



Title	Using Speech to Interrupt Complex Tasks
Authors(s)	Edwards, Justin
Publication date	2023
Publication information	Edwards, Justin. "Using Speech to Interrupt Complex Tasks." University College Dublin. School of Information and Communication Studies, 2023.
Publisher	University College Dublin. School of Information and Communication Studies
Item record/more information	http://hdl.handle.net/10197/25387

Downloaded 2024-05-27 11:00:10

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)



© Some rights reserved. For more information



UCD College of Social Sciences and Law
School of Information and Communication Studies

Using Speech to Interrupt Complex Tasks

*Understanding Human Spoken Interruptions and
Designing Interruptions for Proactive Speech Agents*

Justin Edwards

UCD student no. 18203265

The thesis is submitted to University College Dublin in fulfilment of the requirements for the degree of Doctor of Philosophy in the College of Social Sciences and Law.

Supervisors

Assoc. Prof. Benjamin Cowan, School of Information and Communication Studies
Prof. Julie Berndsen, School of Computer Science

Research Studies Panel members

Prof. Holly P. Branigan
Dr. Sandy J. J. Gould
Asst. Prof. Christian P. Janssen
Asst. Prof. Marguerite Barry

Head of School

Prof. Eugenia Siapera

August 12, 2023

Declaration of Original Authorship

This thesis is submitted to University College Dublin in fulfilment of the requirements for the degree of Doctor of Philosophy. I hereby certify that the submitted work is my own work, that it was completed while registered as a candidate for the degree stated on the title page, and that I have not obtained a degree elsewhere on the basis of the research presented in this submitted work.

Dublin, August 12, 2023

Justin B. Edwards

External Contributions

Dr. Diego Garaialde:

- Coding support (Chapters 3, 4, and 5)
- Second rater - content analysis (Chapter 5)

Dr. Daniel Rough:

- Coding support (Chapters 3, 4, and 5)

Dr. Philip R. Doyle:

- Pre-publication access to and support in administering Partner Model Questionnaire (Chapter 6)

Yunhan Wu:

- Second rater - thematic analysis (Chapter 4)

Anand Hemmady:

- Materials preparation - Tetris gameplay (Chapters 4, 5, and 6)

Allison Perrone:

- Materials preparation - video editing (Chapter 6)

Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

The work herein would not have been possible without support from several people. Chiefly among them, I want to thank my primary supervisor, Benjamin Cowan. Ben has been a model mentor and friend throughout the course of my PhD, supporting my career development, but always doing so within the broader context of supporting me as a person in ways that transcend our roles as supervisor and student. His scientific acumen and professional advice have been quite valuable to me, but his generosity and compassion are invaluable.

I likewise was well-supported throughout this PhD by my second supervisor, Julie Berndsen, as well as my Doctoral Studies Panel members Holly Brannigan, Christian Janssen, Sandy Gould, and Marguerite Barry. With their diverse perspectives and deep expertise across a variety of topics, I always had access to good advice and honest feedback for every decision throughout my thesis. Most importantly, the collaborative environment they each helped to foster through our meetings and correspondence helped me find what I love about science as a profession.

My labmates were an incredible source of both professional and personal support throughout the thesis process. Postdocs Leigh, Dan, and Iona consistently went well beyond their own responsibilities to lend a hand or offer advice when I needed it. The PhD students who came before me - Diego, Odile, and Phil - cleared the trail for me and made my own journey much clearer. They each modelled for me what it means to do research, to be a PhD student, and to complete a thesis, without them going first, I most certainly would have gotten lost along the way. And the PhD students who came after - Yunhan, Anna, Orla, Paola, and Rhys - I can't express how much I appreciate that you always included me in your projects and in your lives. As fulfilling as it has been to call each of you my colleague, it's been far more meaningful to call you all my friends. While we've had plenty of fun debugging JS Psych scripts and complaining about teaching responsibilities, it's the nights at Clonskeagh House, the

Zoom birthday parties, and the time spent doing anything other than work that I'll most remember from this time of my life.

The other staff and students of ICS also have my sincere thanks. Mutual reassurance throughout our PhDs from Louise, fair and compassionate school leadership from Eugenia, sharing snacks with Mai, and inheriting this phenomenal LaTeX template from Giulia all amount to making my time in the school more manageable and joyous. MSc students who I had the pleasure of supervising and collaborating with - most notably Helin, Sophie, and Suchithra - taught me as much or more as I ever taught them, and I'm lucky to have crossed paths with them during our studies. And Claire Nolan's dry wit and tireless efforts keep the whole school running - indeed all the teaching and research that our school produces should include an acknowledgement of her contribution.

Friends back home were a tremendous source of support and stress relief throughout my PhD. Late night calls with Matt, Mario, and Allison throughout the pandemic kept us from complete despair, even as we sorted every imaginable category of people, places and things into tiers. Allison in particular, keeping a (semi) weekly schedule with me of recording a podcast and catching up about each others' lives gave me a structure and a rhythm in an unstructured period of my life, and helped me to keep my personal and professional passions loosely threaded together with bits and banter. To my college friend group - Connor, Anand, Jim, Ezra, and Schuyler - I'm so grateful that we've stayed in each other's lives, and spending time with each of you throughout my PhD was a great way to relive the memes and dreams we shared back in the purple bubble.

My studies and travel never would have been possible in the first place without the support of my family. My grandparents always showed me that the possibilities for my life were limitless, in no small part because I knew they'd always help me to get wherever I was going. This is fundamentally a PhD about technology and quantitative data, so the fascination with computer that my grandfather encouraged in me from as soon as I could talk and the strange familial attachment we both have to spreadsheets set me on this course as much as anything else. The support and encouragement from my parents has likewise been critical in undertaking a PhD and moving to Europe to do it. The lifetime of hard work and sacrifice that they have shown me, all so I could travel around and - as far as they can describe it - do "something related to AI" is a

debt I can only repay by continuing to have the adventures you've always encouraged. Rachel also has my thanks for being my best friend for over 24 years now, speaking our strange sibling language that nobody else would be able to understand. I am glad we got to live together for a year of my PhD, I am proud of your own accomplishments as you earn your veterinary degree, and I am grateful that we're always around for each other, whether we've had too much sleep or not enough.

Finally, I give my most sincere gratitude to my husband Mikael. Your love, support, and encouragement for me throughout my PhD made the whole thing bearable. We had to endure the loneliest parts of the pandemic with a lot of distance between us, but you helped me to never feel alone and to always feel like I mattered. Closing that distance and moving in with you, building our lives together, that is what kept me working at this PhD and getting out of bed at all. Coffee dates, walks with Mutteri, and mökki visits have the magical property of fixing exactly what research and teaching break in me. Thank you for always accepting me, in whatever condition I'm in, and for helping me remember all the other important parts of life outside of a PhD.

Publications

Thesis Research

Edwards, J., Janssen, C.P., Gould, S.J.J., Garaialde, D., Wu, Y. & Cowan, B.R.. (2023). Spoken Interruptions of Complex Tasks: The Effects of Urgency, Interruptibility, and Task Difficulty. *Journal of the Human Factors and Ergonomics Society* (in preparation).

Edwards, J., Janssen, C.P., Gould, S.J.J., & Cowan, B.R.. (2021). Eliciting Spoken Interruptions to Inform Proactive Speech Agent Design. In *CUI 2021-3rd Conference on Conversational User Interfaces*, 1–12.

<https://doi.org/10.1145/3469595.3469618>

Clark, L., Pantidi, N., Cooney, O., Doyle, P.R., Garaialde, D., **Edwards, J.**, Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., and others. (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*, 1–12. *Honorable mention*.

<https://doi.org/10.1145/3290605.3300705>

Clark, L., Doyle, P.R., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., **Edwards, J.**, and others. (2019). The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers* 31, 4, 349–371.

<https://doi.org/10.1093/iwc/iwz016>

Other Research

Peña, P. R., Doyle, P., **Edwards, J.**, Garaialde, D., Rough, D., Bleakley, A., ... & Cowan, B. R. (2023). Audience design and egocentrism in reference production during human-computer dialogue. *International Journal of Human-Computer Studies*, 103058.

<https://doi.org/10.1016/j.ijhcs.2023.103058>

O'Neill, N., Mercille, J., & **Edwards, J.** (2023) Home care workers' views of employment conditions: private for-profit vs public and non-profit providers in Ireland. *International Journal of Sociology and Social Policy*.

<https://doi.org/10.1108/IJSSP-10-2022-0276>

Torggler, A., **Edwards, J.**, & Wintersberger, P.. (2022). Beyond the Halo: Investigation

of Trust and Functional Specificity in Automated Driving with Conversational Agents. In 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 195-203.

<https://doi.org/10.1145/3543174.3546834>

Wu, Y., Porcheron M., Doyle, P.R., **Edwards, J.**, Rough, D., Cooney, R., Bleakley, A., Clark, L. & Cowan, B.R.. (2022). Comparing Command Construction in Native and Non-Native Speaker IPA Interaction through Conversation Analysis In CUI 2022-4th Conference on Conversational User Interfaces, 1-12.

<https://doi.org/10.1145/3543829.3543839>

Cihan, H., Wu, Y., Peña, P., **Edwards, J.**, & Cowan, B.R.. (2022). Bilingual by default: Voice Assistants and the role of code-switching in creating a bilingual user experience. In CUI 2022-4th Conference on Conversational User Interfaces, 1-4.

<https://doi.org/10.1145/3543829.3544511>

Becker, S., Doyle, P.R., & **Edwards, J.** (2022). Embrace your incompetence! Designing appropriate CUI communication through an ecological approach. In CUI 2022-4th Conference on Conversational User Interfaces, 1-5.

<https://doi.org/10.1145/3543829.3544531>

Mercille, J., **Edwards, J.**, & O'Neill, N.. (2022). Home Care Professionals' Views On Working Conditions During The Covid-19 Pandemic: The Case Of Ireland. *International Journal of Care and Caring*.

<https://doi.org/10.1332/239788221X16345464319417>

Bleakley, A., Rough, D., **Edwards, J.**, Doyle, P.R., Dumbleton, O., Clark, L., Sean Rintel, Wade, V., & Cowan, B.R.. (2021). Bridging Social Distance During Social Distancing: Exploring Social Talk and Remote Collegiality in Video Conferencing. *Journal of Human-computer Interaction Special Issue on the Future of Remote Work: Responses to the Pandemic*.

<https://doi.org/10.1080/07370024.2021.1994859>

Edwards, J., Wintersberger, P., Clark, L., Rough, D., Doyle, P.R., Banks, V., Wyner, A., Janssen, C.P., & Cowan, B.R.. (2021). CUI@ Auto-UI: Exploring the Fortunate and Unfortunate Futures of Conversational Automotive User Interfaces. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applica-

tions, 186–189.

<https://doi.org/10.1145/3473682.3479717>

Edwards, J., Clark, L., & Perrone, A.. (2021). LGBTQ-AI? Exploring Expressions of Gender and Sexual Orientation in Chatbots. In CUI 2021-3rd Conference on Conversational User Interfaces, 1–4.

<https://doi.org/10.1145/3469595.3469597>

Edwards, J., Cooney, O., & Edwards, R.. (2021). Alexa, Play Fetch! A Review of Alexa Skills for Pets. In Proceedings of the Eighth International Conference on Animal-Computer Interaction, 1-4.

<https://doi.org/10.1145/3493842.3493902>

Doyle, P.R., Rough, D., **Edwards, J.**, Cowan, B.R., Clark, L., Porcheron, M., Schlögl, S., Torres, M.I., Munteanu, C., Murad, C., and others. (2021). CUI@ IUI: Theoretical and Methodological Challenges in Intelligent Conversational User Interface Interactions. In 26th International Conference on Intelligent User Interfaces, 12–14.

<https://doi.org/10.1145/3397482.3450706>

Baxter, M., Bleakley, A., **Edwards, J.**, Clark, L., Cowan, B.R., & Williamson, J.R.. (2021). “You, Move There!”: Investigating the Impact of Feedback on Voice Control in Virtual Environments. In CUI 2021-3rd Conference on Conversational User Interfaces, 1–9.

<https://doi.org/10.1145/3469595.3469609>

Edwards, J., Perrone, A., & Doyle, P.R.. (2020). Transparency in Language Generation: Levels of Automation. In Proceedings of the 2nd Conference on Conversational User Interfaces, 1–3.

<https://doi.org/10.1145/3405755.3406136>

Wu, Y., **Edwards, J.**, Cooney, O., Bleakley, A., Doyle, P.R., Clark, L., Rough, D., & Cowan, B.R.. (2020). Mental workload and language production in non-native speaker IPA interaction. In Proceedings of the 2nd Conference on Conversational User Interfaces, 1–8. *Honorable mention*.

<https://doi.org/10.1145/3405755.3406118>

Wu, Y., Rough, D., Bleakley, A., **Edwards, J.**, Cooney, O., Doyle, P.R., Clark, L., & Cowan, B.R.. (2020). See what I’m saying? Comparing intelligent personal assistant use for native and non-native language speakers. In 22nd International Conference

on Human-Computer Interaction with Mobile Devices and Services, 1–9.

<https://doi.org/10.1145/3379503.3403563>

Munteanu, C., Clark, L., Cowan, B.R., Schlögl, S., Torres, M.I., **Edwards, J.**, Murad, C., Aylett, M., Porcheron, M., Candello, H., and others. (2020). CUI: Conversational User Interfaces: A Workshop on New Theoretical and Methodological Perspectives for Researching Speech-based Conversational Interactions. In Proceedings of the 25th International Conference on Intelligent User Interfaces Companion, 15–16.

<https://doi.org/10.1145/3379336.3379358>

Edwards, J. & Elaheh Sanoubari. (2019). A need for trust in conversational interface research. In Proceedings of the 1st International Conference on Conversational User Interfaces, 1–3.

<https://doi.org/10.1145/3342775.3342809>

Edwards, J., Liu, H., Zhou, T., Gould, S.J.J., Clark, L., Doyle, P.R., & Cowan, B.R.. (2019). Multitasking with Alexa: how using intelligent personal assistants impacts language-based primary task performance. In Proceedings of the 1st International Conference on Conversational User Interfaces, 1–7.

<https://doi.org/10.1145/3342775.3342785>

Cowan, B.R., Doyle, P.R., **Edwards, J.**, Garaialde, D., Hayes-Brady, A., Branigan, H.P., Cabral, J., & Clark, L.. (2019). What's in an accent? The impact of accented synthetic speech on lexical choice in human-machine dialogue. In Proceedings of the 1st International Conference on Conversational User Interfaces, 1–8. *Honorable mention.*

<https://doi.org/10.1145/3342775.3342786>

Doyle, P.R., **Edwards, J.**, Dumbleton, O., Clark, L., & Cowan, B.R.. (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, 1–12.

<https://doi.org/10.1145/3338286.3340116>

Clark, L., Cowan, B.R., **Edwards, J.**, Munteanu, C., Murad, C., Aylett, M., Moore, R.K., Edlund, J., Szekely, E., Healey, P., and others. (2019). Mapping Theoretical and Methodological Perspectives for Understanding Speech Interface Interactions. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Sys-

tems, 1–8.

<https://doi.org/10.1145/3290607.3299009>

Perrone, A. & **Edwards, J.** (2019). Chatbots as unwitting actors. In Proceedings of the 1st International Conference on Conversational User Interfaces, 1–2.

<https://doi.org/10.1145/3342775.3342799>

Abstract

Interacting with computers using speech promises the benefit of multitasking while one's hands and eyes are occupied by another task. Users of spoken dialogue systems have not seen this technology as living up to its potential however, in part because speech agents interactions largely behave like traditional interface interactions, initiated by the user. In order to harness the full multitasking benefit of speech as an interaction modality, speech agents must interactive proactively, but doing so means that agents will need to interrupt users engaged in other tasks. While general guidelines have been proposed for the design of proactive agents in general, the design of proactive speech which interrupts has not been explored in specific. Interrupting speech from proactive agents can take design inspiration from the ways people use speech to interrupt other people, but human speech interruptions are likewise not well understood. The first study of this thesis uses a mixed methods approach to investigate the effect of the urgency on people's timing and strategies for interrupting with speech. The second study complements that data by comparing the effect of urgency on interruption timings and strategies to the effect of the difficulty of the task which is interrupted. The third study uses a data-driven approach to classify the interruptible moments of a complex task in order to analyse the extent to which participants from the prior studies utilised these dynamic characteristics of the ongoing task to shape their interruptions. Finally, the fourth study applies findings from human speech interruption to the design of a proactive agent to investigate the effects of human-inspired adaptivity to context on people's perceptions of a proactive speech agent. Findings suggest that human speech interruptions are highly diverse and adaptive to context, but such adaptivity may be seen as inappropriate and inconsistent when applied to a speech agent. The implications of this research and its limitations are discussed in the closing chapter.

List of Figures

3.1	An example of a non-urgent trial that participants saw as a practice trial	62
3.2	An example of an urgent trial that participants saw as a practice trial . .	63
3.3	Means and and standard errors for interruption onset times by urgency condition	67
3.4	An example of a possible subtask boundary which some participants identified as a good moment to interrupt. Note that the orange Tetris block on the far right is the currently falling block.	72
4.1	Simulated power for sample sizes from 40 to 100. At 80 participants, the lower confidence interval is marginally below the required threshold of 80%.	90
4.2	Means and and standard errors for interruption onset by urgency and difficulty condition	95
4.3	Standardised β estimates for fixed effects of urgency, task difficulty, and their interaction on interruption onset.	97
4.4	Means and and standard errors for interruption duration by urgency and difficulty condition	98
4.5	Standardised β estimates for fixed effects of urgency, task difficulty, and their interaction on interruption duration.	100
4.6	Log-odds of fixed effects of urgency, task difficulty, and their interaction on access ritual usage. Negative values indicate likelihood that an access ritual will not be produced for a given trial, while positive values indicate likelihood that an access ritual will be produced.	101

LIST OF FIGURES

5.1	Example of the first and last frames of a sequence of Tetris exemplifying the No Spin theme.	129
5.2	Example of three frames of a sequence of Tetris exemplifying the One Spin theme.	130
5.3	Example of three frames of a sequence of Tetris exemplifying the Line Clear theme.	130
5.4	Example of the first and last frames of a sequence of Tetris exemplifying the Line Clear theme.	131
6.1	Example screenshots from the experiment which participants saw as part of pre-test instructions. On the left, there is no red dot, so the agent has not yet been cued to interrupt. On the right, the red dot has appeared, signalling that the agent has been cued to interrupt.	158
6.2	Predicted values of appropriateness questionnaire ratings by condition .	165
6.3	Predicted values of Partner Model Questionnaire total scores by condition	167
6.4	Predicted values of Partner Model Questionnaire <i>partner competence and dependability</i> subscale scores by condition	168

List of Tables

3.1	Table of interruption prompts.	61
3.2	Table of means and standard deviations for interruption onset and interruption duration by urgency condition.	67
3.3	Summary of fixed and random effects for interruption onset - Linear mixed effects model	68
3.4	Summary of fixed and random effects for interruption duration - Linear mixed effects model	68
3.5	Table of counts of trials containing access rituals by urgency condition .	69
3.6	Summary of fixed and random effects for access ritual presence - Logit mixed effects model (Present = 1)	69
4.1	Summary of fixed and random effects for interruption onset - Linear mixed effects model	94
4.2	Table of means and standard deviations for interruption onset by urgency and difficulty condition.	96
4.3	Summary of fixed and random effects for interruption duration - Linear mixed effects model	96
4.4	Table of means and standard deviations for interruption duration by urgency and difficulty condition.	96
4.5	Summary of fixed and random effects for access ritual usage - Logit mixed effects model	99
4.6	Table of counts of trials containing access rituals by urgency and Tetris difficulty condition	99
5.1	Table of interruption prompt questions.	123

5.2	Summary of fixed and random effects for Chapter 3 interruptible window usage - Logit mixed effects model	133
5.3	Summary of fixed and random effects for Chapter 4 interruptible window usage - Logit mixed effects model	133
6.1	Table of means and standard deviations for single-item questionnaire responses by condition	162
6.2	Summary of fixed and random effects for timing single item questionnaire - Linear mixed effects model	162
6.3	Summary of fixed and random effects for appropriateness single item questionnaire - Linear mixed effects model	163
6.4	Table of means and standard deviations for PMQ total score and subscale scores by condition	164
6.5	Summary of fixed and random effects for Partner Model Questionnaire total scores - Linear mixed effects model	164
6.6	Summary of fixed and random effects for Partner Model Questionnaire <i>partner competence and dependability</i> subscale - Linear mixed effects model	166
6.7	Summary of fixed and random effects for Partner Model Questionnaire <i>human likeness</i> subscale - Linear mixed effects model	166
6.8	Summary of fixed and random effects for Partner Model Questionnaire <i>cognitive flexibility</i> subscale - Linear mixed effects model	166

Contents

Declaration of Authorship	ii
External Contributions	iii
Acknowledgements	iv
Publications	vii
Abstract	xii
Abbreviations	xiii
List of Figures	xv
List of Tables	xvii
1 Introduction	1
1.1 Defining the research area	1
1.2 Thesis outline	5
2 Literature Review and Methodology	11
2.1 Introduction	11
2.2 Spoken Dialogue	12
2.2.1 Speech in HCI	12
2.2.2 Human-human dialogue	16
2.2.3 Human-likeness in speech agent design	20
2.3 Multitasking	22
2.3.1 Cognitive resources	22
2.3.2 Interruptions	30
2.3.3 Proactive agents in HCI	34
2.4 Methodology	37
	xix

2.4.1	The scientific method and HCI	37
2.4.2	Application and abstraction continua	41
2.4.3	Experimental design	45
2.4.4	Mixed methods research and qualitative data analysis	47
2.4.5	Tetris as a complex task	50
2.5	Summary	51
3	Eliciting Speech Interruptions to Investigate the Impact of Urgency	53
3.1	Introduction	53
3.1.1	Interruptions and complex tasks	53
3.1.2	Interruption timing	54
3.1.3	Access in speech interruptions	55
3.1.4	Urgent speech	57
3.1.5	Aims, hypotheses, and contribution	58
3.2	Methods	58
3.2.1	Participants	58
3.2.2	Materials	59
3.2.3	Experimental conditions	62
3.2.4	Measures	63
3.2.5	Procedure	64
3.3	Results	66
3.3.1	Quantitative interruption behaviour	66
3.3.2	Qualitative descriptions of interruption strategies	69
3.4	Discussion	75
3.4.1	Explicit cues of urgency impact interruptions of others	76
3.4.2	Interruption strategies are highly diverse	76
3.4.3	Few people use access rituals	78
3.4.4	A paradigm for eliciting spoken interruptions	78
3.4.5	Limitations	79
3.5	Conclusion	81
4	Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions	83
4.1	Introduction	83

CONTENTS

4.1.1	Task difficulty	84
4.1.2	Explicit and estimated cues	86
4.1.3	Aims and hypotheses	88
4.2	Methods	89
4.2.1	Participants	89
4.2.2	Materials	90
4.2.3	Experimental conditions	90
4.2.4	Measures	92
4.2.5	Procedure	92
4.3	Results	93
4.3.1	Quantitative interruption behaviour	93
4.3.2	Qualitative descriptions of interruption strategies	100
4.4	Discussion	107
4.4.1	Interruption urgency affects timing more so than task difficulty	107
4.4.2	Estimated cues affect the structure of interruptions	109
4.4.3	Minimise disruption or interrupt quickly?	111
4.4.4	Limitations	113
4.5	Conclusion	115
5	Characterising Appropriate Moments for Interruptions of a Complex Contin-	
	uous Task	117
5.1	Introduction	117
5.1.1	Classifying task interruptibility	117
5.1.2	Event structure in a Tetris game	119
5.1.3	Aims and research questions	120
5.2	Methods	120
5.2.1	Participants	120
5.2.2	Materials	121
5.2.3	Experimental conditions	124
5.2.4	Measures	124
5.2.5	Procedure	125
5.3	Results	126
5.3.1	Analysis Approach	126
5.3.2	Manipulation Check	127

5.3.3 Cluster generation and stabilisation	127
5.3.4 Content analysis of video segments	128
5.3.5 Usage of interruptible Tetris windows in previous studies	131
5.4 Discussion	133
5.4.1 Characteristics of Tetris interruptibility	133
5.4.2 Harder Tetris games do not cause more breakpoint usage	136
5.4.3 Cluster analysis for data-driven breakpoint identification	138
5.4.4 Limitations	139
5.5 Conclusion	142

6 Comparing Perceptions of Static and Adaptive Proactive

Speech Agents	145
6.1 Introduction	145
6.1.1 Designing proactive agents	146
6.1.2 Partner modelling of machine dialogue partners	148
6.1.3 Aims and hypotheses	150
6.2 Methods	151
6.2.1 Participants	151
6.2.2 Materials	152
6.2.3 Experimental conditions	153
6.2.4 Measures	155
6.2.5 Procedure	157
6.3 Results	159
6.3.1 Analysis approach	159
6.3.2 Quantitative response data	161
6.3.3 Partner model questionnaire	163
6.3.4 Qualitative response data	167
6.4 Discussion	174
6.4.1 Consistency as a salient feature for adaptive agents	175
6.4.2 Appropriateness of adaptive proactive design	177
6.4.3 Individual differences and personalisation	179
6.4.4 Limitations	180
6.5 Conclusion	183

7 General Discussion	185
7.1 Introduction	185
7.2 Contributions and Implications	186
7.2.1 Urgency and strategies in human-speech interruptions	186
7.2.2 Classification of interruptibility of a complex task	189
7.2.3 Perceptions of an adaptive proactive speech agent	194
7.2.4 Human-inspired design of speech agents	196
7.2.5 Individual differences in interruption preferences	198
7.2.6 Limitations	200
7.3 Conclusion	205
Bibliography	209
Appendices	
Appendix A: Materials used in Chapter 3	i
Appendix B: Materials used in Chapter 4	xv
Appendix C: Materials used in Chapter 5	xxxi
Appendix D: Materials used in Chapter 6	xliv

1 * Introduction

The thesis expounded in this work is that speech agents that are designed based on human interruption strategies in dialogue can positively influence multitasking experiences with speech agents, in particular by impacting a user's partner model of the agent. The thesis posits that by understanding how people use speech to interrupt others as they engage in an ongoing complex task, speech agents can be designed to interrupt in a way people find minimally disruptive and maximally appropriate.

1.1 Defining the research area

Speech agents have existed as a technology for many years, and in the last decade, they have become increasingly popular, owing to advances in speech recognition, artificial intelligence, and networking technologies (McTear et al., 2016). Now, as more people have come to use speech agents in their daily lives, it is increasingly possible to understand what people see as their strengths and weaknesses and to understand what users hope for the future of the technology. While speech agent adoption has proliferated (Olson & Kemery, 2019), the settings and tasks people use them in has been quite constrained (Ammari et al., 2019; Dubiel et al., 2018). This relative lack of enthusiasm for speech agents has been explained in part as a result of a gulf of expectations between what speech agent users expect their interactions to be like as compared to their actual experiences using speech agents (Luger & Sellen, 2016).

One of the key perceived benefits of speech as an interaction modality, as noted both by speech agent users (Luger & Sellen, 2016) and industry research (Martelaro et al., 2019), is its superior suitability for multitasking when eyes and hands are busy as compared to using technology with a screen and a manual input. This benefit is not being realised with current speech agents however, as commercially popular

agents like Apple's Siri and Amazon Alexa are more like traditional interfaces than actual agentive technology in that interactions are always initiated by the user. While some research has begun to explore how speech agents can use environmental cues to decide when to initiate interactions (Cha et al., 2020; Semmens et al., 2019), more holistic research about how proactive agent speech should be designed has not yet begun.

One potential avenue for designing speech of proactive agents is by using human speech as a design inspiration. When people interact with spoken dialogue systems, they understand their interaction through two different metaphors: the interface metaphor in which human-machine dialogue is analogous to using a traditional computer with files, menus, and commands, and through the human metaphor, in which human-machine dialogue is analogous to talking to a person who can use and understand natural language (Edlund et al., 2008). Insofar as human conversation is a metaphor through which people set expectations for an interaction with a speech agent, human speech can and has been used as an inspiration for the design of machine speech (Sutton et al., 2019). That said, recent work on the design of speech agents has proposed that agents should not seek to be overly human-like in their design (Aylett, Sutton, et al., 2019; Moore, 2017) and that some users reject speech agents which they see as dishonestly mimicking human-likeness through voices or personalities designed to seem more human than their capabilities can match (Aylett, Cowan, et al., 2019; P. R. Doyle et al., 2019). For this reason it is an ongoing challenge to identify the correct balance between meeting the expectations for speech interactions established by human dialogue and overpromising human-likeness to the point of detriment.

In order to design human-inspired proactive agent speech, it is necessary to understand the characteristics of human speech in the same context. Insofar as multi-tasking is a goal for proactive speech agents, speech interactions would sometimes begin while users are engaged in another task. That is to say, speech from proactive agents will sometimes interrupt another task. While human speech interruptions have not been closely studied in particular, there has been ample research on adjacent topics such as self-interruption behaviour, greetings in human dialogue, and characteristics of human speech under varied circumstances. In the domain of self-interruption, prior research has shown that people tend to deal with interruptions as quickly as pos-

sible, even at the peril of their ongoing task, unless they are explicitly instructed that the interrupting task is more important (Brumby et al., 2011; Horrey & Lesch, 2009). Likewise, people react more rapidly to interrupting speech which conveys urgency about the interruption through an assertive tone (Wong et al., 2019). Along these lines, some work has sought to understand what urgent speech sounds like, albeit outside of the context of interruptions, investigating how it differs from non-urgent speech in characteristics such as pitch and speech rate (Landesberger et al., 2020b). Together, this work helps to illustrate that interruptions which are perceived as urgent lead to different reactions than those which are not, that speech can convey urgency, and that there are some known characteristics of urgent speech. Still, the characteristics of spontaneously produced human speech interruptions have not been studied, so a better understanding of that phenomenon is prerequisite to designing proactive agent speech of the same type.

In order to elicit spontaneous human speech interruptions, it is necessary to select a task which speech interrupts. Prior research on interruptions and multitasking has sometimes focused on driving as an ongoing task (Brumby et al., 2013; Janssen et al., 2014; Martelaro et al., 2019). While some speech agent users express a desire to multitask specifically while driving (Luger & Sellen, 2016), driving remains quite a complex task to model quantitatively (Semmens et al., 2019). Likewise, much speech agent interaction takes place via smart speakers around the home, such as in the kitchen (Zhao et al., 2022) or in a family dining area (Porcheron et al., 2018). As such, this thesis does not use driving as an ongoing task, instead seeking a more basic task which still has the eyes-busy, hands-busy, continuous characteristics which are shared by more applied tasks like driving and cooking. It is also important to identify key fundamental parameters of complex, continuous tasks generally which are shared across a variety of applied contexts. Following prior work which similarly sought to focus on a singular complex task and study it in its entirety rather than using a simulated form (as simulated driving is frequently used to stand in for on-road driving) (Lindstedt & Gray, 2019), this thesis uses the computer game Tetris as a complex task which human and agent speakers must interrupt. Tetris requires a player to make decisions which have consequences upon the success or failure of the overall task and it involves coordination of several cognitive processes including visual monitoring, motor skills, and strategic planning. As such, Tetris is well suited as a complex, continuous

task to focus on in order to better understand interruptions of tasks of this type more generally.

Past research on interruptions has revealed the importance of the timings of interruptions. One key concept in this topic is the notion of natural breakpoints - moments between the conclusion of one subtask and the start of the next subtask within a complex task (Janssen et al., 2010, 2012). These breakpoints have been seen as particularly well-suited to interrupt a task, as interruptions at these moments are less likely to interfere with a person's memory for the state of the task when they return to it (Borst et al., 2015). That said, it is not clear whether equivalent moments exist for continuous tasks like on-road driving or like Tetris which cannot be suspended arbitrarily without severe consequences. Some research has attempted to model interruptible moments for driving as an exemplar task, but the complexity of driving makes this a challenging endeavour (Semmens et al., 2019). It is not clear whether people attempt to use natural breakpoints from other people's tasks to guide how they time interrupting speech, though some research has indicated that people attempt to use cues around the difficulty of a dialogue partner's ongoing task to modify their own speech when speaking to a multitasker (Janssen et al., 2014). This thesis therefore also investigates how people use cues of task difficulty to modify interrupting speech, and it seeks to model the interruptible windows of Tetris games, analogous to natural breakpoints of discrete tasks, in order to understand how and whether they are utilised by human interrupters.

Finally, insofar as this work seeks to inform the design of proactive agents, it is necessary to choose a means of evaluating such an agent. The motivation for this thesis lies in the research proposing a gulf of expectations between speech agent abilities and user expectations for them (Luger & Sellen, 2016) and that, for proactive agents, using human-like strategies for interrupting ongoing tasks might influence these expectations. This phenomenon of user expectations of speech agents as dialogue partners recently been understood through the framework of partner models (P. R. Doyle et al., 2019). Partner models are the internal representations people have about the capabilities of their dialogue partner, be that partner another person or a speech agent (Branigan et al., 2011a; Cowan & Branigan, 2017). Recent research has begun to investigate the salient cues to forming partner models of machine dialogue partners, finding *partner competence and dependability*, *human-likeness*, and

cognitive flexibility as three dimensions (P. R. Doyle et al., 2021). This research has produced the Partner Model Questionnaire (PMQ), a self-report scale for measuring people's perceptions of speech agents' capabilities as dialogue partners (P. R. Doyle, 2022). This work therefore uses the PMQ as a key measure of the proactive agent design it proposes, as the thesis aims to provide inspiration to the design of speech agents which will lead to users viewing them as more competent dialogue partners.

1.2 Thesis outline

This thesis aims to inform understanding of human speech interruptions, apply insights from human speech interruptions to the design of a proactive speech agent, and assess the extent to which these human speech-inspired design choices improve people's perceptions of that agent as compared to prior perceptions and to an agent for which alternative design choices are made. Specifically, this thesis aims to (i) investigate the effect of interruption urgency on human speech interruptions of a continuous, complex task, as exemplified by Tetris; (ii) investigate the effect of the difficulty of the continuous task on human speech interruptions; (iii) describe people's self-reported strategies for interrupting with speech; (iv) model ideal interruption windows for a complex task; (v) evaluate the extent to which spontaneous speech interruptions use these windows and the extent to which urgency (vi) and task difficulty (vii) affect their usage; (viii) apply these insights to the design of a proactive speech agent; and (ix) compare people's perceptions of that speech agent to their perceptions of a baseline proactive agent and to their prior perceptions of speech agents in general.

Following this introduction, Chapter 2 reviews literature on spoken dialogue with both machine and human dialogue partners. It discusses the role of human-likeness in the design of speech agents, critically reflecting on the positive and negative outcomes of emulating human speech when designing machine speech. It then reviews literature on cognitive resources and multitasking in order to introduce key motivations for proactive speech agent usage before outlining prior work on designing proactive agents. Finally, Chapter 2 considers the methodological approach used throughout this thesis, describing the role and aims of scientific research in the field of human-computer interaction (HCI) as aimed at solving conceptual, empirical, and construc-

tivist problems (Oulasvirta & Hornbæk, 2016). The nature and aims of experiments in HCI research are critically considered, and the benefits and rationale for mixed methods research are described. The chapter concludes with a consideration of Tetris as a complex task and its utility in understanding complex tasks in a general sense.

Chapter 3 presents an experiment aimed at gaining an initial understanding of the characteristics of human spoken interruptions. Using Tetris as an ongoing task, the chapter presents an online paradigm by which participants are instructed to repeatedly use speech to interrupt a Tetris player. Informed by literature on self interruptions of complex tasks (Brumby et al., 2011; Horrey & Lesch, 2009), this chapter identifies explicit task priorities as a key variable in how people strategise their interruptions of complex tasks and therefore manipulates interruption urgency as an independent variable. Insofar as urgent speech has been demonstrated to differ from non-urgent speech in terms of its timing and structure (Landesberger et al., 2020a, 2020b) and speech which conveys urgency has been viewed as less polite than speech which does not (Wong et al., 2019), this study focuses on interruption timing and on the use of access rituals (Goffman, 1971) as primary quantitative measures. It likewise investigates self-reported strategies of both interruption timing and interruption structure in order to gain a holistic view of the method of expression (see McFarlane, 1997) of human speech interruptions. Through analysis of 709 interruption trials by 46 participants, this study finds a statistically significant effect of interruption urgency on interruption onset and no significant effect of interruption urgency on either interruption duration nor access ritual use. Qualitative analysis reveals four themes for interruption timing strategies: *prioritising speed*, *prioritising accuracy*, *Tetris task characteristics*, and *message content* as well as three themes for interruption structure strategies: *phrasing*, *delivery*, and *message content*. This chapter likewise contributes a paradigm for eliciting speech interruptions of complex tasks, which is used again in Chapter 4.

Chapter 4 builds upon the paradigm and research questions introduced in Chapter 3, investigating the effects of both explicit and estimated cues on speech interruptions of Tetris. While significant effects of urgency, a cue explicitly known to participants, were demonstrated in the prior chapter, self-report data also indicated that people estimated contextual information such as the cognitive load of the Tetris player and the state of the Tetris game when choosing how to interrupt. Prior research on

complex tasks such as driving has yielded mixed evidence about whether passengers use estimations of the driver's task difficulty to strategically suppress conversation to benefit the driver (Gugerty et al., 2004; Janssen et al., 2014; Nunes & Recarte, 2002). As such, this study further investigates the effect of the estimated cue of Tetris task difficulty in interrupters' strategies for speaking to players. Through analysis of 1400 trials by 90 participants, this study finds a statistically significant effect of urgency, the explicit cue, on interruption onset times and durations, and no significant effects of difficulty, the estimated cue, on timing. Conversely, the study finds a statistically significant effect of difficulty on the use of access rituals with no significant effect of urgency on their usage. Qualitative analysis identifies two competing themes for interruption timing strategies: *Interrupting as soon as possible* and *Interrupting at a good moment* as well as two competing themes interruption structure strategies: *Communicating urgency* and *Communicating calmness*. Participants renewed emphasis on choosing appropriate moments to interrupt contributes further evidence toward the use of natural breakpoints (Janssen et al., 2010) during interruptions of others, which is further investigated in Chapter 5.

Chapter 5 presents a framework of identifying moments of Tetris gameplay as interruptible through an interrupter data-driven classification study. Whereas prior work on identifying interruptible moments of complex tasks have sought to gather data from arrays of video camera, biometric sensors, and task-related sensors, looking at interruptions in real time (Iqbal & Bailey, 2010; Kim et al., 2015; Semmens et al., 2019), this study takes an alternative approach which is less intrusive to participants and less resource intensive to collect. Through an online paradigm using Tetris as the complex task, participants choose precise moments of Tetris games which they found to be most interruptible. K-means cluster analysis (Lloyd, 1982) is conducted to group these moments into a set of windows of interruptible Tetris gameplay. Content and thematic analysis of these clusters are used to group them and describe overarching themes. Four themes of interruptible Tetris gameplay are identified: *No Spin*, *One Spin*, *Line Clear*, and *Calm Episode*. Insofar as participants in Chapters 3 and 4 report seeking good moments to interrupt, these themes are then used to classify interruptions from those studies to measure the effects of urgency and task difficulty on the use of interruptible gameplay windows for timing interruptions. Quantitative analysis reveals that these participants were no more likely than chance to initiate

their interruptions during interruptible windows for either level (easy or hard) of Tetris difficulty. The effect urgency on the use of interruptible windows is likewise shown to have not been statistically significant. The chapter nevertheless contributes a novel approach for interruptible moments of a complex task, uses this approach to propose characteristics of interruptible moments within Tetris, and provides insight into the effects of cues of urgency and task difficulty on their use. This chapter likewise aligns with Chapters 3 and 4 in demonstrating the heterogeneity of people's strategies for interrupting, raising questions about whether preferences for interruptions are likewise heterogeneous, which are explored in Chapter 6.

Chapter 6 applies insights from the three preceding chapters to the design of a proactive speech agent. Prior research on speech agents has identified a "gulf of expectations" between the what users see as the potential benefits of using speech to interact with a machine, such as for multitasking while engaged in another task, and the reality of interactions (Luger & Sellen, 2016). Recent research in this area has focused specifically on the concept of partner models: the mental representations that people construct of the capabilities of a machine as a dialogue partner which the user updates as they gain experience interacting with a particular machine (P. R. Doyle et al., 2021). This chapter presents a prototype proactive speech agent which adapts its speech to both urgency and to the difficulty of the ongoing task it interrupts, modelled on the ways that participants from Chapters 3-5 update their interrupting speech. This agent is compared against a static proactive agent which interrupts in a fixed style, ignorant of both urgency and task difficulty. The study hypothesises that adaptive proactive speech modelled on human speech interruptions will lead to partner models which consider the proactive agent as a stronger conversational partner than a static agent, and that interruptions initiated by an adaptive agent will be judged as better timed and more appropriately asked. These hypotheses are all rejected however, as quantitative analysis reveals that participants view the adaptive agent as a poorer dialogue partner than the static agent and as less appropriate in the style it interrupts. Qualitative analysis sheds light on the source of this surprising finding, as participants see the adaptive agent as less socially appropriate and as less consistent in its interactions than the static agent. These findings align with growing literature on appropriateness for speech agent design which challenges human-machine dialogue should not seek to exactly mimic human-human dialogue and that it is in fact a differ-

ent category of interaction altogether (P. R. Doyle et al., 2021; Moore, 2017). It also represents the first evidence of design decisions for speech agents leading to trade-offs between dimensions of users' partner models for those agents, a phenomenon which has only previously been speculated about (P. R. Doyle, 2022).

Chapter 7 presents a general discussion of the contributions, implications, and limitations of the research presented in this thesis and considers areas for future work. Empirical, conceptual, and constructive contributions are each highlighted. This includes highlighting the empirical findings in Chapters 3 and 4 around the statistically significant effect of interruption urgency on interruption onset and the statistically significant effect of Tetris difficulty on the use of access rituals. It further highlights the contributions of descriptions of people's self-identified strategies for interrupting with speech as well as the methodological contribution of describing and utilising a new paradigm for eliciting spontaneous speech interruptions. Conceptual findings around the characteristics from Chapter 5 that define interruptible windows in Tetris are likewise highlighted, as well as that chapter's empirical findings around the lack of significant effects of urgency and difficulty on the utilisation of interruptible windows. Finally, the constructivist and empirical contributions of Chapter 6 are discussed, including a description of the design of a prototype proactive agent which adapts to context similarly to human interrupters as well as empirical findings that this adaptive agent was seen as a less competent dialogue partner than a baseline agent which does not adapt to context. Along with discussion of these findings and their implications, Chapter 7 also considers the limitations of the research in this thesis, reflecting on both the research design choices which yielded particular limitations and the limitations inherent to the methods used in this thesis. New research questions which this work asks or helps to address are considered throughout the general discussion.

2 * **Literature Review and Methodology**

2.1 Introduction

Interacting with machines through conversational speech has been a feature of science fiction for over a century, with popular films like *2001: A Space Odyssey* (Kubrick, 1968) and *Her* (Jones, 2013) imagining human-like digital agents which people can speak with in just the same ways they talk to other people (Axtell & Munteanu, 2021). Recent decades have seen simultaneous progress in domains from artificial intelligence, speech recognition and synthesis, semantic representation, computing power, and web connectivity, all of which have contributed making conversational speech agents a reality of the present and near future (McTear et al., 2016). Now, 72% of respondents to a Microsoft market survey reporting that they have used a speech assistant like Apple's Siri, Google Assistant, or Amazon Alexa (Olson & Kemery, 2019), with more than 90 million adults in the US owning a smart speaker and more than 50% reporting using it by interacting with a speech agent every day (Kinsella, 2022). Speech interactions with machines have already become ingrained into many people's lives. This moment is therefore a critical one for understanding and improving speech interactions with non-human agents so that this blossoming and long-anticipated interaction style can achieve the potential for interactional utility that has thus far been constrained to works of fiction.

This thesis aims to inform the design of system-initiated speech interruptions for near-future technologies in which an agent and a human work together. Domains like industrial human-robot collaboration and automated driving require information

to be shared between human and machine effectively, sometimes in safety-critical situations (Finzi & Orlandini, 2005; Janssen et al., 2019). Likewise, hands-free, eyes-free multitasking has been identified as a key perceived benefit of everyday speech agent usage by regular users (Luger & Sellen, 2016), so improving speech interactions between machines and busy people stands to benefit users across a variety of interaction contexts.

This review begins by considering the state of research on speech as a modality for HCI. The extent to which human-machine dialogue and human-human dialogue are similar and the ways in which they have differed are considered as the intricacies of human-human spoken communication are discussed. By establishing the present and future need for mixed-initiative human-machine dialogue, the review will then outline cognitive models of multitasking as to illustrate the theoretical bases for multitasking with speech in general and for designing for better speech-based multitasking in complex domains specifically. This will transition the focus of the review back to HCI, where a brief history of proactive agent design can more fully be considered. Finally, the review will outline methodological and philosophical approaches employed in this thesis, beginning with an overview of the philosophy of science and of HCI research adopted here, narrowing in on specific decisions around research design and data analysis which underlie the research carried out throughout the thesis.

2.2 Spoken Dialogue

2.2.1 Speech in HCI

While speech is most typically produced by and directed toward humans, this thesis aims to understand speech produced by or directed at machines. Speech has been a subject of HCI research for over three decades, with focus on each direction of speech - speech input from users, speech output from machines, and bi-directional dialogue interactions (L. Clark, Doyle, et al., 2019). Multitasking as a particular benefit of speech in HCI has been noted by users who mention multitasking as a reason for trying speech technology in the first place (Luger & Sellen, 2016). This has become especially relevant in the high-risk and eyes-busy, hands-busy domain of driving where using speech has been both studied (Caird et al., 2018; Iqbal & Bailey, 2010; Janssen et al., 2019; Martelaro et al., 2019; Semmens et al., 2019), and found to be

practised (Dubiel et al., 2018; Olson & Kemery, 2019), but the multitasking benefits of speech should theoretically be relevant to other primary tasks as well, beyond the highly-studied driving domain. As this thesis aims to integrate knowledge of human conversation with HCI interaction knowledge, the following section will demonstrate how speech has been studied in HCI rather than addressing particular features of speech.

Experimental and user studies

In HCI experiments and user evaluations of commercial products, speech has been studied as an input modality, as an output modality, and as a bi-directional modality (L. Clark, Munteanu, et al., 2019), not unlike how it has been viewed in multitasking research. A 2019 systematic review of speech in HCI found that much of the existing work has featured either computer-based speech systems aimed at supporting accessibility or productivity or telephone based interactive voice response (IVR) systems (L. Clark, Doyle, et al., 2019). The recent advent and surge in popularity of Intelligence Personal Assistants (IPAs) such as Google Assistant and Siri however has led to a rise in research focusing on these more conversationally inspired systems (Ammari et al., 2019; Cowan et al., 2017; Dubiel et al., 2018; Edwards et al., 2019; Porcheron et al., 2018; Wu et al., 2020).

Some of the recent work on IPAs has focused on understanding usage of such assistants. These studies have included user surveys (Dubiel et al., 2018; Olson & Kemery, 2019), device logging (Ammari et al., 2019), ethnographic observation (Porcheron et al., 2018) and interview studies (Porcheron et al., 2018) Some recurring themes from these works have included the popularity of IPAs (Ammari et al., 2019; Olson & Kemery, 2019) and their success in entertaining uses including families with children (Luger & Sellen, 2016; Porcheron et al., 2018; Purington et al., 2017) On the other hand, themes like confinement of usage to limited domains like requesting music and asking for information (Ammari et al., 2019; Dubiel et al., 2018) and disillusionment stemming from concerns with privacy or usefulness (L. Clark, Munteanu, et al., 2019; Luger & Sellen, 2016) have permeated as well.

As IPAs have become ubiquitously available in smartphones and smart speakers, they have seen new users abound. Yet user studies have revealed a tendency toward abandonment of speech agents as they fail to live up to expectations (Dubiel

et al., 2018; Luger & Sellen, 2016). The conversational interaction style of IPAs has in some cases been regarded as unnatural (L. Clark, Munteanu, et al., 2019), artificial (P. R. Doyle et al., 2019), and mere human mimicry (Aylett, Cowan, et al., 2019) with such human-like design simultaneously giving users falsely positive impressions of the capabilities of such systems (Luger & Sellen, 2016; Moore, 2017). Unnaturalness, to IPA users, means that, while human-human dialogue involves a blending of both functional communication chit-chat by which speakers pursue both social and practical objectives (Cheepen, 1988), users see the inclusion of social talk as unnecessary and undesirable from machines (L. Clark, Munteanu, et al., 2019). While some IPA users describe the contrived social talk as an uncanny valley (P. R. Doyle et al., 2019), other research has contended that this is a poor description for the phenomenon, as an uncanny valley implies that increasing the quality of the source of discomfort (e.g. human-like social communication from machines) can reduce the discomfort it causes (Mori et al., 2012), whereas in this case, human mimicry is undesirable altogether and only its avoidance can reduce discomfort (Aylett, Cowan, et al., 2019). Taken together with the limited domains users have settled into with IPAs, current research indicates that while using speech agents for eyes-busy, hands-busy multitasking may be desirable (Luger & Sellen, 2016), few other current or future use cases are identified by users. Even then, current IPAs lack design considerations explicitly aimed to improve multitasking outcomes by using speech as a modality. Likewise, speech technology research has been disjointed and unfocused, with limited unifying theories or design principles (L. Clark, Doyle, et al., 2019; Coupland, 2003) The novelty of IPAs as a technology certainly contributes to the lack of mature research themes and settled use cases, presenting an opportunity for speech technology research to continue to develop the research theme of multitasking, the suitability of which for speech is unpacked further in Section 3.4 of this chapter. A closer focus on multitasking as a key use case for speech technology will allow for a deeper understanding of how the technology needs to further develop in order to benefit users to its full potential.

Human-machine dialogue

Speech as an HCI modality differs from the use of input devices like touchscreens, pointing devices, and keyboards insofar as speech, unlike those modalities, is used in non-HCI contexts. Importantly, speech has practical (i.e. supporting tasks) and social

purposes (Cheepen, 1988; L. Clark, Munteanu, et al., 2019) Much work has focused on the fact that, currently speech with machines is task oriented, highly simplistic and user led. While this sort of social talk has been a goal for speech dialogue system design (Gilmartin et al., 2017), users do not think this has been achieved (P. R. Doyle et al., 2019) nor that it necessarily should be (L. Clark, Munteanu, et al., 2019; Cowan et al., 2017).

When interacting with a machine through speech, people understand their interaction through both an interface metaphor, whereby they are controlling a computer as well as through a human metaphor, whereby they are speaking to a human-like other whom they can interact with as if it were human (Cassell, 2007; Edlund et al., 2008; Edlund et al., 2006). This notion of using metaphors to understand one's interactions with computers has origins in the concept of mental models, the internal models people form and update about technologies based on their prior experiences and interactions (Norman, 1983). Insofar as people approach dialogue with machines as, in part, similar to dialogue with humans, human-likeness is seen as a positive design outcome in some speech agent research, as the human metaphor people employ in these interactions is more likely to be sufficiently accurate to avoid errors (Edlund et al., 2008). Indeed one of the seminal paradigms in HCI, particularly for speech HCI, has been the computer as social actors (CASA) paradigm (Nass et al., 1994). This paradigm holds that people use human social interactions as a heuristic script for their interactions with computers, extending the same social rules and norms to machines that they do to people. This paradigm has historically been applied to understanding speech technology interactions (K.-M. Lee & Nass, 2005). Insofar as human-machine dialogue is understood through an interface metaphor, certain characteristics of human-human dialogue such as its social enjoyability are not clearly beneficial to improving the experience of interacting with an interface, by which efficiency and error avoidance are paramount (Edlund et al., 2008). Indeed qualitative research on speech agent users indicates that people express distaste toward machines that try to mimic the small talk and building of social connections the pervade human-human dialogue (L. Clark, Munteanu, et al., 2019; P. R. Doyle et al., 2019). For this reason, while understanding and emulating human dialogue may be a good starting point for designing machine dialogue, replication of all facets human-human dialogue is not a prudent goal.

Recent work has sought to reframe human-agent conversation as a different sort of conversation with different rules and goals (L. Clark, Munteanu, et al., 2019; Cowan et al., 2017; P. R. Doyle, 2022; P. R. Doyle et al., 2019; Porcheron et al., 2018). This work largely builds on the concept of partner models. Partner models, the specific mental models people develop for the conversational capabilities of particular individuals as dialogue partners (Branigan et al., 2011a), add nuance to the human metaphor for understanding human-machine dialogue. Rather than considering all machines as roughly similar to all humans, partner modelling research contends that people construct and update partner models of each dialogue partner they interact with, whether that partner is a human or a machine (Branigan et al., 2011a; Cowan & Branigan, 2017; P. R. Doyle et al., 2019). This framework is further refined through the notion of global and local partner models, by which superficial cues of identity like gender or regional accent activate a global model for a dialogue partner which is then adapted to a local model for an individual based on interaction experiences (Brennan et al., 2010). Machine dialogue partners may therefore activate different global partner models than do human dialogue partners for the people who speak with them, thereby cuing different sets of expectations (L. Clark, Munteanu, et al., 2019; P. R. Doyle, 2022). That said, human-based designs of spoken dialogue systems may act as an important scaffold to inform these models, and understanding the inadequacies of a given design can only be achieved through careful construction and evaluation of that design.

2.2.2 Human-human dialogue

Insofar as human-human dialogue has been used as a basis for understanding human-machine dialogue for both speech agent users and for researchers, it is therefore necessary to consider some of the key components of human-human dialogue. This section considers two components of human-human dialogue which are closely related to the proactive, machine-initiated speech that this thesis seeks to inform: social access and turn-taking. While each of these has been studied primarily in human-human communication thus far, this thesis aims to utilise prior research of these aspects of human-human spoken dialogue to better understand and inform human-machine spoken dialogue.

Speech based interruptions and social access

Some speech based tasks that act as interruptions in everyday life are for the purpose of initiating dialogues - conversations that begin while people are engaged in other tasks. Because this sort of interrupting (and, for the receiving party, assenting to or declining interruption) is so common in human experience, they follow certain patterns and progress through common stages. Goffman termed these patterns of social interruption access rituals, defining them as the patterned ways by which people request, grant, and terminate social access to one another (Goffman, 1971).

The two categories of access rituals are greetings and farewells, the requests for the beginning and ending of access that mark nearly every conversation (Goffman, 1971). For each of these rituals, qualitative research has sought to map the behaviours - verbal and nonverbal - that most frequently comprise these rituals (Knapp et al., 1973; Krivonos & Knapp, 1975). Verbal signals include things like salutes - words like "hey" or "hi" - and use of people's first name; nonverbal signals include gaze, hand gestures, and physical touch (Krivonos & Knapp, 1975). These studies were foundational in categorising the behaviours that comprise access rituals, focusing on the abrupt start or end of social access in the absence of a secondary task. The studies of hellos and goodbyes did not represent multitasking events however, as participants were not engaged in any other tasks when the access rituals were performed. Instead, both studies had participants engage in a conversation as their only objective, and experimenters reviewed and categorised the access ritual of interest (the hellos in one study, the goodbyes in the other). Nonetheless, the categorisations of behaviours resulting from these studies are a strong foundation for describing the behaviours that people take when initiating spoken communication with another person, so this thesis makes use of those categorisations and of access rituals as a conceptual basis for studying speech as an interrupting task.

The social function of access rituals invites other social factors such as gender to impact the way access is requested and granted (Goffman, 1971). Studies of access rituals have tended to use same-gendered dyads (namely, all men) to minimise effects of gender, biasing observed behavioural data (Knapp et al., 1973; Krivonos & Knapp, 1975). Access is related to hierarchical social power (Goffman, 1971) and the characteristics of requests for access can indicate power differences (Hutte et al., 1972) As such, studies of social access have either used dyads of peers when

power is seen as equal between parties (Corsaro, 1979; Krivonos & Knapp, 1975) or used professor-student dyads when looking at power differentials (Knapp et al., 1973) While this thesis aims to inform the design of interrupting speech from non-human agents toward human users, it does not comment on the power dynamic between humans and speech agents, neither describing what that dynamic is like nor what it ought to be like. While some recent literature has considered this dynamic to be hierarchical, with speech agents in a subordinate social role to users (P. R. Doyle et al., 2019; Luger & Sellen, 2016) , others have described a more equal, peer-like relationship (Purinton et al., 2017) or called for reconsidering the many different social roles that speech agents can be cast in (McMillan & Jaber, 2021). As such, this thesis considers requests for social access between unacquainted peers in Chapters 3, 4, and 5, and uses those interactions as an inspiration for design of agents in Chapter 6 which reflect little in terms of explicit power dynamics.

Speech research has focused on speech-based collaborative work, examining communication as a secondary task supporting a primary task, namely constructing with Lego (H. H. Clark & Krych, 2004). In this experiment, one participant had to build specific structures with Lego, according to instructions given by another participant. This work examined the effect of shared visual information on language efficiency and spatial reference, finding that participants needed fewer words and fewer turns to accomplish a joint task when they mutually could see the Lego workspace. This sort of speech multitasking, in which tasks share relevant goals and problem states, is much less disruptive to a primary task performance than is the sort of irrelevant interruptions typical to HCI notifications according to work on interruption relevance (Gould et al., 2013), relating to cognitive resource allocation in multitasking discussed in section 3.1 of this chapter. The speech considered in this thesis is of the latter type, disruptive and not relevant to the task which it interrupts. This is so that understanding of communication between people gleaned from this thesis is applicable to the most disruptive, and therefore most potentially costly instance of interrupting speech.

Turn taking and nonverbal communication

An important part of conversational speech is the coordination of turns between speakers. In order to coordinate seamless turn-taking, a speaker must plan their utterance

before their partner has finished speaking, minimising the gap between utterances to only a few hundred milliseconds (Holler et al., 2016). This time between the tasks of listening and speaking is consistent with the intervals associated with task-switching times in other concurrent multitasking tasks (Salvucci et al., 2009), providing evidence that the same cognitive resources that bound other sorts of multitasking likewise bound the multitasking that occurs within dialogue. This interval between speaking and listening has been demonstrated to be consistent across several languages and cultures (Stivers et al., 2009), indicating that turn taking is bounded by innate cognitive resources of humans in general rather than the task specifics of a given language. Recent neuroscience work has further supported the theory that listening and utterance planning are bounded by the attentional and memory resources that constraint other multitasking, highlighting the monitoring of an utterance that occurs after word selection as a particular subtask that competes with listening (Fargier & Laganaro, 2019). If turn-taking in speech as thought of as an example of multitasking, speaking and listening turns can separately be modelled as different subtasks which compete with other concurrent tasks in different ways. Indeed some work on multitasking with speech agents has begun to consider the cognitive resource demands of different language tasks as causally related to the effect that speech multitasking has on them (Edwards et al., 2019). This thesis therefore is sensitive to spoken dialogue as an act of multitasking in itself, with listening, planning speech, and delivering an utterance as distinct tasks which can interfere with other tasks to differing degrees.

The features of language that make coordinated turn taking possible have been a topic of investigation in linguistics, with verbal and nonverbal cues both playing important roles. One model of turn taking, the *anticipatory model*, holds that listeners can anticipate the end of a turn by identifying syntactic elements of conversation typical of turn endings and take the floor if the speaker grants it (Sacks et al., 1978). Alternatively, the *signal model* states that signals are not anticipated, and that listeners must react to them upon hearing them if they want to take the floor (Yngve, 1970) Some of the signals of a turn ending that have been studied include changes in loudness, pitch, and intonation patterns, gaze, gestures, and use of specific signalling phrases (Duncan, 1972). More recent experimental data has indicated however that a variety of cues - lexical, syntactic, and prosodic patterns including transition phases (e.g. "you know"), endings of clauses, and shifts in volume or pitch at the end of a phrase - allow

for anticipation of turn endings, and that these cues are more crucial in turn taking coordination than are the proposed signals that underlie the signal model (Riest et al., 2015; Sacks et al., 1978). Chapters 3 and 4 take a mixed-methods approach to investigating which cues people use for taking and yielding the floor with interrupting speech, as turn-taking signals have thus far been studied only in ongoing dialogue in the absence of other tasks. Better understanding of whether and how similar cues manifest in interrupting speech can contribute to the design of non-human spoken interruptions from proactive agents.

2.2.3 Human-likeness in speech agent design

Speech agent interactions are a unique form of human-computer interaction as they require users to engage in dialogue with a non-human conversational partner, making the perceived conversational abilities of that partner central to the interaction (Branigan et al., 2011a). As highlighted above in section 2.1, the global partner models activated for speech agent users when interacting with machine dialogue partners in general and the local partner models activated by particular speech agents each contribute to the perceptions and expectations people have when interacting with a speech agent (P. R. Doyle, 2022). These sources of expectations have been echoed by other research which has highlighted the impact of perceptions of how basic or advanced a system is (Branigan et al., 2011b), human-like markers of identity such as regional accents or perceived nationality (Cowan et al., 2019), propensity of errors (P. R. Doyle, 2022), and the marketing of speech agents (Sin et al., 2022) affect the way that people approach their conversations with machines. These expectations of high human-likeness lead to partner models which construe an agent as having very similar conversational capabilities as a human, which new or infrequent users of speech agents explicitly describe (Cowan et al., 2017; Luger & Sellen, 2016). Ensuring a match between expectations and capabilities is therefore paramount for limiting the size of the gulf of expectations and delivering better experiences for speech agent users.

