Richard Riding and Stephen Rayner. 2013. *Cognitive Styles and Learning Strategies: Understanding Style Differences in Learning and Behavior*. David Fulton Publishers, London. https://doi.org/10.4324/9781315068015

[346] Evan F. Risko and Sam J. Gilbert. 2016. Cognitive Offloading. *Trends in Cognitive Sciences* 20, 9 (September 2016), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

[347] A. Joy Rivera. 2014. A Socio-Technical Systems Approach to Studying Interruptions: Understanding the Interrupter's Perspective. *Applied Ergonomics* 45, 3 (May 2014), 747–756. https://doi.org/10.1016/j.apergo.2013.08.009

[348] S. Ian Robertson. 1999. *Types of Thinking*. Routledge, London. https://doi.org/10.4324/9780203754634

[349] Carol Rodgers. 2002. Defining Reflection: Another Look at John Dewey and Reflective Thinking. *Teachers College Record* 104, 4 (April 2002), 842–866. https://doi.org/10.1111/1467-9620.00181

[350] Neal J. Roese and Kathleen D. Vohs. 2012. Hindsight Bias. *Perspectives on Psychological Science* 7, 5 (September 2012), 411–426. https://doi.org/10.1177/1745691612454303

[351] Yvonne Rogers. 2004. New Theoretical Approaches for Human-Computer Interaction. *Annual Review of Information Science and Technology* 38, 1 (2004), 87–143. https://doi.org/10.1002/aris.1440380103

[352] Yvonne Rogers, Helen Sharp, and Jennifer Preece. 2023. *Interaction Design: Beyond Human-Computer Interaction* (6th ed.). John Wiley & Sons.

[353] Donya Rooein. 2019. Data-Driven Edu Chatbots. In *Companion Proceedings of The 2019 World Wide Web Conference* (*WWW '19*), May 13, 2019. Association for Computing Machinery, San Francisco, USA, 46–49. https://doi.org/10.1145/3308560.3314191

[354] Barak Rosenshine and Carla Meister. 1992. The Use of Scaffolds for Teaching Higher-Level Cognitive Strategies. *Educational Leadership* 49, 7 (1992), 26–33.

[355] Steven I. Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D. Weisz. 2023. The Programmer's Assistant: Conversational Interaction with a Large Language Model for Software Development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (*IUI '23*), March 27, 2023. Association for Computing Machinery, New York, NY, USA, 491–514. https://doi.org/10.1145/3581641.3584037

[356] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-Based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*), May 02, 2019. Association for Computing Machinery, Glasgow, Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300587

[357] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The Cost Structure of Sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (*CHI '93*), May 01, 1993. Association for

Computing Machinery, New York, NY, USA, 269–276. https://doi.org/10.1145/169059.169209

[358] Rainer Sachse. 2002. Zielorientierte Gesprächspsychotherapie. In *Die Vielen Gesichter Der Personzentrierten Psychotherapie*, Wolfgang W. Keil and Gerhard Stumm (eds.). Springer, Germany, 265–284. https://doi.org/10.1007/978-3-7091-6733-5_11

[359] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50, 4 (1974), 696–735. https://doi.org/10.2307/412243

[360] María Consuelo Sáiz-Manzanares, Raúl Marticorena-Sánchez, Luis Jorge Martín-Antón, Irene González Díez, and Leandro Almeida. 2023. Perceived Satisfaction of University Students with the Use of Chatbots as a Tool for Self-Regulated Learning. *Heliyon* 9, 1 (January 2023), e12843. https://doi.org/10.1016/j.heliyon.2023.e12843

[361] Ayshwarya Saktheeswaran, A. Srinivasan, and J. Stasko. 2020. Touch? Speech? Or Touch and Speech? Investigating Multimodal Interaction for Visual Network Exploration and Analysis. *IEEE Transactions on Visualization and Computer Graphics* (2020). https://doi.org/10.1109/TVCG.2020.2970512

[362] Mike Scaife and Yvonne Rogers. 1996. External Cognition: How Do Graphical Representations Work? *International Journal of Human-Computer Studies* 45, 2 (1996), 185–213. https://doi.org/10.1006/ijhc.1996.0048

[363] Johanna Schmidhuber, Stephan Schlögl, and Christian Ploder. 2021. Cognitive Load and Productivity Implications in Human-Chatbot Interaction. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, 2021. IEEE, 1–6.

