Human Teleoperation of a Social Robot: Key Design Challenges

by

Adam Kavanagh Coyne

Thesis

Presented to

Trinity College Dublin, the University of Dublin

for the Degree of

Doctor of Philosophy

Trinity College Dublin, the University of Dublin

September 2024

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

I, the undersigned, agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).



Adam Kavanagh Coyne

September 4, 2024

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.



Adam Kavanagh Coyne

September 4, 2024

Acknowledgments

I'd like to express my deepest gratitude first and foremost to Conor for the invaluable supervision, patience and support that helped me become the academic I am today. I am extremely grateful also to everyone else from the Robotics and Innovation Lab; I'd like to thank Kevin as examiner on my committee, Patrick and Moyin for the much-needed help and peer support, as well as Mike, Andy, Vanessa and all those who extended such a warm welcome to me from start to finish. This thesis would not have been possible without funding graciously provided by the Provost's Award of Trinity College Dublin, and sincerest thanks also go to Tangent and the Digital Hub for providing the spaces that allowed the research to take place.

I would be remiss not to mention the many friends who helped me throughout this journey. In particular, I'd like to thank Jelena, Rian, Ali, Niamh, and others from BSR, VCC and Gamers. I'd also like to thank Maya for important discussions on qualitative analysis.

Finally, words can't express how important family was to the success of this thesis. Mum, Dad and Corey, but Abigail in particular, thank you for standing by me for so long with such unwavering faith. I can't thank you enough for your endless, tireless support.

I dedicate this to Jo. You never stopped encouraging me, even when you didn't understand what it was. I wish you could have seen it finished.

Adam Kavanagh Coyne

Trinity College Dublin, the University of Dublin
September 2024

Abstract

The use of robots outside of factories and industrial automation has been steadily growing in recent years. While the range of functional tasks that these systems can carry out has been steadily increasing, their ability to collaborate and work alongside people has been extremely limited, and has created a barrier prohibiting their broader adoption. The development of effective social interfaces for robots remains an inherently challenging interdisciplinary problem, requiring detailed consideration of both engineering and socio-cultural factors.

Rather than focusing on artificial intelligence directly, something that could accelerate adoption of these robots is social robot teleoperation, allowing remote operators to pilot robots that interact socially. This could overcome workforce limitations in customer-facing jobs in distant or hazardous environments, while also enabling those with mobility constraints or otherwise to engage socially as if in-person from the comfort of their home. Importantly, teleoperation data collected during use of the system could also inform the design of social artificial agents, accelerating progress towards automated social robots.

To enable adequate social interaction, a substantial amount of research and development on teleoperation design is yet required. Task-driven, non-social robot teleoperation has received focus in human-robot interaction studies, but outside of minimally-embodied telepresence, social robot teleoperation is undeveloped in research, particularly in regards to experimental rigour and repeatability. This thesis aims to address key research issues which currently impede the development of social robot teleoperation interfaces. Through analysis of the modalities of social interaction, the task is approached along three key areas of research, with a focus on repeatable, universally-applicable methodologies.

First, a framework for evaluating and analysing perceived robot affect expression is developed and trialled, enabling statistically reliable, data-driven selection of robot affect expression via teleoperation. It is shown that the proposed statistical methods yielded easily understandable graphs on a robot's expressed emotional range, providing valuable insights to developers.

Second, robot speech comprehension is evaluated to determine whether operator prosody should be conveyed in a speech teleoperation interface. By employing methods from the field of audiology, it was shown that speech comprehension did not change by conveying operator prosody for participants local to it, however comprehension was negatively affected by altered prosody for those unfamiliar with the prosody of the operator. Discussion of results provides recommendations and precautions to robot designers who wish to mitigate this issue.

Finally, by adapting standard situation awareness measurement techniques, a tool for objectively measuring *social* situation awareness was developed, and used to evaluate the benefits of spatial audio feedback during robot teleoperation. Analysis of results alongside secondary measurements show that further research may be required to refine and validate the technique, but benefits to spatial audio feedback were nonetheless observed in subjective and qualitative results.

These three investigations yielded valuable, sometimes counter-intuitive insights for robot designers, while also providing robust, re-usable statistics and methodologies that can evaluate teleoperation capabilities for any social robot platform. While the path towards practicable social robot teleoperation is yet long, this thesis hopes to enable through its contributions a more reliable and expedient traversal of this path in future research.

Contents

Acknowledgments **iv**

List of Tables **vi**

List of Figures **vii**

List of Abbreviations **ix**

List of Publications **xi**

Chapter 1 Introduction **1**

1.1 Background 1

1.2 Motivation 3

1.3 Research aims and objectives 5

1.4 Thesis overview 8

Chapter 2 Literature Review **9**

2.1 Teleoperation theory and concepts 9

2.1.1 Human control and automation theory 9

2.1.2 Teleoperation and the "dual ecology" 10

2.1.3 Performance metrics and evaluation 13

2.2 Real-world applications of teleoperated robotics 17

2.2.1 Task-oriented teleoperation 17

2.2.2 Social teleoperation 19

2.3 Key cues and modalities of interaction for teleoperation 20
2.3.1 Understanding multimodal interaction 20
2.3.2 Task-oriented modalities: Navigation, manipulation 21
2.3.3 The vocal modality 22
2.3.4 The visuospatial modality 26
2.3.5 Gaze teleoperation 27
2.3.6 Affect expression 28
2.3.7 Awareness of interactant social cues 32
2.4 Summary 35