Until recently, little research has explored the characteristics of partner models in speech agent interactions. Recent work has begun to explore this question however, investigating the dimensions of partner models which are salient to people engaged in dialogues with machines and with people (P. R. Doyle, 2022; P. R. Doyle et al., 2021;

P. R. Doyle et al., 2019). In order to investigate these dimensions of partner models, Doyle and colleagues invited participants to engage in conversations with each of Siri, Amazon Alexa, and a human researcher and then reflect on how they would describe and differentiate between these interactions (P. R. Doyle et al., 2021; P. R. Doyle et al., 2019). The descriptive terms elicited through this study were then used to generate semantic differentials, which speech agent users then used to categorise the speech agents that they had experienced interacting with (P. R. Doyle et al., 2019). This study resulted in the generation of three themes describing the dimensions of partner models for speech agents: perceptions of partner competence and dependability, assessment of human-likeness, and perceptions of the cognitive flexibility of the system (P. R. Doyle et al., 2021). These semantic differentials and themes were further developed into a validated self-report questionnaire across those factors, the Partner Modelling Questionnaire (PMQ), which can be used to measure the strength of people's partner models for machine dialogue partners (P. R. Doyle, 2022). The PMQ therefore represents a new validated measure for assessing partner models, which have been previously identified as a crucial factor in users' perceptions of speech agents. Understanding speech agent users' partner models of the systems they communicate with is essential for identifying sources of a gulf between expectations and reality and trying to reduce them.

One recent focus for reducing the gulf of expectations in human machine dialogue has been to focus on the appropriateness of the design of the speech agent. The concept of appropriateness in terms of speech agent design has become a popular theme in speech agent research which has highlighted the impact of matching the design of speech interfaces with their capabilities (Moore, 2017; Moore & Morris, 1992). This alignment can be achieved through the selection of non-human names and robotic voices for agents with limited capabilities and the use of an animal-like, toy embodiment for agents designed for play (Moore, 2017). Similar recent research has indicated that pursuing unnatural sounding voices may be beneficial in setting appropriate expectations that a speech agent has different capabilities than a human (Aylett, Sutton, et al., 2019) and that using synthesised speech which does not sound human-like may be a worthwhile path to avoiding setting inappropriate expectations (Le Maguer & Cowan, 2021). These aims are somewhat at-odds however with the pursuit of natural sounding synthesis observed in academic research (Aylett, Sutton,

et al., 2019) and in increasingly human like synthesis demonstrations from companies like Google (Leviathan & Matias, 2018) This tension is somewhat resolved however by recent calls for more ecologically grounded approaches to designing speech agents which signal a capacity for errors just as humans do rather than designing agents which always present themselves as fully competent and confident humans (Becker et al., 2022). To achieve appropriateness, it is therefore necessary to understand both the expectations that people have for their speech agents and the ways in which people set expectations for one another in spoken dialogue. This thesis therefore examines both the ways in which people communicate with other people via speech as well as the impressions that people have of non-human agents designed to speak with them.

2.3 Multitasking

2.3.1 Cognitive resources

In the history of cognitive science, explanations for the allocation of cognitive resources during multitasking and the consequent limitations to people's multitasking abilities have broadly been explained through either *structural theories* or *capacity theories* (Wickens, 1981). Structural theories are those which describe the competition between tasks for mental resources as constrained by competition over the same cognitive structure due to similarities between the tasks and how they are performed. Broadly, these structural theories are focused with either competition over structures of perception or competition over structures of planning and executing responses (Wickens, 1981).

Structural theories

Structural theories of cognitive resources originate from early cognitive psychology experiments like those of Broadbent in which participants would be tasked with listening to multiple sources of speech simultaneously and then asked to provide information about one source or the other (Broadbent, 1954). Experiments like these revealed that listeners could not attend to both sources of speech simultaneously, instead recalling information better from one source rather than the other. This and

similar experiments led to the filter model of attention which describes attention as bottlenecked in terms of processing, by which only one among multiple competing perceptual stimuli is selected for processing (Broadbent, 1958). In this model, physical properties of a signal (e.g. the location or loudness of a sound or the colour or size of an image) are used to filter the attended stimulus from the unattended ones, and semantic information such as the content of speech is only processed for the attended stimulus.

The filter theory of attention was refined by later experimental work finding that participants in simultaneous listening experiments could effectively follow one message even if it had its physical characteristics changed midstream with the other speech source, such as by swapping the locations of the two audio sources (Treisman, 1960). This evidence, along with contemporary research on simultaneous listening that found participants could recognise when their own name was said among the unattended speech signal (Moray, 1959), led to the attenuation model, a modification of the filter model (Treisman, 1969). Rather than seeing attention as binary, by which only one stimuli among those competing would be attended to and others would be ignored for processing, the attenuation model proposed that only one stimulus is strongly attended to while others are still weakly attenuated. In this way, while still proposing a filter on which information is selected for processing, the attenuation model allows for semantic characteristics of stimuli to cause them to be selected for full attention, such as semantically important words like one's own name, or the word fire or words which semantically match the previously attended-to message (Treisman, 1960). The attenuation model therefore expanded the psychological view of the abilities of human multitaskers, as both physical and semantic cues could be seen as useful for selection of information for processing.

In contrast to the filter and attenuation models of attention which stipulated that simultaneous tasks compete for which will be perceived, Deutsch and Deutsch proposed that the structure which information competes for is best framed as decision making (Deutsch & Deutsch, 1963). Presenting evidence from neuroscience, this late-selection model proposed that stimuli are perceived simultaneously and in full, stipulating that differences observed in past multitasking experiments could instead be better explained by a filter upon which stimuli are selected for memory or a response with other stimuli, despite being perceived, then getting ignored. This model, like

previous structural models, still stipulated that a serial system is the cause of the bottleneck by which attention is allocated to one task or another, and like attenuation theory, it proposed that the semantic importance of a stimulus is the quality by which competing stimuli are selected for decision making (Deutsch & Deutsch, 1963). This late-selection model was further refined proposing that both physical properties of a stimulus such as loudness or location as well as semantic properties like the meanings of words are used as filters for selecting a stimulus for response (Norman, 1968). By proposing a later filter, at the stage of reaction to stimuli rather than perception, and by keeping broad the sorts of characteristics by which stimuli are filtered, these structural theories continued to broaden the proposed capacity of human multitaskers.

Capacity theories

As an alternative to structure theories of attention, capacity theories propose that what makes multitasking difficult is not competition in terms of structure of tasks or minds, but instead competition in terms of capacity: that people have a limited amount of attention to allocate, and different tasks, independent of their modality or the stage of processing (Wickens, 1981). Citing evidence of dissimilar simultaneous tasks (e.g. a manual typing task and a verbal response task) causing interference to one another's performance, research proposed that a central executive system was responsible for allocating the pooled resource of attention across tasks without sensitivity to the nature of those tasks (Moray, 1967). Contemporary work on multitasking performance compared people's abilities in a variety of perceptual tasks in isolation with their performance of those tasks simultaneously to develop a quantitative model of capacity for visual and auditory tasks (Taylor et al., 1967). Developing concurrently with structural theories of cognitive resources, capacity models present an alternative understanding of why multitasking causes reduced performance in each task.

Kahneman described his highly influential capacity model of attention and effort, building upon the prior literature on both capacity and structural theories (Kahneman, 1973). In this model, attention became synonymous with mental effort, and its allocation is subject to a person's level of arousal, the demands of a task, and general allocation policy rules. Kahneman explains that arousal influence attentional capacity in a parabolic relationship, with both low arousal states (e.g. fatigue) and high arousal states (e.g. anxiety) leading to less available attentional capacity than states

in between these extremes (Kahneman, 1973). This available capacity is then allocated according to the demands of the task(s) to be performed, with different tasks demanding different levels of attention. This allocation is subject to some fundamental policy principles in the model, such as the principle that completion of one task will be prioritised over other simultaneous tasks, resulting in a primary task and secondary tasks for a multitasker. In keeping with structural theories, another allocation principle of Kahneman's capacity model is that stimuli such as noise, unexpected visual stimuli, or semantically meaningful stimuli like the sound of one's own name can involuntarily capture attention even if that stimulus is unrelated to the prioritised task (Kahneman, 1973). Finally, the capacity model specifies that momentary intentions affect attention allocation. That is to say, even if one task is the focus of attention, a multitasker can decide to direct attention to a different task at their own discretion and then return attention to the primary task. This model further expands both the sources of influence on multitasking performance and the role of the multitasker as a decision maker who chooses how to allocate attention.

Working Memory theory (WM), like capacity theories of attention, proposes a structural-independent and limited-capacity cognitive resource - memory rather than attention - which traces the boundaries for human cognitive function (Baddeley & Hitch, 1974) WM theory primarily consists of a central executive function and three subsystems, the phonological loop, the visuo-spatial scratchpad, and the multisensory episodic buffer (Baddeley, 2003). In WM, sensory input, information perceived or produced across any sensory modality, is either retained in sensory memory via attention from the modality agnostic central executive or ignored and allowed to decay from memory. That sensory memory then passes to sense-specific modal subsystems that hold a limited amount of sensory information. The central executive then may select response by accessing the internal resources of those short-term, modal subsystems and long-term memory via retrieval and overwriting in those memory stores (Baddeley & Hitch, 1974). Marrying the focus on memory and decision making of the attenuation structural model with the modality-insensitive, limited capacity central processing system of other capacity models, WM moves understanding of multitasking closer to a unified and broad theory of cognitive resources.

Multiple Resources Theory

Structural and capacity theories of cognitive resources are further united under Multiple Resources Theory (MRT) which draws on the evidence for and against each of structural and capacity theories (Wickens, 1981). Multiple Resources Theory proposes that cognitive resources are separated across numerous dimensions including modality, code, stage, and visual channel (Wickens, 2002). Modality is described as the perceptual channel by which information is processed, including visual and auditory channels, as well as the response modality such as speech or use of hands. Code denotes the distinction between spatial information and symbolic or verbal processes, drawing on structural theories which indicated that tasks of different codes were less detrimental to multitasking than tasks of the same code, such as speaking and visually tracking an object as compared to speaking and listening to speech (Wickens, 1981). Stage denotes the distinction between perception of stimuli and the planning and execution of response behaviours. Visual channel, a fourth dimension contained within the visual modality and the perception stage, differentiates stimuli perceived in focal vision from stimuli perceived in ambient vision. All together, MRT describes the performance of simultaneous tasks as dependent on both a pool of shared resources but also on structural characteristics of a task, providing a framework for understanding multitasking which accounts for a wide range of empirical findings.

While MRT does not make commitments to the specific identity of the cognitive resources that are the focus of the theory, Wickens proposes effort and memory as two important resources to consider (Wickens et al., 2009). Effort, following from Kahneman (Kahneman, 1973), is synonymous with attention, and viewed as one resource in limited capacity in MRT (Wickens & McCarley, 2007). Attention is seen as a soft constraint in cognitive modelling research, as momentary capacity varies dependent on the state of the cogniser and the demands of the task (Gray et al., 2006). Memory in MRT is viewed through the WM model (Baddeley, 2003) which likewise places emphasis on both structural components like modality as well as capacity (Wickens & McCarley, 2007). Memory, unlike attention, is a hard constraint, a resource which is more or less fixed for an individual and thus sets firm bounds on a person's capacity at tasks which require memory rather than the contextually sensitive bounds around attention (Gray et al., 2006). Together, MRT unites a variety of theories of cognitive resources and provides a framework for understanding task performance across a

variety of human behaviours including multitasking.

To demonstrate these dimensions, take the example of driving (primary task) while talking on a handheld mobile phone (secondary task). Both of these tasks only partially overlap in modality; driving requires visual perception and manual response behaviour while a phone call requires auditory perception and vocal response behaviour with manual behaviour limited to holding the phone up. In terms of code, these tasks differ as well; the information perceived for driving is spatial while the information needed for a phone conversation is verbal. Both tasks require cognitive effort in both the perception and response stages, so the tasks overlap on this dimension. The visual channel dimension is not relevant to talking on a phone, so this dimension is not a source of overlap. Altogether, these tasks have mild overlap and thus, under MRT, we would predict them to have some detrimental impact on one another when multitasked.

MRT holds that concurrent tasks are more disruptive to one another when they have higher overlap across these dimensions. This has been demonstrated experimentally as auditory interruptions are less disruptive than visual interruptions to visual primary tasks (Ho et al., 2004; Ratwani et al., 2008). Importantly, MRT states that while resources are multiple - consuming one does not necessarily impact a different resource - they are nonetheless resources and thus are limited and exhaustible (Wickens, 2002). For instance, while a person may be able to both look at a stimulus and speak at the same time, it is not possible to look at three different stimuli and speak two different words at the same time.

Task switching dynamics

While MRT helps to predict the effect of the style and content of interruptions, it does not provide much insight into the effects of the timing of task switches. Insights around interruption timing instead come from research based on the Memory for Goals and Memory for Problem States models of multitasking. Early work on interruption timing posited that interruptions had different effects at different moments in a task owing to certain moments in tasks having lower cognitive workload than others (Miyata & Norman, 1986). Later experimental research confirmed this notion, that moments of lower mental workload were better for interruptions in terms of both behavioural measures of task performance and subjective assessments of the interrup-

tion, with lower workload moments identified as boundaries between subtasks within tasks (Iqbal & Bailey, 2005). Following the formulation of Memory for Goals theory as a cognitive explanation for the reduced mental workload at task boundaries, further experimental and modelling research confirmed these boundaries are the points around which people orient their task switching when they are free to choose the timing of task switches, leading to researchers terming these boundaries *natural breakpoints* (Janssen et al., 2010, 2012). Taken together, MRT and natural breakpoints provide a framework by which the interruption designer can seek to minimise disruption of interruptions both in their style and in their timing.

Insofar as a tidy problem state results in an easier return to a task, research emphasises that there exist certain tradeoffs that come with task switching. By increasing interruption lag, the time between stopping a primary task and starting a secondary task, multitaskers can better encode the state of the tasks they are pausing, reducing the burden of recalling the problem state upon resumption (Trafton et al., 2003). Likewise, by increasing the resumption lag, the time between finishing and interrupting task and returning to an original task, multitaskers can better remember the original problem state and thus reduce the risk of error - referred to as the speed-accuracy tradeoff (Brumby et al., 2013). This tradeoff reflects the fact that the impact of an interruption depends on the goals of the multitasker and the nature of the task, as some people or tasks may be more time-sensitive while others are more sensitive to avoiding error.

Just as the timing of an interruption can help or hinder returning to an original task, so too can interruption content. Interruptions that occur at natural breakpoints have been demonstrated to be even less disruptive when they are relevant to the primary task that they interrupt (Iqbal & Bailey, 2008). This fits with the memory for problem state hypothesis as a relevant interruption helps a user to retain the goals of their primary task, as the primary task and relevant interruption share goals. Further work has strengthened the case for problem state relevance impacting interruption as laboratory experiments illustrated the benefit of interruptions being relevant as compared to a hindrance when an interruption is similar in content but irrelevant. This work showed that interrupting tasks that share a similar content area but do not have the same goals as the primary task are more disruptive to the primary task as they may interfere with memory of the primary task, introducing new goals rather than

rehearsing primary task goals (Gould et al., 2013).

In laboratory study and in computational modelling, people have been demonstrated to be very good at coordinating their multitasking by switching at these low-cost natural break points (Borst et al., 2010; Janssen et al., 2010; Janssen et al., 2015; Salvucci & Beltowska, 2008) . A prevailing model of how people seamlessly switch between tasks and choose just the right moment is known as threaded cognition (Salvucci & Taatgen, 2008). Threaded cognition imagines tasks as threads running in parallel that can only be processed serially by a cognitive actor who has several limited resources. Once a thread begins, another may only start when the resource it requires is made available. Tasks that can be performed using procedural rules rather than declarative memory - well learned or sufficiently simple tasks, those that do not require conscious rehearsing of procedural steps - may continue occupying a resource while a different task begins to consume a different resource. By knowing which resources are currently available and which tasks are calling for which resource, the cognitive actor thereby serves each task in an efficient balance of resources and needs (Salvucci et al., 2009). While this theory relies on allocation of effort and strategic decision making by the multitasker, it nonetheless helps to demonstrate the fundamental role of memory in multitasking models, even those primarily concerned with the decisions made by the actor rather than the nature of the tasks.

Task categories

In describing the insights which modern computational modelling frameworks like the EPIC architecture can grant our understanding of human cognition during multitasking, Kieras and colleagues first define four categories of tasks for which human cognition might be studied: discrete successive tasks, discrete concurrent tasks, continuous elementary tasks, and compound continuous tasks (Kieras et al., 2000). The first two of these categories, the discrete tasks, represent tasks which require a single input or choice from an actor, then are completed and cease after that choice is made. Discrete tasks are less negatively impacted by task switching as task states need not be encoded when an actor switches tasks nor recalled when the actor returns (Borst et al., 2015), as the moments between tasks act as ideal breakpoints, with the suspended discrete task demanding minimal cognitive resources so long as the task is suspended at the boundaries between tasks (Janssen et al., 2012). The next category

of tasks, elementary continuous tasks, are those which require ongoing input from an actor and do not have pauses between completion of subtasks, like visually tracking a moving object on a screen (Kieras et al., 2000). These tasks are more difficult to interrupt as there are no breakpoints in which the task can be fully suspended and attention can be diverted to a different task. Still, MRT holds that tasks can be performed simultaneously with less competition and disruption when they differ in modality (Wickens, 2002). In this way, a task like visual tracking could be interrupted with less risk to task success if the interruption came via an audio modality. The final category of tasks that Kieras and colleagues describe are compound continuous tasks, tasks which recruit several cognitive processes for perception and interaction and require the monitoring of several constituent subtasks to achieve a broader goal, like driving a car or cooking a meal (Kieras et al., 2000). These tasks are the sort we engage in every day for work and recreation, which require ongoing attention and the coordination of different skills. Due to their varied and ongoing nature, these tasks, which we will heretofore refer to as complex tasks, are particularly difficult to interrupt without incurring a penalty to task performance. That said, in a world in which proactive agents need to get the attention of users who are living their normal lives, it is crucial that we better understand how interruptions of these tasks can be done more safely and efficiently.

2.3.2 Interruptions

Nearly a century of psychology research has identified that tasks which are interrupted and then returned to are completed more slowly but remembered more clearly than tasks which are completed without interruption, which contemporary psychologists attributed to the greater stress incurred by multitasking as compared to monotasking (McKinney, 1935). Later work began to better understand this phenomenon, finding that tasks which are more similar to each other are more detrimental when interleaved than dissimilar tasks owing to difficulty remembering the details of the suspended task (Czerwinski et al., 1991). Research like this provided a foundation for later cognitive accounts of how multitasking works, with the development of the Memory for Goals model (Altmann & Trafton, 2002) and Memory for Problem States (Borst et al., 2010, 2015) model of multitasking, which share a description of task switching as a process of encoding in memory the task which is to be suspended, then switching

tasks, then retrieving from memory the relevant details of the suspended task before returning to it. Through cognitive accounts of multitasking like these, we better understand and can better predict the adverse impacts of multitasking, informing the way we design the necessary or desirable interruptions in our lives.

In order for digital agents to proactively interact with people, they will need to get the attention of that person, potentially interrupting their focus on another task. In other words, proactivity necessarily entails interruption. Interruption by digital agents has been an area of HCI research focus for more than three decades (Card & Henderson, 1987). Most interruptions research in HCI has looked at either (1) forced interruptions (McFarlane, 1999, 2002) by which the task presented to a person changes without their input, though potentially with forewarning (Czerwinski et al., 1991; van der Heiden et al., 2017), (2) notifications (Edwards et al., 2019; Iqbal & Bailey, 2005, 2010) by which an agent alerts a person with some cue that they should switch to another task, or (3) self interruptions (Dabbish et al., 2011; Horrey & Lesch, 2009; Mark et al., 2015) by which people may freely switch between tasks according to their own discretion. The second of those three types, interruptions which are prompted by an agent but executed by the interrupted person, termed negotiated interruptions by McFarlane (McFarlane, 1999) are the primary focus of this thesis, as the interruption is itself interactional, with both agent and human contributing to the initiation of the interruption.

McFarlane describes the way in which interruptions of this type mirror the initiation of human spoken communication, echoing Herb Clark (H. Clark, 1996) in describing the four options people have when another person attempts to initiate a conversation: accepting the joint project as proposed, accepting the joint project with alteration, declining the joint project, or withdrawing from the attempted initiation altogether (McFarlane, 1999). In much the same way, people can respond to an interruption with any of those strategies - by switching tasks immediately, deferring the interruption but switching tasks after some delay, declining the interruption and keeping focus on their original task, or ignoring the interruption altogether without acknowledgement and keeping focusing on their original task. McFarlane further developed a taxonomy of eight factors that are relevant to the outcome of an interruption, including factors describing the person who is interrupted, factors describing the nature of the interrupting agent, factors that describe the content and style of

the interruption itself, and factors describing both the task which is interrupted and the task which constitutes a response to the interruption (McFarlane, 1997). This taxonomy highlights the extent to which interruptions research is highly complex and can be understood from a variety of viewpoints, including understanding humans as cognisers, understanding interruptions as an act of communication, and understanding the outcomes of interruptions as multitasking activity.

Multitasking has been used as a term describing a variety of task-switching activities on a time continuum ranging from tasks performed concurrently to tasks performed successively with many minutes or even hours between task switches (Salvucci et al., 2009). Multitasking is readily observed in our daily habits such as completing work tasks via frequent task switching (Dabbish et al., 2011) and support of multitasking is seen as a desired benefit from the future of technology by some digital technology users (Luger & Sellen, 2016). Indeed multitasking is sometimes necessary as a safety-critical or otherwise urgent task may suddenly emerge during engagement in some other task, like a takeover request during automated driving (Janssen et al., 2019). Multitasking is not without drawbacks however. This insight can be applied to both the selection of tasks which might be appropriate to multitasking and to the design of an interruption, given knowledge of the task which it interrupts. For example, in the domain of driving, recent work has explored both secondary tasks and interruptions which make use of the modalities of audio and speech (Martelaro et al., 2019; Semmens et al., 2019) with the intention of limiting interference with the driving tasks, which primarily uses visuomotor modalities.

Interruption strategies & notifications

Frequently, as a person uses a computer, the computer will need to interrupt the user to deliver information. For example, while a user is typing an email to a particular recipient, an email client may alert the user that a new email from that recipient was just received. Depending on the email client however that interruption may be delivered in a number of different ways - the client may display a visual notification immediately, it may simply update the number of unread messages in a subtle way, or it may do nothing and not alert the user until they refresh the inbox. McFarlane details HCI interruption strategies, classifying them into four groups - immediate interruption, negotiated interruption, mediated interruption, and scheduled interruption (McFarlane,

2002). Immediate interruptions are those that present new information or tasks to the user without warning, immediately when they become available. Negotiated interruptions present a request for the user's attention that the user can accept or deny, either presenting the new information or staying on their current task. Mediated interruptions present this request for attention to a mediating system, like an operating system of a computer or a task scheduling program, to allow an interruption to be accepted or declined without the user being notified at all. Scheduled interruptions are held by the computer until a particular time or for a specified amount of time so interruptions occur at a regular time schedule for the user (McFarlane, 2002). The selection of interruption strategy from these categories can affect user satisfaction and task performance, but which strategy should be selected is dependent on the tasks, users, systems and goals for a given interaction (McFarlane & Latorella, 2002).

Locus of interruption

Interruptions differ in terms of how they are presented: whether a task directly interrupts a person (Li et al., 2008) (through one of the above strategies), the person self-interrupts (Dabbish et al., 2011) and switches task without a cueing from the interrupting task, or if an interruption is forewarned, such as by a notification message (van der Heiden et al., 2017). A negotiated interruption that a person opts in to is a sort of forewarned interruption, but a forewarned interruption may take other strategies as well, as the warning might be in the form of a timer counting down to a scheduled task switch which the user cannot opt out of, as was the case in McFarlane's example of forewarned interruptions (McFarlane, 1999).

The locus of an interruption and people's own decisions about how to interrupt themselves also affect the costliness of the interruption on both tasks. In a simulated driving experiment, participants were told to answer a factual question about a text while driving safely (Brookhuis et al., 1991). After initial trials in which participants used each of a visual interface to read the text or an audio interface to listen to it, participants then were allowed to choose their interface. In a between-subjects design, participants were told to either prioritise driving or prioritise finding the answer to the question. Participants prioritising finding the answer disproportionately chose the visual interface, opting to quickly interrupt their driving and get the answer at the expense of safety. Conversely, driving-focused participants preferred to use the

audio interface, sacrificing control and speed but maximising safety.

Likewise, the timing of a self-interruption influences the cost of that interruption. In a driving study on a test track, participants were asked to complete four tasks on a touchscreen computer mounted on the dashboard over the course of a drive (Gray et al., 2006). Participants were instructed to complete the tasks whenever they wanted to over the course of a multi-lap drive around the track. The track varied between difficult and easy sections, giving participants the opportunity to complete computer tasks on straight roads and focus on the road during turns or obstacles. Instead, participants tried to complete tasks as quickly as possible, engaging in the tasks in quick succession regardless of where they were on the track. In this case, even though participants could have obtained the benefits of deferring interruptions to a natural breakpoint, they did not, instead compromising driving ability to minimise time spent on tasks. Evidence like this presents an argument against HCI interruptions allowing for self-interruptions in safety critical domains, as safety may not be prioritised by a multitasker.

2.3.3 Proactive agents in HCI

Proactive and mixed-initiative interactions have long been seen as the potential benefit of agent-based interactions as compared to interactions involving direct manipulation of user interfaces. Early work on agent-based HCI identified advantages and challenges of agent-based interactions, described agents as user-aware, proactive, autonomous, and adaptive (Shneiderman & Maes, 1997). User-awareness in this case refers to an agent's knowledge of an individual users' behaviour patterns and their interaction preferences; proactivity and autonomy are the abilities of an agent to initiate and to continue acting without user input respectively; and adaptivity is the ability of an agent to alter its own behaviour to account for changes to the user's behaviour and context (Shneiderman & Maes, 1997).

Some early work in the aftermath of the establishment of the research agenda for agent-based interactions sought to describe design principles for mixed-initiative agent-based interactions, sensitive to the principles which had guided the design of direct-manipulation user interfaces before them. Based on this, 12 principles for mixed-initiative interfaces were proposed: developing significant value-added automation; considering uncertainty about a user's goals; considering the status of a

user's attention in the timing of services; inferring ideal action in light of costs, benefits, and uncertainties; employing dialog to resolve key uncertainties; allowing efficient direct invocation and termination; minimising the cost of poor guesses about action and timing; scoping precision of service to match uncertainty, variation in goals; Providing mechanisms for efficient agent–user collaboration to refine results; employing socially appropriate behaviours for agent-user interaction; maintaining working memory of recent interactions; and continuing to learn by observing (Horvitz, 1999). These principles in general call for proactive agents to consider a variety of potential actions, ranging from acting on the users behalf without consultation to taking no action, by weighing the relative value of the action for the user, the relative cost to the user of a wrongly-selected action, and the certainty if the agent that a particular action is correctly selected. The particular set of actions that a proactive interface might take was further refined by Isbell and Pierce's interface-proactivity (IP) continuum (Isbell & Pierce, 2005). The IP continuum illustrated that a mixed initiative system might, for a given task, allow the user to complete the task themselves, tell the user to pay attention to the task domain in general, tell the user what aspect of the task domain to pay attention to, make suggestions to the user about what decision to make for the task, or complete the task on the user's behalf. Isbell and Pierce give the example of a proactive alarm clock which might decide not to act or to suggest that the user checks their alarms, suggest that the user check their alarm for tomorrow morning in relation to weather, suggest that the user sets their alarm for 20 minutes earlier for tomorrow morning due to an anticipated weather-related delay, or set the user's alarm 20 minutes earlier and explain that it was due to the weather. As these actions become more potentially useful to the user, they also become increasingly intrusive and the cost of a wrong selection increases. Following from Horvitz, a proactive agent must be aware of its level of certainty of the user's state and intentions, and it must seek to reduce the cost of intervention while maximising the benefit.

More recent work on proactive agents has been concerned with the design of specific agents, proposing and testing principles for the design of both the types of tasks that an agent proactively performs as well as the specific implementation of those actions, with regards to details such as modality, timing, message content (Cha et al., 2020; Semmens et al., 2019; Yorke-Smith et al., 2012). In one such study on the design of a proactive learning assistant, Yorke-Smith and colleagues developed nine

principles for proactive agent behaviour, specifying that it should be valuable, pertinent, competent, unobtrusive, transparent, controllable, deferent, anticipatory, and safe. Taken together, these characteristics demand that a proactive agent interrupts in a way that takes the user's context, goals, and abilities into account. Interruptions thereby avoid causing negative consequences to both the proactive interaction and to other tasks the user is engaged in, and the interruption be delivered in a way that is understandable and inoffensive to the user (Yorke-Smith et al., 2012). Echoing the design principles laid out by Horvitz (Horvitz, 1999), this set of principles once again highlights the importance of adapting to a variety of contexts for a proactive agent, including contexts of the agent's task, the user's environment, and the social context of a non-human agent initiating interaction with a person.

Some research has aimed to make first steps toward proactive and mixed-initiative agents for multitasking by seeking to model the interruption and task-switching behaviours of people engaged in the complex task of driving and the complex future task of piloting a self-driving car. Kun and colleagues investigated task-switching between driving and verbal tasks with human interlocutors, demonstrating a cognitive load burden imposed by task switches which can be reduced by accommodation behaviours, by which conversants converge on similar speech patterns as they come to understand each others' speaking styles (A. L. Kun et al., 2013), mirroring literature on partner models by which conversation becomes less effortful as partner models become stronger and alignment increases (Branigan et al., 2011a). Similar work on multitasking speech and driving found that certain lexical features including "oh" and "wait" (reminiscent of access rituals (Krivonos & Knapp, 1975)), as well as prosodic features such as changes in pitch are particularly common in interrupting speech as opposed to ongoing multitasking speech (Yang et al., 2011). More recent work looking at piloting self-driving cars has likewise argued for slow, staged task switching (Janssen et al., 2019), and empirically demonstrated that tasks are not switched all at once, but instead through a process of interleaving those tasks before fully switching (Nagaraju et al., 2021), implying safety benefits for interruptions to ongoing tasks which are well-forewarned and potentially negotiable over immediate interruptions, in accordance with the interruption style research of McFarlane (McFarlane, 2002).

Following from the aforementioned general design principles for proactive agents and the literature on interruptions during driving, recent studies have focused on bet-

ter understanding the environment of a user and its suitability for proactive interaction. Experiments involving proactive speech agents have recently sought to explore this goal by asking users repeatedly if particular moments were a good time to talk to a proactive agent while they drove a car (Semmens et al., 2019) or engaged in daily household activities (Cha et al., 2020), correlating participant responses with environmental sensor data in order to model appropriateness of proactive agent interactions across a variety of task and social contexts. In each study, adapting to these contexts was seen as paramount. For in-home interactions, the urgency of a participant's other tasks and their level of concentration on those tasks were seen as key determinants of whether a proactive interruption was appropriately timed (Cha et al., 2020). Likewise, that study found that the social context of a proactive agent's interruption was an important factor, as participants thought interrupting a conversation with another person to talk to a machine was not acceptable, but engaging in a conversation with an agent as entertainment could be appropriate even in the presence of others (Cha et al., 2020). For the driving study, task context was likewise seen as a paramount consideration, with participants less willing to be interrupted if they were engaged in high-concentration driving tasks but more accepting of an interruption if they were in a lower-concentration moment such as waiting at a traffic signal (Semmens et al., 2019). In each of these cases, the general design principles for proactive interactions were born out empirically, highlighting the impact of contextual appropriateness on a potential user's acceptance of a proactive interaction. Taken together, these studies help to demonstrate that proactive agents, including those which proactively use speech to engage a user in a conversation, must adapt to the various contexts that the user is situated in, including both their cognitive and their social contexts. Still, it is not clear what specific adaptations must be made, nor what contextual cues must be considered. This thesis investigates the contextual cues used by human speech interruptions and the adaptation behaviours that interrupters engage in.

2.4 Methodology

2.4.1 The scientific method and HCI

This thesis begins its consideration of science in general with the postpositivist viewpoint expressed by Popper (Popper, 1959) and Kuhn (Kuhn & Hacking, 1962) which

had defined the social sciences in the second half of the twentieth century, during most of the history of HCI (Cairns & Cox, 2008). In Popper's view, the previous aims of positivism and empiricism, seeking to verify the truth of theories through experimentation, as fundamentally outside of the capacity of science (Popper, 1959). Instead, Popper proposed that experiments have only the capacity for falsifications, to disprove conjectures in order to refute theories. This view was extended and contrasted by Kuhn, who proposed that falsification is an equally futile exercise in science as it relies on an absolute standard of evidence just as verification does - that some evidence must be viewed as irrefutable truth for it to falsify a theory (Kuhn & Hacking, 1962). Instead, Kuhn suggests that evidence contrary to a prevailing scientific paradigm is treated as an anomaly rather than as disproof of the paradigm, and that anomaly must be further researched. While Kuhn acknowledges that scientific revolutions can occur when anomalies make a prevailing paradigm untenable, the normal course of science is not to overturn paradigms in light of contradictory evidence, but to instead puzzle out the source of an anomalous observation within an existing paradigm. This thesis takes Kuhn's view, that the role of the scientific research carried out here is neither verification nor falsification, but instead the observation of phenomena and identification of anomalies within the context of existing HCI research.

Science which is bound by neither the aim of verifying nor of falsifying theory nonetheless has a role for empirical research. Chalmers explains that, following Kuhn, a school of new experimentalism viewed experiments as instead being born out of other experiments, investigating further the causes or consequences of past observations (Chalmers, 2013). New experimentalism, in Chalmers's view, nevertheless give theory a role in science as the evidence gathered in experiments is used to construct new theories and to modify existing theories. This process is not fundamentally about seeking the most true theory however, but instead seeking to specify theories that best suit the empirical evidence presented thus far. This view is expanded by Mayo, who suggests that experiments serve as "severe tests" for theories (Mayo, 2018). In this view, experimental evidence can be used both to assess whether a hypothesis is supported by the results of an experiment and whether the results would have been likely to observe in the case that the hypothesis were not true. In this way, Mayo expresses the view that experiments help to evaluate whether a hypothesis is sufficient for understanding the observed phenomenon but also the extent to which the

hypothesis is necessary for explaining it. This thesis carries forward that view, seeing hypothesis testing as a fundamental practice for social science, but not ascribing realist or postpositivist views to hypotheses tested. While the research presented herein proposes hypotheses and tests them experimentally, it is circumspect about whether new hypotheses could be proposed which better explain the same observations, thus viewing its scientific contribution as iterative rather than conclusive.

HCI as problem solving

Philosopher of science Larry Laudan critiqued the then-orthodox view in 1978 of science, and social science in particular, as wrongheaded in its ambition toward truth-seeking (Laudan, 1978). Whereas then-dominant views of science, such as that of Popper, relied on the practice of falsifying hypotheses as the central activity in science (Popper, 1959), Laudan describes the aim of science to be less dependent on determining truth per se, and instead focused more practically on solving problems (Laudan, 1978). In describing what it means to solve problems, Laudan describes both empirical problems and conceptual problems - the former being problems in explaining an observation about the world for which no sufficient prior explanation and the latter being problems arising from inconsistency within or between those solutions to empirical problems (Laudan, 1978). In this way, Laudan diverges from Popper's falsification and which Laudan identifies as nonetheless rooted in a notion of scientific truth, aligning instead with the Kuhnian view of identifying and resolving anomalies within a given paradigm (Laudan, 1978). That is, Laudan sees his philosophy of science as truth-independent, not concerned necessarily with whether the explanations posed for phenomena are true statements about the phenomena themselves, but instead concerned with whether they adequately solve the problem of explaining that phenomenon or of resolving prior inconsistency within an explanation. Scientific progress therefore can be seen in this philosophy not as approaching truth, but as devising new solutions which explain or which resolve inconsistencies between explanations.

The postpositivist viewpoint is adopted by this thesis. While Popperian approaches to solving empirical problems are used here, such as null-hypothesis significance testing which seek to falsify a null hypothesis (such as in Chapters 3, 4, 5, and 6), this thesis conjointly acknowledges the importance of proposing empirical solutions which provide scientific value not only through falsifying other hypotheses (Popper, 1959),

but also in their adequacy in explaining the experience of a phenomenon via constructionist accounts. That is to say, qualitative findings in this thesis do not aim to solve the problem of explaining people's experiences of a given phenomenon within a study solely by describing the phenomenon, but also through describing the experiences as the study participants understood them (such as in Chapters 3, 4, and 6). In this way, these qualitative findings are not anchored to falsifiability, but are nonetheless presented to solve the problem of providing explanation. Likewise, in keeping with the Laudan conception of problem solving, this thesis also seeks to solve conceptual problems, commenting upon how prior theories can be made more compatible with one another (including in Chapters 5 and 6) and by proposing methodologies for addressing particular empirical problems (Chapters 3 and 5) (Laudan, 1978).

The problem-solving view of science has similarly been applied to HCI in particular (Oulasvirta & Hornbæk, 2016). Adding to the problem-solving view articulated by Laudan, this work suggests a third type of problem, constructive problems, which HCI research additionally aims to solve. Constructive problems are those for which the principles of construction of an artefact fit for a particular purpose is insufficiently understood (Oulasvirta & Hornbæk, 2016). This sort of problem is solved in HCI by design and engineering which construct such artefacts but also by the proposition of design principles which guide the construction of such artefacts. The aforementioned paper outlines five criteria for problem-solving capacity by which research can be evaluated: significance, effectiveness, efficiency, transfer, and confidence, building on criteria from Laudan (Oulasvirta & Hornbæk, 2016). Significance here refers to the extent to which solutions are important to particular stakeholders of the research at hand; effectiveness refers to the extent to which a solution addresses the central aspects of the problem which it solves; efficiency refers to the relative relationship between the benefits of a solution and the cost of obtaining it; transfer refers to the ability of a solution to be applied to similar problems; and confidence refers to reliability, consistency, or robustness of a solution to hold up against other possible solutions or new problems of the same type (Oulasvirta & Hornbæk, 2016). No particular metrics are suggested by either that work nor by Laudan for measuring the extent to which research output meets these criteria. Nevertheless, systematic targeting of one criteria or another when planning research is suggested as a way to increase problem solving capacity in HCI, such as sharing materials and code in order

to improve efficiency for other researchers, or replicating existing research in order to improve confidence in the solutions (Oulasvirta & Hornbæk, 2016). This thesis approaches HCI research with this problem-solving approach, contributing constructive solutions in the form of design recommendations (Chapters 3, 4, 5, and 6) as well as prototyping of an artefact which applies these recommendations (Chapter 6).

2.4.2 Application and abstraction continua

Experimental tasks in HCI multitasking research vary across a continuum of application ranging from highly contrived laboratory tests to simulations of real scenarios to highly applied, real-world experiments (Salvucci & Taatgen, 2014). These different levels of application, described below, involve different stakes (e.g. errors on laboratory tests of memory are less costly than errors when driving an actual car) and different levels of control of confounding variables (e.g. laboratory studies allow for more control of dependent variables). As such, each has particular advantages and disadvantages in HCI research. While this thesis conducts research at the middle level of application, balancing application and control, it draws upon knowledge from all levels, each of which make crucial contributions to HCI.

Basic psychology studies

Much of the HCI multitasking research conducted through laboratory studies has looked at basic psychological tasks such as tracking visual stimuli (Janssen et al., 2015; Monk et al., 2008; Nijboer et al., 2013), simple arithmetic problems (Borst et al., 2010; Janssen et al., 2015), counting audio stimuli (Nijboer et al., 2013) or simple tests of memory such as n-back tasks in which a list of letters or numbers is presented one at a time and participants judge whether the character matches the one that appeared N characters previously (Borst et al., 2015; Katidioti et al., 2014; Monk et al., 2008). While studies of this type may involve tasks which are discrete or continuous, the tasks tend to be elementary rather than compound, with response options heavily constrained thus limiting the amount of decision making required. An exception to this tendency is in the use of simple computer games, as by McFarlane (McFarlane, 2002; McFarlane & Latorella, 2002), as an example of a complex tasks with some restricted decision making requirements, thus creating multitasking laboratory experiments which involved complex continuous tasks which was not a simulation of any

other, more complex and applied task. Basic psychology studies are useful in isolating sources of interference between tasks as discrete cognitive processes can be tested with minimal risk of confounds, crucial to testing theoretical models like MRT (Wickens, 2002) or the Memory for Problem States model for interruptions (Borst et al., 2015). In this way, basic psychology studies contribute empirical evidence for solving conceptual problems in HCI by validating and testing theories (Oulasvirta & Hornbæk, 2016).

Prototypes and simulation

At a higher level of application, an abundance of multitasking work has utilised simulation of complex, real world tasks. These studies of multitasking range from low fidelity simulators that capture individual elements of a complex task such as simulating the task of driving by only presenting the task of steering (Brumby et al., 2011; Brumby et al., 2007) to higher fidelity simulators that include a full range of behaviours involved in a complex task, such as full driving simulator studies (Large et al., 2017; Martelaro et al., 2019). Likewise, simulations of other complex task environments have been used to study multitasking among particular professions such as pilots in simulated cockpits (Wickens et al., 2009). Studies of this type allow for research to generalise findings to the applications simulated therein without incurring the dangers that real world error would bring.

Other work at the level of application between basic psychology experiments and real-world studies of technology include studies utilising prototype computer systems, Wizard of Oz paradigms, or scripted interactions with system. These types of laboratory study remove the burden of creating a fully functioning system or using an off-the-shelf system that the experimenter cannot fully control. Wizard of Oz experiments involve telling participants that a computer system works in a particular way while, unknown to the participant, certain operations they believe to be automated are actually controlled by a human experimenter. These sorts of studies have been used particularly to deliver interruptions at specific times (Hudson et al., 2003; Iqbal & Bailey, 2005) or to explore multitasking in a controlled way with systems otherwise prone to technical failure (Edwards et al., 2019). Alternatively, laboratory controlled HCI tasks may be used like simulators, to observe multitasking behaviour in particular domains without exposing participants or others to danger. Fields like medical

device preparation (Gould et al., 2016) and information technology based office tasks like managing emails (Dabbish & Kraut, 2004; Li et al., 2008) have been studied in this way, so that complex tasks within a job can be studied in isolation from other parts of the job. Studies using prototypes or simulation allow greater generalisability about the system and the task contexts that they approximate as compared to basic psychological studies, while affording more experimental control than real-world studies. For this reason, studies at this level of application are critical for exploring potential design elements for new systems and for understanding how particular contextual elements affect interactions. These types of studies therefore tend to combine empirical and constructive problem solving, evaluating designs and providing evidence which motivates future design decisions

Real-world studies

The most applied level of HCI multitasking research includes studies conducted in the real-world, in the actual domain of interest to the study. This may include, for example, on-road studies of driver interruptions or studies of computer use outside of laboratory conditions. In the case of driver-distraction studies, on-road work has ranged from test track studies to live traffic studies. Test track studies have been used to give some protection to members of the public while still allowing experimenters to study multitasking strategies (Horrey & Lesch, 2009) and performance effects (Yager et al., 2015). These studies are a bit more applied than simulator studies in that participants operated a real car on the road, meaning consequences for mistakes could be more severe than in a simulator, but they nonetheless were controlled in that other traffic was not present, so the property or safety of others was not at risk. This level of application allows for a complex task to be studied in its totality while still insulating against the risks of failure of the task.

Other driving studies have examined uncontrolled on-road driving behaviour, as a means of deploying a system for testing in a particular context. One such example installed a speech interface in police cruisers to examine how the interface would be used in complex and unpredictable multitasking scenarios (A. Kun et al., 2004). Another study which examined on-road driving combined prototyping and real-world application by placing a prototype proactive speech agent in a car and asking participants to drive a predetermined route while the agent intermittently asked if it was

a good moment to interrupt them (Semmens et al., 2019). While driving is a popular complex task in this sort of multitasking research, other studies in non-driving contexts have also combined prototypes with real world application. One study gave participants a prototype whispered-speech interface for journaling their daily lives or for use as a personal assistant for four days to understand how they would integrate the artefact into a broad set of complex tasks (Parviainen & Søndergaard, 2020). Another, focusing on cooking as a complex task, allowed participants to use speech to control a recipe video while cooking the meal from the recipe to examine multitasking behaviour in that specific task (Zhao et al., 2022). Studies like these allow for highly ecologically valid observations of complex tasks, but at the expense of control as unintended or unexpected behaviour from either the participant or the environment may create confounds. This type of study, like those with less application, inform constructive problems by introducing designs for technology into the contexts in which they are intended to be used (Oulasvirta & Hornbæk, 2016).

Highly applied HCI multitasking research has involved home, workplace, and public studies of technology in use. These include studies measuring people's interruptibility during their working routine (Hudson et al., 2003) as well as corresponding studies investigating people's interruptibility in their naturalistic home routines (Nagel et al., 2004; Vastenburg et al., 2008). In-home studies of multitasking behaviour have focused on media multitasking (Rigby et al., 2017; Voorveld & Viswanathan, 2015), the concurrent consumption of multiple forms of media. In the workplace, techniques like logging or direct observation have likewise been used to observe the multitasking behaviours of office workers (González & Mark, 2004; Mark et al., 2015; Mark et al., 2012). This level of application can be attained without high risk in some office and home environments, but collecting data in these ways may not be suited for other contexts. Highly detailed logging or observation of people's technology use is a rather invasive data collection strategy and may pose ethical risks when technology is used for sensitive matters or by vulnerable populations like children. Likewise, collecting data at this high level of application may lead to samples biased by idiosyncrasies of individuals or workplaces. And while naturalistic observation may on the one hand be more representative of in-the-wild interaction, this sort of monitoring may create unnatural patterns of behaviour as people may act differently knowing they are being monitored. All levels of application therefore have strengths and weaknesses, requir-

ing HCI multitasking research to combine knowledge from many techniques.

2.4.3 Experimental design

Possible futures

Salovaara and Oulasvirta describe the role of experiments which use prototypes as evaluating possible futures (Salovaara et al., 2017). In this framework, design decisions are seen as aimed at either staging - making the present have some characteristic of a possible future - or controlling - preventing particular characteristics of the present which are not expected to be part of the imagined future from becoming salient (Salovaara et al., 2017). These design techniques are the specific mechanism by which control and application can be traded off along the application continuum (Salvucci et al., 2009) as interaction features are selected to either increase the extent to which a laboratory study is like the real world or to decrease the extent to which real-world variance confounds variables of interest. This thesis aims to inform the design of future technology. Its experiments therefore are aimed particularly at imagining and testing possible futures of interactions with those technologies.

A number of techniques for experiment design are proposed as useful for controlling: namely, narrowing, stabilising and removing, inhibition, and gamification (Salovaara et al., 2017). Narrowing is the process of reducing the scope that an experimental interaction has in representing a specific future by only examining particular future design features. An example in this thesis includes that interactions with an imagined proactive speech agent in Chapter 6 only represent the agent speaking to the user and the user responding, but no action or response from the agent afterward. While the future this interaction represents would likely include some follow-up from the agent, the scope of the study is narrowed to focus on one particular component of the interaction. Stabilising and inhibition are a technique by which the variation of elements of an interaction are limited through scripting of variance or through constricting natural variance (Salovaara et al., 2017). Studies in Chapters 3 and 4 use stabilising and inhibition when examining spoken interactions between participants who are asked to speak to a person who is engaged in another task. In these studies, while participants are free to vary the way they phrase and intone their utterances, their variance is stabilised via particular prompts given to them for each utterance. Likewise, while the task their partner engages in varies from experimental trial to trial,

each task is prerecorded so that participants all see the exact same variations in the state of the task, greatly reducing variance. Removal is an even more extreme version of this technique by which some unwanted variance is eliminated entirely (Salovaara et al., 2017). While this thesis imagines a future in which proactive speech agents initiate interactions with users across a wide variety of complex tasks, the variation of task time is removed entirely, with all studies focusing on a single task, the computer game Tetris, discussed in more detail below. Finally, gamification refers not to design decisions like using a computer game as a task, but instead to creating a competitive, game-like research paradigm in order to motivate participants to behave as if stakes are high or to reduce the social awkwardness of particular tasks (Salovaara et al., 2017). This technique is used in Chapters 3 and 4 in which participant motivation to interrupt the Tetris gameplay task as appropriately as possible is encouraged through gamification by which participants are told they are rated and scored with the best performing participant earning a prize. While each of these techniques aims to increase the control the researcher has over interactions, controlling techniques nonetheless allow for experimental research to be less confounded by noisy elements of the present to instead focus on the features of the potential future interactions they aim to inform.

Staging has the opposite effect of controlling, making laboratory studies more like the applied interactions they seek to represent rather than reducing that application to increase control through techniques like propping, stage selection and feature promotion, repetition, and recruiting (Salovaara et al., 2017). Propping is the use of prototypes or mockups of artefacts that do not yet exist standing in for things that might one day exist. In Chapter 6 of this work, participants observe interactions with such props as two mockup examples of proactive speech agents are presented to represent futures where one or both such agents exist in the real world. Setting selection, while not used in this thesis, is the tailoring of where and when an interaction occurs so that particular environmental features are present (Salovaara et al., 2017), such as deploying a prototype for interactions about culture at a museum (Candello et al., 2019) or deploying a social media prototype at a moment in which conversational volume would be high such as a major sporting event (Jacucci et al., 2005). Feature promotion is the emphasis of a particular element of a prototype by drawing particular attention to that feature throughout an interaction (Salovaara et al., 2017). In all

studies of this thesis, the timing of proactive speech is highlighted for participants to pay attention as a means of better understanding this particular aspect of future interactions. Repetition involves exposing participants to a particular feature frequently as part of the context of the interaction so that repeated exposure to a feature feels natural (Salovaara et al., 2017). Chapters 3 and 4 use a gamified approach to accomplish this, telling participants that proactive speech interactions occur repeatedly as part of the structure of a game, allowing for justification of the many instances of proactive speech taking place throughout the experiment, reducing the unnatural exaggeration of this interaction feature. Finally recruitment involves choosing particular participants with relevant experience or skills to use a prototype so that its features can be utilised to a particularly high degree (Salovaara et al., 2017). This thesis uses a general sample of participants, but recruitment of this type is not uncommon in multitasking studies, such as deploying prototype interfaces for in-car interactions within police cruisers where sample participants already have experience multitasking and driving (A. Kun et al., 2004). Staging techniques, like controlling techniques, allow for experimental studies which are limited to the circumstances of the present moment to imagine and evaluate possible futures. This thesis takes seriously this role of experimentation in its aim to envision and evaluate future HCI interactions, and the experimental design decisions herein draw on the techniques which enable the present-future gap to be narrowed.

2.4.4 Mixed methods research and qualitative data analysis

Research that uses both qualitative methods and quantitative methods - mixed methods research - has a number of advantages over research which only uses one or the other set of methods. Recent meta-scientific research has discussed the need for mixed methods research to present an explicit value-add over and above the value of each of the methods it uses in isolation (Creswell, 2011). A number of potential rationales or sources of value have been proposed for choosing a mixed methods approach, including triangulation, expansion, exploration, and multiple research questions (L. Doyle et al., 2016). Triangulation refers to using results from quantitative and from qualitative methods to mutually corroborate findings and to resolve sources of surprise and ambiguity. This fits within the framework of Laudan (Laudan, 1978) or Kuhn (Kuhn & Hacking, 1962) by which realist aims of finding truths are superseded by

problem solving aims of resolving ambiguity. Expansion and exploration are rationales for mixed methods research which have alternate aims either using qualitative methods to explain qualitative results or using qualitative methods to develop hypotheses for quantitative study (L. Doyle et al., 2016). These aims echo the transfer between solving empirical and constructive problems in HCI, by which artefacts are built and evaluated so that problems of each type can be addressed (Oulasvirta & Hornbæk, 2016). Finally, mixed methods research may be used for addressing distinct research questions, some of which have qualitative aims and others which have quantitative aims (Creswell, 2011; L. Doyle et al., 2016). This thesis primarily uses mixed methods studies for the purposes of triangulation and for posing multiple research questions. Research questions in Chapters 3 and 4 ask how different contextual factors impact the way people speak when interrupting a busy person. By using mixed methods, those studies both quantify the effect of particular variables on the timing and speed of speech, but they also describe the strategies by which those variables influence people's behaviour. In this way, the ultimate causal relationship between variables of interest can be described quantitatively, but the proximate causal pathways - explanations of how and why those relationships exist - can also be described qualitatively thus discovering new problems for future research.

Qualitative data in this thesis is gathered through the use of open-ended questions in online forms presented to crowdworker participants. In HCI research, open-ended survey questions have shown to be effective for answering questions about people's intentions, attitudes, user experience, and awareness of elements of interactions (Müller et al., 2014). Chapters 3 and 4 involve open-ended questions around people's intentions for their interactions in those experiments, while Chapter 6 uses an open-ended question to ask about people's awareness of differences between two different prototypes. The data generated by participants in these experiments is analysed using reflexive thematic analysis (Braun & Clarke, 2006). Reflexive thematic analysis is an approach to describing themes in qualitative data which is flexible to different theoretical and methodological frameworks. For example, coding for reflexive thematic analysis can begin with inductive or deductive coding, and the resulting themes can be analysed from a post-positivist perspective by which themes point toward truths about the underlying phenomena or from a constructionist perspective by which themes describe the realities co-created by the participants and data ana-

lysts but not necessarily toward an objective reality which does not depend on those parties (Braun & Clarke, 2006). Reflexive thematic analysis therefore allows for triangulation of answers to research questions throughout this thesis using a consistent set of qualitative analysis methods from study to study due to its flexibility.

Thematic analysis in this thesis adapts selection of coding philosophies (e.g. inductive or deductive) to each study and its research questions, but thematic analysis here is fundamentally presented as a constructionist exercise. While quantitative findings are reported via null-hypothesis significance testing and confidence intervals, quantitative methods for describing the statistical probability of effects within the sample generalising across the population, this sort of generalisability is not an aim of the thematic analysis used in this thesis. Qualitative analysis here describes the intentions and awareness of particular people within an interaction and their responses are coded and grouped into themes after familiarisation with both the data and with the experiment from which that data was generated. As such, interrater reliability is not reported or measured in qualitative analysis for this thesis, as themes are not presented as if they contain realist content. Instead, themes are described and illustrated with data extracts to convey the constructionist meaning contained within them (Braun & Clarke, 2006). This does not however imply that qualitative themes in this study are not generalisable. Instead of generalising in the quantitative sense, that similar results would be likely to occur across members of the population, this thesis presents qualitative results which may generalise in terms of transferability and analytical generalisability (Smith, 2018). That is to say, the intentions and the awareness of participants from this thesis may not be the same as those that other people would have experienced even in similar contexts. Nonetheless, insofar as this thesis considers how people communicate and how future technologies should be designed, the specific and idiosyncratic data generated by these participants may be valuable in shaping understanding of communication in a general sense or in transferring to design guidelines for proactive speech agents even if they are unique to participants in this study. In this way, this work takes a constructionist approach to using qualitative data as part of a toolbox for addressing conceptual and constructive problems in HCI (Oulasvirta & Hornbæk, 2016) as part of a broader mixed methods approach to HCI research.

2.4.5 Tetris as a complex task

Modelling continuous tasks is an ongoing challenge in cognitive science and HCI, often sidestepped by researchers opting to instead use one or several simple tasks or discrete tasks, generalising insights about human cognition from those. This approach is not adequate however, and some research in the domain of skill acquisition has highlighted the need for social science to focus on actual complex tasks to better understand them (Wulf & Shea, 2002). This echoes the longstanding call in cognitive science, first put forth by Newell in the early 1970s, for a single complex task which the field accepts in its entirety - without abstraction or alteration - as a target phenomenon of study (Newell, 1973). Where Newell himself proposed chess (Newell, 1973), a complex game with a variety of choices available to the player, this proposal was nonetheless a discrete task which negates the penalty of time costs for decisions. Lindstedt and Gray instead proposed Tetris as a candidate task for advancing our understanding of human cognition (Lindstedt & Gray, 2019). Like chess, Tetris is a popular game enjoyed by many people at a variety of levels of expertise. It involves complex, dynamic decision making and recruits several cognitive processes including visual monitoring, motor skills, and strategic planning. Tetris goes beyond chess however in that it is a continuous task, such that Lindstedt and Gray note that even hesitation is a decision which incurs an opportunity cost (Lindstedt & Gray, 2019). Importantly, Tetris has a repeated event structure (Zacks & Swallow, 2007), a sequence of discrete events that constitute the entire complex task and can be learned by the Tetris player. In this way, Tetris is what Lindstedt and Gray describe as a “manageably complex” task (Lindstedt & Gray, 2019) which can be learned and mastered by a player and can be understood and quantified by a researcher owing to the patterned and separable event structure that underlies it.

The event structure of Tetris can be understood as a series of episodes which can be further deconstructed into discrete motions, either initiated by the player or automatically triggered at particular time intervals. Lindstedt and Gray describe the event structure of Tetris in detail in (Lindstedt & Gray, 2019). An episode in Tetris constitutes the travel of a Tetris game piece from the top of the Tetris game board as it falls to either the bottom of the 20-row board or can no longer continue to fall due to other Tetris pieces positioned beneath it. The Tetris piece falls at a continuous rate determined by the game difficulty level. During this falling process, the player can ini-

tiate movements including lateral translations of the Tetris piece, 90 degree rotations of the piece, or manual vertical drops, bringing the piece closer to the bottom of the board. When a row of the gameboard is completely filled by parts of Tetris pieces, that row or rows flash for several frames before disappearing, with all filled cells from rows above falling to fill the vacated row(s). In any other cases, after a piece reaches a point at which it can no longer continue to fall and a falling motion is either automatically triggered or initiated by a player's manual drop, that piece remains in its present position and a new episode begins with the appearance of a new Tetris piece at the top of the game board. This event structure is that by which previous work has sought to understand how Tetris players approach the task of Tetris and gain expertise in it (Lindstedt & Gray, 2019). It is unclear however whether this event structure cleanly comports with the way someone observing a Tetris game but not participating in it would conceptualise the task of Tetris. This event structure of Tetris is premised on a known set of goals and subgoals for the task of playing Tetris (Zacks et al., 2007). The process of interrupting a task requires not only the understanding of the goals of the person carrying out that task, but also the balancing of those goals against the goals of the interrupter, which may compete with the interrupted task for time and the attention of the interrupted party. This thesis therefore uses Tetris throughout all of its studies as a complex, continuous task representative of any number of real-world and high-stakes tasks of the same characteristics, in order to make inferences about interruptions of tasks of that sort without endangering participants or bystanders.

2.5 Summary

Speech interactions with non-human agents have become a popular interaction modality in HCI. Building on technological advances and inspired by theoretical benefits of interacting with computers while eyes and hands are busy, interactions with speech agents are seen as a promising avenue for achieving long imagined HCI and science fiction interactions. These interactions thus far have underwhelmed users however, with the benefits of multitasking not well-realised by currently available technologies (Luger & Sellen, 2016). One potential reason for the disappointing utility of today's speech agents is their lack of true proactivity, hindering the value of multitasking with them. While speech agents promise the benefits of proactive agent interaction which

have inspired optimism for decades (Shneiderman & Maes, 1997), they as-yet deliver the command-and-control interactions which are already realised by traditional interactions with digital technology.

Designing speech agents of the near future requires an understanding of people's expectations of dialogue partners, the cognitive resources employed during multitasking, and the costs and benefits of speech during other tasks. Multiple Resources Theory implies that speech interactions during visual-manual tasks can be beneficial while minimising disruption so long as they also seek to avoid overlap in terms of stages of perception vs execution (Wickens, 2002). Managing this overlap requires monitoring of the ongoing task by the interrupter, seeking to request reactions to speech in low-activity moments such as the natural breakpoints described in past multitasking literature (Janssen et al., 2012). But understanding how these moments are identified and how disruption cost can be minimised is still a major question in HCI, with recent empirical work finding difficulty in doing so purely by modelling ongoing tasks (Cha et al., 2020; Semmens et al., 2019). This thesis therefore leverages human spoken interruptions to better understand the cues and strategies people employ when interrupting complex tasks using speech. By modelling human spoken interruptions, it becomes possible to design non-human speech agents for which people have more accurate models as dialogue partners, reducing the surprise and disappointment caused by a dialogue partner not meeting one's expectations. These human-inspired, better-modelled proactive speech agents may then be a means of better meeting people's expectations for speech agents. This thesis therefore sets out at offering candidate solutions to multiple HCI problems (Oulasvirta & Hornbæk, 2016): the conceptual problem of understanding people's strategies and behaviours when using speech to interrupt complex tasks, the empirical problem of quantifying the effect of particular variables on spoken interruptions, and the constructive problem of designing speech agents which provide value to users by leveraging human behavioural characteristics to meet user expectations.

3 * **Eliciting Speech Interruptions to Investigate the Impact of Urgency**

3.1 Introduction

3.1.1 Interruptions and complex tasks

This study develops a paradigm for eliciting speech interruptions of a complex task to examine the way people structure, time, and strategise these interruptions. To do so, it is necessary to select a complex task and to design a research plan which might be both relevant to the way people interrupt other people and which might be informative to the design of proactive speech agents. Because the complex task at hand would be interrupted with speech, it was necessary that the primary modality of the task would not be auditory or speech, as tasks sharing in modality are more disruptive to one another (Wickens, 2002). Insofar as this research aims to inform the design of speech agents, which are frequently used to multitask while engaged in tasks that call for continuous visual monitoring and intermittent manual reactions like driving, cooking, or monitoring children (Luger & Sellen, 2016; Olson & Kemery, 2019), it was also important that the complex task have these features as well. Computer games were identified as a suitable potential task as they met these criteria while being sufficiently controlled, minimising the unsystematic variance introduced by highly ecologically valid but less controlled tasks like on-road driving. Using computer games as complex tasks for the study of interruptions follows in the tradition

of McFarlane, who studied interruptions of complex tasks through experiments using computer games for many of the same reasons (McFarlane, 1999, 2002). Tetris was specifically selected, following recent work which identified Tetris as a fruitful task for understanding a variety of features of human cognition, terming it “manageably complex” for its simultaneous complexity, familiarity, and its nature of having a highly controlled, designed structure of subtasks (Lindstedt & Gray, 2019). Tetris therefore satisfied several conditions making it a suitable complex task for the study of interruptions of such tasks.