[364] Albrecht Schmidt. 2017. Augmenting Human Intellect and Amplifying Perception and Cognition. *IEEE Pervasive Computing* 16, 1 (January 2017), 6–10. https://doi.org/10.1109/MPRV.2017.8

[365] Albrecht Schmidt. 2020. Interactive Human Centered Artificial Intelligence: A Definition and Research Challenges. In *Proceedings of the International Conference on Advanced Visual Interfaces* (*AVI '20*), September 28, 2020. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3399715.3400873

[366] Albrecht Schmidt, Fosca Giannotti, Wendy Mackay, Ben Shneiderman, and Kaisa Väänänen. 2021. Artificial Intelligence for Humankind: A Panel on How to Create Truly Interactive and Human-Centered AI for the Benefit of Individuals and Society. In *Human-Computer-Interaction – INTERACT 2021* (*Lecture Notes in Computer Science*), 2021. Springer International Publishing, Cham, 335–339. https://doi.org/10.1007/978-3-030-85607-6_32

[367] Maria Schmidt and Patricia Braunger. 2018. A Survey on Different Means of Personalized Dialog Output for an Adaptive Personal Assistant. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (*UMAP '18*), July

02, 2018. Association for Computing Machinery, New York, NY, USA, 75–81. https://doi.org/10.1145/3213586.3226198

[368] Maria Schmidt, Wolfgang Minker, and Steffen Werner. 2020. User Acceptance of Proactive Voice Assistant Behavior. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, 2020. TUDpress, Dresden, 18–25.

[369] Maria Schmidt, Wolfgang Minker, and Steffen Werner. 2020. How Users React to Proactive Voice Assistant Behavior While Driving. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, May 2020. European Language Resources Association, Marseille, France, 485–490. Retrieved October 24, 2023 from https://aclanthology.org/2020.lrec-1.61

[370] Maria Schmidt, Daniela Stier, Steffen Werner, and Wolfgang Minker. 2019. Exploration and Assessment of Proactive Use Cases for an In-Car Voice Assistant. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, 2019. TUDpress, Dresden, Dresden, Germany, 148–155.

[371] Sofia Schöbel, Anuschka Schmitt, Dennis Benner, Mohammed Saqr, Andreas Janson, and Jan Marco Leimeister. 2023. Charting the Evolution and Future of Conversational Agents: A Research Agenda Along Five Waves and New Frontiers. *Information Systems Frontiers* (April 2023). https://doi.org/10.1007/s10796-023-10375-9

[372] Donald A. Schön. 1987. *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions*. Jossey-Bass, San Francisco, CA, US.

[373] Donald A. Schön. 1992. *The Reflective Practicioner: How Professionals Think in Action*. Routledge, Abingdon, Oxfordshire. Retrieved from https://doi.org/10.4324/9781315237473

[374] Gregory Schraw, Michael E. Dunkle, and Lisa D. Bendixen. 1995. Cognitive Processes in Well-Defined and Ill-Defined Problem Solving. *Applied Cognitive Psychology* 9, 6 (1995), 523–538. https://doi.org/10.1002/acp.2350090605

[375] Ralf Schwarzer and Matthias Jerusalem. 1995. Generalized Self-Efficacy Scale. In *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs*. NFER, Windsor, UK, 35–37.

[376] Martin E. P. Seligman and Mihaly Csikszentmihalyi. 2014. Positive Psychology: An Introduction. In *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*, Mihaly Csikszentmihalyi (ed.). Springer Netherlands, Netherlands, 279–298. https://doi.org/10.1007/978-94-017-9088-8_18

[377] Abigail Sellen and Eric Horvitz. 2024. The Rise of the AI Co-Pilot: Lessons for Design from Aviation and Beyond. *Communications of the ACM* 67, 7 (July 2024), 18–23. https://doi.org/10.1145/3637865

[378] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is Now A Good Time? An Empirical Study of Vehicle-Driver Communication

Timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*), May 02, 2019. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300867

[379] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. 2016. Eviza: A Natural Language Interface for Visual Analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (*UIST '16*), October 16, 2016. Association for Computing Machinery, New York, NY, USA, 365–377. https://doi.org/10.1145/2984511.2984588