McFarlane identified eight dimensions of human interruption, synthesising the established theoretical constructs which may be studied in investigation into the phenomenon (McFarlane, 1997). Of these, method of expression, the dimension describing the ways by which an interrupter may vary the way they request attention, including choices like the modality through which they interrupt (e.g. speech or gesture), their use or disuse of politeness, and their adaptations to the timing of their interruption in relation to the other task (McFarlane, 1997) was identified as particularly interesting. This interest in part stems from prior proactive speech work focusing on interruption timing (Cha et al., 2020; Semmens et al., 2019), and a growing literature suggesting that the manner of speaking is likewise an important both for capturing the attention of a multitasker (Wong et al., 2019) and for shaping people’s perceptions of a speech agent (P. R. Doyle et al., 2021). While this thesis focuses exclusively on speech as a modality, the design of the present study sought to elicit interruptions which were free to vary otherwise in method of expression, in order to better understand how and why spoken interruptions vary. This study specifically examines both the timing and strategies of access used in human spoken interruptions, seeking to understand what considerations impact people’s decisions around them.

3.1.2 Interruption timing

To understand the timing of human spoken interruptions, two variables of interest were selected: interruption onset - the amount of time it takes for a person to begin their spoken interruption after being instructed to interrupt - and interruption duration - the total amount of time that an interrupting utterance takes. Interruption duration is one measure of the costliness of an interruption which is relevant to the interrupted party. In a meta-analysis of studies which involved switching between tasks, Wick-

ens and colleagues found that the amount of time that a secondary task takes was a significant factor in multitaskers' decisions about task-switching, with people seeking to avoid switching to tasks which take longer (Wickens et al., 2015). In empirical studies of multitasking, people demonstrate task-switching strategies which seek to minimise time spent on secondary tasks, even when this comes at the cost of safety (Brumby et al., 2011; Horrey & Lesch, 2009). This fits within the Soft Constraints Hypothesis within cognitive science, which stipulates that people choose their strategy for complex tasks by seeking to minimise the time spent on that task, prioritising time over other factors like effort (Gray et al., 2006). Taken together, this evidence makes it clear that multitaskers seek to minimise the duration of secondary tasks when making multitasking decisions for themselves. It is not clear however whether this principle extends to how people strategise interruptions of other people. For that reason, interruption duration was a variable of interest for the present study.

Interruption onset, the amount of time an interrupter waits to decide when to interrupt another person, is another important variable for the present study. On the one hand, interrupters may seek to interrupt as quickly as possible in part to minimise their own time spent on the task of interruption, following the Soft Constraints Hypothesis (Gray et al., 2006). They may likewise seek to minimise interruption onset as a means of minimising the total time that the person they interrupt spends on their combination of tasks, seeking to get the interruption over with as quickly as possible. On the other hand, it may be the case that people choose different factors in deciding on when to interrupt each other. Interruptions are considered less disruptive if they come at subtask boundaries, the natural breakpoints between parts of a task (Borst et al., 2015; Janssen et al., 2010). Whether interrupters use natural breakpoints when interrupting others, is an unexplored question which this chapter raises and which Chapter 5 investigates quantitatively.

3.1.3 Access in speech interruptions

For strategies of social access, access ritual usage was identified as a potentially fruitful measure. Goffman coined the term access rituals in 1971, describing them as "both greetings and farewells: they are ritual displays that mark a change in degree of access." (Goffman, 1971, p.79). Access rituals and other social rituals are broadly grouped together in Goffman's description of supportive interchanges, the ritualised

Eliciting Speech Interruptions to Investigate the Impact of Urgency

communications by which people attest to their own civility and good will to an interlocutor (Goffman, 1971). In this way, access rituals in the form of greetings can be seen as a patterned behaviour signalling a request for social access. Since a spoken interruption is a fresh request for access which must be ceded from another task, access rituals have the potential to mark such a request in any given spoken interruption, or to show a deferred opportunity to make a request, instead demanding access, when they are absent. Whereas Goffman's initial description of the access rituals consisted of a sketch of the phenomenon with a handful of imagined examples, work that followed set about eliciting greetings and farewells and recording the range of access rituals which people use for each of them. Krivonos and Knapp described 11 types of verbal greeting behaviours, including verbal salutes (e.g. "hi" or "hey"), reference to the other person (i.e. address by name or nickname), apologizers (e.g. "excuse me") as well as ten types of nonverbal behaviours including head gestures, eye contact, and smiling (Krivonos & Knapp, 1975). These greeting rituals were demonstrated in face-to-face communication between acquaintances, and little work exists updating this list of greetings over time or for contexts that do not include co-presence or visual access between parties. Nevertheless, access rituals serve as a well-established foothold for examining social access strategies in the initiations of conversation, providing additional insight into the holistic manner of expression (see McFarlane, 1997) of human speech interruptions.

Access rituals also have a function of signalling intent to participate in a joint project. This communicative function of access rituals, signalling an interrupter's intent to take the floor and command attention from their partner functions similarly to turn-taking cues in dialogue without a secondary task. In dialogue, transitions between turns are signalled by the speaker offering to yield their turn rather than by their interlocutor seeking to begin a new turn (Duncan, 1972; Yngve, 1970). A variety of cues including intonation, gesture, changes in pitch or volume, or use of verbal access rituals like "you know?" at the end of a turn (Duncan, 1972). In interruptions however, the primary task does not provide a signal that an unrelated task is ready for attention. Instead of primary tasks signalling an intention to cede the floor, requests to take the floor must be made by the interruption. Interruptions which do attempt to signal their intention to take the floor have been termed pre-alerts (van der Heiden et al., 2017) or negotiated interruptions (McFarlane, 2002). Pre-alerts provide some

signal of the intent to interrupt, following this signalling with the interruption content after some interval of time (van der Heiden et al., 2017). Negotiated interruptions provide a signal of the intent to interrupt followed by a chance for the interrupted party to defer or decline the interruption or otherwise to accept the content of that interruption (McFarlane, 2002). For this reason, the present study seeks to measure the use of access rituals both in their social function and to understand whether interrupters use them as strategies similar to pre-alerts to signal their intention to take the floor before doing so.

3.1.4 Urgent speech

In order to observe variance in these measures and to begin to understand the causes of that variance, this study seeks to manipulate an independent variable which could mutually be utilised by a human interrupter or by a proactive agent which this work aims to inform. Urgency was identified as a fruitful variable for a number of reasons. Urgent speech differs from non-urgent speech in a variety of ways, including prosodic differences like increased speech rate, pitch, and vocal intensity (Landesberger et al., 2020a, 2020b) as well as semantic differences insofar as listeners perceiving some spoken words as conveying more urgency than others, independent of how they are delivered (Hellier et al., 2002). Urgency was also an interesting variable in that prior work on interruption scheduling systems has established that urgency is a critical factor for people who are being interrupted in determining the appropriateness of an interruption's timing, with people being more accepting of urgent interruptions (Iqbal & Bailey, 2010; Vastenburg et al., 2008). Furthermore, urgency might be a variable accessible to a proactive agent, in that certain applications or contacts might be defined as urgent by a user, or particular interruption types (e.g. hand-over requests from automated driving systems) might be defined as urgent by the agent itself. Following recent work which sought to elicit urgent and non-urgent speech through a gamified approach (Landesberger et al., 2020a, 2020b), the present study likewise used a gamified reward structure to manipulate urgency when eliciting speech from participants in the present study. The gamified approach in that work used only a singular task: a game in which participants would ask questions to determine which object on a screen was their target. Like in the present study, urgent trials were explicitly marked as urgent, and participants were instructed that solving them quickly would

result in more points toward a final total score which would determine a cash reward (Landesberger et al., 2020a, 2020b). By manipulating urgency, this study seeks to investigate whether that variable causes interrupters to choose different interruption strategies which are reflected in the timing and style of their spoken interruptions.

3.1.5 Aims, hypotheses, and contribution

This study investigates how people interrupt other people who are engaged in a complex task through spoken interruptions. The study introduces a new experimental paradigm for eliciting spoken interruptions of a complex task. The paradigm manipulates urgency as a within-subject independent variable both to explore the effect of that variable on interruption strategies and to establish whether the paradigm is sufficient for eliciting different interruption strategies at all. By recruiting human participants to serve in an interrupter role, this study seeks to better understand spoken interruptions through a mixed-methods approach, investigating both when and in what way people interrupt other people using speech. Based on the previous research on urgent speech and interruptions outlined above, the present research hypothesises that urgency will have a statistically significant effect on interruption onset (H1) and interruption duration (H2). This work further hypothesises that use of access rituals will statistically significantly vary depending on the urgency of the interruption (H3). Through qualitative data, the study also aims to more deeply explore the various approaches that participants used speech to interrupt people engaged in another task. The research contribution of this study is both methodological, in establishing a paradigm for eliciting spoken interruptions of continuous tasks, as well as empirical, in testing those hypotheses to better understand human spoken interruptions and inform the design of spoken interruptions from non-human agents.

3.2 Methods

3.2.1 Participants

52 crowdworkers (26 women, 24 men, 2 preferred not to specify; $M_{\text{age}} = 29.4$ years, $SD = 7.9$ years) were recruited from a crowdsourcing platform (Amazon Mechanical Turk). All participants were native or near-native English speakers. Participants were

all familiar with the game Tetris, with most indicating that either they had played before, but do not play regularly (N = 44; 84.6% of sample) or that they play regularly (N = 3, 5.7% of sample) (5 point Likert scale; 1 = I am not at all familiar with Tetris; 5 = I regularly play Tetris). The study took approximately 20 minutes and participants were paid \$10 Mechanical Turk credit for participating in the research. The study received ethical approval through the university's ethics procedures for low risk projects (Ethics code: HS-E-20-161).

3.2.2 Materials

Tetris Interruption Paradigm

This study sought to explore how people interrupt a partner who is executing a primary task that requires ongoing attention and cannot be arbitrarily suspended (continuous) and allows for a broad variety of responses rather than a single fixed response (compound) (Kieras et al., 2000; Salvucci, 2005). To satisfy these requirements, an experimental paradigm was devised around Tetris as the primary task. The paradigm was designed to ensure that the interaction context could believably be conducted online. Participants were told that they would be interacting with a remote partner who would be playing Tetris online and that they would have to deliver spoken interruptions to this partner. The Tetris paradigm was built using JSPsych version 6.3 (de Leeuw, 2015). Further details of the paradigm design are outlined below and all code and stimuli are available at ¹. All materials including participant information sheets, consent forms, questionnaires, and debrief sheets are provided in Appendix A.

Tetris Task

The trials within the paradigm used recorded rather than live Tetris gameplay in order to standardise the materials across all participants, controlling for potential variability between the stimuli (e.g. variability within Tetris players and Tetris game states). That said, in order to maintain engagement and to elicit interruptions reflective of how people interrupt other people, participants were told at the start of the experiment that the pre-recorded videos were a live feed of a person they had been matched with who was currently playing Tetris. The experiment involved 2 practice trials followed

¹https://osf.io/uwseq/?view_only=f7e4fb40aafb489bace7bb822c4478d2

Eliciting Speech Interruptions to Investigate the Impact of Urgency

by 16 experimental trials. These trials were generated from 3 minute videos of actual Tetris gameplay conducted by the lead author. Each trial was chosen to ensure that the game state reflected one in which the Tetris player was not at risk of losing when the interruption occurred. Specifically: 1) a Tetris game piece started at the top of the game board; 2) there were at least two rows and no more than half of the rows of the board which already contained Tetris pieces and 3) the falling speed of the game piece was set to the game minimum of 1.25 rows per second. Each trial was presented as a video on a webpage. Videos included a Tetris board and a box in the upper right corner indicating the next piece. Videos were presented at an 800×800 resolution, in colour, on a neutral background, and without sound.

Interrupting Task

Similar to other interruption research (Kubose et al., 2006) participants were tasked with completing a set of interrupting tasks, requesting information from the Tetris player. Once a trial had started, a message would appear on-screen instructing the participant that they needed to request a certain piece of information from their partner. Messages appeared in large black font in a single line on the screen directly below the Tetris video after a random delay between 5 and 15 seconds. In each trial, participants were told what information they needed to request from their partner. To encourage naturalistic generation of utterances, the messages instructing participants on what to ask their partner included only key words rather than full, grammatically complete questions. Specifically, these messages instructed participants to “in your own words, ask your partner:” followed by keywords (e.g “ask your partner last movie watched”). This was to ensure that participants were not led to read aloud or directly use the question prompt when forming their interruption utterance. The prompt was displayed throughout the trial so participants did not need to remember the prompt, allowing them to focus on interruption planning.

Questions focused on requesting details about their partner (task prompts are included in Table 3.1). These were used for two reasons. Firstly, participants would not know the answers to these questions and thus would not be tempted to answer on their partner’s behalf. Secondly, these questions would all be of similarly low difficulty for their partner to answer. This meant the responses could believably be generated after a uniformly short delay, enhancing the realism of the paradigm. Additionally,

questions with simple answers were chosen to ensure that participants did not greatly vary their interruptions to account for varying difficulties of questions.

Interruption prompts: "In your own words, ask partner: _____"	
which hand using	last movie watched
any pets	favorite ice cream flavor
weather	what breakfast this morning
bed time last night	been to Paris
age	favourite fruit
last series watched	favourite colour
any siblings	lucky number
what dinner last night	keyboard colour

Table 3.1: *Table of interruption prompts.*

Partner's Rating of Performance

To keep participants engaged with the task they would interrupt (the Tetris game), participants were told "After each round is finished, your partner will be asked to rate how well you did in terms of how disruptive your question was. Your partner will be asked how much they agree with the following two statements: 'My partner's question came at a good moment.' and 'My partner's question did not distract me.'" Participants were told that these ratings determined a final score and that the participant with the highest total score at the end of the experiment would receive a bonus reward. In reality, one participant was randomly selected to receive this bonus reward, and all participants were equally compensated for their time.

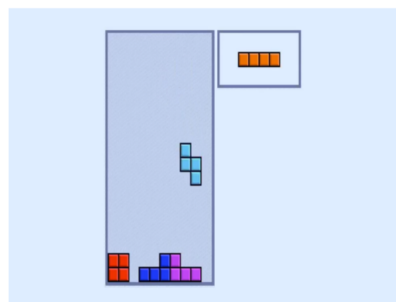
Simulation of Player Responses to Interrupting Task

Prerecorded responses were used to answer the questions posed by the participant. These responses were recorded by a male and female member of the research team who were native speakers of Hiberno-English. The gender of the Tetris player was randomly assigned and balanced across participant gender. Responses were scripted to ensure that they were identical in content and structure. To enhance believability, recordings were made on built-in laptop microphones so audio quality is clear without being so high-fidelity as to lead participants to doubt that their partner is real.

3.2.3 Experimental conditions

The experiment followed a one-way within-subjects design. Interruption urgency was manipulated across two conditions: Urgent vs Non-Urgent. Following (Landesberger et al., 2020a, 2020b), urgency was manipulated by informing participants on urgent interrupting tasks (50% of the trials) that their partner's rating of their performance had a greater impact on their final score by a factor of 10 than the same ratings on non-urgent tasks. This operationalisation matched Landesberger in that urgency was explicitly told to participants and that urgent trials had a 10x greater impact on the score of the game for which participants sought to earn a cash prize, but unlike Landesberger, for which all trials had a time limit and urgent trials had a shorter time limit (Landesberger et al., 2020b), trials in this experiment were untimed. Trials in the present game experiment were not timed so as to avoid biasing participants to interrupt quickly, as the sorts of interruptions it seeks to inform for speech agent design are not limited to time-sensitive interruptions. Interrupting tasks within the trials were either labelled preceding the interruption prompt as urgent - 10x score or not urgent (see Figures 3.1 and 3.1). In this way, urgency was operationally defined as the interrupter's perceived value of interrupting well or cost of interrupting poorly. This operationalisation ensured that urgency was defined explicitly to participants rather than being inferred by message content or confused with interruption relevance.

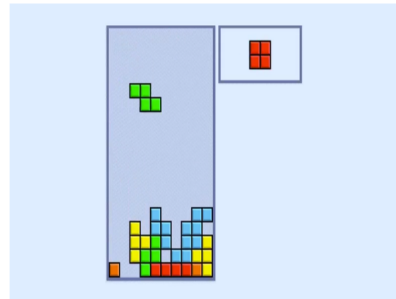
Press and hold the T key to activate your microphone.
After your partner responds, both of you can press the Enter key to move on to the next trial



Your Partner: Leigh
NOT URGENT
In your own words:
ask partner name

Figure 3.1: An example of a non-urgent trial that participants saw as a practice trial

Press and hold the T key to activate your microphone.
After your partner responds, both of you can press the Enter key to move on to the next trial



Your Partner: Leigh
URGENT - 10x SCORE
In your own words:
ask partner if played last month

Figure 3.2: An example of an urgent trial that participants saw as a practice trial

3.2.4 Measures

Interruption Onset

The time it took for someone to commence an interruption (in milliseconds) was measured as the time from the interruption prompt being displayed to the moment the participant began their interrupting utterance. Distinct sounds were labelled automatically in all participant audio, with sound being defined as periods of noise louder than -40db (40db quieter than digital maximum for the recording) and sounds were considered distinct from one another when intervening silence lasted longer than 100ms. These sounds were then manually checked to ensure measurement accuracy and to identify which sounds comprised the interruption utterance (i.e. the interruption message and any preceding access rituals) in order to correctly identify the start and the duration of the interruption.

Interruption Duration

These labelled sounds were also used to identify the total length of time of the interruption (in milliseconds), measured from the interruption onset to the completion of the interrupting utterance.

Access Ritual Frequency

The types of access rituals used by participants to interrupt the Tetris player were categorised based on previous approaches (Krivonos & Knapp, 1975). Audio of participants' verbal responses were used to determine whether each of the access ritual behaviours listed was present in the interruption. This included: Reference to other (i.e., Use of name or impersonal address); Apologizers (e.g., saying "sorry" or "excuse me"); Greeting (e.g., saying "hey", "hi"); Filled openings (e.g., hesitations, disfluencies, "um", "uh", "hmm", occurring at the beginning of an interruption) The presence of these were coded to produce a binary variable (1 = access ritual present; 0 = access ritual absent).

Demographic Questionnaire

Participants were asked a number of questions about themselves such as age, gender, and level of education, their level of experience with Tetris, and whether they believed their partner in the experiment to be another person playing live, a recording of a person, or a computer.

Open Ended Reflective Questions

To gather further context and gain an insight into the interruption strategies used, participants were asked four open-ended questions at the end of the experiment. Reflecting on the urgent and the non-urgent trials separately, participants were asked "how did you decide when to deliver messages to your partner?" and "how did you decide what to say to your partner?".

3.2.5 Procedure

Participants were recruited through Amazon Mechanical Turk. After following the link to the study participants were taken to a set of webpages where they were given information about the aims of the research, the data to be collected, and their data processing rights. Next, participants were shown a webpage where they could test that their microphone and headphones were working properly. Participants were then asked to give consent to take part in the study. Participants then were briefed on the procedure of the experimental task through a series of webpages. On the first

webpage, the general aim of the game - speaking to a Tetris player was described. Next, participants were given procedural details including a description of the Tetris player's goal *"to play Tetris with as few mistakes as possible"*, their own goal *"to ask questions to your partner without distracting them from their game"* and information on how many trials the experiment would last and that they could talk to their Tetris playing partner by pressing a key. The next page described the scoring system, by which their partner would rate their interruption as described above, and that trials marked urgent would have their ratings counted 10x more toward their final score, on which they would be judged for the €20 cash prize. Next, participants were told that they were being matched with a partner from an online Tetris website.

After a randomised delay lasting 10-20 seconds, participants were told they had been connected to their partner and were shown generic partner information, including a unisex first name, a country of residence (e.g., "Leigh", "Ireland") and some statistics indicating that their partner is a regular Tetris player (e.g. "11 hours played this month"). Next, the participants experienced two practice trial tasks, that included one non-urgent prompt and one urgent prompt. After completing each practice trial, the participant saw a screen for a random interval between 2500 and 3500ms informing them that their partner was rating their interruption. Next, participants were instructed that they would engage in 16 trials, after each of which their partner would rate their interruption. The experiment consisted of 16 Tetris trials and 16 interruption prompts. Each interruption prompt was presented only once to each participant. These were ordered randomly, with 8 prompts randomly assigned to each urgency condition across the 16 Tetris trials. The rating screen appeared for 2500 to 3500 ms after each trial. After all trials were completed, participants were asked to complete a brief questionnaire about their own background and their experience with the experiment, comprising the demographic questions and the open ended questions described in Section iv above. After completing the questionnaire, participants were fully debriefed explaining that their partner was actually a recorded member of the research team and that their performance was not being rated. They were informed that they were eligible to receive a bonus prize, but this prize would be awarded randomly through selection of an anonymous Amazon Mechanical Turk ID. Participants were finally thanked for taking part and given instructions on receiving their payment.

3.3 Results

3.3.1 Quantitative interruption behaviour

Data Cleaning and Analysis Approach

A total of 832 trials were recorded across the experiment. Trials in which technical issues rendered audio inaudible (N = 97 trials) or that were classed as extreme values within the measures (+ or - 3 standard deviations from the mean; N = 26 trials) were removed from the dataset. No data needed to be removed by participant request. This resulted in a total of 709 trials by 46 participants being included in the final dataset for analysis. Of these participants, 30 believed after the experiment that their partner had been a recorded human player, 5 believed it had been a live human player, 8 believed it had been a computer, and 3 were not sure.

Linear mixed effects models were used to analyse the effect of urgency on interruption onset and interruption duration. Logit mixed effects models were used to analyse the effect of urgency on use of access rituals. Mixed effects models are extensions of regression that allow data with hierarchical structures to be modelled in a way that accounts for both fixed effects of independent variables as well as participant-level and item-level effects through random intercepts and differences in magnitude of fixed effects through random slopes (Baayen et al., 2008; Barr et al., 2013). Models were fit using the lme4 package version 1.1-26 (Bates et al., 2015) in R version 4.0.3 (R Core Team, 2020). Following best practices, model selection began with the maximal random effect structure for the experiment (i.e. random slopes and intercepts at the subject- and item-level for the main effect of urgency, with item-level effects of Tetris video, interruption prompt, and trial number) with complexity incrementally reduced for a given model until models could converge (Barr et al., 2013). To improve reproducibility, full model syntax and random effect outputs are included for each model (Meteyard & Davies, 2020).

Interruption onset

There was a statistically significant effect of urgency [Unstandardised $\beta = 232.83$, SE $\beta = 112.58$, 95% CI [7.45, 458.30], $t = -2.07$, $p = .04$] with participants delaying significantly longer before non-urgent interruptions (M = 3419ms; SD = 1312ms) as

Table 3.2: Table of means and standard deviations for interruption onset and interruption duration by urgency condition.

Measure	Urgency condition	Mean (ms)	SD (ms)
Interruption Onset	High	3200	1227
	Low	3419	1311
	Overall	3293	1699
Interruption Duration	High	1400	288
	Low	1431	299
	Overall	1419	550

compared to urgent interruptions ($M = 3200\text{ms}$; $SD = 1276\text{ms}$). This supports H1 and is visualised in Figure 3.3. Descriptive statistics for interruption onsets overall and by condition are reported in Table 3.2. Full model syntax and output are included in Table 3.3.

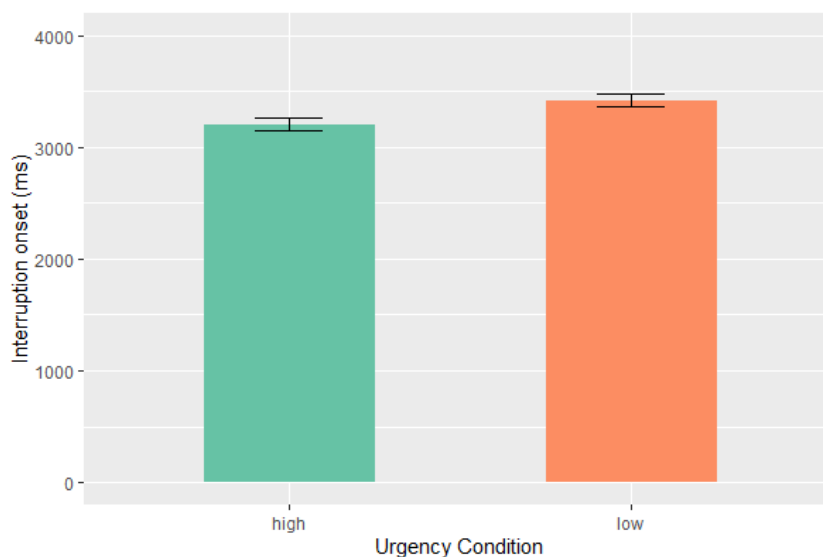


Figure 3.3: Means and standard errors for interruption onset times by urgency condition

Interruption duration

There was no statistically significant effect of urgency [Unstandardised $\beta = 32.25$, SE $\beta = 37.10$, 95% CI [-40.57, 105.07], $t = -0.87$, $p = .39$] on the duration of interruption. This means that H2 was not supported. Descriptive statistics for interruption durations overall and by condition are reported in Table 3.2. Full model syntax and output

Eliciting Speech Interruptions to Investigate the Impact of Urgency

Table 3.3: Summary of fixed and random effects for interruption onset - Linear mixed effects model

Model: *Interruption onset* =
 $urgency + (1 + urgency|subjectID) + (1|prompt) + (1|vidID) + (1|trial\ order)$

Fixed Effect	Unstandardised β	SE β	t statistic	p
Intercept	3198.88	199.88	16.00	.001***
Urgency (Low)	232.83	112.58	2.07	0.04*

Random Effects	
Group	SD
Participant (intercept)	1121.24
Participant*urgency (Low)	458.22
Prompt	298.79
Video	45.08
Trial number	215.12

Table 3.4: Summary of fixed and random effects for interruption duration - Linear mixed effects model

Model: *Interruption duration* = $urgency + (1|subjectID)$

Fixed Effect	Unstandardised β	SE β	t statistic	p
Intercept	1400.68	44.32	31.61	<.001***
Urgency (Low)	32.25	37.10	0.87	0.39

Random Effects	
Group	SD
Participant (intercept)	241.7

are included in Table 3.4.

Access ritual use

There was no statistically significant effect of urgency [Log-odds = -0.29, SE = 0.32, 95% CI [-0.90,0.33], $z = -0.91$, $p = .36$] on the likelihood of using access rituals in interrupting utterances. This means that H3 was not supported. Across the data, 23 out of 46 participants used no access rituals at all, with three participants using access rituals on more than half of their trials. Counts of access ritual usage overall and by condition are reported in Table 3.5. Full model syntax and output are included in Table 3.6.

Table 3.5: Table of counts of trials containing access rituals by urgency condition

	Trials containing an access ritual	Trials without an access ritual
High	57	295
Low	51	306
Overall	108	601

Table 3.6: Summary of fixed and random effects for access ritual presence
- Logit mixed effects model (Present = 1)

Model: $Access\ ritual\ presence = urgency + (1|subjectID)$

Fixed Effect	Unstandardised β	SE β	z statistic	p
Intercept	-3.39	0.65	-5.21	<.001***
Urgency (Low)	-0.20	0.29	-0.69	0.49

Random Effects	
Group	SD
Participant (intercept)	2.93

3.3.2 Qualitative descriptions of interruption strategies

Data Analysis Approach

Answers to open-ended questions were analysed through thematic analysis using a hybrid approach (Fereday & Muir-Cochrane, 2006). Initial codes were generated deductively, guided by prior work on interruptions and speech, with themes also developed inductively through a staged review of the data and initial codes, consistent with a reflexive approach to thematic analysis (Braun & Clarke, 2006). For the questions regarding timing, initial codes were generated to reflect literature on speed-accuracy tradeoffs for interruptions (Brumby et al., 2011), with timing strategies coded as focusing on either the speed of the interruption, accuracy in the interrupting task (i.e. avoidance of error in talking to one's partner), or the accuracy of the primary (Tetris) task. A third code represented responses that gave no indication of a conscious strategy. Note that time spent on the primary task is a direct function of the speed of the interrupting task, in that both tasks end when the interrupting task is completed, so speed of the primary task was not an initial code. For questions regarding what participants said to their partner, initial codes were generated to reflect literature on urgent speech (Hellier et al., 2002; Landesberger et al., 2020a, 2020b), with speaking

strategies coded as phrasing (semantic characteristics) or delivery style (prosodic characteristics). A third code represented responses that gave no indication of a strategy. Because of the hybrid approach used in this thematic analysis (Fereday & Muir-Cochrane, 2006), these inductive codes served as a starting point and do not encompass all of the final themes which were generated inductively through staged review. That is to say, while each utterance was coded using only these inductive codes in the first round of coding, successive rounds then allowed for the introduction of new codes and the combination of codes to deductively describe the data.

Interruption Timing Strategies

Four themes for interruption timing strategies were generated inductively. Participants felt they either timed their interruption in a way that always prioritised accuracy, in a way that always prioritised speed, mixed strategies according to characteristics of the interrupting task (i.e. interrupting message content), or mixed strategies according to characteristics of the Tetris task. Themes are presented below along with counts of how many participants in each condition mentioned a given strategy (out of a total of 52 participants).

Prioritising Speed (Non-urgent: 9 participants, Urgent: 30 participants)

Many participants stated that, when completing the trials, they interrupted as soon as they could. This strategy was mentioned more frequently when discussing strategies in the urgent trials, although it was mentioned when discussing non-urgent trials too. Some participants did not consider the state of the Tetris task when planning their interruption stating that “[I interrupted] as soon as possible, the timing of Tetris didn’t occur to me” (P09) while other explanations were more brief, stating they interrupted “as soon as I could”, “as soon as possible”, or “as soon as they appeared” (Ps 02, 09, 41). The difference in prevalence of the speed strategy between conditions supports the quantitative results highlighting faster interruption onset in the urgent trials.

Prioritising Accuracy (Non-urgent: 6, Urgent: 0)

Especially when discussing the non-urgent condition, participants mentioned the importance of accuracy, trying to prevent errors in interruption delivery, sacrificing speed. Some participants specifically mentioned sacrificing speed across the entirety of a condition, as opposed to timing interruptions based on features of the Tetris task

or of the interrupting task.

"[I] Took my time deciding on how to word and when to deliver the question" (P28)

"[I] just decided to say it casually. not make him feel like he needs to answer too quickly for the low urgency trials." (P44)

The mention of taking one's time in non-urgent trials but not in urgent trials was somewhat surprising, as past research has indicated that people generally prefer to interrupt as quickly as possible when not specifically instructed otherwise (Brumby et al., 2011; Horrey & Lesch, 2009). It may be that participants saw this strategy as more appropriate, but not well-suited to urgent interruptions, and thus were more likely to use this strategy in non-urgent trials. Again this supports the quantitative findings of participants taking longer to start an interruption in non-urgent trials than urgent trials.

Tetris Task Characteristics (Non-urgent: 33, Urgent: 18)

Fifty-one responses mentioned the importance of using characteristics of the Tetris task to decide when to interrupt. From the comments some participants describe themselves as being sensitive to subtask boundaries (Non-urgent: 6, Urgent: 3), to the player's cognitive load (Non-urgent: 25, Urgent: 14), or mention the Tetris task without specifying the characteristics of the task they were sensitive to (Non-urgent: 18, Urgent: 1).

Those who mentioned subtask boundaries as a cue for timing their interruptions seemed to plan interruptions for when a Tetris piece was in its final destination or at the top of the screen - when the subtask of placing a piece had just finished and the next subtask was just beginning (see Figure 3.4). They tend to emphasise that they would interrupt *"When there was a new block so that it was at the top of the screen"* (P10) or *"As soon as a block was placed and a new one was at the top of the screen"* (P12).

There were also those that attempted to identify moments in which their partner was under less cognitive load, unburdened by making a decision for the Tetris task. They focused on moments when *"placing a block was not too difficult"* (P12) or when *"the game was not intense."* (P25) as well as opportune moments when the participants perceived that the player had clearly finished making a decision *"I delivered when I felt she had selected a spot for the falling piece."* (P29)

Others were less specific about the characteristics of the game they prioritised

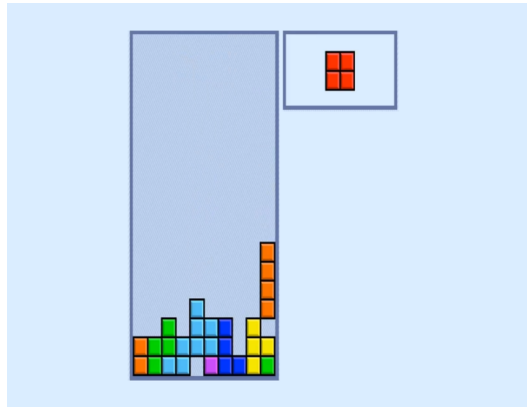


Figure 3.4: An example of a possible subtask boundary which some participants identified as a good moment to interrupt. Note that the orange Tetris block on the far right is the currently falling block.

but still indicated that they used the Tetris task state to assess when was the right time to ask a question: *“I watched the play and then asked the question”* (P01).

There is likely considerable overlap in Tetris task-dependent reasons that these participants picked their moments to interrupt. Natural breakpoints such as subtask boundaries are frequently the lowest cognitive load moment within a task and are thus ideal for interruptions (Bailey & Iqbal, 2008; Brumby et al., 2013). Choosing subtask boundaries as moments of interruption may well be seen as selecting the moments they find to be the least intense or the most convenient. Likewise, selecting moments between decisions construes the game of Tetris as made up of a series of decisions at subtasks. These descriptions of Tetris-task dependent strategies were therefore categorised together in the same theme.

Message Content (Non-urgent: 2, Urgent: 0)

One relatively rare strategy was to time interruptions depending on the content of that interruption. Two participants mentioned that the timing of their utterances depended on what question they were asking their partner. One of these participants explained their exact rationale, saying *“I tried to wait until a piece had been played if it was a longer question, if it was a simple and short question I asked it straight away”* (P51) indicating that the message content was a primary strategy selection criteria, selecting the Tetris task strategy for long questions and the speed strategy for short questions.

No Strategy (Non-urgent: 2, Urgent: 4)

Some participants either explicitly noted that they did not think about how to time their interruptions and as such identified no strategy, suggesting that they *“didn’t really change [their] communication one way or the other.”* (P21).

Interruption Structure

For the questions regarding what participants said to their partner, three clear themes were generated inductively. Participants either focused primarily on the way they phrased their message (i.e. word choice), they focused on how they delivered their message (i.e. prosodic features), or they mixed strategies according to the characteristics of the interrupting task (i.e. interrupting message content). These themes are explored below with comparisons of frequency in the non-urgent and urgent conditions.

Phrasing (Non-urgent: 36, Urgent: 33)

A major theme in how participants structured their interruptions was phrasing. Within this theme, three strategies were identified, delineating what characteristic of their phrasing participants prioritised: word length (Non-urgent: 18, Urgent: 21), naturalness (Non-urgent: 16 Urgent: 9), or other (Non-urgent: 2, Urgent: 3).

Many participants who focused on the phrasing of their interruptions did so by trying to interrupt with as few words as possible, sometimes explicitly acknowledging that this was to reduce cognitive load on their partner: *“I used as few words as possible, so she didn’t have to think about it”* (P15). Others who focused on word length took the opposite approach, seeking to avoid error by *“ask[ing] questions elaborately”* (P01), specifying that they *“Said it in detail so he would give me the correct answer.”* (P44). This phrasing strategy was less prevalent than the former, but both were distributed similarly across urgency conditions.

For some, phrasing was not primarily about length, but about asking questions *“that made sense”* (P42), that were phrased as *“the questions I would normally ask an acquaintance.”* (P23), and questions that *“reflect what needs to be asked.”* (P47). It isn’t clear whether participants perceived natural phrasing as consistent with shorter phrases, longer phrases, or neither, so these strategies were grouped together under the theme of phrasing. There were also participants who prioritised other ways of phrasing such as using *“the most informative way to ask the question.”* (P40). These diverse strategies around phrasing were classified as part of the same broader phras-

ing theme.

Delivery (Non-urgent: 5, Urgent: 11)

Another major theme in how participants structured their interruptions was delivery, focusing in particular on prosody - the way their speech sounded. This theme includes three strategies concerning delivery, each delineated by which characteristic of their their delivery participants mentioned: tone (Non-urgent: 1, Urgent: 1), clarity (Non-urgent: 4, Urgent: 4), or speed (Non-urgent: 0, Urgent: 6).

One participant focused on their tone of voice, seeking to deliver interruptions in *"a calm voice to not startle my partner"* (P24), using this strategy in both urgency conditions: *"Again, I said it calmly"* (P24).

Others who focused on delivery instead prioritised clarity, seeking to deliver interruptions *"clearly so she can understand."* (P47). These participants mention focusing less on choosing their words, instead ensuring that they *"spoke it clearly."* (P45).

A focus on clarity did not always pay off however, as one participant using this strategy expressed regret for not instead focusing on phrasing.

"I tried to make my questions as clear as possible, but in hindsight I think I probably should've made an effort to make my questions shorter as though I started when I thought it was a good time to talk, actually by the time I'd finished asking and it was time for her response it was in the middle of what I'd consider a high risk moment in the game!" (P16)

This expression of regret gives insight into the extent to which themes overlapped and the dynamic nature of strategy selection. Finally, some participants mentioned that they *"tried to speak quickly"* (P29). It should be noted that speaking quickly was considered a delivery strategy in this analysis, but it may overlap with the strategy of minimising phrase length for individual participants, as mentions of speaking speed were typically short vague expressions like *"I spoke quicker"* (P30). For this reason, the use of one strategy should not be viewed as mutually exclusive with the use of other strategies.

Message content (Non-urgent: 5, Urgent: 4)

Some participants mentioned varying their strategies for structuring interruption *"based on the type of question."* (P13). Participants who varied strategies did not give much indication of which features of the content of the message were relevant to them nor how they varied their strategy, vaguely alluding to how they *"relied more*

on the text that was at the bottom of the screen”(P03) in one urgency condition or the other. This theme may not lend much insight to how message content impacts strategy selection, but it nonetheless provides some evidence that message content may impact strategy selection for some people, and that strategies are not rigid functions of urgency or individual preferences.

No strategy (Non-urgent: 6, Urgent: 4)

Just as was the case with timing strategies, some participants either explicitly noted that they did not think about how to structure interruptions or gave short or vague responses like “[I] read the description and made a decision” (P08) that did not fit into any of the above themes, or explicitly stated “I didn’t really change my communication one way or the other.” (P21).

As was the case with timing strategies, a lack of stated strategy is not necessarily an indication of no strategy. The above quote from P21 indicates that some participants may have thought about this question comparatively, noting whether their interruption differed between conditions but not explaining their strategy if it was consistent. Again, no participant in this theme indicated that they randomly altered their interruption structure or that they avoided using a consistent strategy, so this theme is best understood as an absence of an explicit acknowledgement of a strategy rather than an absence of strategy per se.

3.4 Discussion

Through an online experiment, this study found that interruption urgency has a statistically significant effect on interruption onset, with urgent interruptions coming sooner than non-urgent interruptions. Conversely, it found no evidence that interruption urgency has a significant effect on interruption duration nor on the use of access rituals during interruption. Through qualitative analysis of participants self-reported interruption strategies, it found that, for timing interruptions, people choose to either always interrupt as quickly as possible, to always wait before interrupting to minimise errors in their speech or their partner’s play, to vary their timing according to the content of the interruption, or to vary their timing according to the characteristics of the Tetris game state. Likewise, for structuring interruptions, people chose to focus either on the way they phrased their interruption, the way they delivered their interruption,

or they varied their strategy based on the content of the interruption. Together, this work provides initial insights into the method of expression (see McFarlane, 1997) of human speech interruptions.

3.4.1 Explicit cues of urgency impact interruptions of others

This study primarily sought to understand whether the manipulation of an interrupter's goals for interrupting would impact the way they constructed their interruptions of others. The results find that by manipulating urgency, which was operationalised as an explicit variable which impacted the potential reward an interrupter would earn, different strategies for interrupting can be elicited, namely interruptions in high urgency conditions coming at lower average interruption onsets. This fits with prior research on self-interruptions in dual-task environments, in which participants put a higher emphasis on an interrupting task when explicitly directed to do so (Brumby et al., 2013). This study represents the first research to extend this finding to interruptions of other people and the first which focuses on interruptions delivered via speech. Extending that literature to the domain of speech interruptions of other people is critical for the future of research on human-human spoken interruptions, as evidence that human interrupters are sensitive to explicit cues of task priority may indicate that human spoken interruptions are suitable to study through controlled laboratory experiments. This research approach appears promising following this experiment as, in the present study, priority was explicitly manipulated as an independent variable. This enables a variety of future research into human-human spoken interruptions, which may lead to greater understanding of how variables like urgency, primary task type, and the identity and expertise of the interrupter influence a variety of measures of spoken interruptions. This also has major implications for the design of non-human proactive agents as urgency is a factor which must be considered when designing interruptions. Insofar as human spoken interruptions are well-suited for laboratory experiments, they may represent a fruitful and low-cost way to approximate appropriate design of interruptions spoken by non-human agents.

3.4.2 Interruption strategies are highly diverse

Through thematic analysis of participants' descriptions of their strategies, this work provides some key insights into how spoken interruptions are timed and structured.

While some people use characteristics of their partner's primary task (Tetris) to determine when to interrupt, others use characteristics of the interrupting message or interrupt according to fixed strategies irrespective of the tasks. This is consistent with other work on multitasking that found a similar complex mix of strategies for self-interruptions (Cha et al., 2020; Dabbish et al., 2011). As modelling complex situations like driving or daily life is still an ongoing challenge (Cha et al., 2020; Semmens et al., 2019), the insight provided here about the diversity of strategies people use to time interruptions should help to guide speech agent design as task modelling capabilities improve. Future work should investigate whether moments that interrupters identify as natural breakpoints (e.g., when a Tetris piece is at the bottom of the screen) correspond with when they interrupt people. Chapter 5 begins to address this question by investigating the breakpoints in the present Tetris task. Further work around breakpoints and interruptions of others would help unite existing understandings of natural breakpoints (Borst et al., 2010; Janssen et al., 2012) with the ongoing work on communication during multitasking. Furthermore, future work may consider whether an interrupter's expertise in a primary task influences perception of breakpoints and thus impact interruption strategies. The next chapter investigates task difficulty as an additional variable, to further investigate such variability in interruption strategies resulting from variation between complex tasks. This may be particularly important for increasingly complex tasks like driving or workplace environments in which task understanding requires greater expertise than does Tetris.

Themes regarding the structure of interruptions unite present knowledge of urgent speech (Hellier et al., 2002; Landesberger et al., 2020a, 2020b) with research on the effect of explicit goals in multitasking (Brumby et al., 2011), indicating that people alter both their word choice and their prosody depending on the urgency of an interruption. Speech agent designers could implement this feature of human speech production into synthesised speech, allowing users to hear particular notifications in an urgent voice while using a non-urgent voice for other notifications. Recent work has begun to explore this approach, finding that the use of more assertive voices significantly impact the speed of task switching from a complex primary task (Wong et al., 2019). From our findings, it is important to consider that the speech properties people used to communicate urgency varied. Future work should investigate if preferences of expressions of urgency used by speech agents likewise vary between

individuals, which is investigated in part in Chapter 6. As such, the present chapter offers a first attempt at understanding the strategies people use to interrupt another person engaged in a complex task through a framework derived from literature on task-switching.

3.4.3 Few people use access rituals

This work sought to investigate the use of access rituals - short verbal behaviours that signal a request for a listener's attention - in spoken interruptions. Not much is known about how people initiate spoken interruptions, so it was unknown whether people used access rituals at all when interrupting. This study found that urgency did not influence access ritual use. Most participants did not use them across any of the trials, yet some frequently did. The reason for this is unclear. People may have felt they already had social access to their partner due to both taking part in an experiment, and thus did not need to request it. It may also be that the relative importance of interrupting was so high as to diminish the social need for access rituals, or that there is a natural variability in the use of access rituals across the population observed here compared to that in the original research (i.e. American college students who were previously acquainted and interacting face to face) (Krivonos & Knapp, 1975). Nonetheless, that some participants did use access rituals frequently may be of interest to speech agent designers. Future work should investigate whether the use of access rituals by non-human agents is preferred by some users or if, like other human-like personalisations to agents, this is seen as unnatural, fake or unpleasant (L. Clark, Munteanu, et al., 2019; P. R. Doyle et al., 2019). The appropriateness of access rituals across interruptions delivered by non-human speech agents is investigated in a future study (Chapter 6). This work represents the first to investigate the strategic use and disuse of access rituals during interruptions of complex tasks.

3.4.4 A paradigm for eliciting spoken interruptions

To date, little research has explicitly been performed on the structure of speech interruptions. As such, the process of eliciting this kind of speech has not been well-established in the literature. This study therefore contributes a potentially fruitful paradigm by which spoken interruptions can be elicited from participants online in an approach that can be flexible to different independent variables, primary tasks,

and participant populations. The paradigm used here builds on the well-established dual-task paradigm used in multitasking research (Brumby et al., 2013; Edwards et al., 2019; Horrey & Lesch, 2009; Janssen & Brumby, 2010; Mark et al., 2012) by which participants engage in a primary task and are interrupted or interrupt themselves with a secondary task. This paradigm instead casts the participant in the interrupting role, with their alleged partner engaging in dual tasks, allowing the interruption itself to be centred as the object of study. In this way, the present paradigm serves as a tool both for deepening our understanding of human spoken interruptions and for expanding our palette of inspiration for how to design interruptions initiated by machines. This paradigm is used for eliciting further speech interruptions in Chapter 4 and may be flexibly adapted for further research eliciting interrupting speech. Future work may seek to develop a corpus of speech interruptions for continued research on the topic, and the present study presents a methodology through which such data can be gathered.

3.4.5 Limitations

While this work focuses on initiating conversation with people actively engaged in another task, not all agent-initiated interruptions will need a response. Indeed, many interruptions that occur during complex, continuous tasks include information delivery rather than requests of information from the user (e.g. navigation information while driving). Indeed proactive agent interactions can be viewed as a continuum, ranging from a person performing a task without an agent's prompting, to an agent prompting an interaction, to an agent suggesting that it perform a task on the user's behalf, to the agent performing a task without consulting the user (Isbell & Pierce, 2005). Insights from this work may therefore be applicable the design of interruptions that require a spoken response from users, interactions around the middle of this continuum, but they may not be as relevant to other proactive agent interactions.

Likewise, this work looks at the interruption of a low-risk task, and interruption strategies may be more divergent or entirely different for contexts in which errors are more costly. Comparable work exploring how people talk on the phone to a driver when they are aware of the driver's task context similarly found patterns of adaptation based on the state of the ongoing task (Janssen et al., 2014), but the extent to which simulated driving is seen as higher stakes to participants than Tetris is not clear. While

Eliciting Speech Interruptions to Investigate the Impact of Urgency

these results illustrate a complex assortment of interrupting strategies, they emerged from a constrained continuous task and simple interrupting utterances. This work serves as an early step in understanding how agents might coordinate interruptions that vary across dimensions beyond just urgency and in contexts more difficult to model than Tetris. Designing for real world interactions of this sort will require much further work.

Urgency in this study was operationally defined as a reflection of how severely their partner's judgement of their disruptiveness would contribute to their own chance at a reward. While this operationalisation largely matched that used in past research on speech urgency (Landesberger et al., 2020a, 2020b), it critically differed in that urgency in this study was not explicitly associated with increased time pressure, whereas it was in those prior studies. Still, this work found a significant effect of urgency on interruption onset, indicating that participants self-imposed time pressure to urgent trials. Participants may nonetheless have interpreted urgency as indicative that interruptions are time sensitive or that errors during interruption were more costly. In this way, the subtle ambiguity about the meaning of urgency may limit generalisability across other contexts of urgency.

As this work was a first to use this paradigm and exploited hypotheses for which little prior work exists, it was impossible to estimate an appropriate sample size through a priori power estimations. Fortunately, this work allows for the ascertainment of an effect size and thus for power analyses for future studies using the same paradigm (e.g Chapter 4). Finally, participants in this study were crowdworkers interacting with recordings of people rather than dyads of people interacting online or while physically copresent. This may have limited the usage of access rituals, as many greeting access rituals are expressed non-verbally such as through facial expressions, waves, or physical touch (Krivonos & Knapp, 1975), none of which would be available to participants in this study. More work is needed to investigate how social dynamics such as personal relationships between people or physical copresence affect the ways people interrupt others who are engaged in another task.

Finally, while most participants believed that their partner was a human, most reported that they correctly believed that their human partner was pre-recorded. It is unclear whether this view changed over the course of the experiment or whether participants were suspicious throughout the experiment, but modelling trial order as

a random effect helps to capture any effect this may have had on interruption timing or access ritual use. Likewise, self-reported strategies of interruptions reflected that participants took the experiment seriously and sought to interrupt as if their partner was human. For this reason, limitations related to participants not believing they were interrupting another human, while important to acknowledge, are not seen as particularly concerning for the findings of this study overall.

3.5 Conclusion

This chapter sought to explore the method of expression of human speech interruptions by investigating the timing of and use of access rituals in spoken interruptions of the game Tetris, a complex, ongoing task. It likewise sought to identify strategies for timing and for structuring interruptions as self-identified by interrupters. Urgency was identified as a key variable impacting both how people interrupt themselves and how people time and structure speech, so it was manipulated as an independent variable. In order to elicit speech interruptions for this analysis, an online paradigm was created using prerecorded videos of Tetris. Participants interrupted a supposed Tetris player through a gamified paradigm, whereby they believed the Tetris player would rate their interruptions, with urgent interruptions impacting their score to a greater degree. Interruptions were analysed quantitatively and participants' self-reported interruption strategies were analysed qualitatively. The study found that urgency had a statistically significant impact on interruption onset (with urgent interruptions coming at less delay than non-urgent interruptions) but not on interruption duration nor on the use of access rituals. Qualitative analysis revealed four themes for interruption timing strategies: *prioritising speed*, *prioritising accuracy*, *Tetris task characteristics*, and *message content* as well as three themes for interruption structure strategies: *phrasing*, *delivery*, and *message content*.

These findings give a concrete initial empirical insight which can be applied to the constructivist aim of designing of human-inspired proactive agent speech - that urgent interruptions should come at less delay. This finding is therefore applied in the design of a proactive agent in Chapter 6. Likewise, through self-reported strategies, new insights into the importance of the difficulty of the Tetris game in influencing how people time their interruptions draw attention to ongoing task difficulty as a variable

Eliciting Speech Interruptions to Investigate the Impact of Urgency

for further study in Chapter 4. Insofar as this study is the first to seek to elicit speech interruptions of an ongoing complex task, it provides methodological benefits to future research in the form of a paradigm and an observed statistical effect of urgency which can be used for power analysis. As such, both the paradigm and the effect size of urgency are used in the design of the study in Chapter 4. The next chapter therefore seeks to replicate the design of the study from this chapter while adding Tetris task difficulty as a second independent variable and increasing the sample size to improve statistical power. Finally, self-reported interruption strategies in this chapter indicate that some people seek Tetris subtask boundaries or low-load moments for a Tetris player, mirroring literature on natural breakpoints. As such, Chapter 5 reanalyses data from this chapter and from Chapter 4 to quantify this phenomenon.

4 * **Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions**

4.1 Introduction

Chapter 3 established the effect of urgency on spoken interruptions of Tetris. In order to build upon that finding, the present chapter seeks to further explore which other factors impact how spoken interruptions are strategised and timed. In Chapter 3, urgency was framed as a factor which is explicitly known to interrupters but only accessible to an interrupted party explicitly via the interruption itself. Participants from that study mentioned in their self-reported interruption strategies that they did not only utilise the cue of urgency however, but they also took into account characteristics which must be estimated, such as the internal state of the interrupted party or goals they privately hold. Insofar as the present thesis aims to both understand human interruptions of complex tasks and to inform the design of machines which interrupt people engaged in complex tasks, it is necessary to consider these factors which are not explicitly known to interrupters as well as explicit factors like urgency. Participants in Chapter 3 identified the state of the Tetris game as a factor which impacted their interruption strategy in self-reported comments. For this reason, the present study manipulates Tetris game difficulty, a variable which must be estimated by the interrupter through observation, in order to understand how such a cue compares and interacts with urgency, a cue explicitly known to the interrupter.

4.1.1 Task difficulty

Task difficulty is a factor which can be directly and systematically manipulated in a video game like Tetris by setting the game at a higher or lower difficulty level. This feature of Tetris as a designed game task rather than a less controlled real-world task like on-road driving makes Tetris particularly suitable for this research, following from its utility as a complex task in Chapter 3. Manipulating task difficulty has the effect of introducing variance to the Tetris player's risk of failing at the game. Chapter 3 discusses the extent to which participants were constrained in the strategies they selected for interrupting Tetris in that experiment, where task difficulty was uniformly low, as participants did not judge the Tetris player to be at risk of failure. By observing how strategies differ between interruptions of difficult and easy Tetris games, the present study aims to overcome this limitation, seeking to introduce greater variability between the cues that might impact interruption strategies in order to measure the corresponding variability in interruption strategies. Specifically, while the timing and use of access rituals during low-difficulty Tetris gameplay were investigated quantitatively in Chapter 3, this study goes beyond that analysis by also investigating the effect of urgency on those variables during high-difficulty Tetris gameplay.

Task difficulty when designing a game as a primary task for a study comparing interruption methods has been noted as important to consider (McFarlane, 2002). Evidence for this comes from a study where participants were tasked with playing a computer game which required visual monitoring and input from one hand on a keyboard. Intermittently during the task, participants would be interrupted by an interface which employed one of a variety of interruption strategies. McFarlane notes that game difficulty needed to be calibrated through pilot testing to ensure that it was "complex enough to attack participants' vulnerability to interruption, but simple enough not to cause participants to despair of performing well." (McFarlane, 2002, p. 77). The study described in Chapter 3 used a uniform level of task difficulty which did not cause participants to doubt the Tetris player's ability to perform well. It was unclear however whether the difficulty was high enough that participants felt that the Tetris player was made vulnerable by a spoken interruption. The present study seeks to contrast the level of difficulty introduced by the previous study with Tetris tasks of a higher level of difficulty. In this way, the present study aims to overcome a potential threat to generalisability incurred by studying only interruptions of tasks for which in-

ruptions would not likely threaten primary task success, reducing the incentive for selecting an optimal interruption strategy.

Modelling task difficulty for less controlled tasks is an ongoing challenge for designers of proactive interruptions. For instance, recent work has aimed to model the interruptibility of drivers by collecting sensor and video data (Kim et al., 2015; Semmens et al., 2019). In one such study, 15 on-road drivers participated in a study in which physiological sensor, vehicle state sensor, road video, and driver video were used to create a model of the interruptibility of drivers, predicting drivers' likelihood of engaging in periphery tasks such as taking their hands off of the steering wheel or adjusting the radio (Kim et al., 2015). This data-intensive approach was then replicated in a later study seeking to confirm whether drivers considered themselves interruptible in such moments. In this later study, a similar array of sensor and video data was collected as participants drove along a predetermined route for 50 minutes (Semmens et al., 2019). Throughout the drive, participants were asked "is now a good time" to receive information from an in-car speech assistant. Participants were asked to respond either yes or no to indicate their interruptibility at the moment of each request. Data from the vehicle such as acceleration and wheel position as well as audio and video data including location and the driver's utterances were then used to fit models of interruptibility. While modelling interruptibility using only in-vehicle sensors was insufficient, these models were somewhat improved by the inclusion of location data as a comparison against the intended route (Semmens et al., 2019). In this way, Semmens and colleagues found that modelling task difficulty using only implicit task states is insufficient for planning interruptions, but that a comparison between the state of the task and the goal state was more feasible.

For unconstrained tasks like driving from point A to point B, intermediate goals such as turning on a particular road are selected and executed by the actor. Conversely when a task is scripted or controlled, like a driving a predetermined route or a playing a game like Tetris which has a particular designed set of goals, assessing the level of difficulty an actor is facing is more manageable for an outside observer, as they may compare the known, intended state of the task to the current observable state. As highlighted by research on less-controlled tasks (Semmens et al., 2019), the use of an uncontrolled task can lead to difficulties in perceiving the task difficulty and thus may lead to inconsistent manipulation of difficulty as an experimental vari-

able. In order to ensure that the perception of task difficulty is able to be made more clearly by participants, this study again focuses on Tetris as the primary task of the interrupter. Using Tetris as a target of interruption ensures that task difficulty can be estimated even by an interrupter who is not themselves engaged with that task.

4.1.2 Explicit and estimated cues

The study carried out in Chapter 3 revealed that participants tasked with interrupting Tetris players responded to the explicit cue of urgency by altering their interruption strategies during urgent trials, prioritising their own goal of interrupting quickly at the expense of the Tetris player's goal to avoid Tetris failure. The present study aims to compare that variable with one that instead must be estimated by the interrupter through observation. Insofar as participants in Chapter 3 mentioned altering interruption strategies according to their perception of state of the Tetris game or the cognitive load they believed the Tetris player was experiencing, this study aims to indirectly manipulate these perceptions by directly manipulating the difficulty of the Tetris game.

Research on conversational multitasking has mostly focused on driving, previously seeking to understand the extent to which the state of a driving task influences the conversational decisions made by a non-driving interlocutor. Initial studies on this topic have identified a tendency for co-present passengers to speak less and use more simple language, a phenomenon termed conversation suppression, during particularly difficult stretches of driving as compared to remote conversants talking through a mobile phone (Charlton, 2009; Crundall et al., 2005). Further research on this topic has sought to isolate co-presence from shared visual information, through both removing vision of the road from co-present passengers in a driving simulator and through providing driving simulator vision to remote interlocutors. These studies found that co-presence did not induce conversation suppression on its own, but shared visual information did, providing further support to the notion that these conversational partners were using driving task information to consider the driver's goals (Maciej et al., 2011). Another study in this vein found that co-present passengers also tended to select different topics for conversation than remote interlocutors, integrating topics of driving and the state of the driving task into otherwise irrelevant conversation, further aiding the driver's own situational awareness (Drews et al., 2008).

These differences in both conversational topic and conversation structure induced by interlocutors' awareness of the state of the driving task and sensitivity to its difficulty increased driving task performance across a number of simulator studies (Charlton, 2009; Schneider & Kiesler, 2005). These findings together suggest that shared visual context for interlocutors made driving while conversing more safe.

In the above studies, conversants had no particular goal other than to continue conversation. In this way, their consideration of the state and goals of their partner engaged in a driving task, while beneficial to their partner's goals was not to the detriment of any egocentric goal (i.e. suppressing conversation to benefit a partner did not have an associated personal cost to those participants). A number of similar studies have investigated whether remote and co-present interlocutors differ in the way that they speak with a partner engaged in a driving task when the conversation has an explicit goal for the non-driving partner. Studies introducing verbal tasks such as word games or trivia questions as a secondary task found the opposite of the studies involving only freeform chitchat. That is, co-present passengers were no more likely to suppress conversation than passengers without a shared visual context (Amado & Ulupinar, 2005; Gugerty et al., 2004; Nunes & Recarte, 2002). This may indicate that, in the presence of other tasks, making use of cues which must be estimated through observation such as the state of the a partner's ongoing task is less easily carried out.

Further research on this topic considered that these findings in which a passenger does not suppress conversation for the benefit of the driver may stem from difficulty in utilising visual information about the driving task in particular. In another driving simulator experiment, driving task information was shared with an interlocutor not via visual context, but instead through auditory information, with sounds of horns and traffic during the conversation representing difficult driving conditions (Janssen et al., 2014) . This research, which likewise had non-driving participants engaged in task-oriented conversation, found mixed results. While participants engaged in conversation suppression to benefit the driver when all auditory cues were relevant, the inclusion of irrelevant, non-driving related noises to their audio stimuli negated this effect (Janssen et al., 2014). This gives credence to the notion that estimated cues are difficult to utilise as cognitive load increases. A lack of conversational suppression during task-oriented conversation is not due to ignorance of the driver's goals in favour of the speaker's goals, but instead due to difficulty in using cues from an-

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

other person's task when engaged in a task of one's own. Research of this nature has thus far focused on driving as a primary task which is dual-tasked with speech. The present study seeks to investigate the impact of shared contextual information on how speech is timed and strategised when it can be more reliably estimated than it has been demonstrated to be during driving. As such, the present study uses Tetris as an ongoing task, a task which is more constrained and controlled than driving - features which may enable interlocutors to more easily utilise visual cues in order to consider the goals of the player.

4.1.3 Aims and hypotheses

Following literature on conversation suppression during speech multitasking (Charlton, 2009; Crundall et al., 2005; Janssen et al., 2014). The present study hypothesises that interruptions of easy Tetris games will come at a shorter delay (H1) and at a faster duration (H2) than interruptions of difficult Tetris games. In keeping with findings from Chapter 3, this study also predicts that urgent interruptions will come at a smaller delay (H3) and with a faster duration (H4) than non-urgent interruptions. While Chapter 3 showed that the use of access rituals vary across individuals, the present study nonetheless hypothesises that access rituals will be used less frequently during interruptions of easy Tetris games than interruptions of hard Tetris games (H5) as participants may use them as part of conversation suppression. Finally, as there is mixed evidence on the extent to which people use cues from a partner's visuomotor task to modulate their conversation with that person, this study poses the following exploratory research questions: how do the sizes of the effects of interruption urgency and Tetris game difficulty compare for spoken interruptions of Tetris? (RQ1) and is there an interaction between the effects of interruption urgency and Tetris game difficulty in spoken interruptions of Tetris (RQ2). Insofar as the variables of urgency and task difficulty have not been measured together in previous experiments, no hypothesis could be formed about their comparative effect sizes or the extent to which they interact when both are varied. Through analysis of qualitative data, this study also aims to more deeply explore the various approaches our participants used speech to interrupt people engaged in another task, assessing whether participants engage in different strategies for interrupting when the difficulty of the Tetris game varies as compared to those observed in Chapter 3 in which game difficulty was uniformly easy

(RQ3).