[380] Vidya Setlur, Melanie Tory, and Alex Djalali. 2019. Inferencing Underspecified Natural Language Utterances in Visual Analysis. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (*IUI '19*), March 17, 2019. Association for Computing Machinery, New York, NY, USA, 40–51. https://doi.org/10.1145/3301275.3302270

[381] Lubna Bte Iskhandar Shah, Samantha Torres, Premarani Kannusamy, Cecilia Mui Lee Chng, Hong-Gu He, and Piyanee Klainin-Yobas. 2015. Efficacy of the Virtual Reality-Based Stress Management Program on Stress-Related Variables in People with Mood Disorders: The Feasibility Study. *Archives of Psychiatric Nursing* 29, 1 (2015), 6–13.

[382] Amy M. Shapiro. 2008. Hypermedia Design as Learner Scaffolding. *Educational Technology Research and Development* 56, 1 (February 2008), 29–44. https://doi.org/10.1007/s11423-007-9063-4

[383] Suraj Sharma, Joseph Brennan, and Jason Nurse. 2021. StockBabble: A Conversational Financial Agent to Support Stock Market Investors. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (*CUI '21*), July 27, 2021. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3469595.3469620

[384] Stuart Shaw, Martina Kuvalja, and Irenka Suto. 2018. An Exploration of the Nature and Assessment of Student Reflection. *Research Matters: A Cambridge Assessment Publication* (2018), 2–8.

[385] Bayan Abu Shawar and Eric Atwell. 2007. Fostering Language Learner Autonomy through Adaptive Conversation Tutors. In *Proceedings of the the Fourth Corpus Linguistics Conference*, July 27, 2007. Birmingham, UK, 186–193.

[386] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (April 2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[387] John Short, Ederyn Williams, and Bruce Christie. 1976. *The Social Psychology of Telecommunications*. Wiley.

[388] Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick. 1999. Does Automation Bias Decision-Making? *International Journal of Human-Computer Studies* 51, 5 (November 1999), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

[389] Sofie Smedegaard Skov, Josefine Ranfelt Andersen, Sigurd Lauridsen, Mads Bab, Marianne Bundsbæk, and Maj Britt Dahl Nielsen. 2022. Designing a Conversational Agent to Promote Teamwork and Collaborative Practices Using Design Thinking: An Explorative Study on User Experiences. *Frontiers in Psychology* 13, (2022). Retrieved November 15, 2023 from https://www.frontiersin.org/articles/10.3389/fpsyg.2022.903715

[390] Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (*CHI '17*), May 02, 2017. Association for Computing Machinery, New York, NY, USA, 2696–2707. https://doi.org/10.1145/3025453.3025516

[391] Paul R. Smart. 2012. The Web-Extended Mind. *Metaphilosophy* 43, 4 (2012), 446–463. https://doi.org/10.1111/j.1467-9973.2012.01756.x

[392] Pavel Smutny and Petra Schreiberova. 2020. Chatbots for Learning: A Review of Educational Chatbots for the Facebook Messenger. *Computers & Education* 151, (July 2020), 103862. https://doi.org/10.1016/j.compedu.2020.103862

[393] Victoria Smy, Marie Cahillane, and Piers MacLean. 2016. Sensemaking and Metacognitive Prompting in Ill-Structured Problems. In *The International Journal of Information and Learning Technology*, June 06, 2016. 186–199. https://doi.org/10.1108/IJILT-10-2015-0027

[394] Timothy Sohn, Kevin A. Li, Gunny Lee, Ian Smith, James Scott, and William G. Griswold. 2005. Place-Its: A Study of Location-Based Reminders on Mobile Phones. In *Proceedings of the 7th International Conference on Ubiquitous Computing* (*UbiComp'05*), 2005. Springer-Verlag, Berlin, Heidelberg, 232–250. https://doi.org/10.1007/11551201_14

[395] Donggil Song, Eun Young Oh, and Marilyn Rice. 2017. Interacting with a Conversational Agent System for Educational Purposes in Online Courses. In *2017 10th International Conference on Human System Interactions (HSI)*, July 2017. 78–82. https://doi.org/10.1109/HSI.2017.8005002

[396] Arjun Srinivasan, Bongshin Lee, Nathalie Henry Riche, Steven M. Drucker, and Ken Hinckley. 2020. InChorus: Designing Consistent Multimodal Interactions for Data Visualization on Tablet Devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (*CHI '20*), April 21, 2020. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376782

[397] Arjun Srinivasan and John Stasko. 2017. Natural Language Interfaces for Data Analysis with Visualization: Considering What Has and Could Be Asked. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers* (*EuroVis '17*), June 12, 2017. Eurographics Association, Goslar, DEU, 55–59. https://doi.org/10.2312/eurovisshort.20171133

[398] Arjun Srinivasan and John Stasko. 2018. Orko: Facilitating Multimodal Interaction for Visual Exploration and Analysis of Networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (January 2018), 511–521. https://doi.org/10.1109/TVCG.2017.2745219

[399] Ursula M. Staudinger. 2001. Life Reflection: A Social-Cognitive Analysis. *Review of General Psychology* 5, 2 (2001), 148–160. https://doi.org/10.1037/1089-2680.5.2.148

[400] Kalliopi-Evangelia Stavroulia and Andreas Lanitis. 2019. Enhancing Reflection and Empathy Skills via Using a Virtual Reality Based Learning Framework. *International Journal of Emerging Technologies in Learning (iJET)* 14, (April 2019), 18. https://doi.org/10.3991/ijet.v14i07.9946

[401] Petra-Maria Strauß and Wolfgang Minker. 2010. *Proactive Spoken Dialogue Interaction in Multi-Party Environments*. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4419-5992-8

[402] Marcelo Suarez-Orozco and Desirée Qin-Hilliard. 2004. *Globalization: Culture and Education in the New Millennium*. University of California Press.

[403] Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. 2010. Articulate: A Semi-Automated Model for Translating Natural Language Queries into Meaningful Visualizations. In *Smart Graphics*, 2010. Springer Berlin Heidelberg, Berlin, Heidelberg, 184–195.

[404] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (*CHI '19*), 2019. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300833

[405] Vivian Ta, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, and Alexia Loggarakis. 2020. User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. *Journal of Medical Internet Research* 22, 3 (March 2020), e16235. https://doi.org/10.2196/16235

[406] Roderick Tabalba, Nurit Kirshenbaum, Jason Leigh, Abari Bhatacharya, Andrew Johnson, Veronica Grosso, Barbara Di Eugenio, and Moira Zellner. 2022. Articulate+ : An Always-Listening Natural Language Interface for Creating Data Visualizations. In *Proceedings of the 4th Conference on Conversational User Interfaces* (*CUI '22*), September 15, 2022. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3543829.3544534

[407] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2019. Investigating Users' Preferences and Expectations for Always-Listening Voice Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (December 2019), 153:1-153:23. https://doi.org/10.1145/3369807

[408] Hao Tan, Ying Zhao, Shiyan Li, Wei Wang, Ming Zhu, Jie Hong, and Xiang Yuan. 2020. Relationship between Social Robot Proactive Behavior and the Human Perception of Anthropomorphic Attributes. *Advanced Robotics* 34, 20 (October 2020), 1324–1336. https://doi.org/10.1080/01691864.2020.1831699

[409] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (*CHI '24*), May 11, 2024. Association for Computing Machinery, New York, NY, USA, 1–24. https://doi.org/10.1145/3613904.3642902

[410] Stergios Tegos and Stavros Demetriadis. 2017. Conversational Agents Improve Peer Learning through Building on Prior Knowledge. *Journal of Educational Technology & Society* 20, 1 (2017), 99–111.

[411] Peter M. Todd and Gerd Gigerenzer. 2007. Environments That Make Us Smart: Ecological Rationality. *Current Directions in Psychological Science* 16, 3 (June 2007), 167–171. https://doi.org/10.1111/j.1467-8721.2007.00497.x

[412] Toyin Tofade, Jamie Elsner, and Stuart T. Haines. 2013. Best Practice Strategies for Effective Use of Questions as a Teaching Tool. *American Journal of Pharmaceutical Education* 77, 7 (September 2013), 155. https://doi.org/10.5688/ajpe777155

[413] Melanie Tory and Vidya Setlur. 2019. Do What I Mean, Not What I Say! Design Considerations for Supporting Intent and Context in Analytical Conversation. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, October 2019. 93–103. https://doi.org/10.1109/VAST47406.2019.8986918

[414] Paul D. Trapnell and Jennifer D. Campbell. 1999. Private Self-Consciousness and Five-Factor Model of Personality: Distinguishing Rumination from Reflection. *Journal of Personality and Social Psychology* 76, 2 (1999), 284.