4.2 Methods

4.2.1 Participants

93 crowdworkers (48 men, 45 women; $M_{\text{age}} = 33.85$, $SD = 10.19$) were recruited from Amazon Mechanical Turk ($n = 56$) and Prolific Academic ($n = 37$). This number of participants was chosen based on post-hoc power analysis of the effects from Chapter 3, explained further below. Because this study used the same experimental setup and premise as Chapter 3, participants from Chapter 3 were not eligible to participate in this study. All participants were native or near-native English speakers. Participants were all familiar with the game Tetris, with most indicating that either they had played before, but do not play regularly ($N = 70$; 75.3% of sample) or that they play regularly ($N = 11$, 11.8% of sample) (5 point Likert scale; 1 = I am not at all familiar with Tetris; 5 = I regularly play Tetris). The study took approximately 20 minutes and participants were compensated for their time with \$10 of credit on Mechanical Turk or £8 of Prolific Academic credit for participating in the research. The study received ethical approval through the university's ethics procedures for low risk projects (Ethics code: HS-E-20-161).

To calculate the number of participants needed for this study, a power analysis was conducted with the aim of maintaining statistical power of .80 for detecting effects of the variables of interest, keeping with statistical convention (Cohen, 2013; Green & MacLeod, 2016). To estimate the size of the effect, the value for the fixed effect of urgency from Chapter 3 (0.14) was used. A power analysis was conducted using the *simr* package version 1.0.4 (Green & MacLeod, 2016) to estimate the number of participants needed to observe an effect size of 0.14 with 80% power across 1000 simulations. This analysis indicated that a minimum of approximately 85 participants would be required as a minimum for capturing an effect as large of that of urgency on interruption onset observed in Chapter 3. A visualisation of this power analysis is included below in Figure 4.1.

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

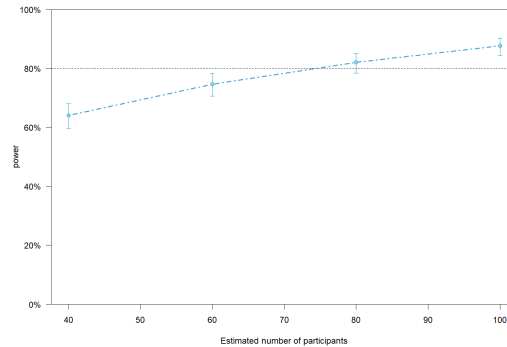


Figure 4.1: Simulated power for sample sizes from 40 to 100. At 80 participants, the lower confidence interval is marginally below the required threshold of 80%.

4.2.2 Materials

The Tetris paradigm was built using JSPsych version 6.3 (de Leeuw, 2015). Further details of the paradigm design are outlined in 3.2.2, and all code and stimuli are available at ¹. All materials including participant information sheets, consent forms, questionnaires, and debrief sheets are provided in Appendix B.

Tetris and Interruption Tasks

The Tetris and Interruption tasks in this study were identical to those described in Chapter 3 with only the addition of an independent variable of Tetris game difficulty differing between studies. Differences between videos in the two difficulty conditions are described below.

4.2.3 Experimental conditions

The experiment followed a two-way within-subjects design with independent variables of interruption urgency and Tetris game difficulty.

Interruption Urgency

Interruption urgency was manipulated across two conditions: Urgent vs Non-Urgent, in the same operationalisation described in Chapter 3.

¹https://osf.io/vkz9d/?view_only=8e7384146b3c40ed992f6616784495ac

Game Difficulty

Tetris game difficulty was manipulated across two conditions: Easy vs Hard. Easy Tetris trials were generated from 3 minute videos of actual Tetris gameplay conducted by the lead author. Each trial was chosen to ensure that the game state reflected one in which the Tetris player was not at risk of losing when the interruption occurred. Specifically: 1) a Tetris game piece started at the top of the game board; 2) there were at least two rows and no more than half of the rows of the board which already contained Tetris pieces and 3) the falling speed of the game piece was set to the game minimum of 1.25 rows per second. There were eight easy experimental trials in total. The eight easy experimental trials and two practice trials were the same videos used in Experiment 1.

Hard Tetris trials were generated from 3 minute videos of actual Tetris gameplay conducted by an experienced Tetris player. Each trial was chosen to ensure that the game state reflected one in which the Tetris player was at risk of losing when the interruption occurred. Specifically: 1) a Tetris game piece started at the top of the game board; 2) at least half of the rows of the board already contained Tetris pieces and 3) the falling speed of the game piece was set to 10 rows per second. There were eight hard Tetris experimental trials in total.

Tetris game difficulty was manipulated across two conditions: Easy vs Hard. To ensure that participants could perceive differences between easy and hard Tetris games, participants were asked after each trial to answer on 7-point Likert-type scales each of three questions: "How complex was the Tetris game you just saw?", "How easy was it to choose a moment to speak during in the Tetris game you just saw?", and "How confident are you that you picked a good moment to speak?". On all three questions, paired-samples t-tests revealed statistically significant differences between easy and hard Tetris games. Participants rated hard games ($M = 4.62$, $SD = 0.96$) as more complex than easy games ($M = 3.38$, $SD = 0.96$) [$t(89) = 12.91$, $p < .001$]. Participants rated hard games ($M = 4.50$, $SD = 1.03$) as less easy to choose a moment to speak than easy games ($M = 5.26$, $SD = 1.02$) [$t(89) = -9.19$, $p < .001$]. Participants rated themselves as less confident that they chose a good moment to interrupt during hard games ($M = 4.69$, $SD = 1.23$) as compared to easy games ($M = 5.27$, $SD = 1.25$) [$t(89) = -7.84$, $p < .001$]. These differences provide evidence that participants readily perceived the differences between easy and hard Tetris games.

4.2.4 Measures

Interruption Onset

The time it took for someone to commence an interruption (in milliseconds) was measured in the same way as is described in 3.2.4.

Interruption Duration

These labelled sounds were used to identify the total length of time of the interruption (in milliseconds), measured from the interruption onset to the completion of the interrupting utterance, as in Chapter 3.

Access Ritual Frequency

The types or access rituals used by participants to interrupt the Tetris player were categorised as in 3.2.4.

Post-trial ratings

After each trial, participants were asked to answer on 7-point Likert-type scales each of three questions: "How complex was the Tetris game you just saw?", "How easy was it to choose a moment to speak during in the Tetris game you just saw?", and "How confident are you that you picked a good moment to speak?". These measures were used to validate that participants perceived differences between levels of Tetris difficulty (see Section 2.3 above).

Demographic and Open Ended Reflective Questionnaire

Demographic and open-ended reflective questions were the same as those described in 3.2.4

4.2.5 Procedure

The procedure of the present experiment was largely the same as that in 3.2.5. Videos were randomly ordered, with each participant seeing eight easy videos and eight hard videos, counterbalanced against urgency conditions so that each participant experienced four trials in each of the four combinations of conditions. Each participant

experienced all 16 prompts and all of the eight videos for each level of Tetris difficulty. After each trial, a rating screen appeared, asking participants the three post-trial rating questions on a single screen with responses as radio buttons. After submitting post-trial ratings, participants either saw a screen that said “Please wait while the Tetris game is being initialized” if their rating took more than 5000ms or a screen that said “Please wait while your partner rates your communication” if their rating took less than 5000ms. The wait screen appeared for 2600 to 3400 ms if the participant’s rating took less than 5000ms or for 2000ms if the participant’s rating took more than 5000ms. After all trials were completed, participants were asked to complete demographic and open ended questionnaires then fully debriefed in the same way as described in 3.2.5.

4.3 Results

4.3.1 Quantitative interruption behaviour

Data Cleaning and Analysis Approach

A total of 1488 trials were recorded across the experiment. Trials in which technical issues rendered audio inaudible (N = 87 trials) or that were classed as extreme values within the measures (+ or - 3 standard deviations from the mean; N = 1 trial) were removed from the dataset. No data needed to be removed by participant request. This resulted in a total of 1400 trials by 90 participants being included in the final dataset for analysis. Of these participants, 60 believed after the experiment that their partner had been a recorded human player, 21 believed it had been a live human player, 5 believed it had been a computer, and 4 were not sure.

Linear mixed effects models were used to analyse the effect of urgency on interruption onset and interruption duration. Logit mixed effects models were used to analyse the effect of urgency on use of access rituals. Models were fit using the lme4 package version 1.1-26 (Bates et al., 2015) in R version 4.0.3 (R Core Team, 2020). Following best practices, model selection began with the maximal fixed and random effect structures for the experiment (i.e. fixed effects of urgency, difficulty, and their interaction and random slopes and intercepts at the subject- and item-level for the each fixed effect, with item-level effects of Tetris video, interruption prompt, and trial

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

Table 4.1: Summary of fixed and random effects for interruption onset - Linear mixed effects model

Model: $Interruption\ duration = urgency * difficulty + (1|subjectID) + (1|prompt)$

Fixed Effect	Std β	Unstd β	SE β	t	p
Intercept	-.18	3708.90	112.44	14.05	<.001***
Urgency (Low)	.27	851.80	1300.18	4.42	<.001***
Difficulty (hard)	.08	259.45	192.18	1.35	.177
Urgency * Difficulty	.00	14.93	272.38	.06	.956

Random Effects	
Group	SD
Participant (intercept)	241.7
Interruption Prompt (intercept)	485

number) with complexity incrementally reduced for a given model until models could converge (Barr et al., 2013). To improve reproducibility, full model syntax and random effect outputs are included for each model (Meteyard & Davies, 2020).

Interruption Onset

As in Chapter 3, there was a statistically significant effect of urgency [Unstandardised $\beta = 851.80$, SE $\beta = 192.81$, 95% CI [473.61, 1229.97], $t = 4.42$, $p < .001$] with participants delaying significantly longer before non-urgent interruptions (M = 4716ms; SD = 3761ms) as compared to urgent interruptions (M = 3865ms; SD = 2418ms). This supports H3 and is visualised in Figures 4.2 and 4.3. Full model syntax and output are included in Table 4.1.

There was no statistically significant effect of Tetris game difficulty [Unstandardised $\beta = 259.45$, SE $\beta = 192.18$, 95% CI [-117.5, 636.50], $t = 1.35$, $p = .18$] nor of the interaction between urgency and game difficulty [Unstandardised $\beta = 14.93$, SE $\beta = 272.38$, 95% CI [-519.32, 549.19], $t = 0.06$, $p = .96$] on interruption onset. H1 was therefore not supported. Descriptive statistics for interruption onsets overall and by condition are reported in Table 4.2.

Interruption Duration

There was a statistically significant effect of urgency [Unstandardised $\beta = 96.03$, SE $\beta = 34.45$, 95% CI [28.45, 163.30], $t = 2.79$, $p < .01$] with participants speaking for significantly longer during non-urgent interruptions (M = 1596ms; SD = 331ms) as

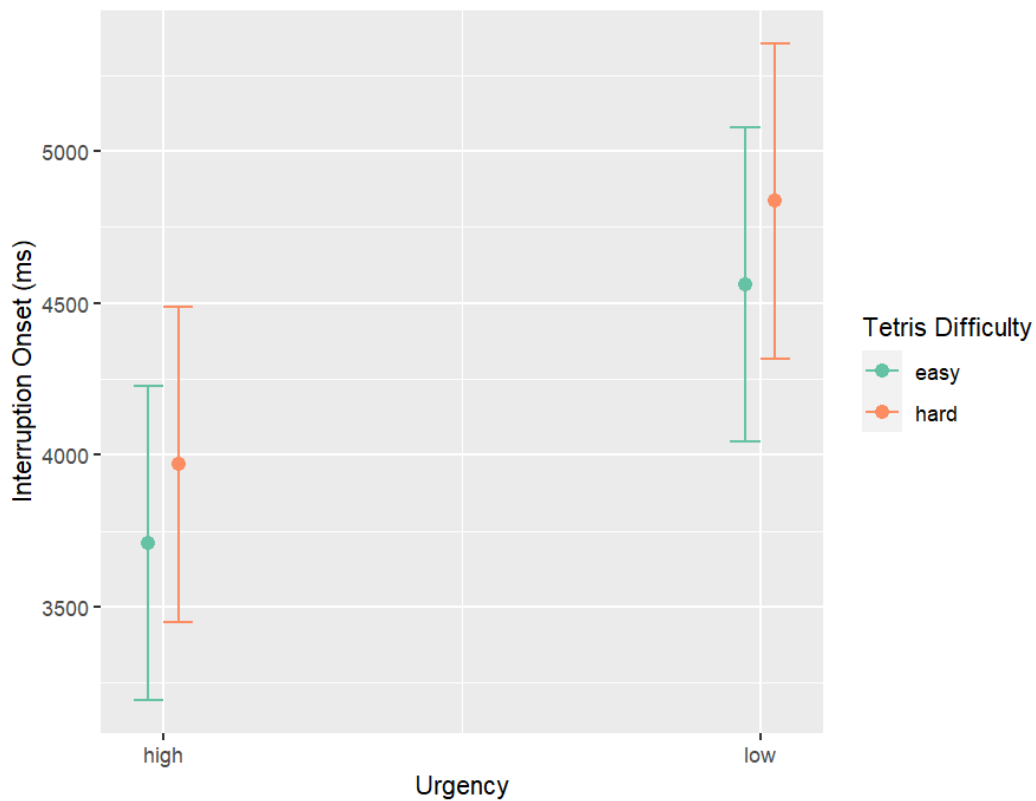


Figure 4.2: Means and standard errors for interruption onset by urgency and difficulty condition

compared to urgent interruptions ($M = 1518\text{ms}$; $SD = 562\text{ms}$). This supports H4 and is visualised in Figures 4.4 and 4.5. Full model syntax and output are included in Table 4.1.

There was no statistically significant effect of Tetris game difficulty [Unstandardised $\beta = 9.74$, $SE \beta = 34.34$, 95% CI [-57.91, 77.10], $t = 0.28$, $p = .78$] nor of the interaction between urgency and game difficulty [Unstandardised $\beta = -17.52$, $SE \beta = 48.67$, 95% CI [-112.97, 77.94], $t = -0.36$, $p = .72$] on interruption duration. H2 was therefore not supported. Descriptive statistics for interruption duration overall and by condition are reported in Table 4.4.

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

Table 4.2: Table of means and standard deviations for interruption onset by urgency and difficulty condition.

Urgency condition	Difficulty condition	Mean (ms)	SD (ms)
High	Easy	3717	2234
	Hard	4013	2584
	Combined	3866	2418
Low	Easy	4550	3554
	Hard	4882	3955
	Combined	4717	3761
Combined	Easy	4132	2992
	Hard	4446	3363
	Combined	4289	3186

Table 4.3: Summary of fixed and random effects for interruption duration - Linear mixed effects model

Model:

$$\text{Interruption duration} = \text{urgency} * \text{difficulty} + (1|\text{subjectID}) + (1|\text{prompt})$$

Fixed Effect	Std β	Unstd β	SE β	t	p
Intercept	-.07	1513.46	83.64	18.1	<.001***
Urgency (Low)	.16	96.03	34.45	2.79	.005**
Difficulty (hard)	.02	9.74	34.34	0.28	.777
Urgency * Difficulty	-.03	-17.53	48.67	-.36	.719

Random Effects	
Group	SD
Participant (intercept)	289.8
Interruption Prompt (intercept)	295.9

Table 4.4: Table of means and standard deviations for interruption duration by urgency and difficulty condition.

Urgency condition	Difficulty condition	Mean (ms)	SD (ms)
High	Easy	1505	580
	Hard	1532	545
	Combined	1519	563
Low	Easy	1589	647
	Hard	1606	674
	Combined	1596	661
Combined	Easy	1547	616
	Hard	1567	614
	Combined	1557	614

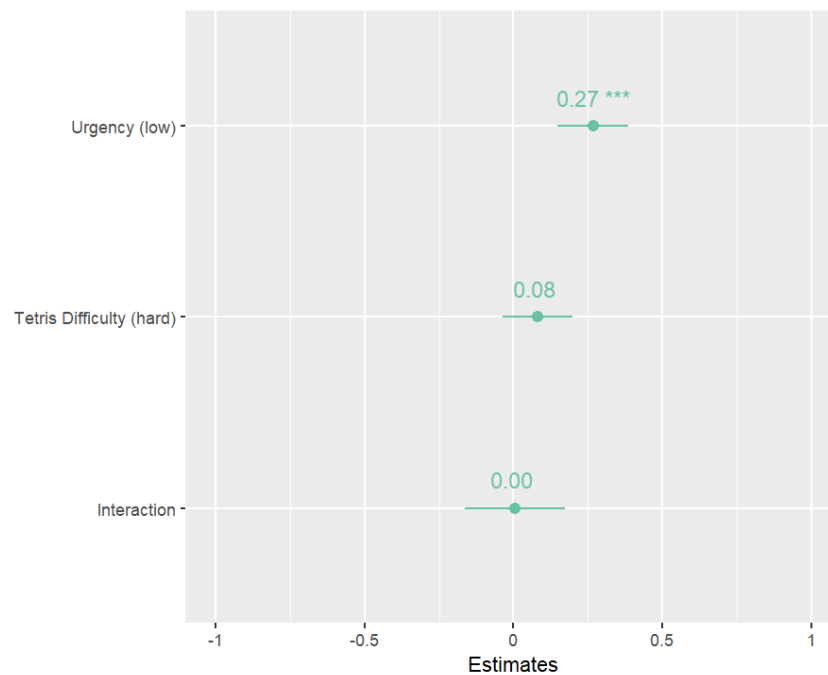


Figure 4.3: Standardised β estimates for fixed effects of urgency, task difficulty, and their interaction on interruption onset.

Access rituals

A logit mixed effects model was used to analyse the effects of urgency and Tetris task difficulty on the presence of access rituals. Again, following best practices, model selection began with the maximal fixed and random effect structure (i.e. fixed effects of urgency, difficulty, and their interaction and random slopes and intercepts at the subject- and item-level for the each fixed effect, with item-level effects of Tetris video, interruption prompt, and trial number) with complexity incrementally reduced until the model could converge (Barr et al., 2013). There was a statistically significant effect of Tetris game difficulty [Log-odds = -1.33, SE = 0.54, 95% CI [-2.43, -0.31], $z = -2.48$, $p = .01$] on the likelihood of using access rituals in interrupting utterances, with participants more likely to use access rituals during easy Tetris games (7.7% of easy trials) as compared to during difficult games (5.4% of easy trials). This supports H5 and is visualised in Figure 4.6. Full model syntax and output are included in Table 4.5.

There was no statistically significant effect of urgency [Log-odds = 0.19, SE = 0.44, 95% CI [-0.67, 1.05], $z = 0.44$, $p = .66$] nor of the interaction between urgency and game difficulty [Log-odds = 0.64, SE = 0.70, 95% CI [-0.70, 2.04], $z = 0.92$, p

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

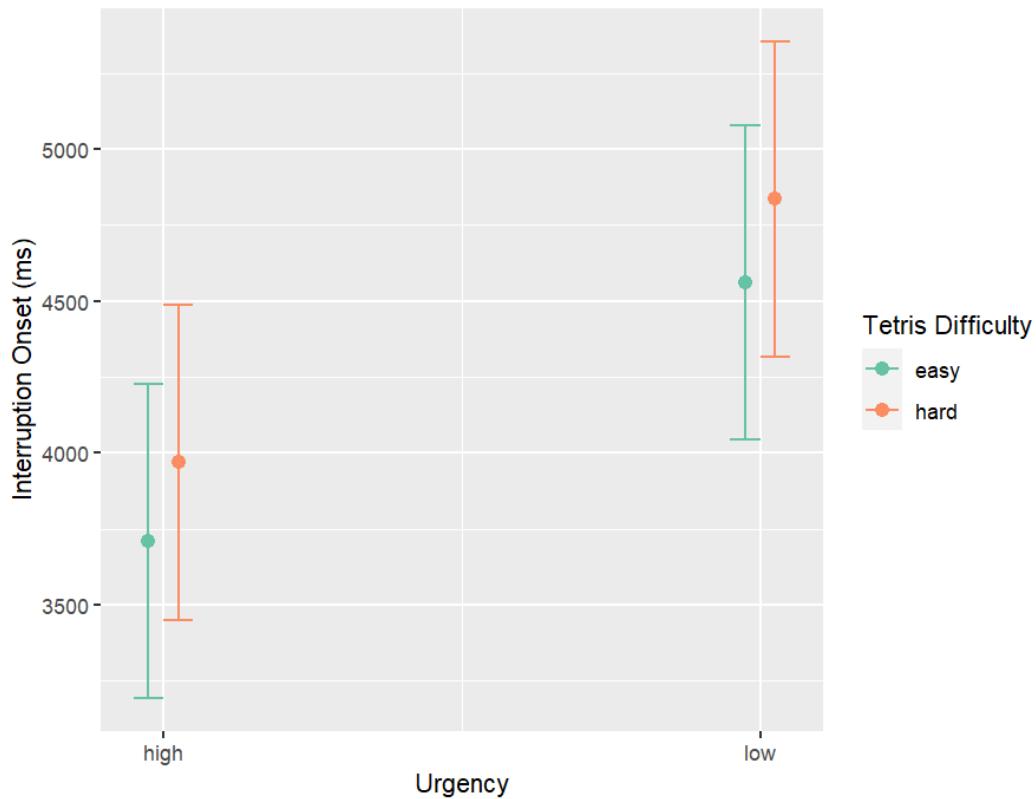


Figure 4.4: Means and standard errors for interruption duration by urgency and difficulty condition

=.36] on access ritual usage. Counts for access ritual usage overall and by condition are reported in Table 4.6.

As effects of urgency on both interruption onset and duration were of greater magnitude than effects of Tetris difficulty, RQ1 can be answered by observing that the effects of the explicit factor of urgency were more impactful on interruption timing than were effects of the estimated factor Tetris difficulty. Insofar as there was no interaction between urgency and Tetris difficulty for any of the measures, RQ2 is answered by the finding that no interaction was observed between these variables.

Table 4.5: Summary of fixed and random effects for access ritual usage - Logit mixed effects model

Model:
 $Access\ ritual\ presence = urgency * difficulty + (1|subjectID) + (1|prompt)$

Predictor	Log-odds	std. Error	z	\textit{p}
Intercept	-8.38	1.14	-7.38	<.001***
Urgency (Low)	.19	.44	0.44	.664
Difficulty (hard)	-1.33	0.54	-2.48	.013*
Urgency * Difficulty	.64	.70	.93	.355

Random Effects	
Group	SD
Participant (intercept)	6.78
Interruption Prompt (intercept)	.47

Table 4.6: Table of counts of trials containing access rituals by urgency and Tetris difficulty condition

Urgency	Difficulty	Trials with access ritual	Trials without access ritual
High	Easy	26	336
	Hard	16	336
	Combined	42	672
Low	Easy	28	320
	Hard	22	327
	Combined	50	647
Combined	Easy	54	656
	Hard	38	663
	Combined	92	1319

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

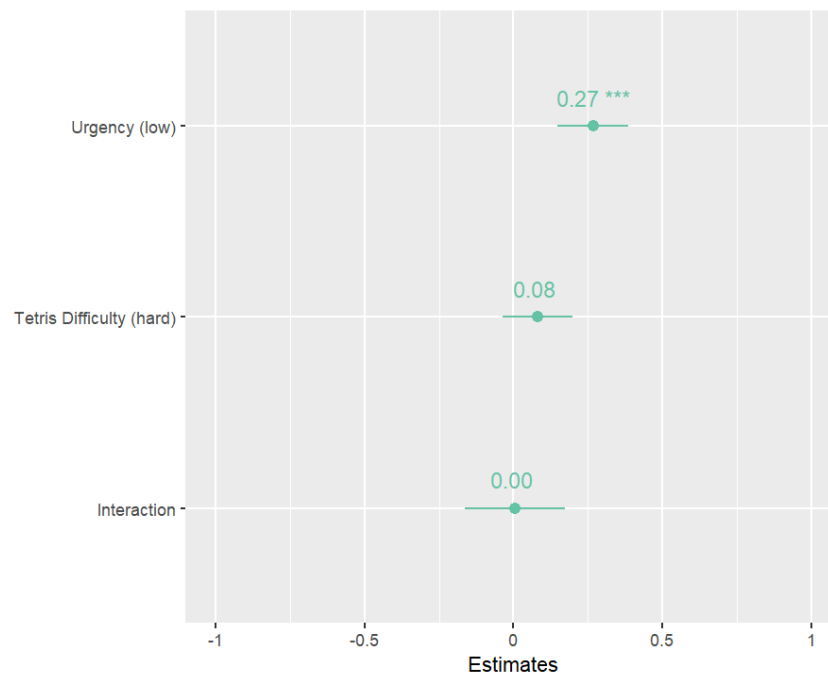


Figure 4.5: Standardised β estimates for fixed effects of urgency, task difficulty, and their interaction on interruption duration.

4.3.2 Qualitative descriptions of interruption strategies

Data analysis approach

Answers to open-ended questions were analysed through thematic analysis using a hybrid approach (Fereday & Muir-Cochrane, 2006). Initial codes were generated deductively, using themes from Chapter 3 as a starting point, with final themes likewise developed inductively through a staged review of the data and initial codes, consistent with a reflexive approach to thematic analysis (Braun & Clarke, 2006). A second coder, familiar with the data and with the codes introduced in Chapter 3, also coded the qualitative responses to enrich the discussion of potential themes through the staged review process. Importantly, this second coder was not included for the purpose of increasing validity of themes, as a reflexive thematic analysis approach holds that themes are actively created by analysts rather than emergent from some ground truth, so validity cannot be derived from agreement between coders (Braun & Clarke, 2006). For that reason, measures of interrater reliability were not calculated for the coding process, nor are they reported here.

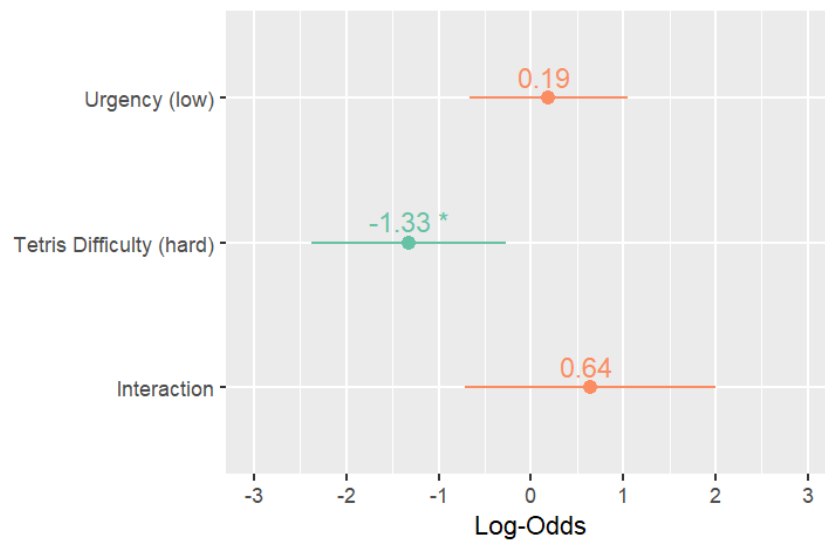


Figure 4.6: Log-odds of fixed effects of urgency, task difficulty, and their interaction on access ritual usage. Negative values indicate likelihood that an access ritual will not be produced for a given trial, while positive values indicate likelihood that an access ritual will be produced.

For the questions regarding timing, initial codes included “Prioritising Speed” for timing strategies focusing on the speed of the interruption, “Prioritising Accuracy” for timing strategies focusing on avoiding errors, “Tetris Task Characteristics” for timing strategies focusing on the state of the Tetris game, and “Message Content” for timing strategies which varied based on the specific message prompt, as well as a fifth code “No Strategy” for responses which either explicitly stated that the participant did not have any strategy in mind for timing their interruption or responses in which it was not possible to infer the participant’s strategy. For questions regarding what participants said to their partner, initial codes included “Phrasing” for speaking strategies focusing on how interruptions were phrased, “Delivery” for speaking strategies focusing on the tone or speech rate of an interruption, “Message Content” for speaking strategies which varied based on the specific message prompt as well as a fourth code “No Strategy” for responses which either explicitly stated that the participant did not have any strategy in mind for how they spoke to their partner or responses in which it was not possible to infer the participant’s strategy. Because of the hybrid approach used in this thematic analysis (Fereday & Muir-Cochrane, 2006), these inductive codes served as a starting point and differ from the themes which were generated deductively through staged review. Note that while urgency was an independent variable

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

that participants were explicitly made aware of, Tetris game difficulty was not, so responses to open-ended questions about strategies were only separated by urgency, but not by Tetris game difficulty. For this reason, themes encompass strategies mentioned across all responses and should be viewed holistically rather than as uniquely describing any particular experimental condition.

Interruption timing strategies

Three themes for interruption timing strategies were generated inductively. Participants reported trying to time their interruptions either by speaking as soon as possible or by choosing appropriate moments which they believed would reduce the risk of error in Tetris gameplay for their partner, or they did not report having a coherent strategy for timing their interruptions. Themes are presented below along with counts of how many participants in each urgency condition mentioned a given strategy (out of a total of 93 participants).

Interrupting as soon as possible (Non-urgent: 12 participants, Urgent: 29 participants) Some participants reported that they aimed to interrupt as soon as they could for all trials within an urgency condition. While many participants did not elaborate on why they chose this strategy, instead merely reporting that they interrupted “As soon as I got the question” (P61) or “after I understood the question” (P92), some participants gave additional insights as to why they chose to interrupt quickly. For some participants, identifying an appropriate moment did not appear to be a feasible strategy, as “there’s never going to be a good time, so [I] just go for it” (P02). In some cases, this was because of the participant’s ability to judge the Tetris game, with one participant indicating “I couldn’t identify when it was best to read a message so I just said it as soon as it appeared” (P87). Others indicated that identifying a good moment was not necessary, as “Most of the non-urgent games were not too complicated, so I delivered most of them as soon as I got them” (P43).

Alternatively, participants who interrupted as soon as possible sometimes indicated that they did so intentionally, to benefit themselves or their partner. One participant indicated that they interrupted “straight away, more time for the other person to respond and pick his best moment” (P66), deferring the task of choosing a good moment to the Tetris player, and construing the Tetris player’s secondary task as generating a response rather than listening to a question. Others saw urgency as a man-

date that they needed to complete their own goal of asking the question as quickly as possible, as one participant explained that the *“question was passed as quickly as possible provided, as it’s urgent”* (P48). Prioritising one’s own goals over the goals of the Tetris player was not limited to urgent trials however, with one participant specifically noting *“I attempted to deliver the messages as soon as they were displayed whether they were urgent or not.”* (P45). In this way, the tendency of some participants to interrupt as quickly as possible reflects a variety of heterogeneous strategies, including seeing Tetris games as low-risk, not having a good sense of when a Tetris game ought to be interrupted, giving their partner more time to plan a response, or prioritising the completion the interruption task over the risk of interrupting at a poor moment.

Interrupting at an appropriate moment (Non-urgent: 79, Urgent: 60) Many participants identified a strategy of waiting for an appropriate moment to deliver their interruption, rather than interrupting immediately. Participants described identifying these good moments through a variety of means however. Some participants described attempting to read the mind of their Tetris partner - to model the agent as cognitive science models might describe it. For some, this was a process of reading the Tetris player’s comfort level, with one participant interrupting *“[w]hen they looked like they had control of the situation and weren’t being pressured by mistakes”* (P48). Others described this as reading the Tetris player’s decisiveness, seeking to *“[w]ait until they had an idea set on where to go”* (P89). Others still described this mind reading process not only as an attempt to assess the Tetris player’s current state, but also trying to predict the way the player might react to an interruption.

“Wait until I could see him becoming overwhelmed. Maybe not to the point of no return, but maybe listening to me talk would have put him over the edge.” (P27)

Other participants instead chose a moment to interrupt by instead reading the state of the Tetris game, modelling the task rather than its agent. This could be a simple process of screening for bad moments and otherwise interrupting, as one participant described asking their question *“After reading it as soon as possible - as long as the board didn’t look too complicated”* (P63). Others used a more nuanced model of Tetris as a task, weighing the difficulty of the Tetris game and anticipating easier moments in gameplay.

“During more challenging games, I watched to see whether there were any mo-

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

ments when placement choice etc. was a little easier, and jump in then. Otherwise asked as soon as I could, more or less.” (P87)

Others still used explicit rules about gamestates, such as speaking only after a Tetris piece as placed in its final location, or using game speed as a single variable for determining the appropriateness of interruptions.

“I picked the point where the other person just laid their last brick down, so they had the most time before the next needed placing.” (P20)

“I tried to assess how quick the game was and then ask the question” (P69)

While many participants gave explicit explanations as to whether they engaged in task modelling or agent modelling, many gave less detailed descriptions, stating *“I just delivered it at the right time” (P55)* without elaborating on how they chose a good time. It is unclear if these participants engaged in modelling of the Tetris player, the Tetris game, both, or neither. For that reason, this theme, like the previous theme, represents a heterogeneous combination of strategies for choosing moments for interruptions, including estimating the cognitive state of the Tetris player, using fixed rules for evaluating the state of the Tetris game, making holistic judgements about the state of the game, and using unspecified strategies for selecting appropriate moments.

No strategy (Non-urgent: 6; Urgent: 4) Some participants either explicitly noted that they did not think about how to time their interruptions and as such identified no strategy, identifying that they had *“no specific pattern anyways” (P85)*. For these participants, even if they did not identify a particular strategy, none reported that they tried to interrupt at random or that they intentionally avoided any pattern to their interruption. Instead, this theme is a catch-all for the participants who did not clearly indicate their specific strategy.

Interruption structure strategies

For the questions regarding what participants said to their partner, three clear themes were generated inductively. Participants either focused primarily on communicating a sense of urgency, communicating a sense of calmness, or they did not have a clear communication strategy. These themes are explored below with comparisons of frequency in the non-urgent and urgent conditions.

Communicating urgency (Non-urgent: 17, Urgent: 55) Many participants reported

structuring their interruption utterances in a way that would communicate urgency to the Tetris player. This strategy was more commonly mentioned in regards to how participants structured interruptions in the urgent condition, but it was not limited to that condition, as participants sought to compel a rapid resolution to the speaking task across both conditions. Generally, descriptions of this strategy were short, with participants noting that they would *“keep it short”* (P82) without much elaboration as to how or why. Some indicated that they would *“try to be quick and precise with [their] questions”* (P11) or choose sentence structures which were *“simple but [a] bit faster”* (P09) so that the communication task would quickly be completed. This strategy was sometimes explicitly noted as an effort to rush the Tetris player's answer, as exemplified by the following participant: *“I tried to word the question so to almost direct them to a quick answer not to give them too much scope for having to think through a variety of possible answers.”* (P84)

That said, participants noted that they nonetheless sought to avoid causing further disruption with an overly urgent communication style, as one participant reported that they *“said what i needed to as briefly as possible but so i knew my partner would understand”* (P26). Another participant alluded to the fact that their brief communication might induce too much urgency in the Tetris player, noting that they phrased their urgent questions *“Simpler, to the point, but try not to alarm”* (P85).

Some participants made clear that their urgent communication style was not intended to rush that Tetris player, but instead a result of feeling rushed themselves, nonetheless reflecting a feeling of pressure on their own goal of completing the communication task rather than on the shared goal of avoiding disruption to the Tetris task.

“It was an urgent trial and I had no time to think carefully about the words to use so I said anything that I could think of after reading the message.” (P35).

Finally, one participant noted the difficulty they faced in an attempt to vary communication strategies by urgency condition, noting *“I tried to be more to the point, although I thik [sic] I got stuck in ‘friendly’ mode.”* (P81). This response highlights the importance of self-reported measures of interruption strategies in addition to quantitative measures, as participants may have sometimes intended to or planned to use a particular strategy for communication, but failed to execute their intended strategy due to the complex and continuous nature of the task. Taken together, this theme

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

highlights the tendency by some participants to communicate a sense of urgency in their interruptions, primarily through concise sentence structures, simple vocabulary, increased speech rates, or some combination of these features. Whether this communication of urgency was intended to hurry the Tetris player or a result of participants themselves feeling hurried varied across participants.

Communicating calmness (Non-urgent: 52, Urgent: 26)

Alternatively to the previous theme, some participants mentioned seeking to communicate a sense of calmness, normalcy, or naturalness to the Tetris player. Conversely to the theme of communicating urgency, this theme was expressed more often with respect to non-urgent trials, but it was not limited to that condition. In general, this saw participants trying to avoid making questions seem rushed or unusual, as one participant put it: *"I simply read the question to her from the bottom of the screen as if I was asking a human in front of me the same question."* (P24). Achieving this goal was sometimes approached with a *"Casual and comfortable tone."* (P30) or alternatively by through decisions around word choice, such as the participant who reported that they *"Took time to phrase the question conversationally."* (P33).

For some participants, the decision to communicate calmly was a way to reflect to the Tetris player that the speaking task was not urgent. A participant described this by stating that for the non-urgent condition *"I decided to be laidback. The message was non-urgent and so there was no reason for my tone to be urgent either."* (P47). This same strategy was used for the opposite reason by another participant, who described that for the urgent condition, they *"Tried not to be too loud and followed the instructions as best as I could and hopefully the easiest way for my partner to answer."* (P71). In this way, calmness for participants in either condition can be seen as an effort to prioritise the Tetris player's success in the game rather than rapidly finishing the speech task and the trial.

As with the previous theme, participants identified that they did not always execute the ideal strategy for interrupting. At least one participant who communicated calmness in the urgent condition ended up regretting this choice and wishing they had communicated urgency: *"I rephrased the question the as I would normally say it although, in hindsight, I should of [sic] been more concise. For example, instead of 'what is your favourite colour?' I should of [sic] said 'favourite colour?' for urgent trials."* (P01)

While participants largely identified using this set of strategies as a way to prioritise the goals of the Tetris player, it is not clear that this theme is purely selfless. While many explicitly noted this goal for the non-urgent condition, several participants simply stated that in the urgent condition, they did not alter their strategy for interruption structure. This may be reflective of an active choice to act in the best interest of the Tetris player, but it might alternatively reflect a lack of sensitivity to the urgency condition or differences between participants in terms of their interpretation of the urgency condition. Indeed for participants who saw the urgent trials as trials for which picking a good moment to interrupt was particularly important, prioritising the needs of the Tetris player is a mutual goal rather than solely allocentric and self-sacrificing.

No Strategy (Non-urgent: 24, Urgent: 12) Similar to timing strategies, some participants either did not mention a specific strategy for structuring their interruptions, or responded ambiguously, stating simply that they *“just asked question”* (P34). For some participants, this meant explicitly having no strategy, like a participant who said *“I didn’t think to much about it”* (P50), but for others, this theme was expressed through a failure to elaborate on their strategy even though they report that *“i thought about it before saying it”* (p73). No participant in this theme stated that they intentionally sought to avoid a consistent strategy, so this theme is best understood as an absence of an explicit acknowledgement of a strategy rather than an explicit absence of strategy per se.

4.4 Discussion

4.4.1 Interruption urgency affects timing more so than task difficulty

Consistent with the findings of Chapter 3, interruption urgency had a significant effect on both interruption onset and interruption duration, with urgent interruption starting sooner and finishing quicker than non-urgent interruptions, supporting H3 and H4. H1 and H2 were not supported however, as Tetris game difficulty did not have a statistically significant effect on interruption timing. Urgency in this experiment was explicitly known to interrupters. While participants interpreted Tetris game difficulty accurately (as shown by validation measures), it was on the other hand was an implicitly known cue, detectable only through an interrupter’s own judgement of a given trial. This may help to explain the much stronger effect that urgency had on interruption

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

timings as compared to Tetris difficulty as the cue of urgency was more salient and readily accessed by all participants. As coordinating the timing of an interruption and planning the content of the utterance itself is effortful work, it is understandable that people might seek to use the most easily accessed cues and forego with information that might be more effortful to process. More work is needed to better understand whether the implicit nature of task difficulty explains its relative unimportance to participants in this study as compared to urgency, or if such a cue would have a limited effect on interruption behaviour even in the absence of an explicit cue. Such work could consider either making measures of the ongoing task state or of the other person's cognitive load explicitly known to interrupters or alternatively by eliminating the provision of explicit cues like urgency altogether.

One alternative explanation for lack of a significant effect of Tetris difficulty on interruption timings might be that participants were not sure which games were more difficult than others, in line with research in the driving domain which suggests that estimated cues may be ignored if they are too difficult to utilise (Janssen et al., 2014). This explanation is limited however, due to the constrained nature of Tetris as a task, and evidenced by participants' ratings of hard games as significantly more complex than easy games. A slightly different version of this explanation might be that participants recognised that certain games were harder than others, but they did not possess the expertise in Tetris to feel confident about this cue. Participants' ratings of their own confidence in picking a time to interrupt and of difficulty in picking a time to interrupt lend support for this explanation. In this case, it may have been that participants hoped to interrupt difficult games sooner and more quickly than easy games, irrespective of urgency, but quickly finding an appropriate time to begin an interruption was more challenging in difficult games, preventing participants from realising this goal. In this case, participants may have been looking for something like natural breakpoints for Tetris games. Natural breakpoints, the moments within a task that come between subtasks, are the moments most suitable for interruptions (Janssen et al., 2012). In order to better assess this explanation, more work would be needed to understand the way interrupters conceive of natural breakpoints in Tetris. It may also be worthwhile for future work to compare the interruption strategies of experts and non-expert in Tetris, as experts in Tetris are distinctive from non-experts in both perceptual and decision making abilities for the task (Lindstedt & Gray, 2019). These

insights may likewise inform the design of non-human proactive agents, as optimal interruption strategies may well depend on an interrupter's proficiency in modelling a given task, with optimal interruption strategies only possible for well-modelled tasks and other tasks reliant on broad heuristics like "interrupt quickly" without consideration of the state of the task which is to be interrupted.

4.4.2 Estimated cues affect the structure of interruptions

Despite the lack of significant effects of game difficulty on the way people timed their interruptions of Tetris players, difficulty did indeed affect the way people structured their interruptions. Specifically, people were more likely to use access rituals during easy Tetris games than during hard Tetris games, irrespective of urgency condition, supporting H5. This difference points to a distinction between the sorts of cues which are relevant to the timing of an interruption as compared to those relevant to the structure of an interruption. While the explicit cue of urgency affected both the onset and duration of interruptions, it had no significant impact on whether that interruption was forewarned with an access ritual. Likewise, while the estimated cue of Tetris difficulty impacted neither the onset nor the duration of an interruption, it was nonetheless a significant cue for how these interruptions were formed.

A potential explanation for this difference could be that, while urgency affected only the interrupter's ratings, Tetris game difficulty directly affected the task performance of the Tetris player as well as having an indirect effect on ratings (e.g. an interruption may be more likely to be viewed as disruptive during a difficult game irrespective of urgency condition). Human-human speech is inherently social, as it conveys not only the information contained by the content of the language, but it also constructs the joint action and context of the speaker and listener (H. Clark, 1996). While the urgency of a given trial was privileged information known only to the interrupter, the difficulty of a trial was shared context, common ground for communication (H. H. Clark, 2020). Prior research has investigated how people design speech for a partner for whom some contextual information is shared and other contextual information is privileged, finding that people design their utterances to reflect common ground (Horton & Gerrig, 2002; Yoon et al., 2012). While that work hasn't yet focused on access rituals in specific nor politeness markers in general, it is possible that the effect of task difficulty on access ritual usage is related to this phenomenon of audi-

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

ence design influenced by contextual information shared in common ground. That is to say, participants in this study may have been more likely to use access rituals during easy trials as compared to hard trials as an acknowledgement of a shared context with the Tetris player, with the common ground of a difficult Tetris game providing an excuse to forego politeness. This finding has important implications for how human interruptions can be used to inform the design of speech interruptions for non-human agents, insofar as politeness may likewise be desirable for speech agents to signal, but acceptable to forego in particular circumstances.

Whereas findings for timing may be limited in how they can be applied to AI-driven non-human agents, which might be different in their cognitive constraints and information processing abilities from human brains, it appears that the cue of task difficulty is usable to humans in a way that can in fact inform non-human agent design. Insofar as people avoided using access rituals during difficult tasks, this observation can be concretely applied by designers. While this study does not show whether or to what extent access rituals enhance interruptions when they are used, the avoidance of their usage in certain contexts points toward an understanding by interrupters about the relative social importance of access rituals as compared to other priorities such as minimising speaking time or avoiding the inclusion of words in an interruption utterance which are not required. While this was not reflected in overall differences in interruption duration, the elimination of access rituals may represent a low-cost strategy in terms of cognitive effort for accomplishing such a goal, without needing more effortful rephrasing of a planned utterance. Whether forgoing access rituals would be considered low-cost for human-machine dialogue is an open question however. Access ritual usage conveys social power, with deferential access rituals (such as a polite knock on a door before entering) more likely to occur when the access requester is socially subordinate to their communication partner (Hutte et al., 1972). Insofar as speech agents are seen as social subordinates (L. Clark, Munteanu, et al., 2019; Luger & Sellen, 2016), a user may find an interruption without an access ritual as less appropriate than they would if a peer used an access ritual in the same context. People apply social rules when interacting with computers (Nass et al., 1994) including upholding rituals of politeness (Nass et al., 1999), but it is not clear if they expect the same from a speech agent. Indeed, some work has indicated that overly-social speech interaction from agents is seen as undesirable or creepy (Aylett, Cowan, et al.,

2019; P. R. Doyle et al., 2019). More work is needed to understand whether speech agents should use access rituals in most interruptions, in no interruptions, or if like people, they should adapt their access ritual use to common ground contextual information. This topic is further investigated in Chapter 6.

To further explain why Tetris difficulty influenced access ritual usage but not interruption onset or duration, some aspects of the cognitive processes relevant to an interruption were considered. The themes generated through qualitative analysis point toward differing strategies in selecting relevant cues in the process of timing interruptions as compared to the process of structuring them. For instance, while some participants reported using the egocentric cue of urgency in isolation for each of these processes, those who described using cues of the Tetris game reported doing so in different ways for timing as compared to structuring interruptions. For instance, participants who tried to find good moments to speak might consider task difficulty at an intra-task level, finding the relatively least difficult moment, but they might use inter-task difficulty to decide whether communicating calmness is necessary for how their interruption is structured. This qualitative finding gives an indication that these two different aspects of forming a spoken interruption are sensitive to different inputs and therefore free to vary differently depending on the context of the interruption. Neuroscience research indicates that, when engaged in conversation, different neural mechanisms are used to predict the content of an interlocutor's message as compared to the timing of their utterance and its ending (Arnal & Giraud, 2012). Indeed this insight underlies current understandings of turn-taking in psycholinguistics, by which people use different cues to predict their conversational partner's utterance content - and thus plan their own message - as compared to the cues they use to predict the end of their partner's turn (Garrod & Pickering, 2015). The present findings point toward a similar separation of cognitive mechanisms for how people plan a spoken interruption's content and its timing in the absence of an interlocutor.

4.4.3 Minimise disruption or interrupt quickly?

Through thematic analysis of self-reported interruption strategies, this chapter builds upon Chapter 3 in exploring the heterogeneity of strategies used by interrupters. Broadly, two coherent strategies were evident for how people timed their interruptions, and two coherent strategies were evident for how people structured their inter-

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

ruptions. In terms of timing, people either interrupted as soon as possible or they attempted to find appropriate moments to interrupt by modelling the state of the Tetris game or the state of the Tetris player. In terms of structure, participants reported seeking either to communicate urgency, minimising the time they spent speaking and attempting to elicit fast responses from their partner, or they sought to communicate calmness, minimising the impact that their interruption might have on the Tetris player's attention and task performance. Like Chapter 3, this mix of strategies aligns with the current state of understanding of how people self-interrupt, with notable differences between individuals (Cha et al., 2020; Dabbish et al., 2011).

Insofar as some participants sought always to minimise the time they spent interrupting, this aligns with findings from Chapter 3 as well as with prior empirical work on task-switching, which has demonstrated people's willingness to risk task performance in order to minimise task completion time in multitasking contexts (Brumby et al., 2011; Horrey & Lesch, 2009). Indeed, the Soft Constraints Hypothesis, a theoretical framework for how people carry out complex tasks, states that people seek to optimise their strategy for a complex task for minimising total time spent rather than any other variable like effort or performance quality (Gray et al., 2006). For this reason, it is unsurprising that interrupting as soon as possible and eliciting fast interruptions were prevalent themes in self-reports of interruption strategies.

Self-reports of interruption strategies were not limited to those which sought to minimise time spent on the task however. Participants also reported another set of strategies: minimising the disruptiveness of their interruptions through choosing appropriate moments to interrupt and through avoiding interruption utterances which would rush the Tetris player. In terms of interruption structure, this finding is in following with Chapter 3, and the use or disuse of language patterns to communicate urgency or to avoid doing so broadly aligns with linguistic evidence for differences between urgent and non-urgent language (Hellier et al., 2002; Landesberger et al., 2020a, 2020b). In terms of timing, participants mentioned that they either sought to identify good moments to interrupt by using cues from the Tetris game itself, such as the current location of a Tetris piece or the presence of an obvious final location, or else they sought to use these cues to infer the cognitive state of the Tetris player. This reference toward task modelling to identify good moments to interrupt further echos research around natural breakpoints (Janssen et al., 2012). While quantitative

data from the present study and qualitative data in both the present study and Chapter 3 both point toward participants using natural breakpoints as cues for timing their interruptions, this cannot be demonstrated empirically due to a lack of established evidence for what constitutes a natural breakpoint in Tetris. The following chapter (Chapter 5) begins to investigate this question.

4.4.4 Limitations

One confounding variable in understanding differences between effects of urgency condition and Tetris game difficulty is the nature of urgency as a variable which had a direct impact on participants' egocentric goals (i.e. their alleged total game score determined by ratings) while Tetris game difficulty only had an indirect effect on this goal and a direct effect on the goals of the Tetris player. For some participants, game difficulty may have been seen as irrelevant to their own goals, so these participants may have ignored it altogether, using only the egocentrically important cue of urgency. Self-reports of strategy cast doubt on this however as participants did not differentiate between whose goals they were seeking to optimise for, and indeed reported using the game state or the player's perceived cognitive load to choose their interruption strategies. Future work should seek to eliminate this confound however, choosing either only explicit or only implicit indications of either egocentric and allocentric goal-related cues.

Similar to the previous limitation, it may conversely be the case that the effects of game difficulty are overstated in the present data, owing to post-trial ratings of game difficulty and of ease in finding a good moment to speak making these features and strategies more salient. Participants may have felt more inclined to find good moments to speak due to this constant reminder, as opposed to ignoring estimated cues and instead always speaking as quickly as possible. Likewise, participants may have identified game difficulty as a feature by which opportune moments for interruption can be identified in part due to the fact that the rating of game difficulty made this cue salient. Insofar as game difficulty failed to produce a significant effect on interruption duration or onset however, it is not clear that this increased salience led to any meaningful difference in how much participants utilised task difficulty as a cue. Likewise, while self-reports of strategies did include many participants mentioning that they tried to find good moments to interrupt and that they used characteristics

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

of the Tetris task to choose those moments, this same tendency was found in Chapter 3 which did not include any rating of game difficulty or ease in finding a moment to speak. Indeed, the decision to include these questions as a rating step was born from observations of data from the study described in Chapter 3 which indicated that these features were already salient to participants. In this way, it is unlikely that the rating steps either made salient cues which otherwise would not have been utilised nor that any increase in salience toward these cues resulted in any meaningful differences in participant behaviour. Nonetheless, future research using this paradigm should consider excluding these ratings, instead seeking to check that different difficulty conditions are noticeable to participants after data has been collected.

This study in part aimed to investigate whether the use of estimated cues by interrupters is affected by the ease with which an interrupter can interpret cues from another person's task, as proposed in prior research on spoken interruptions of driving (Janssen et al., 2014). Whereas that research sought to make cues more salient by presenting them in an auditory modality, the present study sought to use the relatively constrained task of Tetris to make cues easier to understand. While this study did indeed find that participants could tell easy games from hard games, it also found quantitatively that participants were less confident about selecting good moments to interrupt hard Tetris games, indicating that this estimated cue, while interpretable, was still difficult to utilise. This may, in part, owe to differences in Tetris expertise between participants. Indeed, prior research on Tetris has indicated that expert Tetris players differ from novice players in terms of both their perception of a Tetris game board and their decision making around Tetris gameplay (Lindstedt & Gray, 2019). It may be the case that expert Tetris players would have been more able to find and utilise breakpoints in difficult Tetris games, leading to differences in interruption onset by Tetris game difficulty. Still, a lack of differences in interruption duration by game difficulty condition is harder to account for with differences in Tetris expertise, instead pointing toward a lack of willingness to use task difficulty as a cue rather than a lack of ability. Nevertheless, future work should be mindful of the extent to which an interrupter's expertise in the task they interrupt may impact the strategies they select for interrupting.

As in Chapter 3, although a majority of participants believed that their partner was a human, most of these participants correctly judged that their human partner

was pre-recorded. It is unclear whether this view changed over the course of the experiment or whether participants were suspicious throughout the experiment, but modelling trial order as a random effect once again helps to capture any effect this may have had on interruption timing or access ritual use. Likewise, as in Chapter 3, self-reported strategies of interruptions reflected that participants took the experiment seriously and sought to interrupt as if their partner was human. It is not clear how strongly participants held their beliefs of their partner's true identity, nor that this question was salient to them during the experiment. It is therefore unclear whether the use of a pre-recorded partner limits the findings of this study, but future work should consider comparing participants' interruptions of a pre-recorded partner to interruptions of a live partner to better understand how present findings may be limited.

4.5 Conclusion

The present study sought to understand how people use both explicit and estimated cues to select strategies for interrupting complex continuous tasks with speech. To investigate this question, an online experiment was designed which tasked participants with asking questions to a partner playing Tetris, with urgency of the interrupting task and difficulty of the Tetris task manipulated as independent variables. Quantitative analysis found that urgency, the explicit cue, had significant effects on interruption onset times and durations, while difficulty, the estimated cue, did not. Difficulty did however impact the structure of interruptions, with participants using access rituals to interrupt easy games more frequently than difficult ones. Taken together with qualitative findings around participants' desire to communicate a sense of calmness for some interruptions, it is clear that estimated cues are not ignored altogether. Indeed, the prevalence with which participants report seeking to interrupt Tetris games at appropriate moments indicates that interrupters may use natural breakpoints in the tasks they interrupt as opportunities to speak.

The present chapter as well as the previous chapter indicate that people use a variety of strategies for interrupting, taking into account the urgency of their interruption as well as the state of the task they seek to interrupt. In order to investigate whether participants do in fact use natural breakpoints to time their spoken interrup-

Comparing the Impacts of Urgency and Task Difficulty in Speech Interruptions

tions of Tetris, the next chapter seeks to use data from this study and the previous to model the natural breakpoints of the Tetris games which participants watched.

5 * **Characterising Appropriate Moments for Interruptions of a Complex Continuous Task**

5.1 Introduction

The preceding chapters sought to investigate the extent to which urgency and the difficulty of an ongoing complex task influenced the way people interrupted others engaged in that task (operationalised as Tetris). Each of these factors is static across a given Tetris game however. In a continuous task (like Tetris), conditions of the task dynamically change and the demands on the person performing that task consequently fluctuate. For this reason, the present chapter turns toward further investigating the impact of the dynamic aspects of a task, using Tetris as an example of such a task.

5.1.1 Classifying task interruptibility

In multitasking literature, people tend to structure the way they switch between their own tasks by utilising breakpoints (Borst et al., 2010; Iqbal & Bailey, 2008), the moments between discrete tasks or subtasks which are useful for task-switching. Task-switching during these moments is less burdensome seeing as prior goals for the interrupted task have been resolved and new goals have not yet been introduced (Janssen et al., 2012). The boundaries between subtasks in complex tasks are far less clear to identify than those for discrete or simple tasks however. Attempts to model good moments to interrupt a complex task (e.g driving) have proven difficult

however (Semmens et al., 2019). In order to understand the extent to which participants in the studies described in previous chapters used the dynamic characteristics of Tetris games to time their interruptions, this chapter seeks to first understand how observers of Tetris understand those dynamics.

In previous research on spoken interruptions of continuous tasks, tasks like driving or household routines have been modelled in terms of interruptibility by randomly interrupting a person engaged in that task then asking whether an interruption came at an appropriate time (Cha et al., 2020; Semmens et al., 2019). This inductive approach seeks to make many observations of interruptions and their appropriateness to understand markers of interruptibility from the perspective of the person who is being interrupted. Insofar as the decision about interrupting another person is not made by the interrupted party, this study instead seeks to model interruptibility of Tetris from the perspective of the interrupter. Rather than interrupting participants and asking them if the interruption was appropriate, the present study reverses this paradigm, asking participants instead to choose the best moment to interrupt. This approach allows for a deeper investigation of interruption timing data from Chapters 3 and 4, shedding light on whether the spoken interruptions elicited from experiment participants are consistent with what interrupters see as appropriate periods for interruption. The present investigation of ideal moments for interruptions draws upon the insights from literature on natural breakpoints, but it does not do so by breaking Tetris games down into subroutines and labelling moments between these routines as breakpoints. Indeed, as literature on human Tetris skill shows, different people conceive of different subtasks within a given segment of a Tetris game (Lindstedt & Gray, 2019). The present chapter instead seeks to identify the windows of Tetris games that people broadly identify as most interruptible at a holistic level, avoiding decomposing a Tetris game into strategic units which individual interrupters may disagree with. In this way, the present chapter does not specifically identify natural breakpoints in Tetris games, it instead identifies and describes interruptible windows of Tetris gameplay, from a data-driven perspective. By avoiding the natural breakpoints label, this study makes no commitment to the notion that subtask boundaries will be the determinant of which windows of time are seen as interruptible, nor about how interrupters conceive of the tasks and subtasks of Tetris at all. Instead, the chapter follows in the theory-agnostic and inductive tradition of more recent studies of

continuous task interruption (Cha et al., 2020; Semmens et al., 2019) . In doing so, this chapter further seeks to sketch some characteristics of interruptibility for Tetris so that those guidelines can be used to better understand the interruptions of Tetris observed in Chapters 3 and 4.

5.1.2 Event structure in a Tetris game

The event structure of Tetris can be understood as a series of episodes which can be further deconstructed into discrete motions, either initiated by the player or automatically triggered at particular time intervals. Lindstedt and Gray describe the event structure of Tetris in detail in (Lindstedt & Gray, 2019) which is summarised here. An episode in Tetris constitutes the travel of a Tetris game piece from the top of the Tetris game board as it falls to either the bottom of the 20-row board or can no longer continue to fall due to other Tetris pieces positioned beneath it. The Tetris piece falls at a continuous rate determined by the game difficulty level. During this falling process, the player can initiate movements including lateral translations of the Tetris piece, 90 degree rotations of the piece, or manual vertical drops, bringing the piece closer to the bottom of the board. When a row of the gameboard is completely filled by parts of Tetris pieces, that row or rows flash for several frames before disappearing, with all filled cells from rows above falling to fill the vacated row(s). In any other cases, after a piece reaches a point at which it can no longer continue to fall and a falling motion is either automatically triggered or initiated by a player's manual drop, that piece remains in its present position and a new episode begins with the appearance of a new Tetris piece at the top of the game board. This event structure is that by which previous work has sought to understand how Tetris players approach the task of Tetris and gain expertise in it (Lindstedt & Gray, 2019).

It is unclear however whether this event structure cleanly comports with the way someone observing a Tetris game but not participating in it would conceptualise the task of Tetris. This event structure of Tetris is premised on a known set of goals and subgoals for the task of playing Tetris (Zacks et al., 2007). The process of interrupting a task requires not only the understanding of the goals of the person carrying out that task, but also the balancing of those goals against the goals of the interrupter, which may compete with the interrupted task for time and the attention of the interrupted party. In this way, the present research uses the event structure of Tetris as only one

candidate for understanding how interrupters might model that task, but does not deductively impart this structure on analysis of Tetris interruptibility.

5.1.3 Aims and research questions

Measuring the extent to which people use interruptible windows to time their interruption of a task or designing interruptions to make use of interruptible windows requires a model of interruptibility of that task. In order to assess the extent to which interrupters make use of Tetris game states when interrupting a Tetris player, it is therefore necessary to first understand which features of a Tetris game are salient to a non-playing observer. The present study aims to first ask: what features of Tetris gameplay do observers judge to impact the interruptibility of a Tetris game (RQ1)? Upon establishing a framework for interruptibility of Tetris games, this study next seeks to analyse the extent to which interrupters actually utilise cues of interruptibility when interrupting Tetris gameplay in real time. To assess this question, the study further asks: in studies from Chapter 3 and Chapter 4, to what extent did participants initiate interruptions within interruptible windows? (RQ2) and what impact did the variables from those studies (urgency in Chapter 3 and both urgency and Tetris task difficulty in Chapter 4) have on participants' utilisation of those interruptible windows (RQ3)?

5.2 Methods

5.2.1 Participants

70 crowdworkers (40 men, 28 women, 1 non-binary, 1 prefer not to say, $M_{Age} = 40.84$, $SD = 10.42$) were recruited from a crowdsourcing platform (Amazon Mechanical Turk). In order to mitigate familiarity effects, participants from previous studies were not eligible to participate in this study. All participants were native or near-native English speakers. All participants were familiar with the game Tetris, with most indicating they either they had played before, but do not play regularly ($N = 64$; 91.4% of sample) or that they play regularly ($N = 5$, 7.1% of sample) (5 point Likert scale; 1 = I am not at all familiar with Tetris; 5 = I regularly play Tetris). The study took approximately 20 minutes and participants were paid \$6 Mechanical Turk credit for participating in the research. The study received ethical approval through the university's ethics

procedures for low-risk projects (Ethics code: HS-E-21-39-Edwards-Cowan).