[415] Novrman Triplett. 1898. The Dynamogenic Factors in Pacemaking and Competition. *The American Journal of Psychology* 9, 4 (1898), 507–533. https://doi.org/10.2307/1412188

[416] Yin-Te Tsai and Wei-An Lin. 2018. Design of an Intelligent Cognition Assistant for People with Cognitive Impairment. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, June 2018. 1207–1212. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00203

[417] John W. Tukey. 1980. We Need Both Exploratory and Confirmatory. *The American Statistician* 34, 1 (1980), 23–25. https://doi.org/10.2307/2682991

[418] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2015. Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and*

Ubiquitous Computing (*UbiComp '15*), September 07, 2015. Association for Computing Machinery, New York, NY, USA, 801–812. https://doi.org/10.1145/2750858.2807514

[419] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (September 1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

[420] Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S. Kashavan, and John Blake Torous. 2019. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie* 64, 7 (July 2019), 456–464. https://doi.org/10.1177/0706743719828977

[421] Ronald D. Vale. 2013. The Value of Asking Questions. *Molecular Biology of the Cell* 24, 6 (March 2013), 680–682. https://doi.org/10.1091/mbc.E12-09-0660

[422] Kurt Vanlehn, Arthur C. Graesser, G. Tanner Jackson, Pamela Jordan, Andrew Olney, and Carolyn P. Rosé. 2007. When Are Tutorial Dialogues More Effective than Reading? *Cognitive Science* 31, 1 (February 2007), 3–62. https://doi.org/10.1080/03640210709336984

[423] Konstantina Vasileiou, Julie Barnett, Susan Thorpe, and Terry Young. 2018. Characterising and Justifying Sample Size Sufficiency in Interview-Based Studies: Systematic Analysis of Qualitative Health Research over a 15-Year Period. *BMC Medical Research Methodology* 18, 1 (November 2018), 148. https://doi.org/10.1186/s12874-018-0594-7

[424] Laura Villa, Ramón Hervás, Dagoberto Cruz-Sandoval, and Jesús Favela. 2023. Design and Evaluation of Proactive Behavior in Conversational Assistants: Approach with the Eva Companion Robot. In *Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2022)* (*Lecture Notes in Networks and Systems*), 2023. Springer International Publishing, Cham, 234–245. https://doi.org/10.1007/978-3-031-21333-5_23

[425] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (*CHI '21*), 2021. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445536

[426] Lev Vygotsky. 1962. *Thought and Language*. MIT Press, Cambridge, MA, US. https://doi.org/10.1037/11193-000

[427] Lev Vygotsky. 1978. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

[428] Nadine Wagener, Tu Dinh Duong, Johannes Schöning, Yvonne Rogers, and Jasmin Niess. 2021. The Role of Mobile and Virtual Reality Applications to Support Well-

Being: An Expert View and Systematic App Review. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part IV 18*, 2021. Springer, Cham., 262–283. https://doi.org/10.1007/978-3-030-85610-6_16

[429] Nadine Wagener, Jasmin Niess, Yvonne Rogers, and Johannes Schöning. 2022. Mood Worlds: A Virtual Environment for Autonomous Emotional Expression. In *CHI Conference on Human Factors in Computing Systems* (*CHI '22*), April 27, 2022. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3491102.3501861

[430] Thiemo Wambsganss, Sebastian Guggisberg, and Matthias Söllner. 2021. ArgueBot: A Conversational Agent for Adaptive Argumentation Feedback. In *Innovation Through Information Systems*, 2021. Springer International Publishing, Cham, 267–282. https://doi.org/10.1007/978-3-030-86797-3_18

[431] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (*CHI '21*), May 06, 2021. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445781

[432] Thiemo Wambsganss, Naim Zierau, Matthias Söllner, Tanja Käser, Kenneth R. Koedinger, and Jan Marco Leimeister. 2022. Designing Conversational Evaluation Tools: A Comparison of Text and Voice Modalities to Improve Response Quality in Course Evaluations. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (November 2022), 506:1-506:27. https://doi.org/10.1145/3555619

[433] Hong Wang, Weizhi Wang, Rajan Saini, Marina Zhukova, and Xifeng Yan. 2023. Gauchochat: Towards Proactive, Controllable, and Personalized Social Conversation. *Alexa Prize SocialBot Grand Challenge 5 Proceedings* 5, 1 (2023). Retrieved from https://www.amazon.science/alexa-prize/proceedings/gauchochat-towards-proactive-controllable-and-personalized-social-conversation

[434] QianYing Wang and Clifford Nass. 2005. Less Visible and Wireless: Two Experiments on the Effects of Microphone Type on Users' Performance and Perception. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '05*), April 02, 2005. Association for Computing Machinery, Portland, Oregon, USA, 809–818. https://doi.org/10.1145/1054972.1055086

[435] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology* 54, 6 (June 1988), 1063–1070. https://doi.org/10.1037//0022-3514.54.6.1063

[436] Elizabeth Watt, Maria Murphy, Elizabeth Pascoe, Andrew Scanlon, and Sharon Gan. 2011. An Evaluation of a Structured Learning Programme as a Component of the Clinical Practicum in Final Year Bachelor of Nursing Programme: A Pre–Post-Test

Analysis. *Journal of Clinical Nursing* 20, 15–16 (2011), 2286–2293. https://doi.org/10.1111/j.1365-2702.2010.03621.x

[437] Florian Weber, Thiemo Wambsganss, Dominic Rüttimann, and Matthias Söllner. 2021. Pedagogical Agents for Interactive Lernaing: A Taxonomy of Conversational Agents in Education. *ICIS 2021 Proceedings* (December 2021). Retrieved from https://aisel.aisnet.org/icis2021/diglearn_curricula/diglearn_curricula/13

[438] Martin Weber and Colin F. Camerer. 1998. The Disposition Effect in Securities Trading: An Experimental Analysis. *Journal of Economic Behavior & Organization* 33, 2 (January 1998), 167–184. https://doi.org/10.1016/S0167-2681(97)00089-9

[439] Frank Webster. 2014. *Theories of the Information Society* (4th ed.). Routledge, London. https://doi.org/10.4324/9781315867854

[440] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Developing the Proactive Speaker Prototype Based on Google Home. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, May 08, 2021. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3411763.3451642

[441] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding User Perceptions of Proactive Smart Speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (December 2022), 185:1-185:28. https://doi.org/10.1145/3494965

[442] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2021. "Let Me Explain!": Exploring the Potential of Virtual Agents in Explainable AI Interaction Design. *Journal on Multimodal User Interfaces* 15, 2 (June 2021), 87–98. https://doi.org/10.1007/s12193-020-00332-0

[443] Joseph Weizenbaum. 1966. ELIZA - A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* 9, 1 (January 1966), 36–45. https://doi.org/10.1145/365153.365168

[444] Ming-Hui Wen. 2018. A Conversational User Interface for Supporting Individual and Group Decision-Making in Stock Investment Activities. In *2018 IEEE International Conference on Applied System Invention (ICASI)*, April 2018. 216–219. https://doi.org/10.1109/ICASI.2018.8394571

[445] Christopher D. Wickens, Benjamin A. Clegg, Alex Z. Vieane, and Angelia L. Sebok. 2015. Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors* 57, 5 (August 2015), 728–739. https://doi.org/10.1177/0018720815581940

[446] Rebecca A. Wilkinson and Gioia Chilton. 2013. Positive Art Therapy: Linking Positive Psychology to Art Therapy Theory, Practice, and Research. *Art Therapy* 30, 1 (January 2013), 4–11. https://doi.org/10.1080/07421656.2013.757513

[447] Rainer Winkler, Sebastian Hobert, Tizian Fischer, Antti Salovaara, Matthias Soellner, and Jan Marco Leimeister. 2020. Engaging Learners in Online Video Lectures with

Dynamically Scaffolding Conversational Agents. *ECIS 2020 Research Papers* (June 2020). Retrieved from https://aisel.aisnet.org/ecis2020_rp/97