5.2.2 Materials

Tetris paradigm: Following the Tetris paradigm used in the previous chapters, the present study likewise aims to explore how people interrupt a partner who is executing a primary task that requires ongoing attention and cannot be arbitrarily suspended (continuous) and allows for a broad variety of responses rather than a single fixed response (compound) (Kieras et al., 2000; Salvucci, 2005). To ensure that insights from the present study could be used to inform and to better understand data collected in previous chapters, this study likewise uses Tetris as its example complex task. The present research seeks to identify which moments within Tetris games participants see as best suited for interruption, differing from the aim than that of the previous chapters which sought to understand how participants timed and structured interruptions in real time. For this reason, participants in this study were not told that they were watching live Tetris gameplay, instead they were informed that they were watching prerecorded videos of Tetris. The paradigm was made using JSPsych (de Leeuw, 2015) and is described in detail below. All code and stimuli are available at ¹. All materials including participant information sheets, consent forms, questionnaires, and debrief sheets are provided in Appendix C.

Tetris video clips

All trials within the paradigm were recorded videos of Tetris, standardised across all participants. Video clips lasted 8500ms each. This duration was selected as it represented approximately three standard deviations greater than the average interruption onset in the data gathered in Chapter 3. In this way, each video clip represented a wide window of candidate moments at which people might have begun an interruption in previous experiments. All video clips were taken from the same larger videos of Tetris gameplay from which stimuli from Chapters 3 and 4 were generated, though clips in this experiment and stimuli from prior chapters do not overlap exactly, as is further explained below. 36 total Tetris video clips were created, 18 of which matched the characteristics of the easy Tetris videos from Chapter 4 and 18 of which matched the characteristics of the hard Tetris videos from Chapter 4. Each partici-

¹https://osf.io/jchge/?view_only=f44e8041b0724b359ee7548bd99d6ad0

Characterising Appropriate Moments for Interruptions of a Complex Continuous Task

participant saw 18 video clips total, nine from each difficulty condition, counterbalanced so all video clips were seen by an equal number of participants.

It was not possible to create trials which contained Tetris sequences which all participants from either Chapter 3 or Chapter 4 would have seen, owing to the variance between participants in interruption onset in those studies as well as the randomisation of prompt delays. While these all clips used in this study fit the same inclusion criteria used for generating stimuli for the previous studies, the increased variance is intended to yield a more general understanding of perceived interruptibility across Tetris games, reducing effects of unsystematic bias introduced through the reuse of some videos.

Tetris video clips were presented as prerendered gifs at 800×800 resolution displayed on a webpage at 30 frames per second, in colour, on a neutral background, and without sound. Video clips included a Tetris board, a box in the upper right corner indicating the next piece, and a counter displaying the frame number (numbered 000 to 255) of each frame of the clip. The counter was displayed horizontally centre in the clip, toward the bottom half of the clip, just to the right of the Tetris board, in large red text. Each clip was presented once in its entirety, followed by a 1000ms pause during which a grey fixation cross was displayed at the centre of the webpage. After the pause, the clip was presented once more in its entirety.

Interruption prompts and range selection

After a clip was presented twice, the clip would continue to play on loop with additional instructions appearing on the screen. Beneath the clip, a prompt would appear instructing the participant “You need to ask the Tetris player” followed by a question. These interruption prompt questions were the same as those from Chapter 3, but phrased as a complete question rather than as fragments (e.g. “You need to ask the Tetris player ‘How many siblings do you have?’”). The full list of interruption prompt questions can be seen in Table 5.1. Below the interruption prompt, text instructed the participant to “Pick the number that appeared at the best moment to begin speaking.” along with four buttons, labelled “001-064”, “065-128”, “129-192”, and “192-255” respectively, representing the four quartile ranges of frames from the video clip. Buttons were displayed horizontally across one line.

Interruption prompts
What's your name?
Have you played Tetris in the last month?
Which hand do you write with?
Do you have any cats or dogs?
What's the weather like right now?
What time did you go to bed last night?
How old are you?
What's the last series you watched?
How many siblings do you have?
What color are you wearing?
What did you eat for dinner last night?
What was the last movie you watched?
What's your favorite ice cream flavor?
What did you have for breakfast?
Have you ever been to Paris?
What is your favorite fruit?
What is your favorite color?
Do you have a lucky number?

Table 5.1: *Table of interruption prompt questions.*

Tetris frames

Following a participant's selection of a range of frames in which they would choose to begin their interruption, participants were shown a horizontal carousel of images of all of the unique frames within that range. Each frame was presented as a still image at 800×800 displayed on a webpage in colour, on a neutral background, and without sound. In the bottom right corner of each image, a number indicating the frame's order within that range (re-numbered, starting from 1) was displayed in large red text. Beneath the image carousel, a grey slider was visible, enabling participants to slide horizontally between frames on the same webpage. Beneath the slider, text instructed participants to "Choose the best moment to interrupt the player" with a dropdown menu. The menu was set to a default response of "select an option" and contained an option "go back" as well as numbered options for each of the unique frames from the carousel. Participants were able to select only one frame from the dropdown menu, representing the single best moment for initiating an interruption. Beneath the dropdown menu, a button labelled "Continue" became enabled after a response was selected in the dropdown menu.

5.2.3 Experimental conditions

Tetris clips were selected from a variety of Tetris videos across two conditions of Tetris difficulty: easy Tetris games and hard Tetris games. Easy games were defined as games with no Tetris pieces above the middle of the board and falling speed set to the game minimum of 1.25 rows per second. Hard games were defined as games in which the board was approximately half-full, with several columns of the board filled to or beyond the centre, with falling speed set to 10 rows per second.

As a manipulation check, after each trial, participants were asked after each trial to answer on 7-point Likert-type scales each of three questions: “How complex was the Tetris game you just saw?”, “How easy was it to choose a moment to speak during in the Tetris game you just saw?”, and “How confident are you that you picked a good moment to speak?”. Manipulation check results are detailed in section 3.2 below.

5.2.4 Measures

Tetris frame selection

The primary measure of interest was the frame from the video clip which was visible at the moment at which a participant would have decided to begin their interruption of the Tetris player. In order to operationalise the moment of interruption, it was necessary to represent a continuous task, Tetris gameplay, as a sequence of discrete moments. To achieve this, video frames were used rather than, for instance, fixed windows of time (e.g. dividing videos into 100ms sequences). This allowed participants to indicate the precise location of the active Tetris block at the moment of interruption.

Because Tetris pieces can be moved and rotated in addition to their constant vertical drop, unique frames within a clip and between clips are displayed for variable amounts of time. Likewise, each quartile of each clip has a variable number of unique frames. Operationalising discrete moments in a Tetris game as frames rather than periods of time likewise fits with previous research on Tetris as a task in cognitive science research, which decomposes Tetris episodes into “motions” - both those automatically triggered in the game and those initiated by players - as the fundamental unit of analysis (Lindstedt & Gray, 2019).

Post-trial ratings

After each trial, participants were asked to answer on 7-point Likert-type scales each of three questions: “How complex was the Tetris game you just saw?”, “How easy was it to choose a moment to speak during in the Tetris game you just saw?”, and “How confident are you that you picked a good moment to speak?”. These measures were used to validate that participants perceived differences between levels of Tetris difficulty and that perceptions of variance between video clips within each level of Tetris were homogeneous.

Demographic Questionnaire

Participants were asked a number of questions about themselves such as age, gender, and level of education, their level of experience with Tetris, and how they decided which moment would be best to begin speaking, and whether they had any additional comments.

5.2.5 Procedure

Participants were informed of the aims of the research, the data to be collected, and their data processing rights via information sheet webpages. Participants were then asked to give consent to take part in the study. Participants then were briefed on the procedure of the experimental task through a series of instructional pages with screenshots of the experimental task.

Participants were shown video clips of Tetris gameplay and asked to choose the moment at which they would begin an interruption of the Tetris player. Through a multi-stage selection process, participants were able to hone in on a specific moment to interrupt for each video clip by first watching the 8500ms clip twice on its own, then watching the clip on repeat for as long as they wanted. Interruption prompts appeared at this stage, instructing participants about the question they should imagine asking the Tetris player. Buttons also appeared at the bottom of the screen during this stage, with options for selecting the 2125ms quartile which contained the moment they would choose to begin their interruptions. After choosing the quartile, participants moved to the next stage of a trial in which they used a horizontal image carousel to slide through still images of individual frames from that section of the video and

select the individual frame that best represented the moment at which they would begin their interruption. At this stage, participants had the option to select the frame that represented the best moment to begin an interruption of the Tetris player. Participants also had the option to return to the previous stage in order to view the video again or to select a different quartile. After selecting a frame, participants were presented with three post-trial rating questions on one screen. Participants completed 18 trials, nine of which featured hard Tetris games and nine of which featured easy Tetris games, ordered randomly. After completing all 18 trials, participants were asked to complete a brief questionnaire about their own background and their experience with the experiment, comprising the demographic questions and the open ended question listed above. After completing the questionnaire, participants were fully debriefed and thanked for participating and given instructions on receiving their payment.

5.3 Results

5.3.1 Analysis Approach

The aim of this study was to use participants' heterogeneous selections of moments which supported interruption in order to identify groups of moments - interruption windows - for each Tetris video clip. Because the amount of interruption windows in a given trial and the size of these windows was unknown, this analysis uses k-means clustering as a method for identifying interruption windows. K-means clustering is a quantitative optimisation technique which seeks to group points in n-dimensional space into clusters, minimising within-cluster sum of squared distances for a fixed number of k clusters (Lloyd, 1982). For analyses like the present study in which the number of clusters is unknown, approaches vary, but tend to either use heuristic approaches for selecting k by which k is determined using rules of thumb such as visually identifying an elbow in a plot of clusters against explained variance or else by using information criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) so as to balance model fit with parsimony (Kodinariya & Makwana, 2013). As such, for the present study, k is iteratively determined for each Tetris clip by choosing the number of clusters which minimised BIC. While k-means clustering typically involves a heuristic approach to determining cluster membership, cluster analysis of 1-dimensional data, such as the data collected in this study, enables the

use of an algorithmic approach which guarantees optimal clustering for k-clusters (Song & Zhong, 2020). That is to say, while multivariate datapoints are most typically clustered via iteratively refined estimations which are sensitive to a stochastically selected initial clusterings, this study uses a replicably optimal clustering algorithm as its data is univariate. Cluster analysis was carried out in R version 4.1.1 (R Core Team, 2020) using the Ckmeans.1d.dp package version 4.3.3 (Wang & Song, 2011). A total of 1260 frames were selected across the 36 video clips in the experiment. No data needed to be removed due to technical issue or by participant request. All frame selections were therefore retained for the final analysis of the dataset.

5.3.2 Manipulation Check

As a manipulation check, after each trial, participants were asked after each trial to answer on 7-point Likert-type scales each of three questions: “How complex was the Tetris game you just saw?”, “How easy was it to choose a moment to speak during in the Tetris game you just saw?”, and “How confident are you that you picked a good moment to speak?”. On all three questions, paired-samples t-tests revealed significant differences between easy and hard Tetris games. Participants rated hard games ($M = 4.04$, $SD = 1.49$) as more complex than easy games ($M = 3.15$, $SD = 1.46$) [$t(69) = 9.45$, $p < .001$]. Participants rated hard games ($M = 4.69$, $SD = 1.59$) as less easy to choose a moment to speak than easy games ($M = 5.25$, $SD = 1.48$) [$t(69) = -7.12$, $p < .001$]. Participants rated themselves as less confident that they chose a good moment to interrupt during hard games ($M = 4.94$, $SD = 1.58$) as compared to easy games ($M = 5.34$, $SD = 1.50$) [$t(69) = -5.71$, $p < .001$]. These differences indicate that participants readily perceived the differences between easy and hard Tetris games.

5.3.3 Cluster generation and stabilisation

Across the 36 video clips, five clustered optimally into a single group, indicating that no distinct windows of time could be identified. After removing 5 trials, the mean number of optimal clusters (k) was 4.84 and the SD was 2.02, with hard Tetris games ($n = 16$) having a mean of 4.81 clusters ($SD = 1.83$) and easy Tetris games ($n = 15$) having a mean of 4.87 clusters (2.26). In total, this represented 150 total clusters across the 31 trials with multiple clusters. Among these clusters, there was a mean number of observations of 7.23 and an SD of 5.45. Each of these clusters represents a candidate

window of gameplay that several participants identified as suitable for interruption, but given the small sample size of observations per clip and per cluster, it was also necessary to ensure that clusters were stable before drawing any conclusions from them.

Following previous work on cluster validation (Hennig, 2007), bootstrapped re-sampling with added noise was implemented in order to compare similarity between observed clusters and samples that had been resampled with noise. For each clip, 1000 bootstrapped and jittered resamples were run of the same number (35) of observations per trial. These bootstrapped samples were used to generate new clusters using the same observed k for that clip, again following best practices from (Hennig, 2007), using the `fpc` package version 2.2-9 (Hennig, 2020) in R version 4.1.1 (R Core Team, 2020). For each cluster, the mean of the Jaccard coefficient, the measure of similarity between sets, was calculated between the observed cluster and each bootstrapped cluster, following best practices (Hennig, 2007). This coefficient indicates the proportion of similarity between a given clustering from a bootstrapped sample and the observed clusters from the actual data sample. To illustrate the calculation, a bootstrapped sample for which each observation belonged to the same ranked cluster as the observation of the same rank from the observed sample would have a Jaccard coefficient of 1, whereas a bootstrapped sample for which no observation was in the same cluster as the same ranked observation from the observed sample would have a Jaccard coefficient of 0. Following best practices, clusters with a mean Jaccard coefficient of 0.75 or above were regarded as stable (Hennig, 2007). This resulted in 79 stable clusters across 31 of the 36 clips. These clusters of frames can be seen as the windows of time which participants reliably judged to be good moments for interrupting Tetris games within the Tetris videos.

5.3.4 Content analysis of video segments

In order to understand commonalities between these interruptible windows, a content analysis of the segments of video clips represented by stable clusters was performed with a second researcher. First, the range of time represented by each segment was assessed. Segments had a mean of 33.16 frames (1105ms) and a standard deviation of 36.12 frames (1204ms). Segments consisting of more than two standard deviations above the mean number of frames were removed from analysis ($n = 13$)

as these segments represented many Tetris events and likely correspond to several different interruptible windows. This left 66 segments. Ten random segments were coded to generate initial codes of salient features of the Tetris gameplay. Next, along with a second rater all 66 segments were independently coded. Following initial coding, raters met to discuss discrepancies. Through discussion, it was discovered that segments lasting longer than 50 frames (n=5) were major sources of disagreement, as these segments again typically contained the movement of more than two Tetris pieces and several events. The raters had identical codes for 57 of the remaining 61 segments (93.4%) with the remaining discrepancies, mostly resulting from human error, resolved through discussion. The codebook is provided in Appendix C and both raters' codes are available at ².

This content analysis allowed for follow-up thematic analysis of the segments, using an inductive, reflexive approach (Braun & Clarke, 2006). Initial thematic codes were adapted directly from the final content analysis codes. These codes were then reviewed through an iterative process seeking to synthesise commonalities between groups of video segments. This resulted in the creation of four final themes that described the segments of interruptible Tetris gameplay.

The first theme, No Spin, described video segments which included or preceded several spaces of vertical movement (i.e. the Tetris piece falling) of a single Tetris piece without any rotation. Tetris segments categorised often included lateral movement, though some segments in this theme only contained or preceded vertical movement. An example of the first and last frame from a segment categorised as No Spin is visualised in Figure 5.1.

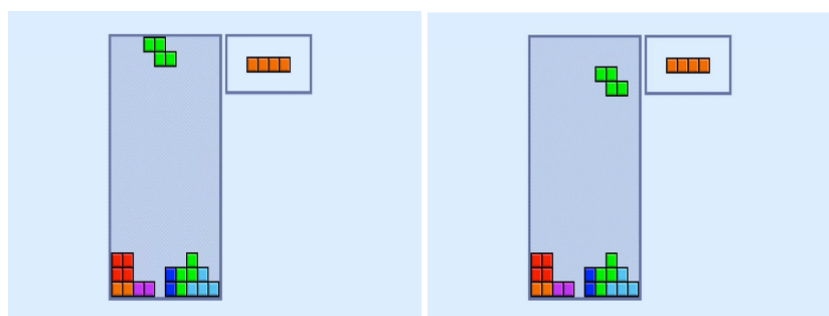


Figure 5.1: Example of the first and last frames of a sequence of Tetris exemplifying the No Spin theme.

²https://osf.io/jchge/?view_only=f44e8041b0724b359ee7548bd99d6ad0

Characterising Appropriate Moments for Interruptions of a Complex Continuous Task

A second theme, One Spin, described similar Tetris gameplay, but these segments included a single rotation across a longer sequence which sometimes spanned the movement of multiple Tetris pieces. An example of the three frames from a segment categorised as Single Spin is visualised in Figure 5.2.

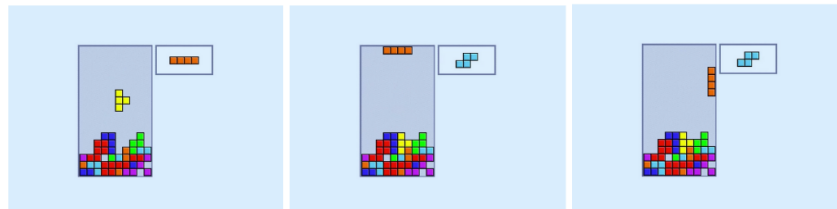


Figure 5.2: Example of three frames of a sequence of Tetris exemplifying the One Spin theme.

Another theme, Line Clear, included segments in which a full line or row of Tetris pieces was completed and the animation of that line getting removed from the Tetris board was visible in its entirety. This clearing animation lasts for approximately ten frames followed by several frames of other rows falling down to take the cleared row's place. The falling part of the animation takes a variable amount of time depending on the number of rows which were cleared and the game difficulty condition, with pieces falling faster in hard games than in easy games. An example of the three frames from a segment categorised as Line Clear is visualised in Figure 5.3.

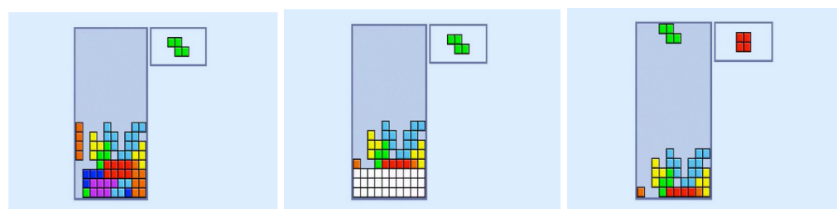


Figure 5.3: Example of three frames of a sequence of Tetris exemplifying the Line Clear theme.

The final theme, Calm Episode, included segments in which a Tetris piece travelled from its initial position to its final position with no more than two rotations or horizontal movements. Segments in this theme were relatively long as they featured entire episodes - the full journey of a given Tetris piece. It should be noted that that segments in this type included the entire episode rather than just the beginning of such an episode, indicating that, even with the benefit of hindsight, participants did not narrow their choice of a moment to initiate an interruption to the beginning of

such a sequence, thus assuring that the interrupting utterance would maximally overlap with that episode. Instead, these segments indicate that the entire episode was seen as suitable for initiating an interruption. The first and last frames from a segment categorised as Calm Episode is visualised in Figure 5.4.

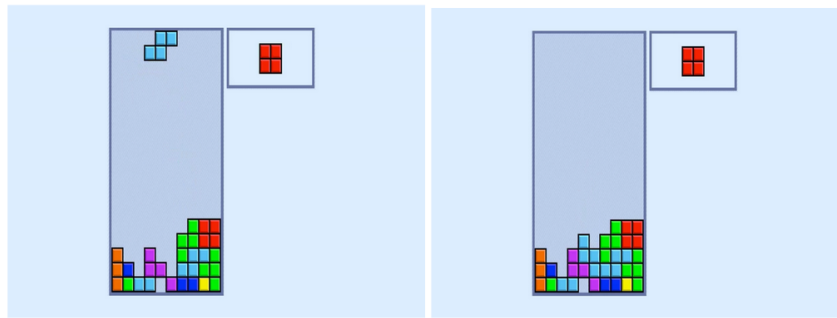


Figure 5.4: Example of the first and last frames of a sequence of Tetris exemplifying the Line Clear theme.

Taken together, these themes help to answer RQ1, highlighting features of Tetris gameplay that observers judge to impact the interruptibility of a Tetris game. The characteristics represented by these themes may be particularly sensitive to the types of Tetris games that participants observed, the length of the Tetris clips, and the characteristics of the participants in this study. That said, by selecting those aspects to match the methods and participants from the studies in Chapters 3 and 4, this framework may be used to address specific questions about the use of interruptible windows within those studies.

5.3.5 Usage of interruptible Tetris windows in previous studies

With a better understanding of the features of a Tetris game which a non-player viewer might view as interruptible, this study aimed to analyse the Tetris video clips from previous chapters to determine whether participants initiated their interruptions around the same features of Tetris games. The 24 Tetris videos from Chapter 3 and 5 (i.e. the 16 easy videos used in both studies and the additional 8 hard videos used in Chapter 4) were analysed, with timestamps labelled for each video which marked the beginning and end of interruptible windows matching themes listed above. Because participants in Chapters 3 and 4 received their interruption prompts no sooner than 5000ms after the start of a Tetris clip, the beginning of analysis for each video was

identified as the the Tetris episode which was in progress at that point in the video. The end of analysis for a given video was identified as the episode end of the which was in progress for the latest interruption for that video from any participant in a prior study. Because interruptions ended trials, the composition of videos after the latest interruption for each video was not analysed, as none of this material was seen by any participant.

Labelled segments of videos averaged 10906ms in duration, with labelled segments of hard videos averaging 11279ms and easy videos averaging 10378ms owing to longer maximums for interruption onset times in hard videos. Of these segments, an average of 58% (6295ms) was labelled as interruptible, with 54.4% (5628ms) of easy segments labelled as interruptible and 65.3% (6806ms) of hard segments labelled as interruptible. This discrepancy was somewhat surprising and is discussed below. Preliminary analysis showed that interruptions from Chapter 4 were no more likely than chance to begin inside an interruptible window given their difficulty condition (57.2% for easy videos [$\chi^2(1) = 2.31, p = .13$], 66.6% [$\chi^2(1) = 0.51, p = .51$] for hard videos). Tetris difficulty was therefore not included as a term in further analysis of interruptible window usage, as the differing baseline rates of the presence of interruptible windows between difficulty conditions would create an illusory difference between usage likelihood in a linear model. RQ2, to what extent did participants initiate interruptions within interruptible windows, is therefore not resolved by this analysis.

Interruption data from each of Chapter 3 and Chapter 4 were used to further investigate the effect of urgency on interruptible window usage to address RQ3. Interruption trials from Chapters 3 and 4 were given binary labels indicating whether they were initiated during an interruptible window (labelled as 1) or not (labelled as 0). Logit mixed effects models were fit for each experiment. Following best practices, model selection began with the maximal fixed and random effect structure (i.e. fixed effect of urgency, and random slopes and intercepts at the subject- and item-level for urgency, with item-level effects of Tetris video, interruption prompt, and trial number) with complexity incrementally reduced until the model could converge (Barr et al., 2013). For Chapter 3, there was no statistically significant effect of urgency [Log-odds = -0.12, SE = 0.16, 95% CI [-0.44,0.19], $z = -0.77, p = .44$] on the likelihood of initiating interruptions during interruptible windows. Full model syntax and output are included in Table 5.2. For Chapter 4, there was no statistically significant effect of

Table 5.2: Summary of fixed and random effects for Chapter 3 interruptible window usage - Logit mixed effects model

Model: $Window\ used = urgency + (1|subjectID) + (1|video)$

Predictor	Log-odds	SE	z	p
Intercept	.19	.21	.091	.365
Urgency (Low)	-.12	.16	-.77	.441
Random Effects				
Group		SD		
Participant (intercept)		.19		
Video (intercept)		.68		

Table 5.3: Summary of fixed and random effects for Chapter 4 interruptible window usage - Logit mixed effects model

Model: $Window\ used = urgency + (1 + urgency|subjectID) + (1|video)$

Predictor	Log-odds	SE	z	p
Intercept	.67	.20	3.35	<.001***
Urgency (Low)	-.21	.12	-1.70	.090
Random Effects				
Group		SD	Corr	
Participant (intercept)		.22		
Participant (slope)		.30	.67	
Video (intercept)		.72		

urgency [Log-odds = -0.21, SE = 0.12, 95% CI [-0.45,0.03], z = -1.70, p = .09] on the likelihood of initiating interruptions during interruptible windows. Full model syntax and output are included in Table 5.3.

5.4 Discussion

5.4.1 Characteristics of Tetris interruptibility

Taken together, the themes defining interruption cluster segments reveal some commonalities about the features people identified as relevant to interrupting Tetris games. The first two themes, No Spin and One Spin, demonstrate that rotation of a Tetris piece is a particularly salient feature in indicating that a moment is not appropriate for interruption. Horizontal movement, even across several positions, was by contrast not particularly deterrent to interruptions. This may indicate that participants saw these

spins as moments in which the Tetris player was still making a decision for a given piece, as opposed to horizontal movement, which merely represented the player carrying out their plan. This reasoning follows from qualitative data from the previous chapters, in which participants described trying to time their interruption around low cognitive load moments or moments after the Tetris player had made their decision about where to place a piece. In this way, players used rotation, and lack thereof, as an indication of the player's status in the subtasks of deciding where to place each Tetris piece. This recognition of the difficult subtask of rotation builds upon previous literature which has demonstrated the difficulty of mental rotation via comparing inexperienced participants with trained participants in planning Tetris rotations (Terlecki et al., 2008). Segments of gameplay after rotation but before a Tetris piece reached its final destination therefore represent a sort of breakpoint (Janssen et al., 2010) between the subtasks in this construction of the task of playing Tetris. While prior work on Tetris as a cognitive task classified optimal Tetris play as play which minimises rotations (Lindstedt & Gray, 2019), it should be noted that this prior research did not differentiate between the difficulty of rotations and horizontal movement, instead merely noting that both are minimised by expert players. Future work which explores Tetris as a cognitive task may seek to differentiate between these two subtasks.

The Line Clear theme, conversely, presents an alternative conceptualisation of Tetris as a task. Rather than structuring an interruption around the implicit goals of the player, interrupting when one or more lines is cleared represents explicit game-state moments, the placement of each Tetris piece, as the subtasks of Tetris. This conceptualisation allows the interrupter to target an explicit boundary between subtasks, as it is unambiguous when a piece has been placed, whereas the moment in which a player makes their decision can only be guessed by the interrupter. Line clear events are particularly important to this conceptualisation of Tetris, as in a typical episode, no line is cleared when a piece reaches its final destination, and the next episode starts immediately afterwards. In these cases, there is only a momentary breakpoint between the subtasks, not enough time to effectively utter an interrupting question before the next subtask is in progress. Line clearances however add a meaningful break between episodes, as the animation occurs, allowing interrupters to begin their utterance when no subtask is active. These events are rare and may indeed be the default selection of interrupters when they occur. Strategies of infer-

ring the goals of the Tetris player may exist as a fallback to the more straightforward explicit strategy. This fallback strategy can be effective given the comparative frequency of segments with minimal rotations, as these events occur several times in every video clip, whereas line clear events are more rare, occurring occasionally if at all over a given clip. Prior work on natural breakpoints has identified different sets of cues for breakpoints that may contrast, such as cognitive cues (e.g. interrupting the task of typing a phone number in the regionally-typically middle where a space or dash is placed) and motor cues (e.g. interrupting the same task after a sequence of repeated digits) that can signal contrasting potential breakpoints to an actor (Janssen et al., 2012). Line clears in this experiment may represent unambiguous breakpoints to interrupters, where both the motor cues and cognitive cues are aligned, as there is no input from the player and an episode has ended without the immediate start of the next episode.

The Calm Episode theme does not represent using a breakpoint between subtasks as a moment to initiate interruptions per se. Instead, interruption initiations in this theme use entire episodes, whole subtasks, as targets for interrupting. In this way, interruptions in this theme did not seek to avoid overlapping their interruptions with a subtask, they instead sought to interrupt a subtask which required little effort, or which had little risk of failure. This third potential strategy guarantees a fallback from even the implicit goals strategy, as an interrupter need not identify moments of decision making, only low-risk subtasks. This may indicate that, in the absence of line clear events, a strategy of picking entire episodes as interruption targets requires either less Tetris expertise or less mental effort from the interrupter, as they need not infer the players' cognitive processes. Instead, this interruption strategy allows an interrupter to narrow down their choices at an episode level, choosing the least risky or least active episode to be sacrificed via an interruption in order to preserve the player's attention for all of the more active episodes. Again, returning to previous work on Tetris, optimal play is seen as that which minimises both rotations and horizontal movement (Lindstedt & Gray, 2019). Calm episodes, those for which the player minimises their actions with the fewest total inputs, may likewise be appropriate moments to interrupt from the player's perspective as well, as they represent subtasks for which this minimisation of input is achieved quickly.

Overall, it is notable that interruption clusters targeted breakpoints or low-load

moments for Tetris players, rather than targeting the moments just before these interruptible windows. This tendency ensured that the interrupting utterance would begin during that window, but leads to a risk of interruptions continuing to demand attention into the beginning of the next subtask. As an alternate strategy, interrupters could have begun their utterance before these events, increasing the chance for the Tetris player to use that window to react to the utterance by answering the question in that window. Even with the benefit of hindsight however, doing this sort of event planning would be quite difficult, as the interrupter would need to estimate the time it would take them to make their utterance and the time it would take for the Tetris player to respond. Likewise, it may have been seen as less costly to allow utterances to interfere with the start of the next subtask rather than with the end of the preceding subtask, as the most cognitively demanding moment for the Tetris player would be seen as coming just as a decision is being made.

5.4.2 Harder Tetris games do not cause more breakpoint usage

It was unexpected to find that hard Tetris games were constituted by a significantly higher proportion of interruptible windows than easy Tetris games. It is possible that the Tetris video clips selected for the hard condition present survivorship bias. That is to say, Tetris games which were consistent with the hard condition and simultaneously long enough in duration to be used in Chapter 4 might be particularly rich with interruptible moments which explain their duration in the first place. Additionally, differences in gameplay styles and strategy between the Tetris player who recorded the easy gameplay videos and the Tetris player who recorded the hard gameplay videos might account for this difference. This could be the case if the player from the hard videos was particularly skilful in manufacturing these windows or in playing these parts more slowly while playing high-load sequences of Tetris more quickly. Finally, it may be the case that harder Tetris games are marked by larger proportions of time in which the player is interruptible in general.

Rather than game difficulty corresponding to the amount of time spent under high cognitive load, game difficulty might instead correspond to the intensity of that cognitive load or to the chance of failure during moments of high load. These different possible interpretations of Tetris gameplay difficulty echo the ongoing debate in the broader field of Cognitive Load Theory and the nature of sources of cognitive load (De

Jong, 2010; DeLeeuw & Mayer, 2008). While the prevailing model of cognitive load includes three sources of load - extrinsic load, intrinsic load, and germane load - the nature of the difference between intrinsic and germane load remains contested (De Jong, 2010). The traditional conception of this tripartite model differentiates intrinsic load as a demand on cognitive resources resulting from the structure and interactivity of the task at hand and germane load as the higher-order demand caused by learning about the task and developing schema around it (Sweller et al., 1998), though some research has deemphasised this difference and has instead indicated that these may be the same types of cognitive load (Kalyuga, 2011). In the context of Tetris, it may be the case that while hard games feature a higher degree of intrinsic load which is evident to observers via the increased amount and pace of activity on the screen or alternatively that harder games feature more germane load in that they do not differ in the amount of time spent under intrinsic load (e.g. time in which a player must interact), but downtime is more demanding in terms of germane load as players must plan and consider the future more deliberately. Differences in conceptions of cognitive load like this across tasks, actors, and research frameworks have led to emphasis in Cognitive Load Theory research on using multiple research methods including physiological methods, self-report, and object task performance measures to triangulate cognitive load in complex tasks (Dehue & van de Leemput, 2014). Further experiments around Tetris gameplay may help to resolve how difficulty is conceived of by Tetris players and observers, and future research on interruptions of continuous tasks must be sensitive in considering the multifaceted nature of task difficulty, reflecting on the different ways that it might be measured.

The fact that participants did not use interruptible windows of Tetris gameplay proportionally more during hard games than during easy games was surprising as well. Again, a number of possible factors might explain this observation. Due to the nature of delivering a spoken interruption during a continuous task, participants may have found it too difficult to use interruptible windows to structure interruptions, knowing that a window might close before their utterance ended. It may likewise be the case that participants were less aware of or less confident about which moments would be interruptible at all without the benefit of hindsight afforded to participants of the present study, minimising their reliance on cues from the Tetris gameplay. This explanation does not account for differences in usage of interruptible windows during

higher urgency trials however, so further studies should seek to replicate that null finding in order to calibrate explanations for null differences between difficulty conditions.

The lack of impact of game difficulty on usage of interruptible windows may also be explained in terms of egocentric goal prioritisation. Participants may have largely ignored the specific features of the Tetris game in practice, focusing on interrupting at a relaxed pace during non-urgent trials and at a faster pace during urgent trials, waiting until they themselves felt ready to speak rather than waiting until they felt their partner was ready to listen. Alternatively, participants may have determined that their partner was not in sufficient danger of losing the Tetris game any time they interrupted, irrespective of breakpoints, and did indeed optimise interruptions to prevent game risks by interrupting as soon as possible and thus exposing their partner to less Tetris play in the uncertain future. Insofar as these final explanations involve rational bounds to optimise behaviour, they may represent testable hypotheses that can be explored through future cognitive modelling work such as the reinforcement learning models of multitasking (Jokinen et al., 2021).

5.4.3 Cluster analysis for data-driven breakpoint identification

One of the primary contributions of this study is methodological. Defining interruptible moments for complex tasks has hitherto been primarily either theory driven (Iqbal & Bailey, 2010) or driven by data which might predict an interruption (Kim et al., 2015; Semmens et al., 2019) rather than driven by observed interruption data directly. The approach demonstrated in this study of gathering experimental data on the interruption of a pre-recorded task is novel and introduces numerous benefits to interruptions researchers. For one, prerecording a task allows interrupters to select optimal moments with the benefit of hindsight owing to a complete knowledge of the set of task states. This allows observed interruptions to represent a strategy closer to what participants see as optimal rather than one bounded by constraints of uncertainty about the future and aversion to risk. Furthermore, this approach allows data to be collected more easily at scale and without confounds of task execution, as a given task need only be performed once and can be presented identically to all participants, including through online platforms like crowdworking websites. The clustering approach allows the analyst to gain data-driven insight from the interruption behaviours without any

a priori understanding of the underlying structure of a given task. Beyond modelling Tetris specifically, this approach may be beneficial for modelling more complex and diverse tasks like driving or workplace-specific tasks which would demand domain-specific expertise of a modeller. Finally, a process of first validating the stability of clusters of observed interruptions then coding and analysing those clusters according to the event structure of the task is a straightforward process for reducing noise and preserving key trends from highly variant unidimensional experimental data like this.

This methodology of observing interruptions of a complex task, clustering observations of interruption onsets, and generating themes which describe those clusters, can be flexibly adapted to different complex tasks and interruption types (e.g. self-interruptions or interruptions of others, spoken interruptions or interruptions signalled through other modalities). As people's work and daily lives become increasingly integrated with digital technology (Finzi & Orlandini, 2005; Janssen et al., 2019), this sort of task modelling will become increasingly important in the design of safe and effective interruptions. A task-sensitive model of interruptibility, especially one derived from observable cues such as those identified by an observer rather than those known to the person engaged in a complex task, may help to model interruptibility in a non-invasive way, reducing reliance on physiological measures like EEG measurements (Züger & Fritz, 2015) in contexts where such measures are not available. This approach of using vision to model ongoing tasks as a strategy to manage proactive interactions with digital systems fits with that proposed by Cha et al. (Cha et al., 2020) by which proactive smart speakers queried user interruptibility during periods of physical movement detected through computer vision. Whereas Cha et al. used participant self-report of their activity to analyse these observations of interruptibility (Cha et al., 2020) future work may consider combining that approach with the one demonstrated here, seeking to either use atheoretical clustering or task groupings defined by either the interrupter or the interrupted party.

5.4.4 Limitations

One limitation of the present work is the fact that interruptible windows of Tetris gameplay were identified from a sample of videos of Tetris then generalised to other parts of those videos to perform additional analysis. Had the present study exclusively

Characterising Appropriate Moments for Interruptions of a Complex Continuous Task

featured the segments of Tetris videos seen by participants in studies from previous chapters, general descriptions and labelling of those segments would not have been necessary, as the frames observed by participants in this study would match those seen across participants of a given trial in previous studies. This approach would have come with some complicating downsides however. Had this study exhaustively shown all of the Tetris gameplay which was interruptible to participants in previous studies, each Tetris clip in this study would have needed to be much longer, in order to account for the full range of both delays in onsetting interruptions and in randomised prompt delays from the previous studies. This may have made picking a single moment from a collection of many hundreds of frames difficult for participants, and reduced the likelihood of obtaining meaningful clusters. Likewise, these very long clips would still not be representative of every given participant's options for interrupting in previous studies, as much of each clip would not have been seen by a given participant who had initiated their interruption and not had seen later parts of the clip. For this reason, while limiting participants to Tetris clips of standardised and manageable lengths reduced the direct comparability between the present study and previous studies, this tradeoff led to an experimental method which did not overwhelm participants and still allowed for analysis relating present findings to previous findings.

Another limitation stemming from differences between this study and studies from previous chapters was that participants in this study were asked to pick a single best moment to interrupt with the benefit of hindsight, whereas participants in previous studies had to interrupt in real time without revisiting previous moments. In this way, participants in the present study were much better equipped to identify good moments than participants in the previous studies, who may have witnessed good moments to interrupt which they soon regretted not making use of. This difference could have been mitigated by allowing participants in the present study to select multiple interruptible moments so that suboptimal but still interruptible moments, such as the moments that regretful prior participants had settled for, would be identified as well. This approach would have been troublesome however, as participants may have identified several frames or windows of time which were interruptible due to the same event. For example, a participant identifying a Line Clear by selecting the first frame of the line clear animation might decide to select all frames from the animation, or frames

just before the impending Line Clear, adding little explanatory information about what makes a Tetris moment interruptible. Fortunately, by having many participants select an ideal moment to interrupt in each clip, a variety of potentially interruptible moments were identified for the majority of clips. While it is possible that allowing participants to select multiple moments would have revealed even more interruptible windows, it is also possible that the added noise or ability of a participant to cast multiple votes for the same event would have reduced the power of this analysis to identify meaningfully differentiated interruptible moments.

A potential alternative approach to the research presented would have been to recruit participants from previous studies to participate in this research. Had participants from previous studies participated in the present study, it may have been possible to model at the participant level whether a given participant's real time interruption came at the best moment in the clip they saw or in a moment that they themselves deemed interruptible. While this approach would make for more precise analysis of whether participants found their own interruptions to have occurred at opportune moments, this analysis would be less generalisable across different people and across Tetris games. That sort of analysis would add little beyond the qualitative results from the previous studies, in which participants already had an opportunity to reflect upon their selection of moments to initiate their interruptions. Instead, by identifying moments which a variety of people judged to be interruptible, the present analysis better identifies whether the moments previous participants chose were in fact interruptible, rather than whether those interrupters thought the moment they chose was interruptible. Along these same lines, the characteristics of interruptible Tetris games identified by this analysis come from a more varied sample of Tetris games than the sample of clips seen by previous participants alone.

A final limitation regarding the participants in this study is that, like in previous studies, a general sample of participants was selected rather than a sample of expert Tetris players. While experts may have been more proficient and more confident in selecting interruptible windows of Tetris, previous literature on Tetris has identified differences between novice and expert players in the way they conceptualise the event structure in Tetris games (Lindstedt & Gray, 2019). For this reason, using Tetris experts as participants may have led to different characteristics of interruptibility than those identified here. Insofar as the previous chapter studies' participants were not

expert Tetris players, it is possible that they selected moments for interrupting that only a non-expert would identify as interruptible. That said, it must be acknowledged that the way an interrupter conceives of a game of Tetris may differ in unknown ways from the way the Tetris player themselves understands the event structure of Tetris. For this reason, while the characteristics of Tetris interruptibility from an interrupter's perspective are described here, it is not certain whether those characteristics would exactly match those identified by either a novice or an expert Tetris player. In order to maximise the extent to which the present research could be used to add insights to the data collected in previous chapters, a non-expert participant pool was necessary here. Future work may seek to replicate this study with an expert sample of participants in order to assess whether different interruptibility characteristics would indeed be observed. Likewise, future work may consider asking Tetris players to perform this exercise on their own Tetris play to investigate the extent to which players and observers agree in their characterisations of interruptibility.

5.5 Conclusion

This chapter sought to understand how people use characteristics of a complex task which they are observing but not engaging with in order to time spoken interruptions of another person who is engaged with that task. In order to gain an understanding of this phenomenon, an online experiment was conducted using Tetris as the complex task, by which participants could choose precise moments of Tetris games which they found to be most interruptible. K-means cluster analysis was conducted to group these moments into a set of windows of Tetris gameplay which represented vignettes of Tetris gameplay which participants reliably identified as interruptible. A content analysis of these vignettes enabled further thematic analysis which identified four themes of interruptible Tetris gameplay: No Spin, One Spin, Line Clear, and Calm Episode. These themes were used to classify interruptions from studies described in previous chapters to generate further insights about the usage of interruptible gameplay windows and the effect of urgency and Tetris game difficulty on that usage. Quantitative analysis revealed that participants in previous studies were no more likely than chance to initiate their interruptions during interruptible windows of Tetris for either easy or hard Tetris games. The effect of urgency on the use of in-

interruptible windows was likewise not statistically significant in each of the two prior experiments.

This chapter aligns with previous chapters which indicate the variety of factors by which people time and structure their spoken interruptions. While some clear patterns could be observed in terms of what makes a Tetris game appear to be interruptible to an observer, it is not clear that the same characteristics of task interruptibility influence how interrupters decide when to interrupt in real time interruption scenarios. This seems in conflict with the emphasis participants in previous chapters have stated they put on choosing good moments to interrupt through their self-reports of interruption strategies. Like other features identified in self-report data in Chapters 3 and 4, the use of Tetris task characteristics may likewise be a heterogeneous cue which is important for some people but unimportant or non-actionable for others. Building from this, the next chapter seeks to investigate whether people find interruptions designed to use these heterogeneous cues and strategies by a voice agent as preferable to interruptions which are not designed to adapt to any particular cues and which use a static strategy for all interruptions.

6 * **Comparing Perceptions of Static and Adaptive Proactive Speech Agents**

6.1 Introduction

As speech agents have become increasingly popular, users have highlighted multitasking during eyes-busy, hands-busy activities as a central motivation to trying these agents out (Luger & Sellen, 2016). That said, users' initial excitement for speech agents is frequently diminished to the point of disappointment and even abandonment, owing to speech agent interactions falling short of their expectations in terms of their abilities as dialogue partners (Cowan et al., 2017; Luger & Sellen, 2016). These expectations and internal models of speech agents as dialogue partners, termed *partner models* (Branigan et al., 2011a; Cowan & Branigan, 2017) play a key role in how speech agent users understand their interactions, so bringing speech agents' behaviour more into line with the expectations users have for them as nearly human-like dialogue partners (Cassell, 2007; Cowan et al., 2017) may help users harness the benefits of speech agent interactions that they seek. But in order for speech agents to meet these expectations and facilitate multitasking to a level comparable to the human personal assistants to whom research participants have unfavourably compared speech agents (Luger & Sellen, 2016), they will need to be able to interact proactively with users rather than waiting for the busy user to turn their attention to a speech interaction.

Proactive and mixed-initiative agent based interactions have long been seen as potentially beneficial to users by human-computer interaction researchers. Interactions with agents have been identified as possessing a variety of benefits, including awareness of the user's context, a capacity for autonomous interaction without requiring user initiation or input, and adaptivity to how the user interacts (Shneiderman & Maes, 1997). Some early work on agent based interaction sought to describe design principles for mixed-initiative agent-based interactions, sensitive to the principles which had guided the design of direct-manipulation user interfaces before them. Horvitz laid out 12 principles for mixed-initiative interfaces with this aim, including among others: considering uncertainty about a user's goals, considering the status of a user's attention in the timing of services, inferring ideal action in light of costs, benefits, and uncertainties, minimising the cost of poor guesses about action and timing, and employing socially appropriate behaviours for agent-user interaction (Horvitz, 1999).

Chapters 3 and 4 investigated the characteristics of human spoken interruptions, and found that people take many of the same considerations that are mentioned in proactive agent design guidelines into account when interrupting another person - seeking to limit the distraction caused by their interruptions by limiting the duration of their speech, attempting to select good moments for their interruptions, and sometimes preceding their interruptions with access rituals to make them more socially appropriate. By combining the well-established design principles for proactive agent interactions and the descriptions of proactive human speech interactions from previous chapters, the present chapter aims to investigate the effect that designing a proactive speech agent to be proactive will have on people's partner models of that agent as compared to their partner models of proactive speech agents which, like existing speech agents, do not adapt speech behaviours to a user's context.

6.1.1 Designing proactive agents

Following from those early design principles, (Horvitz, 1999) recent work on speech based proactive agents has been concerned with their design looked to propose and test principles for the design of both the types of tasks that an agent proactively performs as well as the specific implementation of those actions, with regards to details such as modality, timing, message content (Cha et al., 2020; Semmens et al., 2019;

Yorke-Smith et al., 2012). The present study considers proactive agents which use speech to interrupt a user who is already engaged in a task. As such, it is necessary to further consider the design details of those specific types of proactive interactions. One study on the design of a learning assistant with these characteristics proposed nine principles for proactive agent behaviour, specifying that it should be valuable, pertinent, competent, unobtrusive, transparent, controllable, deferent, anticipatory, and safe (Yorke-Smith et al., 2012). Echoing the general proactive agent design principles laid out by Horvitz (Horvitz, 1999), this set of principles again focuses on adapting interactions based on contextual information, including contexts of the agent's task, the user's environment, and the social context of a non-human agent initiating interaction with a person.

Some proactive agent research has focused on better understanding the environment of a user and its suitability for proactive interaction. Typically, this type of research entails modelling the user's ongoing task context such as by placing sensors and cameras in a car (Kim et al., 2015; Semmens et al., 2019) or by using telemetry to monitor the status of computer-based work tasks (Iqbal & Bailey, 2010). These studies focused mostly on the timing of proactive interaction, seeking to understand the task structure of complex tasks like driving or unstructured computer-based knowledge work, so notifications or requests to a user can be scheduled for these moments. Some of this work has likewise focused on the social context of the user, modelling people's everyday context around the home through camera's and user logging of activities (Cha et al., 2020). That study found that the social context of a proactive agent's interruption was an important factor, as participants thought interrupting a conversation with another person to talk to a machine was not acceptable, but engaging in a conversation with an agent as entertainment could be appropriate even in the presence of others.

While this recent proactive speech agent design research has focused on identifying good moments to schedule interruptions, or else has sought to demonstrate the value of well-timed interruptions in terms of making returning to a prior task easier (Iqbal & Bailey, 2010), little attention has been paid to the broader ways in which a proactive interaction can be made responsive to a user's task context or social context in characteristics other than timing. Chapters 3 and 4 demonstrated the extent to which people consider the urgency of their interruption and the state of a dia-

logue partner's ongoing task in how they time their speech, but also in the words they choose, the tone they try to convey, and the extent to which they mark their speech with access rituals. Some recent work has begun to investigate tone specifically, comparing an assertive-voiced in-car notification to a non-assertive voice (Wong et al., 2019). The present study seeks to build on this and other research into proactive agent design by comparing user impressions of a proactive speech agent which adapts its speech according to urgency and the state of an ongoing task across a variety of characteristics to impressions of a proactive agent which ignores context and interacts in a static way.

6.1.2 Partner modelling of machine dialogue partners

Speech agent interactions are a unique form of human-computer interaction as they require users to engage in dialogue with a machine dialogue partner, making the conversational abilities of that partner central to the interaction (Branigan et al., 2011a). Prior research on spoken interactions, both those with people and with machines, have established the concept of partner models, the models by which people understand the capabilities of their dialogue partners (Branigan et al., 2011a; Cowan et al., 2015). Doyle and colleagues formally define partner models for machine dialogue partners as follows:

The term partner model refers to an interlocutor's cognitive representation of beliefs about their dialogue partner's communicative ability. These perceptions are multidimensional and include judgements about cognitive, empathetic and/or functional capabilities of a dialogue partner. Initially informed by previous experience, assumptions and stereotypes, partner models are dynamically updated based on a dialogue partner's behaviour and/or events during dialogue

(P. R. Doyle et al., 2021)

Recent user studies of speech agent interactions have helped to establish that these partner models play a pivotal role in speech agent users' overall experience of these interactions, with users finding interactions particularly unsatisfying when their experience does not match their partner model (L. Clark, Munteanu, et al., 2019; Luger & Sellen, 2016). In a qualitative study of users of popular speech agents like Siri and Google Assistant, users remarked on the extent to which the promise of human-

likeness, insinuated by marketing, human-like voice synthesis, and designed personalities which mimic a human personality, creating what the researchers called the “gulf of expectations” (Luger & Sellen, 2016) following the more general “gulfs of execution and evaluation” across design disciplines described by Norman (Norman, 1983). Reflecting this research, it is critical for user experience that speech agents which prime human-like partner models to meet this expectation and deliver human-like capabilities.

Until recently, little research has explored the characteristics of partner models in speech agent interactions. Recent work has begun to investigate this question however, investigating the dimensions of partner models which are salient to people engaged in dialogues with machines and with people (P. R. Doyle, 2022; P. R. Doyle et al., 2021; P. R. Doyle et al., 2019). In order to investigate these dimensions of partner models, Doyle and colleagues invited participants to engage in conversations with each of Siri, Amazon Alexa, and a human researcher and then reflect on how they would describe and differentiate between these interactions (P. R. Doyle et al., 2021; P. R. Doyle et al., 2019). The descriptive terms elicited through this study were then used to generate semantic differentials, which speech agent users then used to categorise the speech agents that they had experienced interacting with (P. R. Doyle et al., 2021). This study resulted in the generation of three themes describing the dimensions of partner models for speech agents: perceptions of partner competence and dependability, assessment of human-likeness, and perceptions of the cognitive flexibility of the system (P. R. Doyle et al., 2021). These semantic differentials and themes were further developed into a validated self-report questionnaire across those factors, the Partner Modelling Questionnaire (PMQ), which can be used to measure the strength of people’s partner models for machine dialogue partners (P. R. Doyle, 2022). The PMQ therefore represents a new validated measure for assessing partner models, which have been previously identified as a crucial factor in users’ perceptions of speech agents. This study seeks to apply principles of proactive agent design to our current understanding of partner models in spoken interactions with machines. Specifically, by designing a proactive speech agent which adapts to a user’s context, this study aims to demonstrate a speech agent which is more competent, human-like, and cognitively flexible than existing speech agents which are not adaptive to context.

6.1.3 Aims and hypotheses

This thesis aims to understand the characteristics of human spoken interruptions in order to inform the design of proactive speech agents which may interrupt busy people. Previous chapters have focused on investigating the characteristics of human spoken interruptions, highlighting the ways in which interruptions differ according to the urgency of the interruption and the complexity of task they interrupt. Holistically, these studies found that people adapt their interruptions in terms of timing, word choice, prosody, and the use of particular social markers (i.e. access rituals), taking urgency and task difficulty cues into account.

This study aims to apply these findings to proactive non-human speech agents. Following prior research on the design of proactive agents (Horvitz, 1999) and on partner modelling (P. R. Doyle, 2022) as well as results from Chapters 3-5, the present study hypothesises the following:

- People will rate speech interruptions from an adaptive agent as coming at better moments as compared interruptions from a static (non-adaptive) agent (H1)
- People will rate speech interruptions from an adaptive agent as more appropriately asked as compared interruptions from a static (non-adaptive) agent. (H2),
- People's partner models for an adaptive agent will view it as a more capable dialogue partner than their partner models for a static (non-adaptive) agent (H3).

Additionally, in order to gain an understanding of people's perceptions of the specific adaptive agent used in this study, qualitative analysis explores the following research question:

- What aspects of proactive interactions with an adaptive agent will be most salient to participants as differences from proactive interactions with a static (non-adaptive) agent (RQ1)?

All hypotheses, research questions, and post-hoc analyses were pre-registered before data collection began.¹

Building on the research conducted in chapters 3-5, the study conducted uses Tetris as an ongoing, complex task which a proactive agent must interrupt with speech.

¹osf.io/g8zk6/?view_only=2a1faf7837f443348baf88cd585fc08a

Chapters 3 and 4 both included analysis of participants' self-reported interruption strategies in which they indicated that they tried to choose good moments to interrupt as much as possible by targeting either low cognitive load moments for their Tetris playing partner or by seeking to interrupt during specific moments in the Tetris game, such as when a Tetris piece was at the bottom of the board. Mixed-methods analysis in Chapter 5 described the characteristics of interruptible windows of Tetris gameplay based on participant ratings. Likewise, Chapters 3 and 4 demonstrated through quantitative analysis that participants were quicker to interrupt when interruptions were urgent and more likely to use access rituals (see Section 2.3 of this chapter) during difficult Tetris games, quantifying each of these effects. Insofar as these chapters provided a specific description of the ways people adapt their speech to contextual information in Tetris, the same task is used here so that those adaptations can be directly applied to the design of the proactive agent.

Rather than having participants act as Tetris players in this experiment, participants instead watched videos of interactions between the prototype agents and an unseen Tetris player. This video study technique is commonly used in human-robot interaction research (Ghafurian et al., 2020; Lohse et al., 2008) owing to its benefits in being rapidly deployable to many participants including online participants, greater standardised control over the interaction, and facilitation of the use of early-stage prototypes which may lack features necessary for live interactions (Woods et al., 2006). These aims likewise fit with a general framework for using prototypes in HCI research, by which artefacts which are of research interest but which are not yet available (such as a proactive agent that interrupts live Tetris gameplay at precise moments) can be simulated with mock-up props, whereas features of an interaction which are not germane to research questions (such as participants' Tetris gameplay ability or internet latency) are minimised as to reduce confounding variance (Salovaara et al., 2017).

6.2 Methods

6.2.1 Participants

80 crowdworkers (40 men, 40 women; M age = 38.4 years, SD = 11.9 years) were recruited on Prolific Academic. All participants were native speakers of English living in Ireland or the United Kingdom. 92.5% (N=74) of participants reported having used

speech assistants before, with 66.3% (N=53) of participants reporting that they use a speech assistant once a week or more frequently. Participants were all familiar with Tetris, though most reported that they do not play frequently (81.3% of participants answering 3 or lower on a 7-point Likert-type question asking “If you have played Tetris before, how often do you play Tetris?”), but only 3 participants reported that they had never played Tetris before. Most participants rated their level of expertise with Tetris as moderate (62.5% of participants answering 3, 4, or 5 on a 7-point Likert-type question asking “If you have played Tetris before, how would you rate your level of Tetris expertise?”). The study took approximately 20 minutes and participants were compensated £6 through Prolific Academic for their participation. The study received ethical approval through the university’s ethics procedures for low risk projects (Ethics code: HS-E-22-23-Edwards-Cowan). All materials including participant information sheets, consent forms, questionnaires, and debrief sheets are provided in Appendix D.

6.2.2 Materials

Tetris videos

Twenty-four videos of the game Tetris were created. In each video, a game of Tetris is played by an unseen player. Under the Tetris video, the word Urgent or Non-urgent appeared, indicating whether the video represented an urgent interruption or a non-urgent interruption, following the design of studies in Chapters 3 and 4. A question was written at the bottom of each video, indicating the question that the proactive speech agent would be prompted to ask the Tetris player. After a fixed interval of 10 seconds, a large red dot indicator appeared in the video indicating that a proactive agent had been prompted to interrupt the player. The ten second delay was selected to give participants time to observe the Tetris game before an interruption might occur and reflects the maximum delay used in Chapters 3 and 4 before prompting interruptions. After some delay (described below), a synthesised voice is heard asking the question to the Tetris player. Videos end one second after the audio ends, with each video lasting approximately 20 seconds. Tetris gameplay and interruption prompts were sampled from Chapter 4, with 12 unique videos and prompts being used, resulting in 12 matched trials across two within-subjects conditions.

Half of the videos selected (n = 6) were randomly sampled from the difficult Tetris

games in Chapter 4, in which videos started with a Tetris game piece at the top of the game board, at least half of the rows of the board which already contained Tetris pieces, and the falling speed of the game piece was set to 10 rows per second. The other half of the videos selected ($n = 6$) were randomly sampled from the easy Tetris games in Chapter 4, in which videos started with a Tetris game piece at the top of the game board, at least two rows and no more than half of the rows of the board already contained Tetris pieces, and the falling speed of the game piece was set to the game minimum of 1.25 rows per second. For each difficulty grouping per condition, half of the videos ($n = 3$) are arbitrarily marked “urgent” and the other half ($n = 3$) are arbitrarily marked “non-urgent”. Tetris gameplay, interruption prompts, and urgency are all fixed throughout each block of videos and across participants - for example, Tetris gameplay video 2, which came from the easy condition from Chapter 4, was urgent and had the prompt “What was the last movie you watched?” in both blocks for all participants. In keeping with the other chapters, prompts and urgency are arbitrarily associated and the content of each prompt is unrelated to the Tetris task.

Interruption audio

All interruption audio was synthesised using a free demo of Google WaveNet text to speech. Half of the participants heard voice en-GB-Wavenet-A, a feminine voice, and the other half of participants heard voice en-GB-Wavenet-B, a masculine voice, fully balanced by participant gender. Both voices have a Standard Southern British English accent. In the static agent condition, all synthesised speech was produced at 1.00 speed and 0.00 pitch. In the adaptive condition, synthesised speech was produced at either 1.00 speed and 0.00 pitch or 1.10 speed and 0.00 pitch (further details in Experimental Conditions section below).

6.2.3 Experimental conditions

The experiment followed a one-way within-subjects design. Agent condition was manipulated across two conditions: adaptive and static. Agent condition was manipulated by varying the design of spoken interruptions in a variety of ways:

Use of interruptible windows

For the static agent, interruptions always began 4 seconds after the red dot appeared, or as close as possible while ensuring the interruption did not begin within an interruptible window as identified in Chapter 5. The adaptive agent varied its interruption onset and always began interruptions within an interruptible window as identified in Chapter 5. For urgent interruptions, the adaptive agent interrupted at an onset three seconds after the red dot appeared, or as close as possible to three seconds while interrupting within an interruptible window. For non-urgent interruptions, the adaptive agent interrupted at an onset five seconds after the red dot appeared, or as close as possible to five seconds while interrupting within an interruptible window. The differences in interruption onsets was selected to reflect the difference in mean onsets observed in Chapter 4, in which urgent interruptions came after a mean onset of 3.87 seconds whereas non-urgent interruptions came at a mean onset of 4.72 seconds. This difference was slightly exaggerated in the conditions presented in this experiment with the intention of making differences more salient to an observer. The use of interruptible windows only by the adaptive agent in both urgent and non urgent as well as easy and difficult trials reflected Chapter 5 findings in which participants did not significantly vary their usage of these windows by urgency or by Tetris difficulty conditions. It was also informed by Chapter 3 and 4 qualitative findings that participants did generally seek to interrupt at low-cost moments.

Speech rate

The static agent asked questions exactly as they appeared on screen, with no changes to wording, at the standard 1.00 WaveNet speech rate, and without the use of any access rituals. The adaptive agent spoke at a 1.00 speech rate for non-urgent interruptions and a 1.10 speech rate for urgent interruptions, reflecting the difference in mean interruption durations observed in Chapter 4, in which urgent interruptions lasted for a mean of 1519ms whereas non-urgent interruptions lasted for a mean of 1596ms.

Use of access rituals

The adaptive agent used access rituals such as “hey” and “excuse me” for all six of the trials which featured easy Tetris games and did not use access rituals for any of

the six trials which featured hard Tetris games, reflecting the difference observed in Chapter 4, in which participants were significantly more likely to use access rituals during easy Tetris games as opposed to difficult games.

Interruption phrasing

Finally, the adaptive agent rephrased all of its questions, using concise language (e.g. “Got any pets?” for the prompt “Do you have any cats or dogs?”) or conversational language (e.g. “Are you right or left handed?” for the prompt “Which hand do you write with?”) for all trials, with these styles balanced across Tetris game difficulty and interruption urgency. Each rephrased question was a verbatim recreation of the way a participant phrased the corresponding interruption from Chapter 4. Concise language and conversational language were selected to reflect the two major rephrasing strategies mentioned in qualitative data in Chapters 3 and 4, neither of which was exclusively associated with a single urgency condition in either experiment.

Overall, the static agent condition is meant to be representative of the capabilities of current speech agents like Google Assistant or Amazon Alexa, which do not use access rituals, vary speech rates, or vary the timing of their speech based on contextual cues. The adaptive agent was designed to adapt its speech in a variety of ways representative of the ways people were observed to adapt their speech in previous chapters. While this experimental design does not allow for the analysis of any particular type of adaptation’s causal relationship with interaction outcomes, it nonetheless gives a holistic representation of the overall effect of adaptation, directly informed by the approaches to adapting speech for interruptions demonstrated in the prior chapters.

6.2.4 Measures

Partner Model Questionnaire

Participants were asked to complete the 18 item Partner Model Questionnaire (PMQ). The PMQ is a validated self-report scale consisting of word pairs separated by a 7-point semantic differential scale. The scale comprises three subscales: *partner competence and dependability*, *human-likeness*, and *cognitive flexibility* onto which nine, six, and three items load respectively (P. R. Doyle, 2022). Scores are calculated for

Comparing Perceptions of Static and Adaptive Proactive Speech Agents

each subscale by summing semantic differential ratings for each word pair that loads onto the respective scale, with higher numbers corresponding to responses closer to the word more positively associated with that subscale (e.g. closer to the word "consistent" in the pair "consistent/inconsistent" which loads onto the *partner competence and dependability* subscale). Total PMQ scores are calculated by summing the three component subscale scores.