[448] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 21, 2020. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376781

[449] Rainer Winkler, Matthias Söllner, Maya Lisa Neuweiler, Flavia Conti Rossini, and Jan Marco Leimeister. 2019. Alexa, Can You Help Us Solve This Problem?: How Conversations With Smart Personal Assistant Tutors Increase Task Group Outcomes. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems - CHI EA '19*, 2019. ACM Press, Glasgow, Scotland, UK, 1–6. https://doi.org/10.1145/3290607.3313090

[450] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May 07, 2011. ACM, Vancouver, BC, Canada, 143–146. https://doi.org/10.1145/1978942.1978963

[451] Irmtraud Wolfbauer, Mia Magdalena Bangerl, Katharina Maitz, and Viktoria Pammer-Schindler. 2023. Rebo at Work: Reflecting on Working, Learning, and Learning Goals with the Reflection Guidance Chatbot for Apprentices. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (*CHI EA '23*), April 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3544549.3585827

[452] Irmtraud Wolfbauer, Viktoria Pammer-Schindler, Katharina Maitz, and Carolyn P. Rosé. 2022. A Script for Conversational Reflection Guidance: A Field Study on Developing Reflection Competence With Apprentices. *IEEE Transactions on Learning Technologies* 15, 5 (October 2022), 554–566. https://doi.org/10.1109/TLT.2022.3207226

[453] Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are We There Yet? - A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence* 4, (July 2021). https://doi.org/10.3389/frai.2021.654924

[454] Priscilla N. Y. Wong, Duncan P. Brumby, Harsha Vardhan Ramesh Babu, and Kota Kobayashi. 2019. Voices in Self-Driving Cars Should Be Assertive to More Quickly Grab a Distracted Driver's Attention. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (*AutomotiveUI '19*), 2019. ACM, New York, NY, USA, 165–176. https://doi.org/10.1145/3342197.3344535

[455] David Wood, Jerome S. Bruner, and Gail Ross. 1976. The Role of Tutoring in Problem Solving. *Child Psychology & Psychiatry & Allied Disciplines* 17, 2 (1976), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

[456] Jérémy Wrobel, Ya-Huei Wu, Hélène Kerhervé, Laila Kamali, Anne-Sophie Rigaud, Céline Jost, Brigitte Le Pévédic, and Dominique Duhaut. 2013. Effect of Agent Embodiment on the Elder User Enjoyment of a Game. *ACHI 2013 - The Sixth International Conference on Advances in Computer-Human Interactions*. Retrieved from https://hal.archives-ouvertes.fr/hal-00832097

[457] Jun Xiao, R. Catrambone, and J. Stasko. 2003. Be Quiet? Evaluating Proactive and Reactive User Interface Assistants. In *Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction*, September 01, 2003. IOS Press, Amsterdam, Netherlands. Retrieved October 31, 2023 from https://www.semanticscholar.org/paper/Be-Quiet-Evaluating-Proactive-and-Reactive-User-Xiao-Catrambone/bee7fa1e0aef89ec343e946c48cf6c173e3ab4c3

[458] Ziang Xiao, Sarah Mennicken, Bernd Huber, Adam Shonkoff, and Jennifer Thom. 2021. Let Me Ask You This: How Can a Voice Assistant Elicit Explicit User Feedback? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (October 2021), 388:1-388:24. https://doi.org/10.1145/3479532

[459] Ziang Xiao, Michelle X. Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If I Hear You Correctly: Building and Evaluating Interview Chatbots with Active Listening Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 21, 2020. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376131

[460] Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-Ended Questions. *ACM Transactions on Computer-Human Interaction* 27, 3 (June 2020), 15:1-15:37. https://doi.org/10.1145/3381804

[461] Wei Xu. 2019. Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *Interactions* 26, 4 (June 2019), 42–46. https://doi.org/10.1145/3328485

[462] Hatice Yildiz Durak. 2022. Conversational Agent-Based Guidance: Examining the Effect of Chatbot Usage Frequency and Satisfaction on Visual Design Self-Efficacy, Engagement, Satisfaction, and Learner Autonomy. *Education and Information Technologies* (July 2022). https://doi.org/10.1007/s10639-022-11149-7

[463] Neil Yorke-Smith, Shahin Saadati, Karen L. Myers, and David N. Morley. 2012. The Design of a Proactive Personal Agent for Task Management. *International Journal on Artificial Intelligence Tools* 21, 01 (February 2012), 1250004. https://doi.org/10.1142/S0218213012500042

[464] Nicola Yuill and Yvonne Rogers. 2012. Mechanisms for Collaboration: A Design and Evaluation Framework for Multi-User Interfaces. *ACM Transactions on Computer-Human Interaction* 19, 1 (May 2012), 1:1-1:25. https://doi.org/10.1145/2147783.2147784

[465] Nicola Yuill, Yvonne Rogers, and Jochen Rick. 2013. Pass the iPad: Collaborative Creating and Sharing in Family Groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '13*), April 27, 2013. Association for Computing Machinery, Paris, France, 941–950. https://doi.org/10.1145/2470654.2466120

[466] Syed Aliya Zahera and Rohit Bansal. 2018. Do Investors Exhibit Behavioral Biases in Investment Decision Making? A Systematic Review. *Qualitative Research in Financial Markets* 10, 2 (May 2018), 210–251. https://doi.org/10.1108/QRFM-04-2017-0028

[467] Emily H. van Zee, Marletta Iwasyk, Akiko Kurose, Dorothy Simpson, and Judy Wild. 2001. Student and Teacher Questioning during Conversations about Science. *Journal of Research in Science Teaching* 38, 2 (2001), 159–190. https://doi.org/10.1002/1098-2736(200102)38:2<159::AID-TEA1002>3.0.CO;2-J

[468] Bolin Zhang, Zhiying Tu, Yangqin Jiang, Shufan He, Guoqing Chao, Dianhui Chu, and Xiaofei Xu. 2021. DGPF:A Dialogue Goal Planning Framework for Cognitive Service Conversational Bot. In *2021 IEEE International Conference on Web Services (ICWS)*, September 2021. 335–340. https://doi.org/10.1109/ICWS53863.2021.00051

[469] Jiaje Zhang and Donald A. Norman. 1994. Representations in Distributed Cognitive Tasks. *Cognitive Science* 18, 1 (January 1994), 87–122. https://doi.org/10.1016/0364-0213(94)90021-3

[470] Rui Zhang, Stephen North, and Eleftherios Koutsofios. 2010. A Comparison of Speech and GUI Input for Navigation in Complex Visualizations on Mobile Devices. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services* (*MobileHCI '10*), September 07, 2010. Association for Computing Machinery, Lisbon, Portugal, 357–360. https://doi.org/10.1145/1851600.1851665

[471] Yutao Zhu, Jian-Yun Nie, Kun Zhou, Pan Du, Hao Jiang, and Zhicheng Dou. 2021. Proactive Retrieval-Based Chatbots Based on Relevant Knowledge and Goals. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '21*), July 11, 2021. Association for Computing Machinery, New York, NY, USA, 2000–2004. https://doi.org/10.1145/3404835.3463011

[472] Naim Zierau, Edona Elshan, Camillo Visini, and Andreas Janson. 2020. A Review of the Empirical Literature on Conversational Agents and Future Research Directions. *ICIS 2020 Proceedings* (December 2020). Retrieved from https://aisel.aisnet.org/icis2020/hci_artintel/hci_artintel/5

[473] Barry J. Zimmerman. 2001. Theories of Self-Regulated Learning and Academic Achievement: An Overview and Analysis. In *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives* (2nd ed.). Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 1–37.

[474] Barry J. Zimmerman. 2002. Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice* (May 2002). https://doi.org/10.1207/s15430421tip4102_2

[475] Manuela Züger and Thomas Fritz. 2015. Interruptibility of Software Developers and Its Prediction Using Psycho-Physiological Sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (*CHI '15*), April 18, 2015. Association for Computing Machinery, New York, NY, USA, 2981–2990. https://doi.org/10.1145/2702123.2702593

[476] Janet Mannheimer Zydney. 2012. Scaffolding. In *Encyclopedia of the Sciences of Learning*, Norbert M. Seel (ed.). Springer US, Boston, MA, 2913–2916. https://doi.org/10.1007/978-1-4419-1428-6_1103