Participants were asked to complete the PMQ with the instructions "Thinking about the speech assistant you just watched, how would you rate its communicative ability on a scale between each of the following poles?". As a control, before the experiment began, participants were also asked to complete the PMQ with the instructions "Please complete the following questionnaire based on your previous interactions with speech interfaces. Speech interface may include a broad range of technologies such as Amazon's Alexa, Apple's Siri, Google Assistant and Microsoft's Cortana, along with speech-based chatbots and telephony systems (i.e. like those used in telephone banking and ticket booking). You may have accessed these using a smartphone, smart speaker, laptop or desktop and/or in-car. Thinking about the speech interface you interact with most frequently, how would you rate its communicative ability on a scale between each of the following poles?". PMQ semantic differential item orders were randomised between participants, and 9 items were reversed (e.g. lower-scoring poles appeared on the left of the screen rather than the right) per participant, with 5 randomly selected items reversed for the *partner competence and dependability* subscale, 3 randomly selected items reversed for the *human-likeness* subscale, and one randomly selected item reversed for the *cognitive flexibility* subscale.

Single Item Questionnaires

After each trial, participants were asked to answer on 5-point Likert-type scales how much they agreed with each of two statements: "The question came at a good moment" and "The assistant asked the question in an appropriate way." These items mirrored the themes described in Chapter 4, timing and delivery, which participants identified as important features of the structure of a spoken interruption.

Demographic Questionnaire

Participants were asked a number of questions about themselves including their age, nationality, level of expertise with Tetris, how recently they played Tetris, their level of experience with speech agents, and which speech agents they use.

Open Ended Reflective Question

To gather further insight into the differences which participants perceived between the two agent conditions, participants were asked an open-ended question at the end of the experiment: "What differences, if any, did you notice between the two versions of the speech assistants?"

6.2.5 Procedure

Participants were directed to a webpage where they read an information sheet describing the study and the data rights of participants. They were then asked to indicate their consent to participating in the experiment and sharing their anonymised data. Participants were told that they would watch 12 short videos of a person playing Tetris, during which the Tetris player would be interrupted by a proactive speech agent asking them a question. Participants were told that after each video, they would be asked to answer 2 questions about the interruption that they just watched and, after all 12 videos, they would be asked to complete a 18 item questionnaire about the agent they just listened to. The informational screens explained that after completing this routine once with one agent, they would then be asked to do the same again with a different agent. Participants were told that each agent was engaged in an exercise in which it needed to ask the Tetris player a variety of questions, and its goal was to minimise disruption to the Tetris player while asking its set of questions as quickly as possible. Informational screens explained that for some questions, minimising disruption to the player is urgent as the Tetris player was rated on their play during the game shown, rated games were used to choose the winner of a cash prize, and that the Tetris player did not know which games were rated.

After this information was presented and the participants consented to take part in the study, they were asked to complete an initial PMQ questionnaire to get a baseline understanding of their views of speech agents in general. After the initial PMQ,

Comparing Perceptions of Static and Adaptive Proactive Speech Agents

participants saw example screenshots from a video. In the example screenshots, one image displayed a game of Tetris with a question prompt and the other screenshot displayed the same game of Tetris and question prompt with a large red dot overlaid above the Tetris game board (Figure 6.1). Participants were told that this visual indicator is not visible to the Tetris player, but it indicates to the observer (i.e. to the participant) that the agent has been prompted to ask a question to the Tetris player. Participants were informed that the agent could see the Tetris game and could decide when to begin its interruption any time after the interruption was prompted. After the participant viewed the example screenshots, they were shown a practice video in which a Tetris game is played by an unseen player, the visual indicator appears after some time, and a synthesised voice asks the Tetris player a question.

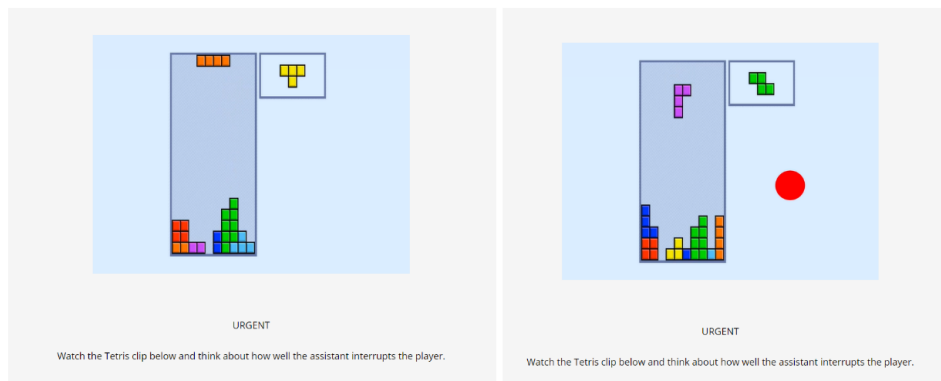


Figure 6.1: Example screenshots from the experiment which participants saw as part of pre-test instructions. On the left, there is no red dot, so the agent has not yet been cued to interrupt. On the right, the red dot has appeared, signalling that the agent has been cued to interrupt.

After the participant watched the practice video, they were asked to click a button to indicate that they were ready to continue and begin their first block of trials. Blocks of trials contained 12 videos of a single agent condition, with condition order counterbalanced across participants. Within a block of 12 trials, the order of videos was randomised for each participant. Following each video, participants rated how much they agree with each of the following statements on a 7 point Likert scale: “The question came at a good moment” and “The question was asked in a disruptive way”. After each trial (video and Likert items), a plain white screen with a black central fixation cross appeared for a short interval before the next trial began. Each trial lasted between 10 and 20 seconds.

After completing a block of trials, participants again completed an online version of the PMQ on a single webpage. After completing the PMQ, participants were asked to confirm that they were ready for the second block of 12 trials by clicking the continue button. After completing their second block of trials, participants completed another PMQ, being asked to reflect on the agent they just saw.

After completing the final PMQ, participants were asked to complete a short demographic questionnaire, asking about their age, nationality, level of expertise with Tetris, how recently they played Tetris, their level of experience with speech agents, and which speech agents they use. Finally, participants were asked a single open-ended question: "What differences, if any, did you notice between the two versions of the speech assistants?" Participants were then thanked for their participation, given an opportunity to submit any other questions or comments, and debriefed on the aims of this study, including letting them know which block of trials was adaptive and which was the static agent. Finally, participants were given information for receiving payment. The full source code and materials for the experiment is provided² and experimental materials including consent forms, participant information sheets, and instruction screens visible to participants are included in Appendix D.

6.3 Results

6.3.1 Analysis approach

A total of 1920 interruption trials were viewed across the experiment by 80 participants, with participants responding to single-item questionnaires after each trial and to the Partner Model Questionnaire before the experiment and after each of the two agent conditions. Therefore, 1920 single-item questionnaire responses and 240 PMQ responses were recorded across all participants. No data needed to be removed for technical issues or by participant request. For PMQ responses, total and subscale scores within each condition were assessed for extreme values (± 3 standard deviations from the condition means) and none were detected. For each single item questionnaire, condition means were calculated for each participant and condition means were assessed across participants for extreme values (± 3 standard deviations from the between-participant condition mean) and none were detected. This

²osf.io/g8zk6/?view_only=2a1faf7837f443348baf88cd585fc08a

Comparing Perceptions of Static and Adaptive Proactive Speech Agents

resulted in all 1920 responses for each single-item questionnaire and all 240 full PMQ responses being included in the final analysis.

PMQ total scores and subscales were checked for internal reliability, with strong Cronbach's alpha measurements in each of the three conditions for total scores (pretest $\alpha = .83$, static $\alpha = .78$, adaptive $\alpha = .85$) as well as Cronbach's alpha measurements for each condition in the *partner competence and dependability* subscale (pretest $\alpha = .85$, static $\alpha = .82$, adaptive $\alpha = .88$) and the *human likeness* subscale (pretest $\alpha = .76$, static $\alpha = .81$, adaptive $\alpha = .84$). The *cognitive flexibility* saw relatively weak reliability (pretest $\alpha = .32$, static $\alpha = .57$, adaptive $\alpha = .48$). These reliability measures largely matched prior work using the PMQ, which demonstrated Cronbach's alpha around .80 for total PMQ scores and for the first two subscales and Cronbach's alpha around 0.6 for the third subscale, which is comprised of only three items.

Linear mixed effects models were used to analyse the effect of agent condition on PMQ scores, single-item timing scores, and single-item appropriateness scores. Models were fit using the lme4 package version 1.1-26 (Bates et al., 2015) in R version 4.1.1 (R Core Team, 2020). Because PMQ responses were not measured for each video stimulus, the linear mixed model of PMQ responses fits the fixed effect of agent condition (pretest, static, and adaptive) with intercepts per participant. The linear mixed model of each single item questionnaire score fits fixed effects of agent condition (static and adaptive) with random by-participant and by-item slopes and intercepts (by-item effects include effects of stimulus, condition order, and trial order). Each model therefore represents the maximal model for that variable. Note that the urgency and Tetris difficulty of a given trial are not modeled individually as each stimulus is fixed in terms of Tetris clip (and thus Tetris difficulty) as well as interruption content and interruption urgency condition. For PMQ models which have three levels of agent condition, the adaptive condition was used the reference level as H3 predicts PMQ differences between the adaptive and static conditions (but not differences between PMQ scores for either condition and the pretest scores). To improve reproducibility, full model syntax and random effect outputs are included for each model (Meteyard & Davies, 2020). Additional linear mixed models were fit for each PMQ subscale as exploratory analysis to identify sources of differences between total PMQ scores. All analyses were preregistered before data collection began³.

³[https://osf.io/g8zk6/?view_only=2a1faf7837f4433\(48baf88cd585fc08a](https://osf.io/g8zk6/?view_only=2a1faf7837f4433(48baf88cd585fc08a)

Because the single-item questionnaires for timing and for appropriateness are not validated measures and because the other dependent variable, the PMQ, is multidimensional covering many aspects of speech agent interactions, additional steps were taken to check for discriminant validity between these dependent measures. To ensure discriminant validity, repeated measures correlation tests were performed between each dependent variable. Repeated-measures correlations allow for correlations between variables to be assessed across within-subject conditions on a per-subject level in order to determine the strength of association between paired variables across a sample (Bakdash & Marusich, 2017). Because single-item questionnaire scores were measured once per trial and PMQ scores were measured once per condition, single-item questionnaire measures were aggregated on a per-condition basis for each participant. There was a weak positive correlation between the timing and appropriateness scores, $r(79) = .33$, 95% CI: [.12,.51], no statistically significant correlation between PMQ scores and timing question scores, $r(79) = .17$, 95% CI: [-.07, .36], and no statistically significant correlation between PMQ scores and appropriateness question scores, $r(79) = .21$, 95% CI: [-.00, .41]. Insofar as none of these measures are correlated to a very strong degree, there appears to be appropriate discriminant validity between these variables (Nunnally & Bernstein, 1994). The timing and appropriateness measures used here can therefore be seen as distinct measures from both the PMQ and from one another.

6.3.2 Quantitative response data

Single-item questionnaires

Timing: For the first single-item questionnaire, “The assistant asked the question at a good moment”, there was no significant fixed effect of agent condition on participant ratings in a 5-point Likert-type scale [Unstandardised $\beta = -0.57$, SE $\beta = 0.75$, 95% CI -0.30, 0.00], $p = .974$]. H1 is rejected. Full model syntax and output are included in Table 6.2. Means and standard deviations of single-item questionnaire responses by condition are presented in Table 6.1.

Appropriateness: For the first single-item questionnaire, “The assistant asked the question in an appropriate way”, there was a significant fixed effect of agent condition on participant ratings in a 5-point Likert-type scale [Unstandardised $\beta = -0.58$, SE $\beta = 0.17$, 95% CI -0.93, 0.-0.24], $p = .998$]. This indicates that participants rated questions

**Comparing Perceptions of Static and Adaptive Proactive
Speech Agents**

Table 6.1: Table of means and standard deviations for single-item questionnaire responses by condition

Measure	Condition	Mean	SD
Timing	Static	2.33	1.13
	Adaptive	2.22	1.18
Appropriateness	Static	2.75	0.93
	Adaptive	2.21	1.18

Table 6.2: Summary of fixed and random effects for timing single item questionnaire - Linear mixed effects model

Model: *Timing rating* =
Agent Condition + (1|*subjectID*) + (1 + *Condition*|*stimulus*) + (1|*trialOrder*)

Fixed Effect	Std β	Unstd β	SE β	t	p
Intercept	.08	2.38	.14	16.85	.001***
Adaptive Agent	-.15	-.17	.09	-1.98	.07

Random Effects			
Group	SD	Corr	
Participant (intercept)	.53		
Stimulus (intercept)	.42		
Stimulus (slope)	.21	-.51	
Trial order	.07		

Table 6.3: Summary of fixed and random effects for appropriateness single item questionnaire - Linear mixed effects model

Model:

$$\text{Appropriateness rating} = \text{Agent Condition} + (1 + \text{Condition}|\text{subjectID}) + (1 + \text{Condition}|\text{stimulus}) + (1|\text{trialOrder}) + (1|\text{conditionOrder})$$

Fixed Effect	Std β	Unstd β	SE β	t	p
Intercept	.19	2.72	.13	21.65	<.001***
Adaptive Agent	-.46	-.50	.17	-2.93	.010**

Random Effects		
Group	SD	Corr
Participant (intercept)	.71	
Participant (slope)	.64	-.65
Stimulus (intercept)	.09	
Stimulus (slope)	.51	-.55
Trial order	.02	
Condition order	.12	

asked by the static agent as being more appropriately asked than those asked by the adaptive agent. H2 is therefore rejected as the opposite result was found. This result is visualised in Figure 6.2. Full model syntax and output are included in Table 6.3.

6.3.3 Partner model questionnaire

There was a significant fixed effect of agent condition on Partner Model Questionnaire scores, with participants having significantly stronger partner models of speech agents before the experiment as compared with after interacting with the adaptive model [Unstandardised $\beta = 6.86$, SE $\beta = 2.03$, 95% CI [2.86, 10.87], $t = 3.38$, $p = .003$] and stronger partner models of the static agent as compared to the adaptive agent [Unstandardised $\beta = 7.36$, SE $\beta = 2.03$, 95% CI [3.36, 11.37], $t = 3.63$, $p = .001$]. H3 is therefore rejected as the opposite result was found, which is visualised in Figure 6.3. PMQ and subscale means and standard deviations by condition are presented in Table 6.4. Full model syntax and output are included in Table 6.5. There was no difference between participants' partner models of the static agent as compared with their pretest partner model of speech agents. This indicates that the manipulation was successful insofar as the static agent condition matched people's preconceived notions of speech agents.

To better understand the source of Partner Model Questionnaire differences be-

**Comparing Perceptions of Static and Adaptive Proactive
Speech Agents**

Table 6.4: Table of means and standard deviations for PMQ total score and subscale scores by condition

Scale	Condition	Mean	SD
Total PMQ	Pretest	72.2	12.0
	Static	72.6	11.1
	Adaptive	67.0	13.9
Competence & Dependability	Pretest	42.2	7.57
	Static	42.8	6.66
	Adaptive	37.3	8.96
Human-Likeness	Pretest	20.1	5.68
	Static	19.8	6.54
	Adaptive	19.9	6.95
Cognitive Flexibility	Pretest	9.90	2.74
	Static	9.93	3.29
	Adaptive	9.85	2.88

Table 6.5: Summary of fixed and random effects for Partner Model Questionnaire total scores - Linear mixed effects model

Model: $PMQ = Agent\ Condition + (1|subjectID)$

Fixed Effect	Std β	Unstd β	SE β	t	p
Intercept	-.28	67.05	1.38	48.68	<.001***
Pretest	.40	5.40	1.70	3.00	.003**
Static	.40	5.55	1.70	0.28	.001**
Random Effects					
Group	SD				
Participant (intercept)	5.99				



Figure 6.2: Predicted values of appropriateness questionnaire ratings by condition

tween the agent conditions, further models were fit to compare participants's scores across each of the three subscales of the PMQ. There was a significant fixed effect of agent condition on *partner competence and dependability* subscale scores, with participants identifying speech agents as rating higher on this factor before the experiment as compared with after interacting with the adaptive model [Unstandardised $\beta = 5.38$, SE $\beta = 1.31$, 95% CI [2.79, 7.96]] and rated the static agent as stronger on this factor as compared to the adaptive agent [Unstandardised $\beta = 7.14$, SE $\beta = 1.31$, 95% CI [4.55, 9.73]]. This result is visualised in Figure 6.4. There was no difference between participants' *partner competence and dependability* subscale ratings of the static agent as compared with their pretest partner model of speech agents. Full model syntax and output are included in Table 6.6.

There were no significant fixed effects of agent conditions on either *human likeness* or *cognitive flexibility* subscales, indicating that overall PMQ differences between conditions are largely explained by differences in perceived competence and dependability. Full model syntax and output are included in Table 6.7 and 6.8 respectively.

**Comparing Perceptions of Static and Adaptive Proactive
Speech Agents**

Table 6.6: Summary of fixed and random effects for Partner Model Questionnaire partner competence and dependability subscale - Linear mixed effects model

Model: $PMQ F1 = Agent\ Condition + (1|subjectID)$

Fixed Effect	Std β	Unstd β	SE β	t	p
Intercept	-.28	37.28	.87	48.68	<.001***
Pretest	.40	4.88	1.07	4.54	<.001***
Static	.44	5.55	1.07	5.18	<.001***
Random Effects					
Group	SD				
Participant (intercept)	3.70				

Table 6.7: Summary of fixed and random effects for Partner Model Questionnaire human likeness subscale - Linear mixed effects model

Model: $PMQ F2 = Agent\ Condition + (1|subjectID)$

Fixed Effect	Std β	Unstd β	SE β	t	p
Intercept	.00	19.93	.71	27.98	<.001***
Pretest	.03	.18	.84	.21	.835
Static	-.01	-.09	.84	-.10	.917
Random Effects					
Group	SD				
Participant (intercept)	3.52				

Table 6.8: Summary of fixed and random effects for Partner Model Questionnaire cognitive flexibility subscale - Linear mixed effects model

Model: $PMQ F3 = Agent\ Condition + (1|subjectID)$

Fixed Effect	Std β	Unstd β	SE β	t	p
Intercept	-.01	9.85	.33	29.43	<.001***
Pretest	.02	.05	.42	.12	.904
Static	.02	.08	.42	.18	.857
Random Effects					
Group	SD				
Participant (intercept)	1.43				

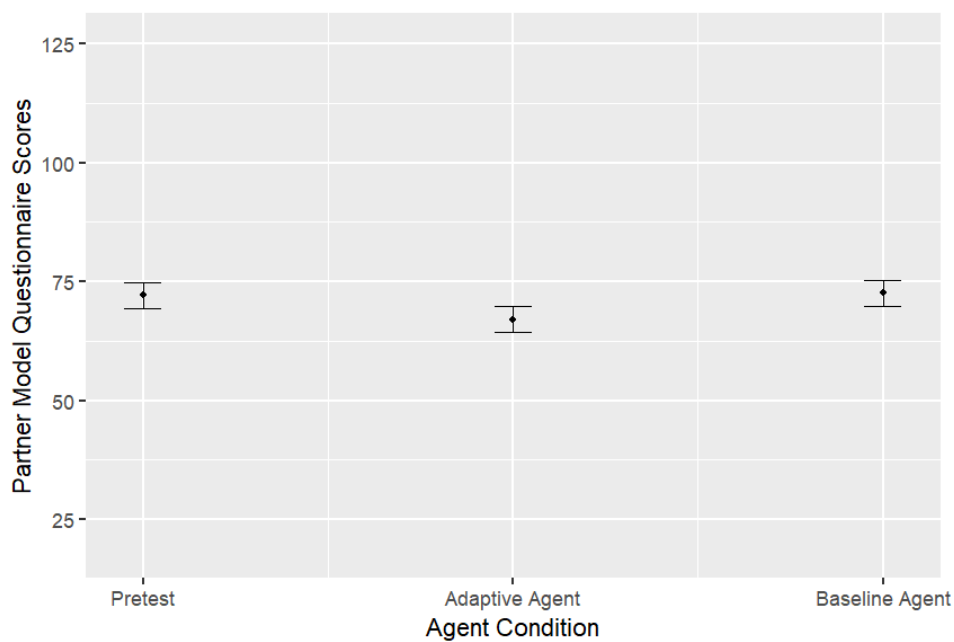


Figure 6.3: Predicted values of Partner Model Questionnaire total scores by condition

6.3.4 Qualitative response data

Data analysis approach

Answers to the open-ended question “What differences, if any, did you notice between the two versions of the speech assistants?” were analysed through thematic analysis (Braun & Clarke, 2006). The analysis took a reflexive approach, whereby codes were generated inductively and revised into themes through an iterative process of familiarisation and staged coding (Braun & Clarke, 2019). Unlike in previous chapters, in which deductive codes from prior literature formed a starting point for a hybrid analysis process (Fereday & Muir-Cochrane, 2006), the present analysis started inductively, generating codes directly from participant responses. This decision reflected the nature of prior research on how people understand non-human agents as dialogue partners which identified 18 semantic differentials across three factors which comprise people’s partner models (P. R. Doyle et al., 2021). These factors were seen as an overly broad starting point for deductive coding. Instead initial codes were generated inductively so that the meaning of each participant response

Comparing Perceptions of Static and Adaptive Proactive Speech Agents



Figure 6.4: Predicted values of Partner Model Questionnaire partner competence and dependability subscale scores by condition

would be preserved while allowing for the generation of central conceptual themes which nonetheless reflect this prior work.

Five themes were identified inductively in participants' open-ended responses. These themes are described below along with counts of how many participants' responses comprised a theme and illustrative examples of those responses. Three of the 80 participants did not provide a response to the open-ended question and therefore did not produce data for the qualitative analysis, so themes were generated through analysis of the 77 participant responses for which data was provided. Note that in the experiment, agents were identified as Agent A and Agent B with order randomised, and participants largely referred to agents either this way or as "the first one" or "the second one." For reading clarity, data extracts presented below have been edited so that agents are consistently referred to as "adaptive" or "static." Full participant responses as well as iterations of codes and themes are provided at https://osf.io/g8zk6/?view_only=2a1faf7837f443348baf88cd585fc08a

Clarity, directness, politeness (32)

A common theme which participants identified as a difference between the two agents was differences in any or all of the clarity, directness, and politeness which marked the tone and word choice of their utterances. While these characteristics were frequently mentioned, they were not solely attributed to one agent or the other, with different participants disagreeing over which agent was more clear, direct, or polite. “The [adaptive] was much more pleasant and polite, the [static] was a lot more cold and blunt” (P08) “the [static] seemed more relaxed and human like and more polite and not intrusive the [adaptive] was abrupt [sic] almost rude and just butting in with unpolite ways of asking things” (P77)

While participants largely identified the same characteristics of agent speech design - conversational phrasing for non-urgent interruptions and terse phrasing for urgent interruptions - their interpretation of how these decisions impacted the tone of the adaptive agent differed between participants. “The [static] speech assistant asked the questions more clearly and human like” (P11) “The [adaptive] version was more human in terms of the way sentences would be said socially - but not to the point it made the questions clearer” (P14) “[Static] spoke more clearly and directly than [adaptive] who seemed somewhat querulous [sic] in its tone” (P31)

The lack of agreement between the way participants characterised each agent within this theme added ambiguity to responses in which participants did not clearly mention which agent they were referring to. It is unclear for instance whether the participant who said “One was more hesitant, the other more direct” (P78) found the adaptive agent hesitant, owing to its occasional use of access rituals, or direct, owing to its occasional use of a faster speech rate and terse phrasing. Likewise while a participant mentioning that “one used a little more human interaction” (P07) , it is unclear whether this human interaction is an allusion to the aforementioned access rituals or to the clarity in communication which other participants attributed to the static agent.

Rushed speech (11):

Within the theme of clarity, directness, and politeness, there was a subtheme of rushed speech which several participants noted specifically. Like the broader theme which encompasses this subtheme, participants did not reach consensus about which agent sounded rushed, even though only the adaptive agent used a faster speech

Comparing Perceptions of Static and Adaptive Proactive Speech Agents

rate. “the [adaptive] one was more informal and rushed. it also spoke in less cohesive sentences.” (P23) “the [adaptive] version appeared friendlier and less rushed” (P13) This disagreement may result from the adaptive agent’s use of access rituals and conversational phrasing in non-urgent interruptions, giving these particular trials a less rushed character. Likewise, participants who found the adaptive agent less rushed may have noticed the adaptive agent waiting for an opportune moment to interrupt on non-urgent trials, rather than always interrupting as soon as possible. In any case, this lack of consensus likewise makes some participant responses ambiguous as “Clearer, less rushed words” (P41) fits with some participants’ descriptions of the static agent and other participants’ descriptions of the adaptive version.

The particular challenge of understanding rushed language while multitasking was noted by some participants. The nature of listening to interrupting speech as opposed to listening to speech without any other task demanding attention may affect the extent to which participants characterised it as too fast or rushed, and it may therefore colour the responses of participants who did not explicitly mention the attentional demands of the Tetris task. At least one participant did explicitly note this demand however, taking the perspective of the Tetris player and stating that “the speed of the questions being asked could be missed if the player is focusing” (P69) for this reason, the extent to which speech was perceived rushed should be considered contextually - the rate and verbosity of an agent’s speech may feel rushed in the context of interrupting a particular Tetris game, but it is not clear it would be perceived the same way in a different context.

Appropriateness (13)

Another prominent theme in describing the difference between the agents was the extent to which participants found the agent’s speech to be contextually appropriate. For some, the adaptive agent was seen as more contextually appropriate, in that its speech suited the level of urgency of its interruption or that it used access rituals to mitigate the potential social inappropriateness of speaking to a busy person. “The [adaptive] version was always appropriate to urgency, it would talk quickly when urgent and normally when not. The [static] version was speaking at a normal pace regardless of urgency.” (P39) “The [adaptive] voice assistant sometimes interjected with communicative phrases such as ‘Excuse me’ which I personally preferred. Espe-

cially when interrupted it would feel nice to have the moment at least seem to have been considered.” (P52)

For others, the static agent felt more appropriate, attributing the tone of speech and the phrasing of questions to this perception. Rather than finding the agent more appropriate to the context of the interruption task, participants who identified the static agent as more appropriate seemed to imply that it was more appropriate for human-agent interaction, owing to its tone and phrasing. “The [static] version of the speech assistant was generally worded better for engagement” (P01) “I found the [static] more appropriate with tone, and asked the questions in a slower more human way.” (P02)

Responses in this theme reveal a nuance in how appropriateness is construed throughout the interactions of this study. While some participants focus on the context of the interrupted task or on the social context on an interruption, others focus on the social context of the identity of the speaker and listener as an agent and a person respectively. This helps to reveal the extent to which the design of a particular agent may be appropriate for some contextual aspects of a spoken interaction while being inappropriate for other aspects of that interaction’s context. Likewise, differences between participants in which contexts were salient to them in their open-ended responses helps to underscore the subjectivity between people in evaluating appropriateness.

Human mimicry (6):

Discussion of human mimicry was a subtheme of the appropriateness theme. This subtheme included responses which agreed conceptually with participants who found the static agent more appropriate for a human-agent interaction, but responses in this subtheme particularly noted that the adaptive agent was inappropriate due specifically to its aim of mimicking human speech. Participants noted both changes in speech rate and use of access rituals as explicit attempts at sounding human, and it was this deliberate attempt by a non-human agent to sound human which participants identified as inappropriate. “The [adaptive] speech assistant seemed to be making an attempt at sounding more human. I think it failed in this and the attempts became a tad annoying.” (P10) “The [static] one was much more believable. The [adaptive] one started the question with things such as ‘erm and sorry’” (P60)

Some participants even noted the aforementioned nuance in appropriateness,

that adaptation may have been appropriate for the nature of the task, but the way this adaptation was achieved felt inappropriate for a non-human agent to attempt. “[The adaptive] one would go faster for urgent questions, but using things like ‘umm’ when asking a predetermined question in a robotic voice is jarring and unneeded.” (P19)”

This subtheme helps to further highlight the extent to which cues about the identity of an agent - design elements which make the agent seem more human like or more machine like - can supersede the function of those cues for some people. Where some found access rituals pleasant and appropriate in offsetting the intrusiveness of an interruption, others found them jarring and inauthentic when uttered by a non-human agent.

Inconsistency (6)

One theme for which participants agreed upon a characterisation fitting one agent in particular was the theme of inconsistency. In this theme, participants noted the nature of the adaptive agent varying its interruptions. Unlike in the appropriateness theme in which this adaptive nature was seen as ensuring interruptions were contextually appropriate, responses in this theme construe the adaptive agent as inconsistent. “The [adaptive] one was inconsistent, some questions were read very quickly at times” (P73) “The [adaptive] was unclear and inconsistent. The [static] was far better” (P21)

This theme fits well with the results of the Partner Model Questionnaire, in which participants rated the adaptive agent as lower on the Partner Competence and Dependability factor than the static model. Indeed consistent/inconsistent is one of the items in the questionnaire which loads onto that factor (P. R. Doyle, 2022). The concurrence with this theme and the quantitative data results may be a reflection of the consensus between participants in this theme, that all participants who noted the inconsistency of a particular agent identified the adaptive agent as the inconsistent one. This stands out as a major difference between this theme and the theme of appropriateness, for which participant descriptions were mixed in terms of which agent was more human like and whether this human likeness was perceived as inauthentic. Just as responses were mixed in that theme, so too were Partner Model Questionnaire results where no fixed effect of agent condition was found on the human-likeness factor. The alignment of qualitative and quantitative results across these two aspects help to

reveal a greater agreement between participants in the perception of inconsistency as compared to the perception of other themes from the qualitative data.

Describing without judging (23)

Some participants described the differences between agents without commenting on which agent they preferred or which characteristics were beneficial or harmful to the interaction. Participants in this theme largely commented upon the faster speech rate that the adaptive agent used in urgent interactions or the access rituals the adaptive agent used in non-urgent interactions. "The [adaptive] batch seemed to have a different speeds of speech" (P37) "[The adaptive] version varied speed of question depending on urgency and also inserted extra comments like "excuse me"." (P80) "[The adaptive] one varied in speed of speech [sic] and tried to use human like terms such as erm and hey etc" (P62)

While some participants like the above made note that the adaptive agent varied its speech rate, other participants merely noted that the adaptive agent used a faster speech rate, not clearly differentiating whether they thought the speech rate for the adaptive agent was always higher than the static agent's or whether they thought the speech rate varied. "The [static] spoke more slowly" (P68) "the [static] spoke more slowly and so was easier to understand whilst concentrating on something else" (P42)

Other participants in this theme further noted that the way the adaptive agent phrased its questions differed from the static agent, noting "The speed of the speech varied between the two, as well as the preciseness of the questions asked." (P33) or "The [adaptive] speech assistant was a lot more varied, both in terms of its approach to asking questions as well as the speed in which those questions were asked." (P59). All participants who described differences, even those who did not describe how they felt those differences impacted the interaction, accurately identified real differences between the two agent conditions, with no participant mentioning any features which were not present. While this theme provides limited insight into the way particular characteristics of a proactive agent impact interactions, it nonetheless reveals that the manipulations used with the adaptive agent are salient enough to participants that they can be accurately described after observing the interactions.

No differences (3)

Three participants explicitly stated that they did not detect any difference between the two agents. One participant particularly stated “I actually thought they were the same. I think i was concentrating more on the task.” (P35). This theme may indicate that for some participants, the manipulation was not apparent or that monitoring the Tetris task was too demanding of attention to notice particular characteristics of the agents.

6.4 Discussion

This chapter aimed to apply insights about the method of expression (see McFarlane, 1997) of human speech interfaces to the design of a proactive speech agent, investigating the effects of adapting speech to cues of urgency and of ongoing task difficulty in the ways humans try to do when they interrupt. Prior work on speech agents has identified a gulf between user expectations and interaction realities (Luger & Sellen, 2016) owing to speech agents giving cues to users that they are more capable dialogue partners than they are revealed to be through interactions (P. R. Doyle et al., 2019). The present study therefore hypothesised that participants would rate speech interruptions from an adaptive agent as coming at better moments (H1) and as more appropriately asked (H2) as compared interruptions from a static agent. It further hypothesised that participants’ partner models of an adaptive agent would be rated as stronger on the Partner Model Questionnaire (P. R. Doyle, 2022) than their partner model for the static agent (H3). The study also sought to investigate which aspects of interactions with a proactive adaptive agent would be most salient as differences from interactions with the static agent (RQ1).

These questions were investigated through an online experiment using pre-recorded interactions between agent prototypes and a Tetris player. There was no significant difference between ratings of how well the agent timed its interruptions by condition, so H1 was rejected. Interruptions from the static agent were rated as statistically significantly more appropriately asked than those from the adaptive agent, so H2 was rejected. Likewise, participants’ partner models of the proactive agent were statistically significantly weaker than their partner models of the static agent or their pretest control partner model of speech agents in general, as measured by

the PMQ, so H3 was rejected. Qualitative analysis of open-ended descriptions of agent differences constructed five themes of participant responses: *Clarity, directness, politeness; Appropriateness; Inconsistency; Describing without judging; and No differences detected.*

6.4.1 Consistency as a salient feature for adaptive agents

Contrary to expectation, the adaptive speech agent was rated lower on the PMQ by participants than was the static agent or people's pretest perception of speech agents. Post-hoc analysis revealed that differences in PMQ scores resulted from differences in perceptions of partner competence and dependability. This finding was echoed in qualitative analysis which identified inconsistency as a theme participant's descriptions of differences between agent conditions. Reflecting on the items which load onto the partner competence and dependability PMQ factor, it becomes more clear why an adaptive agent would lead to a weaker partner model across this dimension. Items such as "Dependable/Unreliable", "Consistent/Inconsistent", and "Reliable/Uncertain" (P. R. Doyle, 2022) illustrate the importance of consistent, predictable behaviour in the formation of partner models. It may be the case that participants in this study did not have sufficient exposure to the adaptive agent to learn what contextual adaptations they could expect from the agent, leading to a poor understanding of those adaptations which cause them to seem arbitrary or inconsistent.

Insofar as commercially available speech agents are not adaptive, participants' mental models for the agents in this study would not likely lead to expectations of adaptivity. In other words, participants may have been expecting agents to behave the same way across all interactions rather than consistently behaving in particular ways given particularly contextual conditions. Indeed early work on mental models in HCI suggests that people apply previously constructed models mindlessly to machines, even if they don't actively believe that machines are human-like (Nass & Moon, 2000). Further, some research on adaptive interfaces has pointed toward the benefit of explicitly describing the sorts of adaptive features that an interfaces has and the errors that it may cause particularly for the purpose of setting appropriate expectations (Beggiato & Krems, 2013). Insofar as the novelty of adaptation diverges from prior experiences with speech agents which people have drawn on to form their mental models, extended exposure to an adaptive agent or explicit instruction around

Comparing Perceptions of Static and Adaptive Proactive Speech Agents

how its interactions differ from other agents may be beneficial in order to overcome perceptions of inconsistency. Returning to Horvitz's principles of mixed-initiative interface design, agent behaviour should be socially appropriate (Horvitz, 1999). While more research is needed to determine what people consider socially appropriate interactions between speech agents and people, it may be unsurprising when interactions with novel properties like contextual adaptivity are seen as socially inappropriate when they diverge from people's prior experiences of similar interactions. Extended exposure and explicit descriptions of adaptive features should therefore be explored as ways of introducing potentially beneficial conversational features like adaptivity without introducing perceptions of inconsistency or inappropriateness.

The potential for design decisions leading to stronger partner models in one dimension while sacrificing partner model strength in another dimension was considered by Doyle (P. R. Doyle, 2022). Specifically mentioning the unknown potential impact of proactivity on partner models, that work speculated that proactive interactions may be an avenue by which speech agents become perceived as more human-like, but this may have unintended consequences on perceptions of competence and consistency or perceptions of cognitive flexibility (P. R. Doyle, 2022). This is supported to some extent by the present study, with the adaptive agent here seen as less consistent than people's prior conceptions of speech agents, though neither the static proactive agent nor the adaptive proactive agent in this study were regarded by participants as more human-like than their prior model for speech agents overall. A trade-off between perceived partner dependability and human-likeness may indeed persist across speech interactions with machines, as some prior work has highlighted potential benefits to user satisfaction and efficiency when interacting with a machine designed to intentionally sound robotic rather than humanlike (Aylett, Sutton, et al., 2019; Moore, 2017). While the adaptive behaviours tested in this study were human-like in a desirable way to some participants, they were not so to all participants. Conversely, the inconsistency that adaptivity implied had a negative impact on the strength of partner models which people formed about the adaptive agent across participants.

Consistency in speech agent interactions should therefore be seen as a paramount consideration for user experience. Several qualitative studies have identified consistency and reliability as critical to the impressions of speech agents held by both frequent and infrequent users (L. Clark, Munteanu, et al., 2019; Cowan et al., 2017; Luger

& Sellen, 2016). While studies which created and validated the PMQ helped to illustrate the importance of perceptions of dependability, central to the largest of three factors in explaining people's partner models (P. R. Doyle, 2022), this study is the first to quantitatively demonstrate the sensitivity of that factor, showing that even design decisions aimed at influencing the human-likeness factor may instead only serve to negatively impact the partner competence and dependability factor. For this reason, the design of future speech agents must proceed cautiously. Insofar as people form their models of dialogue partners based on past experiences (Branigan et al., 2011a; Cowan & Branigan, 2017), introducing novel design elements like proactivity or adaptivity may do more harm than good in terms of setting accurate expectations for a speech agent, at least at first.

It may be the case that novel interactional elements wane over time in the extent to which they are perceived as inconsistent. It is not clear how partner models develop over time with repeat exposure to a new partner, and the longitudinal work required to make that determination has been identified as a challenge for over a decade (Branigan et al., 2011a; Cowan et al., 2015; P. R. Doyle, 2022). Even if prolonged exposure to adaptive proactive agent would improve people's partner models of those agents, with better understanding of adaptive agents behaving consistently relative to particular contextual cues (rather than seeing them as inconsistent from utterance to utterance), this benefit is of little value when an early disappointment leads to abandonment of a system (Cowan et al., 2017; Luger & Sellen, 2016). With sensitivity toward the negative impact of novelty on partner models and the effect of poor partner models on technological abandonment, further design of adaptive proactive speech agents may need to be more incremental than this study. Introducing adaptive features piecemeal across product lines or across time spent with an agent may be less jarring to a user than interacting with a speech agent with many novel design features introduced simultaneously. In order to make progress toward adaptive speech agents, the importance of consistent interactions must be considered.

6.4.2 Appropriateness of adaptive proactive design

The lack of improvement on PMQ scores for the adaptive agent as compared to static agent and pretest conceptions was not only surprising because of a decrease in partner competence and dependability, but also because of a lack of increase in human-

Comparing Perceptions of Static and Adaptive Proactive Speech Agents

likeness. Similarly, while it was hypothesised that the adaptive agent would be rated as asking questions more appropriately than the static agent, the opposite was found. Each of these findings can be better understood through the lens of appropriateness in human-machine dialogue. In addressing the gulf of expectations in speech agent interactions (Luger & Sellen, 2016), some recent work has focused on the idea of appropriateness in these interactions (Aylett, Cowan, et al., 2019; Aylett, Sutton, et al., 2019; Le Maguer & Cowan, 2021; Moore, 2017). This trend toward appropriateness has argued that increased human-likeness should not be a goal of itself in the design of speech agents. Instead, speech agents should be designed, in terms of voice (Aylett, Cowan, et al., 2019; Le Maguer & Cowan, 2021) and in physical appearance in the case of embodied agents (Aylett, Sutton, et al., 2019; Moore, 2017), to suit the role of the agent. Indeed qualitative work comparing human-human dialogue and human-machine dialogue has indicated that people see these two interactions as different in roles and in characteristics (Porcheron et al., 2018; Reeves, 2019), with people expressing a dislike of speech agents which try to act human-like (L. Clark, Munteanu, et al., 2019; P. R. Doyle et al., 2019). In this context, it may be clearer why participants in the present study rated the adaptive agent as no more human-like and as less appropriate in asking questions than the static agent. While adaptive interruptions may more appropriately utilise the context of an ongoing task or the user's level of attention, appropriateness also entails awareness of the social context. This was made clear by the diverse ways that participants described what they saw as appropriate in the appropriateness theme identified through participants' open ended responses to the agents. For many participants, the extent to which the agent behaved appropriately to its social role as an agent, rather than mimicking human behaviour, was the most salient difference between the two agents. For people who see human-like contextual adaptation as socially inappropriate, it might not matter how well-tailored to other contextual factors an adaptive agent tailors its speech, as adapting like a human feels inappropriate on its face.

Within the appropriateness theme identified in people's qualitative reaction to the two proactive agents was the perception of human mimicry, particularly in descriptions of the adaptive agent. This finding echoes past qualitative research in which people have expressed discomfort at voice assistants which they perceived as being designed to seem human (L. Clark, Munteanu, et al., 2019; Cowan et al., 2017; P. R.

Doyle et al., 2019). This notion, summarised by Aylett and colleagues by the phrase “mimicry is creepy” (Aylett, Cowan, et al., 2019), repeatedly surfaces among people describing interactions with commercially available speech agents, but little quantitative work using Wizard of Oz or simulated speech agents like the one in this study has demonstrated similar discomfort with design proposals for future agents. Finding a distaste for mimicry here reinforces Aylett’s claim that this distaste for mimicry should not be described as an uncanny valley (Mori et al., 2012), as a valley implies that further pursuit of the same goal will lead to improvement in the target measure (Aylett, Cowan, et al., 2019). Instead, participants in this study feeling uneasy at an imagined future proactive agent mimicking human-like mannerisms such as colloquial speech or using access rituals helps reinforce the idea that appropriate design for speech assistants need not mimic the behaviour of human speakers. Making proactive speech agents adaptive to users’ contexts may still be a worthwhile pursuit, but this study builds upon a collection of literature which suggests that these adaptations should perhaps not seek to mimic human speech behaviour. Whereas Chapters 3, 4, and 5 established some of the ways that human interrupters adapt their speech to a partner’s context by altering speech rates, varying their use of access rituals, and adjusting their phrasing, this study showed that people did not find it appropriate when a proactive agent adapted its speech in the same ways. Instead, more work is needed to explore other ways machines can adapt appropriately, such as using non-speech sounds like beeps or chimes as access rituals, or using distinct voices for different contexts like recent studies have explored (S. C. Lee et al., 2021; Torggler et al., 2022) rather than seeking to vary the prosody or phrasing of a single voice.

6.4.3 Individual differences and personalisation

While neither the static proactive agent nor the adaptive proactive agent were seen as significantly more human-like than participants’ prior conceptions of speech agents overall, this may be a result of differences between participants in opposing directions rather than a lack of difference in perceptions across participants. This potential explanation is bolstered by qualitative findings, as most themes in describing differences between agents included contradictory impressions across participants, with some finding particular features like the use of access rituals by the adaptive agent to be beneficial and human-like and others finding them off-putting and unnecessary.

Research on personalisation of speech agents has found a high degree of variation between people in how they would like speech agents to be designed. Some strongly preferring agents which fulfil social functions, whilst others prefer agents to only perform functional, tool-like roles (Völkel, Kempf, et al., 2020). Likewise, prior research comparing the roles of conversations with machines to those with humans revealed tension between people's desires to have speech agents learn more about them to personalise interaction with those who saw speech agents building this sort of common ground with users as undesirable (L. Clark, Munteanu, et al., 2019). Tailoring speech agent design to individual users may prove especially tricky due to high variance between individuals. Recent research on how people understand the personalities of speech agents, with the popular Big Five personality types used in human personality research proving less effective for classifying machines than a more gradated model of ten personality types (Völkel, Schödel, et al., 2020). While more research is needed to determine differences among people's preferences for these different personality types, it is clear that the design space for machine personalities is wide and that different designs are differentiable by the people that interact with these agents. These large individual differences in the perceptions of speech agent design decisions support the notion that personalisation of agents is both necessary and difficult.

6.4.4 Limitations

Individuals in this study varied not only in their preferences toward speech agents, but also in their Tetris expertise. While this study mostly involved participants with some Tetris experience (i.e. neither experts nor total novices), there is sure to be variation in skill across participants. This may impact participants' perceptions of the adaptive agent due to the differences in how expert and non-expert Tetris players perceive Tetris games (Lindstedt & Gray, 2019). Participants who have weaker understanding of Tetris gameplay and strategy were likely less sensitive to the state of the Tetris game and to the mental demand particular game states might put on the player. It may be for this reason that the adaptive agent's consistent use of interruptible windows of Tetris for initiating its interruptions went unnoticed across the study, with participants finding neither agent as significantly better at timing its interruptions. Likewise, if the adaptive agent was not perceived as timing its interruptions any better than the static agent, this may further help to explain why the adaptive agent was

seen as less competent and dependable than the static agent - for most participants, it was seen as no more competent and as less dependable. While people have been demonstrated in prior research as being somewhat skilful in identifying natural break-points for discrete tasks (Janssen et al., 2012), their abilities to do so for a complex task like Tetris may be much more dependent on their expertise in that task. Future research should investigate both the effect of expertise on identifying interruptible moments in complex tasks and on whether well-timed interruptions can be beneficial to people who do not perceive the interruptions as being well-timed.

A notable limitation to the interpretation of this study is its casting participants as observers of proactive agent interactions rather than as the interrupted party. While this approach was beneficial in terms of controlling for interaction length, timing, and task success, it also limits the extent to which this study can comment on the overall impact of adaptation to context with proactive agents. It may be the case, like in recent work on the design of spoken take-over request during automated driving, that participants perform tasks more effectively when interacting with a speech agent that they dislike (Wong et al., 2019). This study primarily investigates people's perceptions of adaptive agents, but those perceptions may be relatively unimportant as compared to other considerations like safety impact for the designs of particular agents such as those in cars. For this reason, the present study represents only one of a variety of experiments which are needed to gain a comprehensive understanding of the impact of particular design decisions for proactive speech agents.

Another limitation of the current study is the holistic manipulation of adaptivity rather than isolating particular adaptive behaviours or contexts around which to adapt. While isolating individual behavioural or contextual variables would have allowed for a more precise description of causes to changes in participant perceptions of an agent, this study was the first to look at adaptivity as an independent variable in the design of a proactive speech agent. As such, there was little theoretical basis for choosing one finding over another when considering results from previous chapters which demonstrated the highly varied cues and decisions people consider when producing interrupting speech. This study thus presents an initial investigation into the salience and broad impact of adaptivity in this context without establishing particular causal links between particular behaviours and outcomes. Further work is needed to iterate upon the design of this study in order to better understand the questions already

Comparing Perceptions of Static and Adaptive Proactive Speech Agents

raised, such as the specific impact of access ritual use on perceptions of human likeness or artificiality, and the salience of using interruptible moments to deliver proactive speech during complex tasks. This study should be seen as an introduction to the question of how people want proactive agents to speak rather than a conclusive or prescriptive set of design guidelines.

Insofar as the PMQ was developed using perceptions of non-adaptive and non-proactive speech agents, the items and factors of the questionnaire may inadequately reflect how people understand speech agents with those properties. The items of the PMQ were developed using terms elicited from people who had recently interacted with speech agents and with a human across a variety of tasks in which they asked their partner a question and received a response (P. R. Doyle, 2022; P. R. Doyle et al., 2021). In those interactions, it may have been salient to participants that a human could consistently give a coherent answer to a question while a speech agent could not. In a mixed-initiative spoken interaction, in which a partner speaks to a person without being requested, this feature may be less salient, in turn contributing less to one's partner model for a proactive dialogue partner. Further development of instruments like the PMQ is needed to assess how partner model formations differ between unidirectional and mixed-initiative spoken interactions.

Finally, this study focuses solely on proactive speech interactions of a particular kind - personal questions - which interrupt a particular task - Tetris. These tasks were selected in order to maintain a high level of control over the structure of tasks and to maximally match the design of previous chapters so as to directly apply the findings of those studies to the design of this study. Salovaara and Oulasvirta describe the role of prototype experiments like this one as a way of evaluating possible futures (Salovaara et al., 2017). In their framework, they describe design decisions as aimed at either staging - making the present have some characteristic of a possible future - or controlling - preventing particular characteristics of the present which are not expected to be part of the imagined future from becoming salient (Salovaara et al., 2017). While Tetris differs in a variety of ways from the sorts of tasks people report wanting to use speech to multitask like cooking or driving (Luger & Sellen, 2016), it nonetheless matches the eyes-busy, hands-busy and complex, continuous nature of each of those tasks. In this way, the Tetris task maintains the aim of experimentally staging certain features of a possible future that this study represents. That said, it

is not clear that proactive agents asking questions which are irrelevant to the user's ongoing task represent a likely future use case for proactive systems. Nonetheless, the narrowing of speech tasks and stabilising of the tasks both across participants within this study and across the studies of this thesis serve the goal of controlling unsystematic variance, likewise helping to position this study as one which informs understanding of a possible future (Salovaara et al., 2017). In order to explore different potential future interaction scenarios, further research will need to make different choices with regards to how to stage that future and how to control for unwanted present circumstances. It is only through the exploration of a variety of potential futures that clear predictions can be made around effective design decisions for novel interaction types.

6.5 Conclusion

This chapter aimed to apply the insights about human spoken interruptions garnered from previous chapters to the design of a proactive speech agent in order to assess people's perceptions of such an agent. Applying prior work on the design of proactive non-human agents, this study identified adaptivity as a key variable for proactive agent interactions. As prior chapters identified varied contextual cues that people consider when interrupting with speech and the different modifications they make in light of those contexts, this study manipulated adaptivity by designing one agent condition around those findings with the other, static agent, insensitive to context. Specifically, the adaptive agent used the difficulty of the ongoing Tetris task as a cue for deciding whether or not to use access rituals, it used the state of the Tetris game as a cue for selecting a moment to initiate its interruption, and it used urgency of the interrupting task as a cue for varying speech rate and word choices. These agent conditions were compared across three measures: the PMQ, and single-item measures of how well interruptions were timed and how appropriately they were delivered. Quantitative results revealed that participants had a stronger partner model of the static agent as compared to the adaptive agent, owing to lower ratings of the adaptive agent's competence and dependability. Participants likewise found the adaptive agent's interruptions to be less appropriate than the static agent's and detected no differences in the quality of the timing of interruptions between agents. Qualitative

Comparing Perceptions of Static and Adaptive Proactive Speech Agents

analysis of open-ended descriptions of agent differences constructed five themes of participant responses: Clarity, directness, politeness; Appropriateness; Inconsistency; Describing without judging; and No differences detected. Taken together, this study found that impressions of the adaptive agent varied greatly between participants, but overall it was seen as inconsistent without being seen as clearly more human-like or appropriate as a partner. This study echoes previous literature questioning the appropriateness of using human dialogue as a model or metaphor for non-human speech. Like other chapters, it also highlights the large individual differences between people's preferences for spoken communication. While this early step toward the design of an adaptive, human-inspired, proactive speech agent revealed minimal overarching benefit to this approach, it may nonetheless serve as a guidepost for future investigation into this domain, by revealing limitations to the aim of designing speech agents through studying human behaviour.

7 * **General Discussion**

7.1 Introduction

This thesis presents four studies aimed at better understanding the way people use speech to interrupt another person who is engaged in a complex task and investigating how machines which speak proactively should be designed. Motivated by research which has indicated that speech agents fail to live up to the expectations of their users, particularly in terms of not facilitating multitasking as users hope (Luger & Sellen, 2016), this thesis contributes new insights into the role of the urgency of an interruption (Chapters 3 & 4) and the difficulty of an interrupted task (Chapters 4 & 5) in human speech interruptions. These insights are then applied to the design of prototype proactive speech agents (Chapter 6) in order to compare people's perceptions of a proactive agent designed to adapt its interrupting speech similar to the way human interrupters do with a proactive agent designed to interrupt in a static manner similar to existing non-proactive speech agents. This thesis likewise contributes novel methodologies for eliciting speech which interrupts a complex task (Chapters 3 & 4) as well as for classifying interruptibility of a complex task via clustering of observer ratings (Chapter 5). As such, withing a problem-solving framework (Oulasvirta & Hornbæk, 2016), this thesis contributes to solutions for empirical, conceptual, and constructive problems in human-computer interaction. Human interrupters are highly varied in the strategies they use to time and to structure spoken interruptions of complex tasks. Designing appropriate spoken interruptions for proactive agents requires sensitivity not only to characteristics of the interruption and of the interrupting task, but also to individual preferences of the people machines will be speaking to and the expectations people have for interactions of this type.

7.2 Contributions and Implications

7.2.1 Urgency and strategies in human-speech interruptions

The primary contribution of the study in Chapter 3 is an empirical contribution to the understanding of human spoken interruptions of other people engaged in a complex task. This chapter found that people are sensitive to the urgency with which they are instructed to interrupt others, with urgent interruptions occurring with less delay than non-urgent interruptions. This finding on the interruption of others fits with previous research on self-interruption during multitasking which found that explicit cues around the priority of an interrupting task affects the timing of interruptions (Brumby et al., 2013). This finding contributed to the design of the adaptive proactive agent in Chapter 6 as it indicated that interrupting task urgency is a cue that interrupters are sensitive to.

Chapter 3 additionally contributed a deeper understanding of the diversity of interruption strategies that human interrupters choose through qualitative analysis of participants self-reported strategies. Participants' qualitative descriptions of their strategies revealed the salience of other cues in the interrupting process including the difficulty of the task which participants sought to interrupt as well as interrupters' perceptions of the other party's cognitive load. The diversity of both cues and responses around interrupting complex tasks echoed recent research on the development of proactive speech agents which likewise found that a variety of cues influence people's perceptions about the interruptability of complex tasks, and that these cues vary depending on the specifics of the task and the person (Cha et al., 2020; Semmens et al., 2019). One such task-specific cue, task difficulty, was further studied in Chapter 4 and further contributed to the design of the adaptive proactive agent in Chapter 6. Likewise, large individual differences in the use of access rituals (see Krivonos and Knapp, 1975) led to further study of access rituals as a component of spoken interruptions in Chapter 4 and to the inclusion of access rituals in the design of the adaptive proactive speech agent in Chapter 6.

Participants' mention in Chapters 3 and 4 of their partner's cognitive load and focus on finding good moments to interrupt conceptually mirrored prior research on natural breakpoints (Borst et al., 2010; Janssen et al., 2010). While natural breakpoints have been studied in the context of self interruptions (Janssen et al., 2012)

and machines interrupting people (Borst et al., 2015), these studies contribute initial evidence that people consider natural breakpoints in tasks that others are engaged in when choosing when to interrupt. This finding helped inspire the direction for the study in Chapter 5 which explicitly sought to model interruptible windows of Tetris games like those used across these theses.

Finally, Chapter 3 contributed to the study of interruptions by demonstrating a paradigm for eliciting spoken interruptions of a complex task through a gamified approach. While prior work has used the game Tetris to study complex tasks (Lindstedt & Gray, 2019) or used a gamified approach for eliciting urgent human speech (Landesberger et al., 2020b), the study presented in Chapter 3 combines these approaches for continued research into human interruptions of this type. The utility of this paradigm for investigating this phenomenon was such that the paradigm was used for Chapter 4. The Online, gamified paradigm demonstrated by this study may be seen as a low-cost and low-risk way to advance research into proactive speech agents, particularly when compared to costly and invasive methods such as installing sensors in people's homes (Cha et al., 2020) or studying higher-risk environments such as on-road driving (Semmens et al., 2019).

The primary contribution of the study in Chapter 4 is the empirical comparison between the effects of the urgency of an interrupting task and the difficulty of the complex task being interrupted by human speech. Specifically, that study found that while interruption task urgency significantly affected interruption timing, the difficulty of the ongoing Tetris game which participants interrupted did not have a significant effect on timing. This finding echoed prior work around the strategies people use while speaking to a person engaged in a complex task (Janssen et al., 2014). While that study found that, in the absence of other cues, people will use the cues of their dialogue partner's ongoing task to adapt their speech, the study in Chapter 4 of this thesis demonstrated that an explicit cue of urgency for the interrupter had an effect on how interrupters timed their speech when these cues were presented together, with interruptions coming at less of a delay when they were urgent, irrespective of the difficulty of the Tetris game they interrupted.

Similarly, the study in Chapter 4 contributed an empirical finding around the structure of interruptions with regards to access rituals. While Tetris task difficulty had no significant effects on the timing of interruptions in terms of onset or duration, inter-

ruptions of less difficult Tetris games were significantly more likely to begin with an access ritual than were interruptions of more difficult games, irrespective of urgency. This finding came in spite of the fact that, like in Chapter 3, many participants never used access rituals while interrupting, further pointing to a high degree of individual variation in how people interrupt with speech.

Chapter 4 likewise replicated the qualitative finding from Chapter 3 regarding the high diversity in self-reported interruption strategies, contributing a refined set of themes for describing these different strategies. Strategies for timing interruptions largely represented either of two themes: *interrupting as soon as possible* or *interrupting at appropriate moments*. The former theme aligns with previous research on self-interruptions, which found that people tend to minimise the time spent to complete an interrupting task by switching tasks as soon as possible (Brumby et al., 2011), even when the primary task they interrupt is a high-stakes task like driving a car (Horrey & Lesch, 2009). This theme and prior work fits with the Soft Constraints Hypothesis, a model for how people coordinate complex tasks, which predicts that people will seek to minimise the total time spent on a task even at the cost of other variables like safety or cognitive load (Gray et al., 2006). The latter theme on the other hand, choosing appropriate moments to interrupt, again mirrored literature on natural breakpoints. While much research has demonstrated the use of natural breakpoints in task-switching behaviour which is self-directed (Borst et al., 2015; Janssen et al., 2010; Janssen et al., 2015), this finding expands understanding of spoken interruptions of tasks carried out by others, as it points toward interrupters' use of cues from other people's tasks in order to identify and utilise natural breakpoints in these interruptions as well. As such, this theme led to the design of the study in Chapter 5, seeking to classify natural breakpoints in Tetris games to better understand the extent to which participants utilised them in the prior studies.

In terms of strategies for the delivery of a spoken interruption, the study in Chapter 4 also contributes two themes, iterating on those presented in Chapter 3. These themes, *communicating urgency* and *communicating calmness*, represent distinct strategies for invoking a particular response strategies from the player. These alternative strategies broadly mirroring the distinct timing strategies that interrupters may select, in that the former seeks to induce a response as soon as possible and the latter seeks to minimise the disruption and time pressure placed upon the Tetris player.

These distinct strategies around urgency follow prior research identifying linguistic differences between speech elicited in urgent laboratory tasks compared to speech elicited in non-urgent tasks of the same type (Hellier et al., 2002; Landesberger et al., 2020a, 2020b). Specifically, this study found that participants sought to use fewer words, faster rates of speech, and were less likely to use access rituals when communicating urgency, in some cases because they explicitly sought to rush their partner's response and in others because they themselves felt rushed. By contrast, when communicating calmness, participants sought to use a casual tone, more conversational diction, and a more relaxed speech rate so as to not rush their partner's response and make responding to their speech less effortful. Research on the design of take-over requests via synthesised speech in self-driving cars has shown that a voice designed to be more assertive, with features such as a more serious tone and less polite phrasings of utterances, was seen as sounding more urgent and more likely to distract from another task as compared to a voice designed to sound friendly (Wong et al., 2019). Reflecting on that finding, the strategies selected by interrupters here may be seen as effective for accomplishing their goals, as communicating urgency or minimising distraction by communicating calmness are empirically supported. As such, these adaptations to spoken interruptions in response to urgency shaped the design of the adaptive proactive agent in Chapter 6.

7.2.2 Classification of interruptibility of a complex task

Following the findings of Chapters 3 and 4, Chapter 5 presented a study aimed at modelling interruptibility in Tetris. This study makes both a conceptual contribution of a model of that specific task and also an empirical contribution around the use of breakpoints by interrupters when speaking to someone engaged in a complex task through analysis of data collected in prior chapters. The specific model of Tetris interruptibility contributes four themes that describe interruptible moments in Tetris games: *No Spin*, *One Spin*, *Line Clear*, and *Calm Episode*.

The first two themes, indicating segments of Tetris gameplay which include horizontal movement and vertical dropping but no more than a single rotation, demonstrate the difficulty that perceived difficulty of coordinating rotations in Tetris for observers watching another person play the game. This followed prior work which established the difficulty of mental rotation of Tetris pieces experienced by untrained

Tetris players (Terlecki et al., 2008). This builds upon previous study of Tetris as a cognitive task which identified both rotation and horizontal movement as both minimised during optimal Tetris play (Lindstedt & Gray, 2019) and raises new questions about whether minimising rotation may be more important than minimising horizontal movement in terms of reducing cognitive burden to the player. These themes of interruptibility may reflect motor cues of breakpoints in a task (Janssen et al., 2012) insofar as interruptibility is signalled by the presence or absence of a movement by the player rather than a boundary between subtasks of the game Tetris (e.g. the end of a Tetris episode signalled by a piece reaching its final destination).

The next theme, Line Clear, indicates segments of the Tetris game in which a full horizontal line of Tetris pieces is completed and an animation of that line disappearing (and any pieces above that line falling down to replace the completed line) is seen before the next piece begins falling. Insofar as Tetris is a continuous task, these moments are the only times in the game in which no input from the player is possible and there is an actual break in the task. These moments therefore contain both motor and cognitive cues for a breakpoints between subtasks (see Janssen et al., 2012), as there is no motor input from the player and the moment is between distinct episodes with the previous Tetris piece having reached its final destination and the next piece not yet having started its descent. In this way, Line Clears are unambiguously interruptible moments that do not require an observer to make predictions about upcoming gameplay or inferences about the player's current thought process. While these occurrences may represent the best moments to interrupt a Tetris game, there is not a Line Clear event at the end of every episode. Indeed many episodes may take place between Line Clear events, and predicting the next occurrence may be no easier than predicting any other event in an observed Tetris game, so Line Clear events could not be relied on by participants as the only option for interrupting.

The final theme, Calm Episode, indicated Tetris episodes in which a no more than two total player inputs (i.e. horizontal movements or rotations) occurred in the full episode. Because Tetris is a continuous task without any time between episode subtasks (with the exception of Line Clear events when they occur), interrupters must interrupt during a subtask. Calm Episodes instead construe an entire subtask, a full Tetris episode, as a breakpoint between two, potentially more difficult Tetris episodes. In this way, while there are motor cues that a subtask is in progress, cog-

nitive cues around the difficulty of the episode may indicate to an interrupter that a task is nonetheless interruptible. The mixed use of cognitive cues, motor cues, or the concurrence of these cues echoes research on self interruptions in which either or both types of cue may be used depending on the characteristics of the primary task (Janssen et al., 2012). That Calm Episodes are concerned with the combined number of player inputs but do not differentiate between horizontal movements and rotations follows prior work on modelling Tetris gameplay which considers these inputs together as minimised during optimal play (Lindstedt & Gray, 2019). Altogether, these four themes provide a firm foundation for future research on Tetris as a complex cognitive task, building on past work which considered the episode as a subtask in Tetris to include different sorts of subtasks with identities specific to emergent characteristics of the game and an individual player's choices rather than a fixed subtask dictated by the design of the game.

Beyond contributing a description of Tetris as a complex cognitive task, Chapter 5 further contributed an empirical finding around people's use of natural breakpoints when interrupting Tetris games with speech by reanalysing data from Chapters 3 and 4 with consideration of the aforementioned descriptions of interruptible windows. This reanalysis contributes two distinct empirical findings. First, despite participants reporting that they attempted to choose good moments to interrupt - particularly in low-urgency trials - in Chapters 3 and 4, there was no significant effect of urgency on the use of interruptible windows upon reanalysis. Likewise, there was no effect of Tetris game difficulty on the use of interruptible windows across interruptions from Chapter 4. This pair of null findings may point to a disconnect between participants' understanding of their own strategy and their actual behaviour during real-time interruptions. On the one hand, participants may have been unable to identify and utilise interruptible windows in earlier studies in which the interrupting task was continuous and in which they did not have the benefit of hindsight (unlike participants in Chapter 5). On the other, they may have instead not feel incentivised to choose good moments to interrupt, either because they prioritised the interruption task over their partner's Tetris gameplay or because they did not perceive their Tetris playing partner as being in real risk. Further research is needed to determine to what extent either of these possibilities explains participants' relative lack of use of interruptible windows. Studies of different complex tasks and of interrupters with expertise in the task they seek

to interrupt may help further explain these findings.

The second major empirical contribution of Chapter 5 is the finding that the hard Tetris gameplay segments used in Chapter 4 unexpectedly contained more interruptible than did easy Tetris gameplay segments. This unexpected finding may have broader implications for the laboratory study of complex tasks. On the one hand, because Tetris gameplay videos needed to be sufficiently long in order to be useful for the Chapter 4 experiment, it may be the case that these long segments of high difficulty Tetris gameplay are selected through survivorship bias such that only continuous segments of difficult Tetris gameplay which have high proportions of interruptible moments were long enough for selection given the requirement of duration. Alternatively, it may be the case that the relationship between Tetris game difficulty and interruptibility is not straightforwardly in the expected direction. In the tripartite model of Cognitive Load Theory, cognitive load comes from three sources: intrinsic load, extrinsic load, and germane load (Sweller et al., 1998). While extrinsic load - resource demand from irrelevant factors such as a participant's external environment or physiological state unrelated to the experiment - would not have been different between Tetris difficulty conditions, it is somewhat unclear whether difficult games differ from hard games in terms of intrinsic or germane load. Intrinsic load, attentional demands from interactions within the task itself (DeLeeuw & Mayer, 2008), may have been higher in hard games than easy games even if there were more interruptible windows in hard games owing to non-interruptible moments requiring more inputs from the player (e.g. more rotations and horizontal movements) than non-interruptible moments in easy games. Conversely, it may be that differences in germane load within the interruptible moments render the hard games harder than easy games. Germane load is resource demand stemming from learning the patterns and operational requirements of a task (Sweller et al., 1998). It may be that, because hard games required more complex planning due to more lines of Tetris pieces existing at the bottom of the screen on those trials, they were seen as more difficult even though they had more low-intrinsic-load moments. Either explanation opens up new questions for future study of complex tasks. Due to the multifaceted nature of cognitive load, it is not clear that a hard continuous task should have more or fewer interruptible moments than an easy one. Likewise, while certain types of cognitive load may be perceived as more interruptible than others (e.g. perhaps participants believed that moments of

germane load are acceptable to interrupt whereas moments of intrinsic load are not), each may still contribute to the perceived overall difficulty of the task. This finding can therefore be a cautious warning to future studies of interruption of complex tasks, as the difficulty of a task and the interruptibility of a task cannot be treated as one in the same or even as straightforwardly related.

In addition to the conceptual and empirical contributions of Chapter 5, the methodology used throughout that chapter also contribute constructivist solutions to the problem of classifying the interruptibility of complex tasks. While prior research has aimed to collect a variety of data streams during the ongoing execution of complex tasks and asking engaged participants about their interruptibility (Cha et al., 2020; Semmens et al., 2019), this study presents an alternative approach. Rather than randomly presenting interruptions and correlating their appropriateness with task characteristics, Chapter 5 used a pre-recorded task to allow participants to select a moment to interrupt with the benefit of viewing an entire sequence of the task before selecting the moment to interrupt. This allowed for data-driven task modelling using interruptible windows actually selected by interrupters, rather than relying on the conditions present at arbitrarily selected moments. Furthermore, unlike experiments which use a variety of sensors to collect data about a task and its performer (Cha et al., 2020; Semmens et al., 2019), this approach was low cost and easy to implement online for rapid modelling of Tetris interruptibility. This approach is not aimed at replacing task performer-centric modelling of complex tasks, but instead as an interrupter-centric alternative which can be used across any number of complex tasks beyond Tetris using the same methodology. By specifically targeting Tetris as a complex task of interest, this study has the further contribution of reusing a particular complex task across cognitive science, using Lindstedt and Gray's proposal of Tetris of that task (Lindstedt & Gray, 2019) as a means of meeting the longstanding challenging of reusing a game as a complex task which the field can seek study in its entirety (Newell, 1973; Wulf & Shea, 2002). Chapter 5 therefore equips future researchers of complex tasks with both a methodology for rapidly modelling a diverse selection of complex tasks and an affirmation that Tetris can be used as a task which is considered in its fully-applied form at low cost to researchers. Future work may seek to cross-validate the classification methodology proposed here with performer-centric methods of modelling task interruptibility for both Tetris and for other complex tasks like driving (Semmens et al.,

2019) and household chores (Cha et al., 2020) so as to refine the models of those tasks and the methodologies used in addressing these problems.

7.2.3 Perceptions of an adaptive proactive speech agent

The primary contribution of Chapter 6 was an empirical comparison of two designs of proactive speech agents: an adaptive agent and a static (non-adaptive) agent. The study compared participants' partner models of the two agents, finding significant differences in an unexpected direction. While the adaptive agent was designed to follow patterns for adapting spoken interruptions according to the interruption's urgency and to the characteristics of the task they interrupt, it nonetheless did not lead to a stronger partner model than the static agent did. Indeed, participants had weaker partner models for the adaptive agent than they did for speech agents in general before the experiment and for the static agent. This finding makes a novel contribution to the area of proactive speech agents and to partner models of speech agents, as neither partner models of proactive speech agents nor of adaptive speech agents have previously been examined, owing to the nascence of speech agent partner model measurement (P. R. Doyle, 2022). As such, this is the first empirical work to demonstrate that proactive speech agents which largely follow the design of existing non-proactive agents do not lead to significantly different partner models than people's prior beliefs about speech agents. It is also the first empirical work to demonstrate that participants have a weaker partner model of an adaptive speech agent than of a comparable speech agent which is not adaptive.

While weaker partner models for an adaptive speech agent was not an expected finding, Chapter 6 proposes an explanation for this finding based on further quantitative analysis of the individual factors of the Partner Model Questionnaire (P. R. Doyle, 2022) and qualitative analysis of participants' descriptions of the speech agents. The mixed-methods data from that study point to interactional consistency as the source of weakened partner models. Among the three subscales of the PMQ, only the *partner competence and dependability* yielded significant differences between the adaptive agent and each of the static agent and pretest PMQs. Likewise, among themes identified in qualitative analysis of participants descriptions of differences between the agents, only *inconsistency* saw agreement between the participants who mentioned it with regard to which agent was stronger in that theme. Taken together,

these findings contribute a clear message to speech agent designers that a consistent, predictable interaction experience has a large effect on people's perceptions of speech interactions. Prior research on adaptive interfaces has likewise highlighted a trade-off between potential interactional benefits of adaptivity and the potential cost of unexpected inconsistency in the behaviour of the interface, suggesting that adaptive features should be explicitly explained to users before interaction (Beggiato & Krems, 2013). Chapter 6 of this thesis demonstrates that this trade-off of adaptivity is present for proactive speech agents, and future studies of proactive agents which adapt their speech according to contextual factors should investigate whether explicit teaching or extended exposure may help reduce the extent to which adaptivity is seen as inconsistent.

The empirical finding in Chapter 6 around the impact of perceptions of inconsistency on partner models also contributes to the conceptual framework of partner models of machine dialogue partners. The adaptive proactive agent in that study was designed to interrupt in a way more closely modelled on human interaction behaviour as compared to the static agent modelled on speech agent behaviour. In this way, its design targeted a stronger *human-likeness* dimension of partner models of dialogue partners, in the three factor model used in prior study of human-machine dialogue (P. R. Doyle et al., 2021). Despite the design of the agent seeking to strengthen partner models along that factor (as compared to pre-tests partner models of speech agents and to the static model), Chapter 6 instead found no differences on that dimension and weakened partner models for the adaptive agent along a different dimension. This sort of trade-off between dimensions of partner models has been speculated about as a possible challenge to design (P. R. Doyle, 2022), but this study is the first which demonstrated just such an effect occurring. It likewise echoes prior research on the opposite effect, in which machine dialogue partners were perceived as more efficient for interaction when designed to intentionally sound robotic rather than humanlike (Moore, 2017). The dynamics of partner models and the extent to which changes to one dimension may exert opposite changes upon other dimension require further study, but Chapter 6 highlights directly that design choices aimed at impacting users' partner models of machine dialogue partners must be considerate of potential tradeoffs.

Chapter 6 likewise contributes empirically to an understanding of appropriate-

ness in human-machine dialogue. Qualitatively, the adaptive proactive agent in that study was rated as no more human-like than the static agent, and it was rated as asking questions less appropriately than the static agent. Likewise, qualitative analysis of participants' descriptions of the agents identified appropriateness as a theme describing their differences, with participants largely identifying the proactive agent as engaged in human mimicry. The notion that human mimicry is inappropriate for a speech agent has been expressed in prior research (Aylett, Cowan, et al., 2019; L. Clark, Munteanu, et al., 2019; Cowan et al., 2017), which this finding reinforces. Insofar as ratings of human-likeness did not significantly differ between the adaptive and static agents in the aggregate and participants' open-ended responses did not unanimously identify one agent as more appropriate than the other, perceptions of human-likeness and attitudes toward mimicry may be somewhat varied between individuals. Indeed, just as there may be trade-offs between perceptions of human-likeness and dependability, there may also be trade-offs for individuals between their perceptions of human-likeness and appropriateness, such that particular design decisions are seen as appropriate and human-like by some but as inappropriate mimicry by others.

7.2.4 Human-inspired design of speech agents

Chapters 3, 4, and 5 all featured studies in which human participants used speech to interrupt an ongoing task. For human interrupters, an interruption requires effort both in terms of planning their utterance, and in monitoring cues which allow them to determine how to adapt their speech. The use of cues such as the urgency of the interruption and the difficulty of the ongoing task in adapting interrupting speech were explored in this thesis as well as in prior work on communication during ongoing tasks (Janssen et al., 2014). These findings may provide insight into the cognitive bounds for human-human interruptions, and they were a design inspiration for the adaptive proactive speech agent in Chapter 6. That said, it may not be the case that this is the optimal design strategy for a non-human agent, as the bounds of human cognition do not necessarily bound the behaviour of nonhuman agents.

A proactive speech agent may instead optimise its behaviour to suit the goals and needs of the person that it interrupts, not burdened by the same resource demands incurred by human interrupters. In this case, the design of the adaptive proac-

tive agent in Chapter 6 may be an insufficient model for nonhuman agents, as its behaviour unnecessarily reflects limitations of human cognitive resources which machines do not share. Still, tasks like visual monitoring, by which an agent would infer cues of the user's ongoing task (such as in Cha et al., 2020) are computationally intensive, particularly for more complex tasks (Escobar-Alvarez et al., 2018). Some artificial intelligence research has begun to consider bioinspired artificial intelligence design: designing AI agents which conserve computational resources by carrying out complex tasks in ways modelled on how resource-conserving animals or humans carry them out (Escobar-Alvarez et al., 2019). The human-inspired design of the adaptive agent in Chapter 6 can therefore be viewed as a sort of bioinspired, resource-conserving design for a proactive agent. Future work is needed to compare the effectiveness of proactive agents designed around a model of human interrupters to proactive agents aimed at instead modelling the users they seek to interrupt. The advantages to designing proactive agents based on the perspective of human interrupters are yet unknown, but this work presents a vision for how such designs can be carried out.

Insofar as the adaptive agent in Chapter 6 was largely seen by participants as less appropriate than a static agent inspired by existing speech agents, it raises larger questions about the appropriateness of human-human dialogue as a model for human-machine dialogue in general. Prior work on user perceptions of speech agents has indicated a gulf of expectations between the human-likeness that speech agents engender through natural sounding synthesis and through marketing as compared to feature-limited and error-prone interactions in practice (Luger & Sellen, 2016). Research in this area has highlighted the importance of appropriateness, typically indicating that speech agents should be designed such that their style of speech and their functionality is in line with the role they are cast in (Aylett, Sutton, et al., 2019; Le Maguer & Cowan, 2021; Moore, 2017). As such, Chapters 3, 4, and 5 sought to sketch the properties of speech used by human interrupters so that an agent fulfilling the role of interrupting a person engaged in a complex task would have appropriately aligned speech. That such an agent was seen as less appropriate by participants raises further questions about whether human-human dialogue is an appropriate basis for designing machine speech at all. Indeed a growing body of qualitative research has indicated that people are uncomfortable with speech agents explicitly designed to seem human-like (L. Clark, Munteanu, et al., 2019; Cowan et al., 2017; P. R. Doyle et al.,

2019). Chapter 6 specifically noted mimicry as a theme in describing how the adaptive agent differed from the static agent, echoing prior research which warned against the inappropriateness of human mimicry in the design of speech agents (Aylett, Cowan, et al., 2019).

But if a misalignment between style and role is seen as appropriate, and a style tightly aligned to how humans carry out a role is likewise inappropriate, how then can appropriate speech agent design be achieved? One solution may be to aim to carry out the functions of a role with an explicitly non-human style, such as by using robotic speech synthesis or an animal-like physical form (Moore, 2017) or else by using sound other than human speech, such as synthesised non-human speech (Le Maguer & Cowan, 2021), audio icons (Dingler et al., 2008), or music and sound effects (Aylett, Cowan, et al., 2019). Further study is needed to determine if human mimicry is the source of the perceived inappropriateness of the adaptive agent in Chapter 6. This can be achieved by a 2×2 designed replication comparing the agents from Chapter 6 to simulated human interrupters who follow the exact same speech patterns but using natural human speech. In any case, while this study did not demonstrate benefits of human-inspired speech agent design for proactive agents, it adds to a body of research questioning the appropriateness of human-likeness as a speech agent design goal altogether.

7.2.5 Individual differences in interruption preferences

An insight that united studies from each of Chapters 3, 4, 5, and 6 is the high degree of individual differences between people's preferences for spoken communication. In Chapters 3 and 4, this manifested in the extreme distribution of access ritual usage, with some participants using access rituals practically every time they interrupted and most others never using them at all. Likewise, differences in interruption strategies, such as seeking to either communicate calmness or to communicate urgency during urgent Tetris trials, demonstrated that human speech patterns are highly sensitive to the individual speaker. In Chapter 5, even without the pressure of interrupting in real-time during an ongoing Tetris game, participants varied greatly in when they chose as the best moment to interrupt a Tetris game, with 8.5 second Tetris video clips yielding an average of more than 4 clusters of best moments per clip. Finally, free responses around the differences between agents in Chapter 6 highlighted a wide variety of

preferences, with participants identifying different agents as more polite, more clear, more human-like, or more appropriate. This wide variance in behaviours, strategies, and preferences within spoken communication echoes research on preferences for different sorts of speech agents, which likewise observed large individual difference in preferences between more functional and more social agents (Völkel, Kempf, et al., 2020). This makes the challenge of designing speech agents all the more difficult, as what it means to design human-like speech becomes highly dependent on the human whom that speech is like, and suiting the preferences of a user requires an understanding of the individual user's preferences and personalising an interaction to suit them. This sort of personalisation may require approaches to better model users at an individual level, such as personality measurements (Völkel, Schödel, et al., 2020), longitudinal conversation logging (Bentley et al., 2018), and co-design with users (Woodward et al., 2018). Insofar as speech behaviour and preferences vary so greatly, one-size-fits-all models for speech agents and their users will insufficiently reflect these individual differences.

Chapters 3 and 4 indicated large individual differences in people's tendencies to use access rituals when interrupting a partner and Chapter 6, in which an adaptive proactive agent sometimes used access rituals when interrupting a human found large individual differences in which agent was seen as more appropriate in its communication style. The use of access rituals specifically draws attention to the social role that a proactive speech agent takes when interrupting a person as compared to the role than an unfamiliar other person takes when interrupting. Prior work on access rituals have noted that the acceptability of requesting and of either granting or withholding access is closely related to power dynamics between the two parties (Goffman, 1971; Hutte et al., 1972). Access rituals categorised in Chapters 3 and 4 were based on access rituals observed in studies of peers in which power was seen as equal between parties (Krivonos & Knapp, 1975) as participants were informed that the Tetris player was another research participant like themselves rather than a member of the research team who may have been viewed as hierarchically above them in terms of power dynamics. Applying the access rituals used by participants in these studies to the design of a proactive agent took a neutral stance on how participants in Chapter 6 might view the agent vis-à-vis the Tetris player in that study in terms of power dynamics, implicitly assuming that this was a neutral relationship. There is

evidence for both the notion that people see speech agents as subordinates (Luger & Sellen, 2016) or as peers (Purinton et al., 2017), either of which may come with different social expectations for what sort of speech is appropriate. Future work that aims to use human-human dialogue as an inspiration for the design of human-machine dialogue should carefully consider the power relationships within the human-human speech it observes and consider whether it is congruent with the power relationship between the agent that it seeks to inform and its user. This is all the more crucial as future research aims to cast speech agents in a variety of roles including as leaders, subordinates, and friends (McMillan & Jaber, 2021).

7.2.6 Limitations

In considering the empirical, conceptual, and constructivist contributions of this thesis, a number of limitations must be acknowledged. First, while this thesis focuses on Tetris as an example of a complex task in order to consider complex tasks generally, it must be noted that the stakes for Tetris are low. Although some work has highlighted gamification as a means of increasing participant focus on a low-stakes task (Landesberger et al., 2020a, 2020b; McFarlane, 1997), it is not clear how the interruption behaviours people engage in or see as appropriate when an agent engages in may differ between Tetris and high stakes tasks like driving or cooking, in which mistakes can be quite costly. Further research around the behaviours that people engage in when interrupting more dangerous tasks with speech can shed further light on the generalisability of these findings and illustrate differences between high and low stakes tasks as targets of interruption.

Participants throughout this thesis varied in their experience with Tetris as a task, which may contribute to the variance of behaviours and preferences demonstrated in each study. This choice was intentional, as findings around the behaviour of Tetris experts would only be applicable to the design of a speech agent which could model Tetris to the same ability as an expert. Given the difficulty prior research has demonstrated in modelling driving as an interruptible task (Semmens et al., 2019), this thesis did not aim to rely on expert-level task modelling for insights toward the design of proactive speech agents. That said, insofar as Tetris experts and non-experts understand the structure of Tetris as a task quite differently (Lindstedt & Gray, 2019), differences between judgements of interruption timings in particular may be influenced

by levels of Tetris expertise across all studies. Further exploration of this topic using expert participants may therefore yield different categorisations of interruptible windows of Tetris games and produce different perceptions of agents which are sensitive to these windows.

Conversely, Tetris skill variance in this thesis may be less varied than distributions of expertise in other, more complex domains. That is to say, while even participants who were novices in Tetris may have been able to make informed choices about when a good moment to interrupt the game might be, particularly in Chapter 5 in which there was no time pressure on this judgement, the same might not be true for a highly complex task such as operating specialist machinery. As such, while methods used for interruption elicitation demonstrated in Chapters 3 and 4 and methods for interruptibility categorisation demonstrated in Chapter 5 are presented as flexible to other prerecorded complex tasks beyond Tetris, there may be tasks for which full methodological replication is not suitable, as a general sample of participants would be unable to produce meaningful data. This makes elicitation of interruptions for such tasks more difficult, as restrictions to a participant sample may yield sample with less variance in communication behaviour than the participants from this thesis. Future work on interruptions of complex task must therefore be sensitive to balancing task expertise with sample diversity, viewing the present work as seeking high variance in communication at the expense of expertise in Tetris.

One of the central variables used throughout this study to elicit adaptation in interruption strategies was the cue of interruption urgency. In considering how participants reacted to this cue, it is critical to consider the way urgency was operationalised throughout this study. In Chapters 3 and 4, participants were told that urgent trials had a greater impact on the final score they would receive in the evaluation of their task performance, which was based on their alleged partner's rating of the extent to which the participant asked a question at a good moment and the extent to which the question did not distract the Tetris player. In this way, urgency was an entirely explicitly manipulated variable, unrelated to the Tetris task and to the content of the question. In Chapter 6, for each proactive agent condition, half of the video clips of Tetris play were arbitrarily labelled as urgent, likewise unrelated to either the Tetris task or to the content of the interrupting question. In that study, participants were told that the Tetris player's performance would be rated on only urgent trials, and

that urgency was known only to the agent but not the player. Urgency was not investigated in Chapter 5 as participants were instructed to select the best moment for interruption across a Tetris segment with no time constraint. In this way, urgency was a variable which was explicitly known by the interrupter and related to the costliness of disruptions (i.e. the interrupting participant would be judged more harshly for disruption in urgent trials in Chapters 3 and 4, and the Tetris player would be judged only on their gameplay in trials where interruptions were urgent in Chapter 6). As such, urgency was not confounded with interruption relevance nor did it need to be inferred by participants from indirect cues. Still, this operationalisation may have led to subtle differences between participants who interpreted urgency differently. Participants may have forgotten the way urgency was operationalised in any of these studies and assumed that interruptions in these conditions must come as soon as possible. While this is a valid strategy for minimising disruption (as discussed in Chapters 3 and 4), participants were not instructed to interrupt faster in these conditions. Future studies may seek to reduce this ambiguity by operationalising urgency differently, such as by making urgency a cue inferred indirectly via a countdown timer or instead calling trials of this type "safety critical" trials. Insofar as this operationalisation of urgency only captures a particular operationalisation of interruption urgency - in this case, the costliness of interruption as known only to the interrupter - it is not clear to what extent interruption behaviour and preferences observed in these studies would overlap with interruptions in which urgency was defined in some other way.

The explicit labelling of urgency in Chapters 3, 4, and 6 may likewise have made this cue more salient to participants in those studies. Similarly, in Chapters 4 and 5, participants were asked after each trial to rate the complexity of the Tetris game they saw as well as the difficulty they had in selecting a moment to interrupt and their confidence in selecting a good moment. In Chapter 6, participants were asked after each trial to rate whether interruptions came at a good moment and whether they were asked in a disruptive way. In each of these studies, there is a risk that directly drawing attention to cues such as Tetris game complexity or urgency or to interruption features like selecting a good moment may have exaggerated the effects of those cues or the attention on those behaviours through priming effects (Bargh et al., 2001). Insofar as the questions asked in Chapters 4 and 5 were manipulation checks of the Tetris task difficulty variable, future work may avoid priming by reusing experimental

materials from this thesis or by separating manipulation check pilot studies from the experiment proper. Similarly, just as exploring different operationalisations of urgency can investigate the external validity of findings around urgency from this study, they may also be useful as a means to determine whether the effects observed in Chapters 4 and 5 were due to urgency per se or the result of urgency being a salient cue which was explicitly expressed to participants.

Reflecting generally on the operationalisation of urgency in this study, it is important to consider how different participants within and between studies of this thesis may consider the variable, and how those conceptualisations may differ too from how the construct is understood in other studies of urgency. In emotional psychology research, Fiske describes the so-called "lexical fallacy" by which the labels used to describe psychological constructs may be less stable across time and cultures than the constructs themselves due to differences in language and usage (Fiske, 2020). This phenomenon is not confined to the study of emotions however, as constructs like urgency likewise have vernacular meanings which may be understood differently by participants and by researchers in different contexts. As such, explicit operationalisation of urgency as in this thesis or of any construct in human behavioural research is sensitive to threats of generalisability borne of this ambiguity. As such, it is critical that applications of research like the studies described here are reflective in their own operationalisation of the same constructs. For example, while the present work may support the design decision for a humanlike proactive speech agent to behave differently when speech is urgent, the appropriateness of that decision may depend on the user's understanding of urgency. As such, personalisation of agents demands not only co-creation between user and agent of the rules for which variables yield which behaviours, but also a more fundamental co-creation of how variables like urgency are understood by the dyad.

The particular participants who took part in this study may likewise be a source of limitation to generalisability of findings as a result of their participation being through online crowdworker platforms. The decision to involve crowdworker participants was in part to collect data quickly and with a sample more diverse in terms of age and socioeconomic status than a typical university student participant sample (Naderi & Naderi, 2018). Likewise, as data was collected between 2020 and 2022, the decision to invite crowdworkers to be participants was partly a constraint of the Covid-19 pan-

demically. While prior research on time-sensitive interruption studies has demonstrated the viability of having crowdworkers as research participants (Gould et al., 2015), it may nonetheless be the case that participant behaviour in these studies would have been different had participants been accessed in-person. Likewise, while care was taken to ensure that participants understood experimental tasks, remained engaged with them throughout the experiments, and data which was not suitable was removed from analysis, crowdworker experimental environments are inherently less controlled than a laboratory. As such, it is possible that unsystematic variance caused by crowdworkers' home environments may have biased experimental results in this thesis in unknown ways. For this reason, replication of methods from this thesis across different samples of participants including in-person lab participants would be welcome in ensuring the validity of findings.

Just as having crowdworkers as participants may be a limitation of this thesis, so too might be the decision to invite different crowdworkers to each study. Chapters 3 and 4 largely followed the same method and used the same deception around the identity of the Tetris player, so these studies would not have benefited from having the same participants. Chapter 5 however asked participants to choose the best moment to interrupt a variety of Tetris games which were similar to the games presented in Chapter 4. In this case, it may have been beneficial to access the same participants again, so that participants ratings of interruptible moments could be compared against their own interruption behaviour during realtime interruptions. Likewise, insofar as different participants expressed different strategies to interrupt in Chapters 3 and 4, and then participants commented on the use of a variety of interruption strategies in Chapter 6, accessing the same participants for Chapter 6 as those in Chapters 3 and/or 4 could have enabled analysis of whether participants have a preference for the same interruption behaviours that they themselves use. While accessing the same participants across studies could have been beneficial in terms of additional avenues for analysis, it likely would have added significant time costs to data collection in that the population of potential participants for later studies would be limited to only the relatively few people who participated in previous studies. This dropout risk was seen as too great to justify the potential benefits of accessing the same participants, particularly given the limited access experimenters have to contacting crowdworker participants. Future work may consider combining multiple experiments from this the-

sis into combined sessions to better facilitate comparison of individual participants' data across tasks. Alternatively, researchers who have more longitudinal access to a particular participant pool such as university researchers accessing a student population may consider inviting participants back across sessions to compare speech interruption behaviour with speech interruption preferences within participants.

Unexpected findings in Chapter 6 in regards to the Partner Model Questionnaire may be partly explained by differences between the context for which the questionnaire was developed and the context in which it was used for that study. While the PMQ was developed using interaction experiences with a variety of commercially available speech agents, all of the speech agents used in its development are purely reactive (P. R. Doyle, 2022). That is to say, none of the speech agent interactions that influenced the items and dimensions included in the PMQ involved the proactivity featured by the agents in Chapter 6. As such, dimensions such as reliability may be more critical to people's understanding of a typical, reactive speech agent's conversational capabilities, but this might not be a suitable measurement for proactive agents' capabilities. The area of partner modelling in human-machine dialogue is nascent however, and speech agent users do not yet have experiences of proactive agent conversations to draw on for the development of new measures of partner models in that context. As such, while the PMQ may have limitations in how much insight it gives to people's perceptions of proactive agents, it is unlikely that any other currently existent measure would be more effective. As the areas of proactive speech agents and of partner modelling in human-machine dialogue both mature, measures of conversational capabilities of proactive agents can be better refined to reveal the extent to which partner models of proactive and reactive speech agents differ.

7.3 Conclusion

Speech agents promise a means of interacting with computers by which eyes and hands can remain free to engage in other tasks. While speech agents have become increasingly popular in the last decade, users see them as failing to live up to their high expectations. In part, the gap between expectation and reality is caused by the unidirectional nature of speech agent interactions, by which a person must always speak first, and an agent can only react to their request. Proactivity is therefore a promising

avenue for expanding the functionality of speech agents. The thesis expounded in this work is that the design of proactive agent speech which occurs while a user engaged in another task can be improved by understanding how people use speech to interrupt others. Through mixed-methods and quantitative experiments, this thesis presents solutions to empirical, conceptual, and constructive problems around understanding both human spoken interruptions and interruptions spoken by proactive agents. Specifically, this thesis demonstrates the significant effect of interruption urgency in decreasing the delay before the onset of human speech interruptions, the significant effect of ongoing task difficulty on decreasing the likelihood of the use of access rituals, and it outlines the varied goals and strategies employed by people when they use speech to interrupt. It likewise demonstrates a methodological paradigm for eliciting speech which interrupts an ongoing complex task from people participating in research online and a methodology for identifying interruptible windows in a complex task using data from online participants taking the role of an interrupter. These interruptible windows are used to demonstrate that, while interrupters self-identify a strategy of choosing good moments to begin interruptions, particularly when their interruption is not urgent, their ability to select interruptible moments in real-time interruptions is not significantly affected by urgency.

These findings are then applied to the design of a proactive agent which adapts its speech interruptions in the same ways that human interrupters did. A study comparing this adaptive agent to a static agent with speech designed to follow standard, non-adaptive patterns used by popularly available speech agents demonstrates the significant negative impact of this kind of adaptation on the strength of people's partner models of proactive agents. This finding is explained through examination of partner competence and dependability, a dimension of partner models of machine dialogue partners, in which the adaptive agents is seen as less dependable than the static non-adaptive agent. Mixed-methods data is likewise used to demonstrate that participants found the proactive agent as less appropriate than the static agent in the way it uses speech. These findings are interpreted as an indication that the design of proactive speech agents should not seek to exactly copy human speech interruption patterns, lest they be seen as inappropriately mimicking human speech. Instead, this thesis offers design suggestions around increasing human-likeness through the understanding of people's goals when using speech to interrupt while avoiding copying

their strategies for achieving those goals. Taken together, this thesis provides a basis for better understanding the way people use speech to interrupt ongoing complex tasks and experimentally tests an initial design of a proactive speech agent which incorporates that understanding. In doing so, this thesis asks new questions around human speech interruptions and the design of proactive agents which may be addressed by future work, and provides replicable methods for investigating these questions.

Bibliography

- Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science*, 26(1), 39–83. https://doi.org/10.1207/s15516709cog2601_2
- Amado, S., & Ulupinar, P. (2005). The effects of conversation on attention and peripheral detection: Is talking with a passenger and talking on the cell phone different? *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(6), 383–395. <https://doi.org/10.1016/j.trf.2005.05.001>
- Ammari, T., Kaye, J., Tsai, J. Y., & Bentley, F. (2019). Music, search, and iot: How people (really) use voice assistants. *ACM Trans. Comput. Hum. Interact.*, 26(3), 17–1.
- Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, 16(7), 390–398. <https://doi.org/10.1016/j.tics.2012.05.003>
- Axtell, B., & Munteanu, C. (2021). Tea, earl grey, hot: Designing speech interactions from the imagined ideal of star trek. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Aylett, M. P., Cowan, B. R., & Clark, L. (2019). Siri, Echo and Performance: You have to Suffer Darling. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–10. <https://doi.org/10.1145/3290607.3310422>
- Aylett, M. P., Sutton, S. J., & Vazquez-Alvarez, Y. (2019). The right kind of unnatural: Designing a robot voice. *Proceedings of the 1st International Conference on Conversational User Interfaces*, 1–2. <https://doi.org/10.1145/3342775.3342806>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature reviews neuroscience*, 4(10), 829–839.
- Baddeley, A., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.). Academic Press. [https://doi.org/https://doi.org/10.1016/S0079-7421\(08\)60452-1](https://doi.org/https://doi.org/10.1016/S0079-7421(08)60452-1)

- Bailey, B. P., & Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction*, *14*(4), 1–28. <https://doi.org/10.1145/1314683.1314689>
- Bakdash, J. Z., & Marusich, L. R. (2017). Repeated Measures Correlation. *Frontiers in Psychology*, *8*, 456. <https://doi.org/10.3389/fpsyg.2017.00456>
- Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A. Y., Barndollar, K. A., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of personality and social psychology*, *81* 6, 1014–27.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Becker, S., Doyle, P. R., & Edwards, J. (2022). Embrace your incompetence! designing appropriate cui communication through an ecological approach. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–5.
- Beggiato, M., & Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information [Publisher: Elsevier]. *Transportation research part F: traffic psychology and behaviour*, *18*, 47–57.
- Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the long-term use of smart speaker assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, *2*(3). <https://doi.org/10.1145/3264901>
- Borst, J. P., Taatgen, N. A., & van Rijn, H. (2010). The problem state: A cognitive bottleneck in multitasking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 363–382. <https://doi.org/10.1037/a0018106>
- Borst, J. P., Taatgen, N. A., & van Rijn, H. (2015). What Makes Interruptions Disruptive?: A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2971–2980. <https://doi.org/10.1145/2702123.2702156>
- Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011a). The Role of Beliefs in Lexical Alignment: Evidence from Dialogs with Humans and Computers. *Cognition*, *121*(1), 41–57. <https://doi.org/10.1016/j.cognition.2011.05.011>
- Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011b). The role of beliefs in lexical alignment: Evidence from dialogs with hu-

- mans and computers. *Cognition*, 121(1), 41–57. <https://doi.org/https://doi.org/10.1016/j.cognition.2011.05.011>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology [Publisher: Taylor & Francis (Routledge)]. *Qualitative Research in Psychology*, 3(2). <https://doi.org/10.1191/1478088706qp063oa>
- Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4), 589–597. <https://doi.org/10.1080/2159676X.2019.1628806>
- Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. In *Psychology of learning and motivation* (pp. 301–344). Elsevier.
- Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, 47(3), 191–196. <https://doi.org/10.1037/h0054182>
- Broadbent, D. E. (1958). *Perception and Communication* [Google-Books-ID: ZCOLBQAAQBAJ]. Pergamon.
- Brookhuis, K. A., de Vries, G., & De Waard, D. (1991). The effects of mobile telephoning on driving performance. *Accident Analysis & Prevention*, 23(4), 309–316.
- Brumby, D. P., Cox, A. L., Back, J., & Gould, S. J. J. (2013). Recovering from an interruption: Investigating speed–accuracy trade-offs in task resumption behavior. *Journal of Experimental Psychology: Applied*, 19(2), 95–107. <https://doi.org/10.1037/a0032696>
- Brumby, D. P., Davies, S. C., Janssen, C. P., & Grace, J. J. (2011). Fast or safe?: How performance objectives determine modality output choices while interacting on the move. *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 473. <https://doi.org/10.1145/1978942.1979009>
- Brumby, D. P., Salvucci, D. D., & Howes, A. (2007). Dialing while driving? a bounded rational analysis of concurrent multi-task behavior. *Proceedings of the 8th international conference on cognitive modeling*, 121–126.
- Caird, J. K., Simmons, S. M., Wiley, K., Johnston, K. A., & Horrey, W. J. (2018). Does Talking on a Cell Phone, With a Passenger, or Dialing Affect Driving Performance? An Updated Systematic Review and Meta-Analysis of Experimental Studies. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 60(1), 101–133. <https://doi.org/10.1177/0018720817748145>
- Cairns, P. E., & Cox, A. L. (2008). *Research methods for human-computer interaction*. Cambridge University Press.
- Candello, H., Pinhanez, C., Pichiliani, M., Vasconcelos, M., & Conde, H. (2019). Can direct address affect user engagement with chatbots embodied in physical spaces? *Proceedings of the 1st International Conference on*

- Conversational User Interfaces*, 1–9. <https://doi.org/10.1145/3342775.3342787>
- Card, S. K., & Henderson, A. (1987). A multiple, virtual-workspace interface to support user task switching. *Proceedings of the SIGCHI/GI conference on Human factors in computing systems and graphics interface - CHI '87*, 53–59. <https://doi.org/10.1145/29933.30860>
- Cassell, J. (2007). Body language: Lessons from the near-human. *Genesis Redux*, 346, 374.
- Cha, N., Kim, A., Park, C. Y., Kang, S., Park, M., Lee, J.-G., Lee, S., & Lee, U. (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), 28. <https://doi.org/10.1145/3411810>
- Chalmers, A. F. (2013). *What is this thing called science?* (Fourth edition). Hackett Publishing Company, Inc.
- Charlton, S. G. (2009). Driving while conversing: Cell phones that distract and passengers who react. *Accident Analysis & Prevention*, 41(1), 160–173. <https://doi.org/10.1016/j.aap.2008.10.006>
- Cheepen, C. (1988). *The predictability of informal conversation*. Pinter Publishers.
- Clark, H. H. (2020). Common ground. In *The international encyclopedia of linguistic anthropology* (pp. 1–5). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118786093.iela0064>
- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81. <https://doi.org/10.1016/j.jml.2003.08.004>
- Clark, H. (1996). *Using Language*. Cambridge University Press. <https://books.google.ie/books?id=DiWBGOP-YnoC>
- Clark, L., Doyle, P. R., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., Edwards, J., & R Cowan, B. (2019). The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*, 31(4), 349–371. <https://doi.org/10.1093/iwc/iwz016>
- Clark, L., Munteanu, C., Wade, V., Cowan, B. R., Pantidi, N., Cooney, O., Doyle, P. R., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., & Murad, C. (2019). What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12. <https://doi.org/10.1145/3290605.3300705>
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.
- Corsaro, W. A. (1979). 'we're friends, right?': Children's use of access rituals in a nursery school. *Language in society*, 8(2-3), 315–336.
- Coupland, J. (2003). Small talk: Social functions. *Research on language and social interaction*, 36(1), 1–6.

- Cowan, B. R., & Branigan, H. (2017). They Know as Much as We Do: Knowledge Estimation and Partner Modelling of Artificial Partners. *Proceedings of CogSci '17*, 6.
- Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., & Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human-computer dialogue. *International Journal of Human-Computer Studies*, 83, 27–42. <https://doi.org/10.1016/j.ijhcs.2015.05.008>
- Cowan, B. R., Doyle, P. R., Edwards, J., Garaialde, D., Hayes-Brady, A., Branigan, H. P., Cabral, J., & Clark, L. (2019). What's in an accent?: The impact of accented synthetic speech on lexical choice in human-machine dialogue. *Proceedings of the 1st International Conference on Conversational User Interfaces - CUI '19*, 1–8. <https://doi.org/10.1145/3342775.3342786>
- Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., & Bandeira, N. (2017). " what can i help you with?" infrequent users' experiences of intelligent personal assistants. *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–12.
- Creswell, J. W. (2011). Controversies in mixed methods research. *The Sage handbook of qualitative research*, 4(1), 269–284.
- Crundall, D., Bains, M., Chapman, P., & Underwood, G. (2005). Regulating conversation during driving: A problem for mobile telephones? *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(3), 197–211. <https://doi.org/10.1016/j.trf.2005.01.003>
- Czerwinski, M., Chrisman, S., & Schumacher, B. (1991). The Effects of Warnings and Display Similarity on Interruption in Multitasking Environments. *ACM SIGCHI Bulletin*, 23(4), 38–39. <https://doi.org/10.1145/126729.1056014>
- Dabbish, L., & Kraut, R. E. (2004). Controlling Interruptions: Awareness Displays and Social Motivation for Coordination. *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, 6(3), 10.
- Dabbish, L., Mark, G., & González, V. M. (2011). Why Do I Keep Interrupting Myself?: Environment, Habit and Self-Interruption. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3127–3130.
- De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional science*, 38(2), 105–134.
- Debue, N., & van de Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.01099>
- de Leeuw, J. (2015). Jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47, 1–12.

- DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223–234. <https://doi.org/10.1037/0022-0663.100.1.223>
- Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review, 70*(1), 80–90. <https://doi.org/10.1037/h0039515>
- Dingler, T., Lindsay, J., Walker, B. N., München, L.-M.-U., & Medieninformatik, F. (2008). Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech.
- Doyle, L., Brady, A.-M., & Byrne, G. (2016). An overview of mixed methods research – revisited [Publisher: SAGE Publications Ltd]. *Journal of Research in Nursing, 21*(8), 623–635. <https://doi.org/10.1177/1744987116674257>
- Doyle, P. R. (2022). *The Dimensions and Adaptation of Partner Models in Human-Machine Dialogue* (PhD Thesis). University College Dublin. School of Information and Communication Studies.
- Doyle, P. R., Clark, L., & Cowan, B. R. (2021). What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–14*. <https://doi.org/10.1145/3411764.3445206>
- Doyle, P. R., Edwards, J., Dumbleton, O., Clark, L., & Cowan, B. R. (2019). Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '19, 1–12*. <https://doi.org/10.1145/3338286.3340116>
- Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied, 14*(4), 392–400. <https://doi.org/10.1037/a0013119>
- Dubiel, M., Halvey, M., & Azzopardi, L. (2018). A survey investigating usage of virtual personal assistants. *arXiv preprint arXiv:1807.04606*.
- Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*(2), 283–292. <https://doi.org/10.1037/h0033031>
- Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech communication, 50*(8-9), 630–645.
- Edlund, J., Heldner, M., & Gustafson, J. (2006). Two faces of spoken dialogue systems.
- Edwards, J., Liu, H., Tianyu, Z., Gould, G., Sandy J. J., Clark, L., Doyle, P. R., & Cowan, B. R. (2019). Multitasking with Alexa: How Using Intelligent Personal Assistants Impacts Language-based Primary Task Performance [Accepted]. *Proceedings of the 1st International Conference on Conversational User Interfaces*.

- Escobar-Alvarez, H. D., Johnson, N., Hebble, T., Klingebiel, K., Quintero, S. A., Regenstein, J., & Browning, N. A. (2018). R-advance: Rapid adaptive prediction for vision-based autonomous navigation, control, and evasion. *Journal of Field Robotics, 35*(1), 91–100.
- Escobar-Alvarez, H. D., Ohradzansky, M., Keshavan, J., Ranganathan, B. N., & Humbert, J. S. (2019). Bioinspired approaches for autonomous small-object detection and avoidance. *IEEE Transactions on Robotics, 35*(5), 1220–1232. <https://doi.org/10.1109/TRO.2019.2922472>
- Fargier, R., & Laganaro, M. (2019). Interference in speaking while hearing and vice versa. *Scientific reports, 9*(1), 1–13.
- Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development [Publisher: SAGE Publications Inc]. *International Journal of Qualitative Methods, 5*(1), 80–92. <https://doi.org/10.1177/160940690600500107>
- Finzi, A., & Orlandini, A. (2005). Human-Robot Interaction Through Mixed-Initiative Planning for Rescue and Search Rovers. In S. Bandini & S. Manzoni (Eds.), *AI*IA 2005: Advances in Artificial Intelligence* (pp. 483–494). Springer. https://doi.org/10.1007/11558590_49
- Fiske, A. P. (2020). The lexical fallacy in emotion research: Mistaking vernacular words for psychological entities [Place: US Publisher: American Psychological Association]. *Psychological Review, 127*(1), 95–113. <https://doi.org/10.1037/rev0000174>
- Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology, 6*. <https://doi.org/10.3389/fpsyg.2015.00751>
- Ghafurian, M., Lakatos, G., Tao, Z., & Dautenhahn, K. (2020). Design and evaluation of affective expressions of a zoomorphic robot. In A. R. Wagner, D. Feil-Seifer, K. S. Haring, S. Rossi, T. Williams, H. He, & S. Sam Ge (Eds.), *Social robotics* (pp. 1–12). Springer International Publishing.
- Gilmartin, E., Collery, M., Su, K., Huang, Y., Elias, C., Cowan, B. R., & Campbell, N. (2017). Social talk: Making conversation with people and machine. *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents - ISIAA 2017, 31–32*. <https://doi.org/10.1145/3139491.3139494>
- Goffman, E. (1971). *Relations in public: Microstudies of the public order* [OCLC: 699515377]. Basic Books.
- González, V. M., & Mark, G. (2004). "Constant, constant, multi-tasking craziness": Managing multiple working spheres. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 113–120*. <https://doi.org/10.1145/985692.985707>
- Gould, S. J., Brumby, D. P., & Cox, A. L. (2013). What does it mean for an interruption to be relevant? an investigation of relevance as a memory

- effect. *Proceedings of the human factors and ergonomics society annual meeting*, 57(1), 149–153.
- Gould, S. J., Cox, A. L., Brumby, D. P., & Wickersham, A. (2016). Now Check Your Input: Brief Task Lockouts Encourage Checking, Longer Lockouts Encourage Task Switching. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 3311–3323. <https://doi.org/10.1145/2858036.2858067>
- Gould, S. J., Cox, A. L., Brumby, D. P., & Wiseman, S. (2015). Home is where the lab is: A comparison of online and lab data from a time-sensitive study of interruption. *Human Computation*, 2(1), 45–67.
- Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, 113(3), 461–482. <https://doi.org/10.1037/0033-295X.113.3.461>
- Green, P., & MacLeod, C. J. (2016). Simr: An r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493–498.
- Gugerty, L., Rakauskas, M., & Brooks, J. (2004). Effects of remote and in-person verbal interactions on verbalization rates and attention to dynamic spatial scenes. *Accident Analysis & Prevention*, 36(6), 1029–1043. <https://doi.org/10.1016/j.aap.2003.12.002>
- Hellier, E., Edworthy, J., Weedon, B., Walters, K., & Adams, A. (2002). The Perceived Urgency of Speech Warnings: Semantics versus Acoustics. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 44(1), 1–17. <https://doi.org/10.1518/0018720024494810>
- Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1), 258–271. <https://doi.org/10.1016/j.csda.2006.11.025>
- Hennig, C. (2020). *Fpc: Flexible Procedures for Clustering*. <https://CRAN.R-project.org/package=fpc>
- Ho, C.-Y., Nikolic, M. I., Waters, M. J., & Sarter, N. B. (2004). Not now! Supporting interruption management by indicating the modality and urgency of pending tasks. *Human Factors*, 46(3), 399–409.
- Holler, J., Kendrick, K. H., Casillas, M., & Levinson, S. C. (Eds.). (2016). *Turn-Taking in Human Communicative Interaction*. Frontiers Media SA. <https://doi.org/10.3389/978-2-88919-825-2>
- Horrey, W. J., & Lesch, M. F. (2009). Driver-initiated distractions: Examining strategic adaptation for in-vehicle task initiation. *Accident Analysis & Prevention*, 41(1), 115–122. <https://doi.org/10.1016/j.aap.2008.10.008>
- Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, 47(4), 589–606. [https://doi.org/https://doi.org/10.1016/S0749-596X\(02\)00019-0](https://doi.org/https://doi.org/10.1016/S0749-596X(02)00019-0)

- Horvitz, E. (1999). Principles of mixed-initiative user interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*, 159–166. <https://doi.org/10.1145/302979.303030>
- Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., & Yang, J. (2003). Predicting human interruptibility with sensors: A Wizard of Oz feasibility study. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 257–264. <https://doi.org/10.1145/642611.642657>
- Hutte, H. A., et al. (1972). The perception of door-knocks in terms of authority and urgency. *European Journal of Social Psychology*, 2(1), 98–99.
- Iqbal, S. T., & Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. *CHI '05 extended abstracts on Human factors in computing systems - CHI '05*, 1489. <https://doi.org/10.1145/1056808.1056948>
- Iqbal, S. T., & Bailey, B. P. (2008). Effects of intelligent notification management on users and their tasks. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 93–102.
- Iqbal, S. T., & Bailey, B. P. (2010). Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Transactions on Computer-Human Interaction*, 17(4), 1–28. <https://doi.org/10.1145/1879831.1879833>
- Isbell, C. L., & Pierce, J. S. (2005). An IP continuum for adaptive interface design. *Proc. of HCI International*, 10.
- Jacucci, G., Oulasvirta, A., Salovaara, A., & Sarvas, R. (2005). Supporting the shared experience of spectators through mobile group media. *Proceedings of the 2005 ACM International Conference on Supporting Group Work*, 207–216.
- Janssen, C. P., & Brumby, D. P. (2010). Strategic adaptation to performance objectives in a dual-task setting. *Cognitive science*, 34(8), 1548–1560.
- Janssen, C. P., Brumby, D. P., & Garnett, R. (2010). Natural Break Points: Utilizing Motor Cues when Multitasking [Publisher: SAGE Publications Inc]. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54(4), 482–486. <https://doi.org/10.1177/154193121005400444>
- Janssen, C. P., Brumby, D. P., & Garnett, R. (2012). Natural Break Points: The Influence of Priorities and Cognitive and Motor Cues on Dual-Task Interleaving. *Journal of Cognitive Engineering and Decision Making*, 6(1), 5–29. <https://doi.org/10.1177/1555343411432339>
- Janssen, C. P., Gould, S. J., Li, S. Y., Brumby, D. P., & Cox, A. L. (2015). Integrating knowledge of multitasking and interruptions across different perspectives and research methods. *International Journal of Human-Computer Studies*, 79, 1–5. <https://doi.org/10.1016/j.ijhcs.2015.03.002>
- Janssen, C. P., Iqbal, S. T., & Ju, Y.-C. (2014). Sharing a driver's context with a caller via continuous audio cues to increase awareness about driver

- state. *Journal of Experimental Psychology: Applied*, 20(3), 270–284. <https://doi.org/10.1037/xap0000020>
- Janssen, C. P., Iqbal, S. T., Kun, A. L., & Donker, S. F. (2019). Interrupted by my car? Implications of interruption and interleaving research for automated vehicles. *International Journal of Human-Computer Studies*, 40.
- Jokinen, J. P. P., Kujala, T., & Oulasvirta, A. (2021). Multitasking in Driving as Optimal Adaptation Under Uncertainty [Publisher: SAGE Publications Inc]. *Human Factors*, 63(8), 1324–1341. <https://doi.org/10.1177/0018720820927687>
- Jonez, S. (2013). Her.
- Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.
- Kalyuga, S. (2011). Cognitive Load Theory: How Many Types of Load Does It Really Need? *Educational Psychology Review*, 23(1), 1–19. <https://doi.org/10.1007/s10648-010-9150-7>
- Katidioti, I., Borst, J. P., & Taatgen, N. A. (2014). What happens when we switch tasks: Pupil dilation in multitasking. *Journal of Experimental Psychology: Applied*, 20(4), 380–396. <https://doi.org/10.1037/xap0000031>
- Kieras, D., Meyer, D., Ballas, J., & Lauber, E. (2000). *Modern Computational Perspectives on Executive Mental Processes and Cognitive Control: Where to from Here?*
- Kim, S., Chun, J., & Dey, A. K. (2015). Sensors know when to interrupt you in the car: Detecting driver interruptibility through monitoring of peripheral interactions. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 487–496. <https://doi.org/10.1145/2702123.2702409>
- Kinsella, B. (2022). *The Rise and Stall of the U.S. Smart Speaker Market* (Market Report). Voicebot AI.
- Knapp, M. L., Hart, R. P., Friedrich, G. W., & Shulman, G. M. (1973). The rhetoric of goodbye: Verbal and nonverbal correlates of human leave-taking. *Communications Monographs*, 40(3), 182–198.
- Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advanced Research in Computer Science and Management Studies*, 1(6), 90–95.
- Krivosos, P. D., & Knapp, M. L. (1975). Initiating communication: What do you say when you say hello? *Central States Speech Journal*, 26(2), 115–125. <https://doi.org/10.1080/10510977509367829>
- Kubose, T. T., Bock, K., Dell, G. S., Garnsey, S. M., Kramer, A. F., & Mayhugh, J. (2006). The effects of speech production and speech comprehension on simulated driving performance. *Applied Cognitive Psychology*, 20(1), 43–63. <https://doi.org/10.1002/acp.1164>
- Kubrick, S. (1968). 2001: A Space Odyssey.
- Kuhn, T. S., & Hacking, I. (1962). *The structure of scientific revolutions* (Fourth edition). The University of Chicago Press.

- Kun, A., Miller, W., & Lenharth, W. (2004). Computers in police cruisers. *IEEE Pervasive Computing*, 3(4), 34–41. <https://doi.org/10.1109/MPRV.2004.3>
- Kun, A. L., Shyrokov, A., & Heeman, P. A. (2013). Interactions between human-human multi-threaded dialogues and driving. *Personal and Ubiquitous Computing*, 17(5), 825–834. <https://doi.org/10.1007/s00779-012-0518-1>
- Landesberger, J., Ehrlich, U., & Minker, W. (2020a). "What is it?" How to Collect Urgent Utterances using a Gamification Approach. *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 19–22. <https://doi.org/10.1145/3409251.3411713>
- Landesberger, J., Ehrlich, U., & Minker, W. (2020b). Do the Urgent Things first! - Detecting Urgency in Spoken Utterances based on Acoustic Features. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 53–58. <https://doi.org/10.1145/3386392.3397598>
- Large, D. R., Clark, L., Quandt, A., Burnett, G., & Skrypchuk, L. (2017). Steering the conversation: A linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied Ergonomics*, 63, 53–61. <https://doi.org/10.1016/j.apergo.2017.04.003>
- Laudan, L. (1978). *Progress and its problems: Towards a theory of scientific growth* (1st paperback print). Univ. of Calif. Press.
- Le Maguer, S., & Cowan, B. R. (2021). Synthesizing a human-like voice is the easy way. *CUI 2021 - 3rd Conference on Conversational User Interfaces*, 1–3. <https://doi.org/10.1145/3469595.3469614>
- Lee, K.-M., & Nass, C. (2005). Social-psychological origins of feelings of presence: Creating social presence with machine-generated voices. *Media Psychology*, 7(1), 31–45.
- Lee, S. C., Jeong, S., Wang, M., Hock, P., Baumann, M., & Jeon, M. (2021). "To Go or Not To Go? That is the Question": When In-Vehicle Agents Argue with Each Other. *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 223–224. <https://doi.org/10.1145/3473682.3481876>
- Leviathan, Y., & Matias, Y. (2018). Google duplex: An ai system for accomplishing real-world tasks over the phone.
- Li, S. Y., Blandford, A., Cairns, P., & Young, R. M. (2008). The effect of interruptions on postcompletion and other procedural errors: An account based on the activation-based goal memory model. *Journal of Experimental Psychology: Applied*, 14(4), 314.
- Lindstedt, J. K., & Gray, W. D. (2019). Distinguishing experts from novices by the Mind's Hand and Mind's Eye. *Cognitive Psychology*, 109, 1–25. <https://doi.org/10.1016/j.cogpsych.2018.11.003>

- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137. <https://doi.org/10.1109/TIT.1982.1056489>
- Lohse, M., Hanheide, M., Wrede, B., Walters, M. L., Koay, K. L., Syrdal, D. S., Green, A., Huttenrauch, H., Dautenhahn, K., Sagerer, G., & Severinson-Eklundh, K. (2008). Evaluating extrovert and introvert behaviour of a domestic robot - a video study. *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 488–493. <https://doi.org/10.1109/ROMAN.2008.4600714>
- Luger, E., & Sellen, A. (2016). "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 5286–5297. <https://doi.org/10.1145/2858036.2858288>
- Maciej, J., Nitsch, M., & Vollrath, M. (2011). Conversing while driving: The importance of visual information for conversation modulation. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(6), 512–524. <https://doi.org/10.1016/j.trf.2011.05.001>
- Mark, G., Iqbal, S., Czerwinski, M., & Johns, P. (2015). Focused, Aroused, but so Distractible: Temporal Perspectives on Multitasking and Communications. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 903–916. <https://doi.org/10.1145/2675133.2675221>
- Mark, G., Volda, S., & Cardello, A. (2012). A pace not dictated by electrons: An empirical study of work without email. *Proceedings of the SIGCHI conference on human factors in computing systems*, 555–564. <https://doi.org/10.1145/2207676.2207754>
- Martelaro, N., Teevan, J., & Iqbal, S. T. (2019). An Exploration of Speech-Based Productivity Support in the Car. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12. <https://doi.org/10.1145/3290605.3300494>
- Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars* [OCLC: 1018457393]. Cambridge University Press.
- McFarlane, D. C. (1997). *Interruption of People in Human-Computer Interaction: A General Unifying Definition of Human Interruption and Taxonomy*: (tech. rep.). Defense Technical Information Center. Fort Belvoir, VA. <https://doi.org/10.21236/ADA333587>
- McFarlane, D. C. (1999). Coordinating the Interruption of People in Human-Computer Interaction. *INTERACT*.
- McFarlane, D. C. (2002). Comparison of Four Primary Methods for Coordinating the Interruption of People in Human-Computer Interaction. *Human-Computer Interaction*, 17(1), 63–139. https://doi.org/10.1207/S15327051HCI1701_2

- McFarlane, D. C., & Latorella, K. A. (2002). The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *Human-Computer Interaction, 17*(1), 1–61. https://doi.org/10.1207/S15327051HCI1701_1
- McKinney, F. (1935). Studies in the retention of interrupted learning activities. *Journal of Comparative Psychology, 19*(2), 265–296. <https://doi.org/10.1037/h0056005>
- McMillan, D., & Jaber, R. (2021). Leaving the Butler Behind: The Future of Role Reproduction in CUI. *CUI 2021 - 3rd Conference on Conversational User Interfaces*, 1–4. <https://doi.org/10.1145/3469595.3469606>
- McTear, M. F., Callejas, Z., & Griol, D. (2016). *The conversational interface* (Vol. 6). Springer.
- Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092. <https://doi.org/10.1016/j.jml.2020.104092>
- Miyata, Y., & Norman, D. A. (1986). Psychological issues in support of multiple activities. In *User Centered System Design: New Perspectives on Human-Computer Interaction* (pp. 265–284). Lawrence Erlbaum Associates, Hillsdale, NJ.
- Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied, 14*(4), 299–313. <https://doi.org/10.1037/a0014402>
- Moore, R. K. (2017). Appropriate Voices for Artefacts: Some Key Insights. *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, 5.
- Moore, R. K., & Morris, A. (1992). Experiences collecting genuine spoken enquiries using woz techniques. *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.
- Moray, N. (1967). Where is capacity limited? A survey and a model [Place: Netherlands Publisher: Elsevier Science]. *Acta Psychologica, 27*, 84–92. [https://doi.org/10.1016/0001-6918\(67\)90048-0](https://doi.org/10.1016/0001-6918(67)90048-0)
- Moray, N. (1959). Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. *Quarterly Journal of Experimental Psychology, 11*(1), 56–60. <https://doi.org/10.1080/17470215908416289>
- Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field] [Publisher: IEEE]. *IEEE Robotics & automation magazine, 19*(2), 98–100.
- Müller, H., Sedley, A., & Ferrall-Nunge, E. (2014). Survey Research in HCI. In J. S. Olson & W. A. Kellogg (Eds.), *Ways of Knowing in HCI* (pp. 229–266). Springer. https://doi.org/10.1007/978-1-4939-0378-8_10
- Naderi, B., & Naderi, B. (2018). Who are the crowdworkers? *Motivation of workers on microtask crowdsourcing platforms*, 17–27.
- Nagaraju, D., Ansah, A., Ch, N. A. N., Mills, C., Janssen, C. P., Shaer, O., & Kun, A. L. (2021). How Will Drivers Take Back Control in Automated Vehi-

- cles? A Driving Simulator Test of an Interleaving Framework. *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 20–27. <https://doi.org/10.1145/3409118.3475128>
- Nagel, K. S., Hudson, J. M., & Abowd, G. D. (2004). Predictors of availability in home life context-mediated communication. *Proceedings of the 2004 ACM conference on Computer supported cooperative work - CSCW '04*, 497. <https://doi.org/10.1145/1031607.1031689>
- Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? responses to computer-based interviewing systems1. *Journal of Applied Social Psychology*, 29(5), 1093–1109. <https://doi.org/https://doi.org/10.1111/j.1559-1816.1999.tb00142.x>
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 72–78.
- Newell, A. (1973). You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium. In *Visual Information Processing* (pp. 283–308). Elsevier. <https://doi.org/10.1016/B978-0-12-170150-5.50012-3>
- Nijboer, M., Taatgen, N. A., Brands, A., Borst, J. P., & van Rijn, H. (2013). Decision Making in Concurrent Multitasking: Do People Adapt to Task Interference? (K. Watanabe, Ed.). *PLoS ONE*, 8(11), e79583. <https://doi.org/10.1371/journal.pone.0079583>
- Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological review*, 75(6), 522.
- Norman, D. A. (1983). Some observations on mental models. In *Mental models* (pp. 15–22). Psychology Press.
- Nunes, L., & Recarte, M. A. (2002). Cognitive demands of hands-free-phone conversation while driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 5(2), 133–144. [https://doi.org/10.1016/S1369-8478\(02\)00012-8](https://doi.org/10.1016/S1369-8478(02)00012-8)
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed). McGraw-Hill.
- Olson, C., & Kemery, K. (2019). *2019 Voice report: Consumer adoption of voice technology and digital assistants* (tech. rep.). Microsoft.
- Oulasvirta, A., & Hornbæk, K. (2016). HCI Research as Problem-Solving. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4956–4967. <https://doi.org/10.1145/2858036.2858283>
- Parviainen, E., & Søndergaard, M. L. J. (2020). Experiential Qualities of Whispering with Voice Assistants. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3313831.3376187>

- Popper, K. (1959). *The Logic of Scientific Discovery*. Psychology Press.
- Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. *proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12.
- Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). " alexa is my new bff" social roles, user satisfaction, and personification of the amazon echo. *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, 2853–2859.
- R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ratwani, R. M., Andrews, A. E., Sousk, J. D., & Trafton, J. G. (2008). The effect of interruption modality on primary task resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52, 393–397.
- Reeves, S. (2019). Conversation considered harmful? *Proceedings of the 1st International Conference on Conversational User Interfaces*, 1–3. <https://doi.org/10.1145/3342775.3342796>
- Riest, C., Jorschick, A. B., & de Ruiter, J. P. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00089>
- Rigby, J. M., Brumby, D. P., Gould, S. J., & Cox, A. L. (2017). Media Multitasking at Home: A Video Observation Study of Concurrent TV and Mobile Device Usage. *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video - TVX '17*, 3–10. <https://doi.org/10.1145/3077548.3077560>
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A Simplest Systematics for the Organization of Turn Taking for Conversation. In J. Schenkein (Ed.), *Studies in the Organization of Conversational Interaction* (pp. 7–55). Academic Press. <https://doi.org/10.1016/B978-0-12-623550-0.50008-2>
- Salovaara, A., Oulasvirta, A., & Jacucci, G. (2017). Evaluation of Prototypes and the Problem of Possible Futures. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2064–2077. <https://doi.org/10.1145/3025453.3025658>
- Salvucci, D. D., & Beltowska, J. (2008). Effects of Memory Rehearsal on Driver Performance: Experiment and Theoretical Account. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(5), 834–844. <https://doi.org/10.1518/001872008X354200>
- Salvucci, D. D. (2005). A Multitasking General Executive for Compound Continuous Tasks. *Cognitive Science*, 29(3), 457–492. https://doi.org/10.1207/s15516709cog0000_19
- Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, 115(1), 101–130. <https://doi.org/10.1037/0033-295X.115.1.101>

- Salvucci, D. D., & Taatgen, N. A. (2014). *The multitasking mind*. Oxford University Press.
- Salvucci, D. D., Taatgen, N. A., & Borst, J. P. (2009). Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, 1819. <https://doi.org/10.1145/1518701.1518981>
- Schneider, M., & Kiesler, S. (2005). Calling while driving: Effects of providing remote traffic context. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 561–569. <https://doi.org/10.1145/1054972.1055050>
- Semmens, R., Martelaro, N., Kaveti, P., Stent, S., & Ju, W. (2019). Is Now A Good Time?: An Empirical Study of Vehicle-Driver Communication Timing. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–12. <https://doi.org/10.1145/3290605.3300867>
- Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. interface agents. *Interactions*, 4(6), 42–61. <https://doi.org/10.1145/267505.267514>
- Sin, J., Chen, D., Threatt, J. G., Gorham, A., & Munteanu, C. (2022). Does Alexa Live Up to the Hype? Contrasting Expectations from Mass Media Narratives and Older Adults' Hands-on Experiences of Voice Interfaces. *4th Conference on Conversational User Interfaces*, 1–9. <https://doi.org/10.1145/3543829.3543841>
- Smith, B. (2018). Generalizability in qualitative research: Misunderstandings, opportunities and recommendations for the sport and exercise sciences. *Qualitative Research in Sport, Exercise and Health*, 10(1), 137–149. <https://doi.org/10.1080/2159676X.2017.1393221>
- Song, M., & Zhong, H. (2020). Efficient weighted univariate clustering maps outstanding dysregulated genomic zones in human cancers (R. Schwartz, Ed.). *Bioinformatics*, 36(20), 5027–5036. <https://doi.org/10.1093/bioinformatics/btaa613>
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592.
- Sutton, S. J., Foulkes, P., Kirk, D., & Lawson, S. (2019). Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–14.
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Load During Problem Solving: Effects on Learning. *Educational Psychology Review*, 10(3), 251–296. <https://doi.org/10.1023/A:1022193728205>

- Taylor, M. M., Lindsay, P. H., & Forbes, S. M. (1967). Quantification of shared capacity processing in auditory and visual discrimination. *Acta Psychologica*, 27, 223–229. [https://doi.org/10.1016/0001-6918\(67\)99000-2](https://doi.org/10.1016/0001-6918(67)99000-2)
- Terlecki, M. S., Newcombe, N. S., & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied Cognitive Psychology*, 22(7), 996–1013. <https://doi.org/10.1002/acp.1420>
- Torggler, A., Edwards, J., & Wintersberger, P. (2022). Beyond the Halo: Investigation of Trust and Functional Specificity in Automated Driving with Conversational Agents. *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 195–203. <https://doi.org/10.1145/3543174.3546834>
- Trafton, J., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, 58(5), 583–603. [https://doi.org/10.1016/S1071-5819\(03\)00023-5](https://doi.org/10.1016/S1071-5819(03)00023-5)
- Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review*, 76(3), 282–299. <https://doi.org/10.1037/h0027242>
- Treisman, A. M. (1960). Contextual cues in selective listening [Publisher: Routledge _eprint: <https://doi.org/10.1080/17470216008416732>]. *Quarterly Journal of Experimental Psychology*, 12(4), 242–248. <https://doi.org/10.1080/17470216008416732>
- van der Heiden, R. M., Iqbal, S. T., & Janssen, C. P. (2017). Priming Drivers before Handover in Semi-Autonomous Cars. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 392–404. <https://doi.org/10.1145/3025453.3025507>
- Vastenburger, M. H., Keyson, D. V., & de Ridder, H. (2008). Considerate home notification systems: A field study of acceptability of notifications in the home. *Personal and Ubiquitous Computing*, 12(8), 555–566. <https://doi.org/10.1007/s00779-007-0176-x>
- Völkel, S. T., Kempf, P., & Hussmann, H. (2020). Personalised Chats with Voice Assistants: The User Perspective. *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–4. <https://doi.org/10.1145/3405755.3406156>
- Völkel, S. T., Schödel, R., Buschek, D., Stachl, C., Winterhalter, V., Bühner, M., & Hussmann, H. (2020). Developing a Personality Model for Speech-based Conversational Agents Using the Psycholexical Approach. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3313831.3376210>
- Voorveld, H. A. M., & Viswanathan, V. (2015). An Observational Study on How Situational Factors Influence Media Multitasking With TV: The Role of Genres, Dayparts, and Social Viewing. *Media Psychology*, 18(4), 499–526. <https://doi.org/10.1080/15213269.2013.872038>

- Wang, H., & Song, M. (2011). Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. *The R Journal*, 3(2), 29–33. Retrieved April 14, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5148156/>
- Wickens, C. D. (1981). *Processing Resources in Attention, Dual Task Performance, and Workload Assessment*. (Technical Report ADA102719) [Section: Technical Reports]. Defense Technical Information Center. Retrieved February 11, 2023, from <https://apps.dtic.mil/sti/citations/ADA102719>
- Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2), 159–177. <https://doi.org/10.1080/14639220210123806>
- Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., & Talleur, D. A. (2009). Attentional Models of Multitask Pilot Performance Using Advanced Display Technology [Num Pages: 21]. In *Human Error in Aviation*. Routledge.
- Wickens, C. D., Gutzwiller, R. S., & Santamaria, A. (2015). Discrete task switching in overload: A meta-analysis and a model. *International Journal of Human-Computer Studies*, 79, 79–84. <https://doi.org/10.1016/j.ijhcs.2015.01.002>
- Wickens, C. D., & McCarley, J. S. (2007). *Applied attention theory*. CRC Press.
- Wong, P. N. Y., Brumby, D. P., Babu, H. V. R., & Kobayashi, K. (2019). Voices in Self-Driving Cars Should be Assertive to More Quickly Grab a Distracted Driver's Attention. *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 165–176. <https://doi.org/10.1145/3342197.3344535>
- Woods, S. N., Walters, M. L., Koay, K. L., & Dautenhahn, K. (2006). Methodological issues in hri: A comparison of live and video-based methods in robot to human approach direction trials. *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 51–58. <https://doi.org/10.1109/ROMAN.2006.314394>
- Woodward, J., McFadden, Z., Shiver, N., Ben-Hayon, A., Yip, J. C., & Anthony, L. (2018). Using co-design to examine how children conceptualize intelligent interfaces. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14.
- Wu, Y., Rough, D., Bleakley, A., Edwards, J., Cooney, O., Doyle, P. R., Clark, L., & Cowan, B. R. (2020). See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers. *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–9. <https://doi.org/10.1145/3379503.3403563>
- Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review*, 9(2), 185–211. <https://doi.org/10.3758/BF03196276>

- Yager, C., Dinakar, S., Sanagaram, M., & Ferris, T. K. (2015). Emergency vehicle operator on-board device distractions. *Texas A&M Transportation Institute Technical Report, 2015*, 1–50.
- Yang, F., Heeman, P. A., & Kun, A. L. (2011). An Investigation of Interruptions and Resumptions in Multi-Tasking Dialogues. *Computational Linguistics, 37*(1), 75–104. https://doi.org/10.1162/coli_a_00036
- Yngve, V. H. (1970). On getting a word in edgewise. Papers from the Sixth Regional Meeting of the Chicago Linguistic Society. *Papers from the sixth regional meeting Chicago Linguistic Society*.
- Yoon, S. O., Koh, S., & Brown-Schmidt, S. (2012). Influence of perspective and goals on reference production in conversation. *Psychonomic bulletin & review, 19*, 699–707. <https://doi.org/10.3758/s13423-012-0262-6>
- Yorke-Smith, N., Saadati, S., Myers, K. L., & Morley, D. N. (2012). The Design of a Proactive Personal Agent for Task Management. *International Journal on Artificial Intelligence Tools, 21*(01), 1250004. <https://doi.org/10.1142/S0218213012500042>
- Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin, 133*(2), 273–293. <https://doi.org/10.1037/0033-2909.133.2.273>
- Zacks, J. M., & Swallow, K. M. (2007). Event Segmentation. *Current Directions in Psychological Science, 16*(2), 80–84. <https://doi.org/10.1111/j.1467-8721.2007.00480.x>
- Zhao, Y., Jaber, R., McMillan, D., & Munteanu, C. (2022). ‘Rewind to the Jiggling Meat Part’: Understanding Voice Control of Instructional Videos in Everyday Tasks. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–11. <https://doi.org/10.1145/3491102.3502036>
- Züger, M., & Fritz, T. (2015). Interruptibility of software developers and its prediction using psycho-physiological sensors. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

Appendices

Appendix A: Materials used in Chapter 3

- Experiment recruitment informational page (1 page)
- Participant local storage notification screen (1 page)
- Participant information sheet (1 page)
- Consent form (1 page)
- Demographic exclusion criteria questionnaire (1 page)
- Experiment instructional screens (6 pages)
- Post-experiment questionnaire (1 page)
- Participant debriefing page (1 page)



Tetris Communication Game

This study is being conducted by Justin Edwards and Dr. Benjamin Cowan as part of a Science Foundation Ireland (SFI) funded research project looking at social communication. Justin Edwards is a PhD student at University College Dublin's (UCD) School of Information and Communication Studies and is leading this study. Dr. Benjamin Cowan is an assistant professor at University College Dublin's (UCD) School of Information and Communication Studies and is supervising this study.

Description of the proposed study

A research study looking at how people communicate with a partner who is playing a computer game.

Explanation of what participation entails

Participation in this study is completely voluntary and you may withdraw at any time, for any reason, without penalty. If you wish to withdraw from the study, any data collected up to that point will be deleted.

You will be matched with a volunteer from www.free Tetris.org. Your partner will be assigned the task of playing Tetris games while you will have the task of communicating messages to your partner.

The aim of this game is to communicate all of the messages that are assigned to you while your partner tries to clear as many Tetris lines as possible.

Prior to beginning the game you will be given information about the task, asked to provide consent, and asked to provide some basic demographic information. Participation is expected to take 30-45 minutes in total.

Requirements to take part

- To take part in the study you must be **at least 18 years of age**.
- You must also be a **native English speaker**.
- Unfortunately, if you have a cognitive or speech based impediment you will not be able to participate.
- You must also have **normal to corrected vision and hearing**.
- We also ask that **you must have either speakers, earphones, or headphones connected to the computer** so that you can hear the sounds involved in the study, **as well as a microphone** so as to record your descriptions. Note that this data will be de-identified and stored on a secure University server and on encrypted hard drives.
- Please make sure you have 60 minutes free from distractions** before taking part in the experiment.
- It is very important that you DO NOT leave the experiment screen once you begin. Payment may be withheld if evidence of multitasking is identified.**
- The experiment has only been tested on the latest versions of Chrome, Edge, and Firefox, so we ask that you use one of those browsers to prevent any issues.**
- Your device will require a physical keyboard.** Do not attempt to take part if you are on a touch-only device.

Remuneration

Everyone who completes the tasks will receive \$10.00 for their time. This will be redeemable through MTurk by using the code provided at the end of the experiment.

How the data will be kept confidential

The data you supply, including microphone recordings of speech, will be de-identified and will be entered into a database that can only be accessed by authorised personnel involved with this project. This database will be stored on an encrypted UCD server that can only be accessed by authorised personnel involved with this project. The information will be retained by UCD and will only be used for the purposes of research and audit. By supplying this information, you are consenting to the university storing your data for these purposes. The information will be processed by UCD in accordance with the provisions of the Data Protection Act (1998) and the General Data Protection Regulation (2016). If you wish to withdraw your data, we ask that you contact us (ucd.ics.research@gmail.com) with the unique ID provided. The data will be used in academic research only and will not be shared with any commercial entities. Deidentified data may appear in publications stemming from this work. The data will be stored for 5 years on a password protected and encrypted computer owned by the university, with access only provided to researchers involved in the study, and will be archived thereafter.

For further information on your GDPR data rights please consult <https://www.dataprotection.ie/sites/default/files/uploads/2018-12/Rights-of-Individuals-under-the-General-Data-Protection-Regulation-04-2018.pdf>

Possible risks involved

There are no perceived risks from participating in this study, above what would be expected from an educational setting. However, if you are not comfortable with any part of the task, you have the option to withdraw at any time with no penalty to you.

Dissemination of the results

The results of this study will be disseminated through peer-reviewed publications. If you wish to receive information of the findings of this study or subsequent publications, please use the contact details provided below.

Research funding bodies

The current study is funded by Science Foundation Ireland (SFI) award number: 13/RC/2106 ADAPT.

Contact details

If you would like to contact us please email ucd.ics.research@gmail.com quoting "Tetris Study" in the subject line.

How to take part

Please click on the survey link below to open a new window and take part in the study. Remember to leave this window open so you can submit your confirmation code once the study is completed. Please copy and paste the code to avoid mistakes. Please contact us directly if you have any issues with submission.

Study link:

Provide the confirmation code here:

Figure 1: Appendix A: Experiment recruitment informational page

Local storage consent

This website uses local storage to check if you've already completed the experiment. No additional personal data is collected until you've given your informed consent about that data. Your personal data is not shared with any third-parties.

Continue

Figure 2: Appendix A: Participant local storage notification screen

Participant Information

Tetris Communication Game

What is the study about?

You have been invited to take part in a research study looking at how people communicate with a partner who is playing a computer game. The study is being conducted by University College Dublin's School of Information and Communication Studies. The study is being led by Justin Edwards under the supervision of Dr. Benjamin Cowan.

Research Activity:

You will be matched with a volunteer from www.freetetris.org. Your partner will be assigned the task of playing Tetris games while you will have the task of communicating messages to your partner.

The aim of this game is to communicate all of the messages that are assigned to you while your partner tries to clear as many Tetris lines as possible.

Continue

Figure 3: *Appendix A: Participant information sheet*

Consent Form

Title of Study

Tetris Communication Game

Fair Processing Statement

Participation data is being collected as part of a research project concerned with communication and playing the game Tetris. The research is being conducted by the School of Information and Communication Studies in University College Dublin (UCD) and is funded by Science Foundation Ireland (SFI). The data you supply, including microphone recordings of speech, will be de-identified and will be entered into a database that can only be accessed by authorised personnel involved with this project. This database will be stored on an encrypted UCD server that can only be accessed by authorised personnel involved with this project. The information will be retained by UCD and will only be used for the purposes of research and audit. By supplying this information, you are consenting to the university storing your data for these purposes. The information will be processed by UCD in accordance with the provisions of the Data Protection Act (1998) and the General Data Protection Regulation (2016).

Ethical approval details

This project has been deemed Low-risk and has received Ethics Exemption code HS-E-20-161-Edwards-Cowan as per UCD HREC Guidelines.

Statements of Consent

- I agree to take part in this study.
- I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason. If I withdraw during the experiment, the data will be deleted and removed from the study.
- I understand that my personal data will be processed for the purposes detailed above in accordance with the Data Protection Act (1998) and the General Data Protection Regulation (2016).

Figure 4: *Appendix A: Consent form*

Demographics Questionnaire

Are you a native speaker of English?

Do you have normal to corrected vision?

Do you have normal to corrected hearing?

Do you suffer from a diagnosed speech or cognitive impairment?

Figure 5: *Appendix A: Demographic exclusion criteria questionnaire*

Instructions

In this experiment, we are interested in how you and your partner work together to accomplish different goals.

Your partner's goal is to play Tetris with as few mistakes as possible.

Your goal is to ask questions to your partner without distracting them from their game.

You will play 16 rounds. In each round, your partner will start playing Tetris in the middle of a game that is already in progress. You will be able to see your partner's screen. After a few seconds, you will be shown a question to ask your partner. Instructions will be minimal, so you need to use your own words to ask these questions. The instructions will appear on your screen only. Some questions are marked as urgent and others are not. On every round, you should do your best to ask your question quickly whilst avoiding disrupting your partner.

When you want to talk to your partner, **press and hold the T key to activate your microphone**. After your partner has responded to the message, **both partners can press the Enter key to advance to the next round**.

Continue

Figure 6: *Appendix A: Experiment instructional screen*

Participant scoring

After each round is finished, **your partner will be asked to rate how well you did in terms of how disruptive your question was.** Your partner will be asked how much they agree with the following two statements: "My partner's question came at a good moment." and "My partner's question did not distract me."

At the end of the experiment you will be given a score based on the ratings your partner gave you as well as your partner's Tetris score.

Again, some questions are urgent and others are not. The ratings you receive for urgent questions count TEN TIMES MORE toward your score.

Please refrain from clicking away to other programs while you are doing this experiment. Doing so will result in a score penalty.

The participant with the best score will receive a €20 bonus prize after all participants have taken part.

Continue

Figure 7: *Appendix A: Experiment instructional screen*

Participant Information

Tetris Communication Game

What is the study about?

You have been invited to take part in a research study looking at how people communicate with a partner who is playing a computer game. The study is being conducted by University College Dublin's School of Information and Communication Studies. The study is being led by Justin Edwards under the supervision of Dr. Benjamin Cowan.

Research Activity:

You will be matched with a volunteer from www.freetetris.org. Your partner will be assigned the task of playing Tetris games while you will have the task of communicating messages to your partner.

The aim of this game is to communicate all of the messages that are assigned to you while your partner tries to clear as many Tetris lines as possible.

Continue

Figure 8: *Appendix A: Experiment instructional screen*

First, you will play some practice rounds with your partner. Remember, when you want to talk to your partner **press and hold the T key to activate your microphone**. After your partner responds to your question, you can both **press the enter key to move to the next round**. When a round begins, you will watch your partner play Tetris for a few seconds before a message appears instructing you what question to ask your partner. Your partner will not see the message. Messages are deliberately vague as we want you to use your own words.

In the first practice round, the question will not be urgent. After the trial, your partner will rate whether your question was distracting and whether it came at a good moment.

Continue

Figure 9: *Appendix A: Experiment instructional screen*

Good job completing the first practice round. In the next round, the question for your partner will be urgent, so your partner's rating will count ten times more toward your total score.

Continue

Figure 10: *Appendix A: Experiment instructional screen*

You've finished the practice session, these rounds did not count toward your total score. Next, you and your partner will play several real rounds just like those. Afterwards, we will ask you a few questions about yourself and the game and you will be debriefed about the purpose of the experiment.

Continue

Figure 11: *Appendix A: Experiment instructional screen*

Post-trial Questionnaire

Age:

Gender:

What is the highest level of education completed?

What is your level of experience with Tetris?

For non-urgent trials, how did you decide when to deliver messages your partner?

For non-urgent trials, how did you decide what to say to your partner?

For urgent trials, how did you decide when to deliver messages your partner?

For urgent trials, how did you decide what to say to your partner?

Do you have any other comments about your experience today?

I thought my partner I played the game with today was:

Figure 12: Appendix A: Post-experiment questionnaire

Debrief Form

Thank you for taking part in this study. We are interested in understanding how people speak when communicating with a partner. Specifically, we are interested in how people interrupt someone who is busy with another task. We are researching whether people use different strategies for how and when to speak depending on whether the message they have is urgent and how busy they think the other person was.

Your partner for this game was pre-recorded. Their responses and Tetris play were not affected by your messages.

We were not actually rating your performance and no score was collected. We told you there were urgent trials with higher scores to see whether you would adopt a different strategy for those trials.

Even though there is no prize for having the best score, one random Turker will be awarded the €20 reward we promised, so you might still win that prize!

If you would like to hear more about this work or if you have any further questions, please contact us at ucd.ics.research@gmail.com

Okay

Figure 13: *Appendix A: Participant debriefing page*

Appendix B: Materials used in Chapter 4

- Experiment recruitment informational page (1 page)
- Participant local storage notification screen (1 page)
- Participant information sheet (2 pages)
- Consent form (1 page)
- Demographic exclusion criteria questionnaire (1 page)
- Experiment instructional screens (8 pages)
- Post-experiment questionnaire (1 page)
- Participant debriefing page (1 page)



Tetris Communication Game

This study is being conducted by Justin Edwards and Dr. Benjamin Cowan as part of a Science Foundation Ireland (SFI) funded research project looking at social communication. Justin Edwards is a PhD student at University College Dublin's (UCD) School of Information and Communication Studies and is leading this study. Dr. Benjamin Cowan is an assistant professor at University College Dublin's (UCD) School of Information and Communication Studies and is supervising this study.

Description of the proposed study

A research study looking at how people communicate with a partner who is playing a computer game.

Explanation of what participation entails

Participation in this study is completely voluntary and you may withdraw at any time, for any reason, without penalty. If you wish to withdraw from the study, any data collected up to that point will be deleted.

You will be matched with a volunteer from www.free Tetris.org. Your partner will be assigned the task of playing Tetris games while you will have the task of communicating messages to your partner.

The aim of this game is to communicate all of the messages that are assigned to you while your partner tries to clear as many Tetris lines as possible.

Prior to beginning the game you will be given information about the task, asked to provide consent, and asked to provide some basic demographic information. Participation is expected to take 30-45 minutes in total.

Requirements to take part

- To take part in the study you must be **at least 18 years of age**.
- You must also be a **native English speaker**.
- Unfortunately, if you have a cognitive or speech based impediment you will not be able to participate.
- You must also have **normal to corrected vision and hearing**.
- We also ask that **you must have either speakers, earphones, or headphones connected to the computer** so that you can hear the sounds involved in the study, **as well as a microphone** so as to record your descriptions. Note that this data will be de-identified and stored on a secure University server and on encrypted hard drives.
- Please make sure you have 60 minutes free from distractions** before taking part in the experiment.
- It is very important that you DO NOT leave the experiment screen once you begin. Payment may be withheld if evidence of multitasking is identified.**
- The experiment has only been tested on the latest versions of Chrome, Edge, and Firefox, so we ask that you use one of those browsers to prevent any issues.**
- Your device will require a physical keyboard.** Do not attempt to take part if you are on a touch-only device.

Remuneration

Everyone who completes the tasks will receive \$10.00 for their time. This will be redeemable through MTurk by using the code provided at the end of the experiment.

How the data will be kept confidential

The data you supply, including microphone recordings of speech, will be de-identified and will be entered into a database that can only be accessed by authorised personnel involved with this project. This database will be stored on an encrypted UCD server that can only be accessed by authorised personnel involved with this project. The information will be retained by UCD and will only be used for the purposes of research and audit. By supplying this information, you are consenting to the university storing your data for these purposes. The information will be processed by UCD in accordance with the provisions of the Data Protection Act (1998) and the General Data Protection Regulation (2016). If you wish to withdraw your data, we ask that you contact us (ucd.ics.research@gmail.com) with the unique ID provided. The data will be used in academic research only and will not be shared with any commercial entities. Deidentified data may appear in publications stemming from this work. The data will be stored for 5 years on a password protected and encrypted computer owned by the university, with access only provided to researchers involved in the study, and will be archived thereafter.

For further information on your GDPR data rights please consult <https://www.dataprotection.ie/sites/default/files/uploads/2018-12/Rights-of-Individuals-under-the-General-Data-Protection-Regulation-04-2018.pdf>

Possible risks involved

There are no perceived risks from participating in this study, above what would be expected from an educational setting. However, if you are not comfortable with any part of the task, you have the option to withdraw at any time with no penalty to you.

Dissemination of the results

The results of this study will be disseminated through peer-reviewed publications. If you wish to receive information of the findings of this study or subsequent publications, please use the contact details provided below.

Research funding bodies

The current study is funded by Science Foundation Ireland (SFI) award number: 13/RC/2106 ADAPT.

Contact details

If you would like to contact us please email ucd.ics.research@gmail.com quoting "Tetris Study" in the subject line.

How to take part

Please click on the survey link below to open a new window and take part in the study. Remember to leave this window open so you can submit your confirmation code once the study is completed. Please copy and paste the code to avoid mistakes. Please contact us directly if you have any issues with submission.

Study link:

Provide the confirmation code here:

Figure 14: Appendix B: Experiment recruitment informational page

Local storage consent

This website uses local storage to check if you've already completed the experiment. No additional personal data is collected until you've given your informed consent about that data. Your personal data is not shared with any third-parties.

Continue

Figure 15: *Appendix B: Participant local storage notification screen*

Participant Information

Tetris Communication Game

What is the study about?

You have been invited to take part in a research study looking at how people communicate with a partner who is playing a computer game. The study is being conducted by University College Dublin's School of Information and Communication Studies. The study is being led by Justin Edwards under the supervision of Dr. Benjamin Cowan.

Research Activity:

You will be matched with a volunteer from www.freetetris.org. Your partner will be assigned the task of playing Tetris games while you will have the task of communicating messages to your partner.

The aim of this game is to communicate all of the messages that are assigned to you while your partner tries to clear as many Tetris lines as possible.

Continue

Figure 16: *Appendix B: Participant information sheet 1*

Participant Information

Time Commitment:

The study should take no longer than 30 minutes to complete in its entirety.

You may withdraw from the study at any point. If you do decide to withdraw from the study, the data collected to that point will be deleted. If you have any questions as a result of reading this information sheet, please contact the researchers before proceeding.

Data Collection/Protection:

Your data will be de-identified and only associated with your participant ID code. Data, including audio recordings, will be securely stored by the researchers on a secured UCD server and then on secured and encrypted hard drives for data analysis. It will be used only for the purposes of this academic study.

The purpose of using your audio data:

Audio data from the experiment will be pitch shifted to ensure de-identification. De-identified written transcripts will be produced, with the written transcripts being statistically analysed for this study only. Analysis of this transcription data may be used in academic publications and the de-identified transcripts may be shared with other academics or through open science repositories as per publication guidelines. Following analysis, audio recordings will be destroyed.

For further information on GDPR data rights please consult:

<https://www.dataprotection.ie/sites/default/files/uploads/2018-12/Rights-of-Individuals-under-the-General-Data-Protection-Regulation-04-2018.pdf>

Contact Details:

The researchers will be happy to answer any questions you may have about this study before you start. You may contact the project team through ucd.ics.research@gmail.com

Continue

Figure 17: Appendix B: Participant information sheet 2

Consent Form

Title of Study

Judging Interruptions from Speech Assistants

Fair Processing Statement

Participation data is being collected as part of a research project concerned with communication and playing the game Tetris. The research is being conducted by the School of Information and Communication Studies in University College Dublin (UCD) and is funded by Science Foundation Ireland (SFI). The data you supply, including microphone recordings of speech, will be de-identified and will be entered into a database that can only be accessed by authorised personnel involved with this project. This database will be stored on an encrypted UCD server that can only be accessed by authorised personnel involved with this project. The information will be retained by UCD and will only be used for the purposes of research and audit. By supplying this information, you are consenting to the university storing your data for these purposes. The information will be processed by UCD in accordance with the provisions of the Data Protection Act (1998) and the General Data Protection Regulation (2016).

Ethical approval details

This project has been deemed Low-risk and has received Ethics Exemption code HS-E-22-23-Edwards-Cowan as per UCD HREC Guidelines.

Statements of Consent

- I agree to take part in this study.
- I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason. If I withdraw during the experiment, the data will be deleted and removed from the study.
- I understand that my personal data will be processed for the purposes detailed above in accordance with the Data Protection Act (1998) and the General Data Protection Regulation (2016).

Figure 18: Appendix B: Consent form

Demographics Questionnaire

- Are you a native speaker of English?
- Do you have normal to corrected vision?
- Do you have normal to corrected hearing?
- Do you suffer from a diagnosed speech or cognitive impairment?

Figure 19: Appendix 4: Demographic exclusion criteria questionnaire

Microphone Testing

To take part in this experiment you will require a working microphone. Please test your voice using the record button to make sure that your microphone works.

Information Logs

Recordings

Please make sure to listen to a few recordings before you press continue. Make sure that you can clearly hear your voice at a regular volume.

Figure 20: *Appendix B: Experiment instructional screen*

Instructions

In this experiment, we are interested in how you and your partner work together to accomplish different goals.

Your partner's goal is to play Tetris with as few mistakes as possible.

Your goal is to ask questions to your partner without distracting them from their game.

You will play 16 rounds. In each round, your partner will start playing Tetris in the middle of a game that is already in progress. You will be able to see your partner's screen. After a few seconds, you will be shown a question to ask your partner. Instructions will be minimal, so you need to use your own words to ask these questions. The instructions will appear on your screen only. Some questions are marked as urgent and others are not. On every round, you should do your best to ask your question quickly whilst avoiding disrupting your partner.

When you want to talk to your partner, **press and hold the T key to activate your microphone**. After your partner has responded to the message, **both partners can press the Enter key to advance to the next round**.

Continue

Figure 21: *Appendix B: Experiment instructional screen*

Participant scoring

After each round is finished, you will be asked a few questions about the Tetris game you just saw and your experience picking a moment to speak. You will be asked how complex the Tetris game was, how easy it was to choose a moment to speak, and how confident you are that you chose a good moment.

Meanwhile, **your partner will be asked to rate how well you did in terms of how disruptive your question was.** Your partner will be asked how much they agree with the following two statements: "My partner's question came at a good moment." and "My partner's question did not distract me."

At the end of the experiment you will be given a score based on the ratings your partner gave you as well as your partner's Tetris score.

Again, some questions are urgent and others are not. **The ratings you receive for urgent questions count TEN TIMES MORE toward your score.**

Please refrain from clicking away to other programs while you are doing this experiment. Doing so will result in a score penalty.

The participant with the best score will receive a €20 bonus prize after all participants have taken part.

Continue

Figure 22: *Appendix B: Experiment instructional screen*

Searching for your partner

Thank you once again for agreeing to take part in research.

Please wait to be matched with a Tetris volunteer. This may take a few minutes. When you are matched with a partner, the button below will say Continue.



Searching for your partner

Figure 23: Appendix B: Experiment instructional screen

Connected to Partner

Your Partner

Name:

Leigh

Age:

28

Occupation:

Customer service representative

Country:

Republic of Ireland

freetetris.org Activity

Hours played this week: 2

Hours played this month: 11

Hours played this year: 47

Member since: 14 Jan 2018

Figure 24: Appendix B: Experiment instructional screen

First, you will play some practice rounds with your partner. Remember, when you want to talk to your partner **press and hold the T key to activate your microphone**. After your partner responds to your question, you can both **press the enter key to move to the next round**. When a round begins, you will watch your partner play Tetris for a few seconds before a message appears instructing you what question to ask your partner. Your partner will not see the message. Messages are deliberately vague as we want you to use your own words.

In the first practice round, the question will not be urgent. After the trial, your partner will rate whether your question was distracting and whether it came at a good moment.

Continue

Figure 25: *Appendix B: Experiment instructional screen*

Good job completing the first practice round. In the next round, the question for your partner will be urgent, so your partner's rating will count ten times more toward your total score.

Continue

Figure 26: *Appendix B: Experiment instructional screen*

You've finished the practice session, these rounds did not count toward your total score. Next, you and your partner will play several real rounds just like those. Afterwards, we will ask you a few questions about yourself and the game and you will be debriefed about the purpose of the experiment.

Continue

Figure 27: *Appendix B: Experiment instructional screen*

Post-trial Questionnaire

Age:

Gender: -- select an option --

What is the highest level of education completed?

-- select an option --

What is your level of experience with Tetris?

-- select an option --

For non-urgent trials, how did you decide when to deliver messages your partner?

For non-urgent trials, how did you decide what to say to your partner?

For urgent trials, how did you decide when to deliver messages your partner?

For urgent trials, how did you decide what to say to your partner?

Do you have any other comments about your experience today?

I thought my partner I played the game with today was:

-- select an option --

Figure 28: Appendix A: Post-experiment questionnaire

Debrief Form

Thank you for taking part in this study. We are interested in understanding how people speak when communicating with a partner. Specifically, we are interested in how people interrupt someone who is busy with another task. We are researching whether people use different strategies for how and when to speak depending on whether the message they have is urgent and how busy they think the other person was.

Your partner for this game was pre-recorded. Their responses and Tetris play were not affected by your messages.

We were not actually rating your performance and no score was collected. We told you there were urgent trials with higher scores to see whether you would adopt a different strategy for those trials.

Even though there is no prize for having the best score, one random participant will be awarded the €20 reward we promised, so you might still win that prize!

If you would like to hear more about this work or if you have any further questions, please contact us at ucd.ics.research@gmail.com

Okay

Figure 29: *Appendix B: Participant debriefing page*

Appendix C: Materials used in Chapter 5

- Local storage information screen (1 page)
- Participant information sheet (2 pages)
- Consent form (1 page)
- Demographic exclusion criteria questionnaire (1 page)
- Experiment instructional screens (4 pages)
- Experiment inter-trial rating screen (1 page)
- Post-experiment questionnaire (1 page)
- Participant debriefing page (1 page)
- Content analysis codebook (1 page)

Local storage consent

This website uses local storage to check if you've already completed the experiment. No additional personal data is collected until you've given your informed consent about that data. Your personal data is not shared with any third-parties.

Continue

Figure 30: *Appendix C: Participant local storage notification screen*

Participant Information

Tetris Interruption Timing

What is the study about?

You have been invited to take part in a research study looking at how people communicate with a partner who is playing a computer game. The study is being conducted by University College Dublin's School of Information and Communication Studies. The study is being led by Justin Edwards under the supervision of Dr. Benjamin Cowan.

Research Activity:

You will be watching clips of people playing Tetris, and you will be asked to imagine that you need to ask the Tetris player a question. You will see a Tetris clip twice, then you will be asked to pick the moment you think is best to begin speaking to the player. Each clip is less than 10 seconds long, and a **counter will be visible throughout the clip to help you pick the best moment to begin speaking.** After you make your selection, you will be asked how difficult it was to make a selection, and how complex you thought the Tetris game was.

The aim of this experiment is to determine the best moments to begin speaking to someone who is playing Tetris. You should assume the player would be distracted if they were asked a question, so you should pick the moment that will disrupt their game the least. At the end of the experiment, you will be asked a few questions about yourself and your experience with this experiment.

Continue

Figure 31: *Appendix C: Participant information sheet*

Consent Form

Title of Study

Tetris Communication Game

Fair Processing Statement

Participation data is being collected as part of a research project concerned with communication and playing the game Tetris. The research is being conducted by the School of Information and Communication Studies in University College Dublin (UCD) and is funded by Science Foundation Ireland (SFI). The data you supply, including microphone recordings of speech, will be de-identified and will be entered into a database that can only be accessed by authorised personnel involved with this project. This database will be stored on an encrypted UCD server that can only be accessed by authorised personnel involved with this project. The information will be retained by UCD and will only be used for the purposes of research and audit. By supplying this information, you are consenting to the university storing your data for these purposes. The information will be processed by UCD in accordance with the provisions of the Data Protection Act (1998) and the General Data Protection Regulation (2016).

Ethical approval details

This project has been deemed Low-risk and has received Ethics Exemption code HS-E-21-39-Edwards-Cowan as per UCD HREC Guidelines.

Statements of Consent

- I agree to take part in this study.
- I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason. If I withdraw during the experiment, the data will be deleted and removed from the study.
- I understand that my personal data will be processed for the purposes detailed above in accordance with the Data Protection Act (1998) and the General Data Protection Regulation (2016).

Figure 32: *Appendix C: Consent form*

Thank you for taking part in the experiment today. Before you begin, please answer the following demographic questions.

Demographics Questionnaire

Are you a native speaker of English?

Do you have normal or corrected vision?

Do you have normal or corrected hearing?

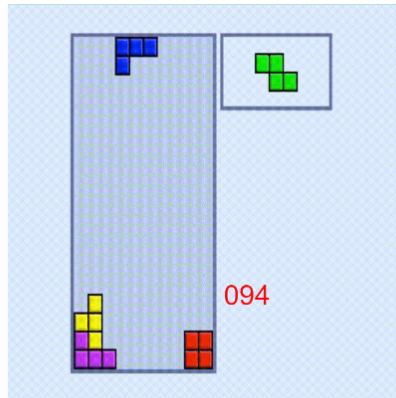
Do you suffer from a diagnosed speech or cognitive impairment?

Figure 33: Appendix C: Demographic exclusion criteria questionnaire

Instructions

Step One - Viewing a clip

The first screen you see will show a clip of Tetris gameplay, like the one shown below. Each clip is 8.5 seconds long and is displayed at 30 frames per second, for a total of 255 frames. Each frame is labeled with a number in red.



In the first step in a round, you will see the full clip, then a quick pause, then the full clip will play again. **During this step, watch the clip and think about what moment would be best to start speaking to the player.**

After the clip plays twice, there will be a pause before continuing onto the next step of the experiment.

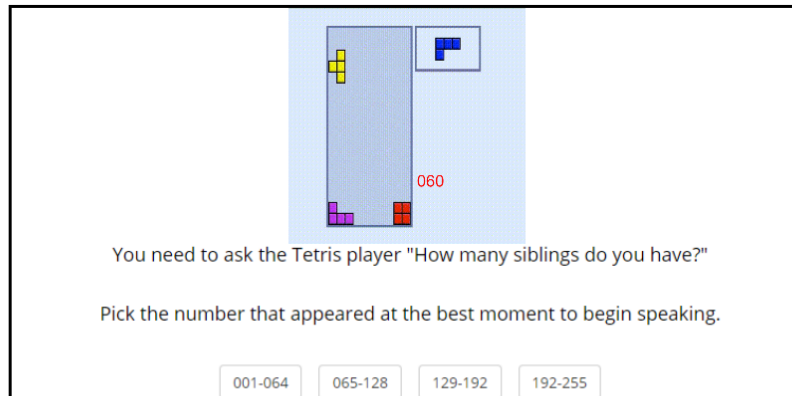
Next

Figure 34: Appendix C: Experiment instructional screen

Instructions

Step Two - Picking when to speak

The next step in each round of the experiment is choosing when you would begin speaking to the Tetris player. The same clip from Step One will play on repeat, and you will see a question that you will imagine asking the player, as shown below.



You need to ask the Tetris player "How many siblings do you have?"

Pick the number that appeared at the best moment to begin speaking.

001-064 065-128 129-192 192-255

You will be asked to pick the moment in which you would begin speaking. To identify the moment, you can estimate which frame number was on the screen during that moment. The buttons beneath the clip will give a range of frames, so you can select the range that best matches the moment you picked. **If you mistakenly pick the wrong range, you will have an opportunity to go back.**

After you have selected the range of frames, you will move on to step 3.

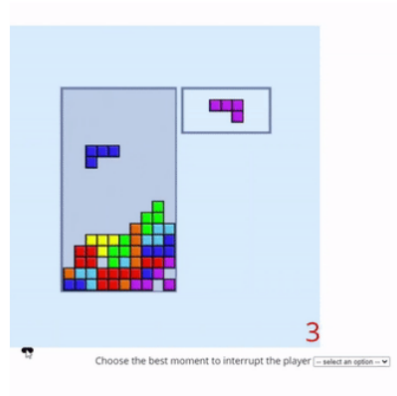
Next

Figure 35: Appendix C: Experiment instructional screen

Instructions

Step Three - Picking a single moment

The third step in each round of the experiment is choosing the single moment in which you would begin speaking to the Tetris player. To pick a moment, you will get a slider of images to scroll through, representing the frames from the range you selected. Each frame is labeled with a new number. The video below gives a demonstration of the image slider.



You will be asked to pick the frame that represents the single best moment to begin speaking. Use the dropdown menu to pick the number that matches the frame you've picked. **If you need to go back and see the clip again or pick a different range of frames, select "Go back" in the dropdown menu and press the continue button.**

After you pick a frame, you'll be asked a few short questions about the round you just completed.

Next

Figure 36: Appendix C: Experiment instructional screen

Instructions

Step Four - Questions about the round

The final step of each round is answering a few questions about that round. Those questions are:

- "How complex was the Tetris game you just saw?"
- "How easy was it to choose a moment to interrupt in the Tetris game you just saw?"
- "How confident are you that you picked the best moment to interrupt?"

Each question will have a 7 point scale to answer the question, with labels on each end. After you have answered the questions, you can click the Continue button to finish a round. When you finish the round, the next round will begin at Step One.

This experiment has 16 rounds. When you are ready to begin, press the Start Experiment button below to begin round 1.

Start Experiment

Figure 37: *Appendix C: Experiment instructional screen*

How complex was the Tetris game you just saw?

Very simple | | | | | Very Complex

How easy was it to choose a moment to interrupt in the Tetris game you just saw?

Very difficult | | | | | Very easy

How confident are you that you picked the best moment to interrupt?

Not at all confident | | | | | Very confident

Continue

Figure 38: Appendix C: Experiment inter-trial rating screen

Post-trial Questionnaire

Age:

Gender:

What is your highest level of education completed?

What is your level of experience with Tetris?

If you have played Tetris before, when was the most recent time you played Tetris?

How did you determine the which moment would be best to begin speaking?

Do you have any other comments about your experience today?

Figure 39: *Appendix C: Post-experiment questionnaire*

Debrief Form

Thank you for taking part in this study. We are interested in understanding how people speak when communicating with a partner. Specifically, we are interested in how people interrupt someone who is busy with another task. We are researching whether certain moments within tasks are better for interrupting than other moments, and whether these moments are apparent to the interrupter.

This study is part of our ongoing research to understand how people use speech to interrupt Tetris players. We hope that it will be a first step in understanding how people interrupt others across a variety of different tasks. We hope this will inform the design of technology that can safely get people's attention through speech, even when people are busy.

If you would like to hear more about this work or if you have any further questions, please contact us at ucd.ics.research@gmail.com

Please continue to the final screen to get a code to receive payment.

Figure 40: *Appendix C: Participant debriefing page*

Sequence name	One piece? (Only one Tetris piece is the active piece during the sequence)	Lateral? (A Tetris piece moves laterally (from left to right) during the sequence)	Rotation? (A Tetris piece rotates during the sequence)	Start of piece? (A Tetris piece is seen at the start of its drop from the top of the screen during the sequence)	End of piece? (A Tetris piece reaches its final destination at the bottom of the board or on other pieces during the sequence)	Line clear? (A Tetris piece fills in a line in the Tetris board causing the line to clear from the board during the sequence)
2-1						
2-2						
2-3						
2E-1						
2E-2						
3-1						
3-2						
4-1						
4E-1						
4E-2						
4E-3						
5-1						
5-2						
6E-1						
6E-2						
6E-3						
7-1						
7-2						
7-3						
7E-1						
7E-2						
7E-3						
8E-1						
8E-2						
8E-3						
9-1						
9-2						
9-3						
9E-1						
9E-2						
9E-3						
9E-4						
10-1						
10-2						
10-3						
10E-1						
10E-2						
10E-3						
10E-4						
11-1						
11E-1						
11E-2						
12-1						

Figure 41: Appendix C: Content analysis codebook

Appendix D: Materials used in Chapter 6

- Cookie consent screen (1 page)
- Participant information sheet (1 page)
- Consent form (1 page)
- Experiment instructional screens (3 pages)
- Pre-test PMQ survey page (1 page)
- Experiment instructional screens (1 page)
- Experiment inter-trial rating screen (1 page)
- Experiment instructional screens (1 page)
- Experimental PMQ survey page (1 page)
- Experiment instructional screens (3 pages)
- Post-experiment questionnaire (1 page)
- Participant debriefing page (1 page)

Cookie Consent Information

This website uses local storage to check if you've already completed the experiment. No additional personal information is collected until you have given your consent.

This information is repeated at the bottom of the screen where you can also click 'Learn more' for a more detailed explanation about cookie policies.

Continue

Figure 42: *Appendix D: Cookie consent screen*

Participant Information

Judging Interruptions from Speech Assistants

What is the study about?

You have been invited to take part in a research study looking at how speech assistants similar to Alexa or Siri should be designed in a future in which they can initiate interactions. The study is being conducted by University College Dublin's School of Information and Communication Studies. The study is being led by Justin Edwards under the supervision of Dr. Benjamin Cowan.

Research Activity:

You will be watching and listening to video clips of people playing the computer game Tetris in which they are interrupted by a speech assistant. You are asked to imagine that the Tetris player is multitasking and needs to answer questions that the speech assistant asks. You will watch the Tetris game for a few seconds, then you will hear the speech assistant ask a question. After the question is asked, you will watch a few more seconds of the Tetris game before the video ends. The Tetris player will not respond in the video clips. Your task is to assess how well the speech assistant asked its question, both in terms of choosing a **good moment** the player, and asking the question in an **appropriate way**. After each video clip, you will be asked to rate both of those factors.

The aim of this experiment is to understand how speech assistants of the future should best interrupt a busy person. You should assume the Tetris player would be distracted when they are asked a question, so try to be mindful of **how distracting** each interruption might be to each Tetris game. At the end of the experiment, you will be asked a few questions about yourself and your experience with this experiment.

Time Commitment:

The study should take no longer than 30 minutes to complete in its entirety.

You may withdraw from the study at any point. If you do decide to withdraw from the study, the data collected to that point will be deleted. If you have any questions as a result of reading this information sheet, please contact the researchers before proceeding at ucd.ics.research@gmail.com

Data Collection/Protection:

Your data will be de-identified and only associated with your Prolific ID. Data will be securely stored by the researchers on a secured UCD server and then on secured and encrypted hard drives for data analysis. It will be used only for the purposes of this academic study.

For further information on GDPR data rights please consult <https://www.dataprotection.ie/sites/default/files/uploads/2018-12/Rights-of-Individuals-under-the-General-Data-Protection-Regulation-04-2018.pdf>

Contact Details:

The researchers will be happy to answer any questions you may have about this study. You may contact the project supervisor through ucd.ics.research@gmail.com

Thank you once again for agreeing to take part in research.

Continue

Figure 43: Appendix D: Participant information sheet

Consent Form

Title of Study

Judging Interruptions from Speech Assistants

Fair Processing Statement

Participation data is being collected as part of a research project concerned with communication and playing the game Tetris. The research is being conducted by the School of Information and Communication Studies in University College Dublin (UCD) and is funded by Science Foundation Ireland (SFI). The data you supply will be de-identified and will be entered into a database that can only be accessed by authorised personnel involved with this project. This database will be stored on an encrypted UCD server that can only be accessed by authorised personnel involved with this project. The information will be retained by UCD and will be used for the purposes of research and audit. Deidentified data will be made available to other researchers via open science repositories. By supplying this information, you are consenting to the university storing your data for these purposes. The information will be processed by UCD in accordance with the provisions of the Data Protection Act (1998) and the General Data Protection Regulation (2016).

Ethical approval details

This project has been deemed Low-risk and has received Ethics Exemption code HS-E-22-23-Edwards-Cowan as per UCD HREC Guidelines.

Statements of Consent

- I agree to take part in this study.
- I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason. If I withdraw during the experiment, the data will be deleted and removed from the study.
- I understand that my personal data will be processed for the purposes detailed above in accordance with the [Data Protection Act \(1998\)](#) and the [General Data Protection Regulation \(2016\)](#).

Figure 44: *Appendix D: Consent form*

Instructions

Urgency

Each video clip will begin in the middle of a Tetris game. Below the game, you will see instructions reminding you of your task. You will also see an indication of whether the speech interruption you will hear is **Urgent** or **Not Urgent**. For urgent interruptions, the speech assistant is instructed that its question is more important than the Tetris game. For interruptions that are not urgent, the speech assistant is instructed that the Tetris game is more important than the question. Please note that **the specific question the speech assistant asks does not impact urgency** - urgency only concerns the relative priority that the speech assistant has been instructed to give each task. The image below shows what you will see at the beginning of a video clip.

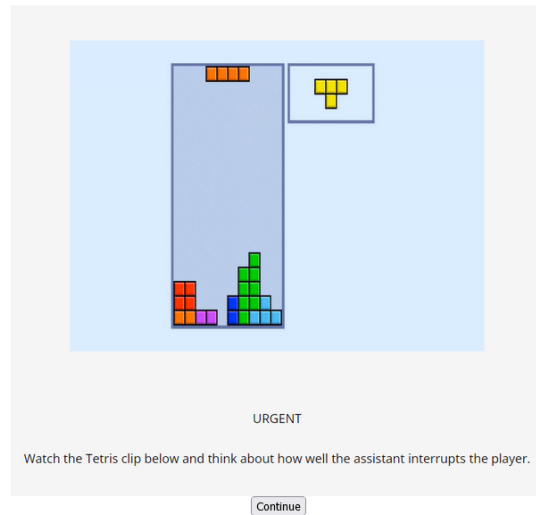
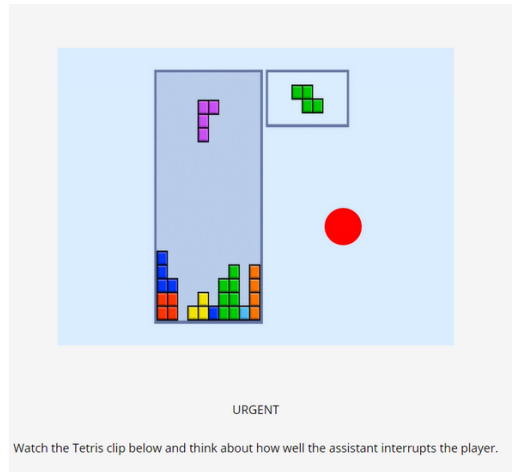


Figure 45: Appendix D: Experiment instructional screen

Instructions

Choosing when to interrupt

After a few seconds, the speech assistant receives its instruction about what it needs to ask the Tetris player. When the assistant gets its instruction a **red dot will appear on the screen**. This red dot is not visible to the Tetris player, it is just a visualization to help you judge the speech assistant's decision about how to interrupt. **The speech assistant can choose to wait for any amount of time after the red dot appears, or to interrupt immediately.** The speech assistant can see the Tetris game and may decide when to interrupt based on the game. A video clip with a red dot will look like the image below.



After the speech assistant asks its question, the Tetris game will continue for a few more seconds, then you will be asked to judge the interruption.

Figure 46: Appendix D: Experiment instructional screen

Instructions

PMQuestionnaire

At the beginning of this study, we will ask you to complete a short survey to understand your impressions of speech assistants in general. Next, you will be asked to watch a series of videos of **one version of a speech assistant** interrupting a Tetris player. You will then be asked to fill out the survey again, thinking about the assistant you just watched. Next, you will see a series of videos of another version of a speech assistant interrupting a Tetris player. You will then be asked to fill out the survey one more time thinking about the assistant you just watched. Finally, we will ask you some demographic questions about yourself and your experience of this experiment.

Thank you for taking part in this study. In the next screen, you will be asked to fill out the first survey of the study.

Continue

Figure 47: *Appendix D: Experiment instructional screen*

Please complete the following questionnaire based on your previous interactions with speech interfaces*

*Speech interface may include a broad range of technologies such as Amazon's Alexa, Apple's Siri, Google Assistant and Microsoft's Cortana, along with speech-based chatbots and telephony systems (i.e. like those used in telephone banking and ticket booking). You may have accessed these using a smartphone, smart speaker, laptop or desktop and/or in-car.

Thinking about the speech interface you interact with most frequently, how would you rate its communicative ability on a scale between each of the following poles?

Please read each pair of words carefully.

Please respond as quickly and accurately as possible and try to only use neutral responses when absolutely necessary.

Clear	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Ambiguous
	<<<	<<	<	N	>	>>	>>>	
Authentic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Fake
	<<<	<<	<	N	>	>>	>>>	
Uncooperative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Cooperative
	<<<	<<	<	N	>	>>	>>>	
Inconsistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Consistent
	<<<	<<	<	N	>	>>	>>>	
Inefficient	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Efficient
	<<<	<<	<	N	>	>>	>>>	
Cold	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Warm
	<<<	<<	<	N	>	>>	>>>	
Spontaneous	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Predetermined
	<<<	<<	<	N	>	>>	>>>	
Literal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Interpretive
	<<<	<<	<	N	>	>>	>>>	
Machine-like	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Human-like
	<<<	<<	<	N	>	>>	>>>	
Flexible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Inflexible
	<<<	<<	<	N	>	>>	>>>	
Expert	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Amateur
	<<<	<<	<	N	>	>>	>>>	
Life-like	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Tool-like
	<<<	<<	<	N	>	>>	>>>	
Transactional	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Social
	<<<	<<	<	N	>	>>	>>>	
Dependable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unreliable
	<<<	<<	<	N	>	>>	>>>	
Incapable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Capable
	<<<	<<	<	N	>	>>	>>>	
Misleading	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Honest
	<<<	<<	<	N	>	>>	>>>	
Precise	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Vague
	<<<	<<	<	N	>	>>	>>>	
Stop-start	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Interactive
	<<<	<<	<	N	>	>>	>>>	
Direct	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Meandering
	<<<	<<	<	N	>	>>	>>>	
Incompetent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Competent
	<<<	<<	<	N	>	>>	>>>	
Personal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Generic
	<<<	<<	<	N	>	>>	>>>	
Empathetic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Apathetic
	<<<	<<	<	N	>	>>	>>>	
Reliable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Uncertain
	<<<	<<	<	N	>	>>	>>>	

Continue

Figure 48: Appendix D: Pre-test PMQ survey page

Post-survey information

Version A

Thank you for completing the survey. Next, you will see a series of video clips. For this series of clips, you will hear a speech assistant interrupting a Tetris player.

Figure 49: *Appendix D: Experiment instructional screen*

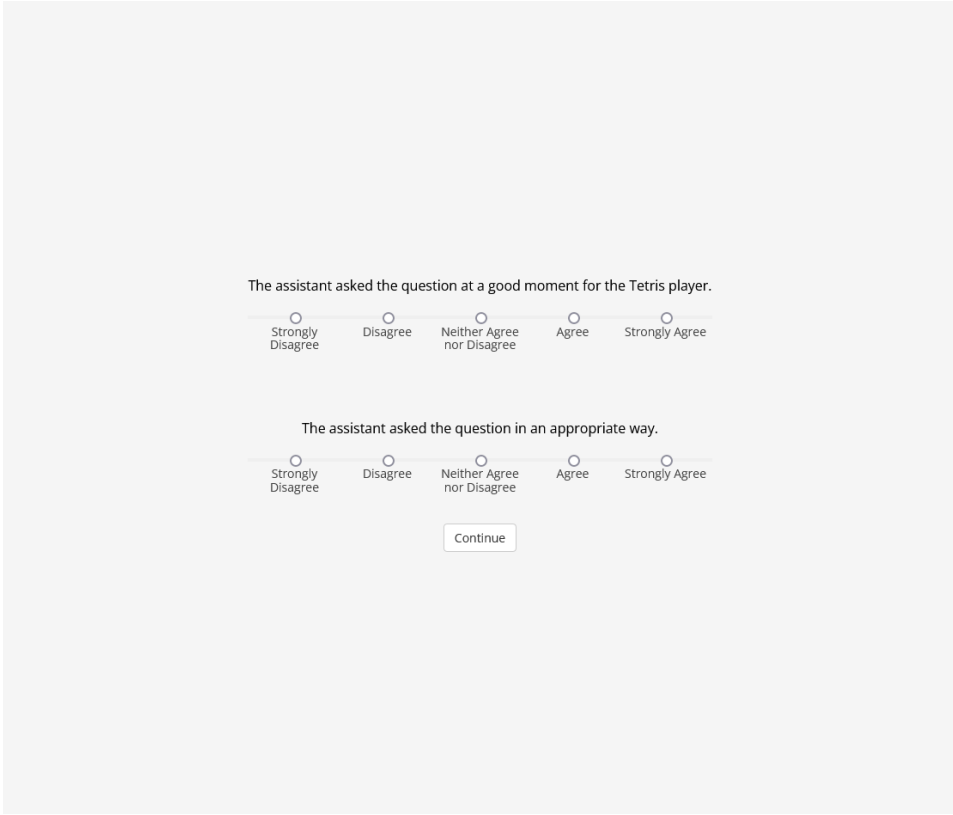


Figure 50: Appendix D:Experiment inter-trial rating screen

Questionnaire

Thank you for watching and rating those video clips. Next, we would like you to complete the PMQ survey again to reflect on your impression of Version A of the speech assistant.

[Continue](#)

Figure 51: *Appendix D: Experiment instructional screen*

Please complete the following questionnaire based on the speech assistant you just watched.

Thinking about the speech assistant you just watched, how would you rate its communicative ability on a scale between each of the following poles?

Please read each pair of words carefully.

Please respond as quickly and accurately as possible and try to only use neutral responses when absolutely necessary.

Meandering	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Direct
Tool-like	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Life-like
Unreliable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Dependable
Inconsistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Consistent
Reliable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Uncertain
Social	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Transactional
Competent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Incompetent
Efficient	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Inefficient
Generic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Personal
Authentic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Fake
Misleading	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Honest
Amateur	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Expert
Spontaneous	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Predetermined
Human-like	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Machine-like
Clear	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Ambiguous
Apathetic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Empathetic
Interpretive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Literal
Stop-start	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Interactive
Uncooperative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Cooperative
Precise	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Vague
Warm	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Cold
Inflexible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Flexible
Capable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Incapable

Continue

Figure 52: Appendix D: Experimental PMQ survey page

Post-survey information

Version B

Thank you for completing the survey. Next, you will see a series of video clips just like in the previous part of the trial. For this series of clips, you will hear another speech assistant interrupting the Tetris player.

Figure 53: *Appendix D: Experiment instructional screen*

Questionnaire

Thank you for watching and rating those video clips. Next, we would like you to complete the PMQ survey again to reflect on your impression of Version B of the speech assistant.

[Continue](#)

Figure 54: *Appendix D: Experiment instructional screen*

Post-survey information

End of Main Study

Thank you for completing the survey. This concludes the main part of the study. Next, we will ask you some questions about yourself your experience today. Following this, you will be debriefed about details of this study and you will receive instructions for payment.

[Continue](#)

Figure 55: *Appendix D: Experiment instructional screen*

Demographics Questionnaire

Prolific ID:

Age:

Gender:

Do you have strong English reading comprehension?

What is the highest level of education completed?

Nationality:

How often do you use computer based speech agents like Alexa, Siri, or Google Assistant?

Which of the following speech agents do you use most often:

Other assistant used (if applicable):

Which device do you use most frequently to access these speech agents?

Other device used (if applicable):

What is your level of experience with Tetris?

If you have played Tetris before, how often do you play Tetris?

If you have played Tetris before, how would you rate your level of Tetris expertise?

What differences, if any, did you notice between the two versions of the speech assistants?

Do you have any other comments about your experience today?

Figure 56: Appendix D: Post-experiment questionnaire

Debrief

THANK YOU FOR YOUR PARTICIPATION!

The data collected in this study will be used to make recommendations about the design of future speech assistants which might initiate interactions. One of the versions of the assistant you listened to was designed to interrupt people using characteristics similar to how people use speech to interrupt other people by considering the urgency of the interruption and the state of the Tetris game. The other version was a baseline design which did not take urgency or the Tetris game into account when making its interruptions. We predict that people will find the human-inspired system to be better in terms of timing and style, and that this system will be seen as more human-like in the partner model questionnaires (PMQs) that you completed.

All data will be used and stored in accordance with details outlined in the instructions sheet and in accordance with GDPR.

You will receive a completion code to collect payment from Prolific on the next page. Copy and paste this code into Prolific to avoid mistakes.

You are reminded that all the data has been anonymised. Should you wish to access data relevant to your participation, or if you have any other questions about the study, contact Justin Edwards at ucd.ics.research@gmail.com making sure to include your Prolific ID in the subject field.

Thank you once again!

The time and effort you have taken to participate in this study is greatly appreciated.

Okay

Figure 57: Appendix D: Participant debriefing page