Chapter 3 Affect expression in social robot teleoperation 37

3.1 Introduction 37
3.2 Development of statistics for the Geneva Emotion Wheel 39
3.2.1 Quantifying Findings from the GEW 41
3.2.2 Bayesian modelling of the GEW 43
3.2.3 Visualising Findings from the GEW 49
3.2.4 Adapting the GEW as a HRI design tool 51
3.3 Methodology 53
3.4 Results 54
3.5 Discussion 57

Chapter 4 Speech expression in social robot teleoperation 61

4.1 Introduction 61
4.2 Methodology 63
4.2.1 Stimuli creation and administration 63
4.2.2 Experimental design 66
4.3 Results 69
4.4 Discussion 69

Chapter 5 Spatial audio in social robot teleoperation 75

5.1 Introduction . . . 75

5.2 Developing an assessment technique of social situation awareness . . . 77

5.2.1 Social interaction and awareness traits of interest . . . 78

5.2.2 Existing technique selection . . . 79

5.2.3 Adapting SAGAT for social contexts . . . 79

5.2.4 Practical considerations . . . 81

5.2.5 Summary . . . 82

5.3 Methodology . . . 83

5.3.1 Stimuli creation and administration . . . 83

5.3.2 Experimental design . . . 86

5.4 Results . . . 92

5.5 Discussion . . . 98

Chapter 6 Conclusion 102

6.1 Reflection on research objectives . . . 102

6.2 Academic impact . . . 106

6.3 Limitations and future work . . . 106

Appendix A Geneva Emotion Wheel data visualisations 109

Appendix B Situation awareness testing data 111

Bibliography 115

List of Tables

2.1 Sheridan-Verplank Levels of Automation of Decision and Action Selection . . 10
2.2 Partial breakdown of social interaction modalities 21

3.1 Prior selection and likelihood statistics for each evaluated affect expression. . 56
3.2 Other emotion names specified 56

4.1 List of the shorter and longer sentences used in the experiment 67

5.1 Pool of all potential queries delivered during SAGAT pauses 88
5.2 List of questions used from Witmer and Signer's Presence Questionnaire . . . 90
5.3 SAGAT scores detailed by each individual SAGAT query 94
5.4 Statistical summary of objective results from SA testing 95
5.5 Statistical summary of results of Presence Questionnaire 95
5.6 Pearson correlation matrix of all quantitative measures from SA testing . . . 97
5.7 Categorised open feedback summary after spatial audio testing 98

B.1 Quiz questions asked during the experiment for both videos 112
B.2 Positive feedback provided in answer to the question *"Is there anything in particular that you liked about the experience?"* 113
B.3 Negative feedback provided in answer to the question *"Is there anything in particular that you disliked about the experience?"* 114

List of Figures

2.1 Diagram representing the fundamentals of robot (vehicle) teleoperation, by Fong and Thorpe . . . 11
2.2 The dual ecologies of robot teleoperation . . . 12
2.3 Presence measurement methods in 41 presence studies from the years 2016 and 2017 . . . 14
2.4 A sample aerospace-related SAGAT query from 1995 [1] . . . 16
2.5 Examples of real-world applications of robot teleoperation . . . 17
2.6 Examples of emotion models . . . 29
2.7 The Geneva Emotion Wheel . . . 30
2.8 Examples of emotion portrayal (here, the Sad emotion) by robots in literature. 32

3.1 The Geneva Emotion Wheel (ver. 3) . . . 40
3.2 The proposed components and calculated prior distribution for a non-neutral expression on the Geneva Emotion Wheel . . . 46
3.3 The proposed prior distribution for a neutral expression on the Geneva Emotion Wheel (Equation 3.11 after normalization). . . . 46
3.4 Raw data visualisations for two perceived emotions. . . . 50
3.5 Emotion expressions e_1-e_9 for the robot. . . . 53
3.6 GEW heatmaps for each tested emotion. . . . 55
3.7 Continuous map visualisations created from participant data. . . . 56

4.1 Screenshots of the testing software. . . . 68

4.2 Overall paired comprehension ratios for ‘default’ and ‘human’ prosody conditions 70
4.3 Paired comprehension ratios along both levels of participant English dialect locality (local and nonlocal) and sentence length (short, long and overall) . . 71

5.1 Side view of the set used for recording spatial audio stimuli 84
5.2 Sample view of the teleoperation screen during a SAGAT pause 88
5.3 View of the simulator screen shortly after a quiz question was asked 91
5.4 Box plot summaries of SAGAT scores and overall presence scores 93
5.5 Density of SAGAT scores observed for both audio feedback groups 93

A.1 Visualisations for all evaluated affect expressions 109
A.1 (cont.) 110

List of Abbreviations

ASD	autism spectrum disorder.
BKB	Bamford-Kowal-Bench (sentence list).
CGI	computer generated imagery.
FACS	Facial Action Coding System.
GEW	Geneva Emotion Wheel.
HINT	Hearing in Noise Test.
HRI	human-robot interaction.
HRTF	head-related transfer function.
IMU	inertial measurement unit.
IPQ	Igroup Presence Questionnaire.
ITD	interaural level difference.
NASA-TLX	NASA Task Load Index.
NUI	natural user interface.
PQ	Presence Questionnaire (Witmer & Singer).
PRoP	Personal Roving Presence.
SA	situation awareness.
SAGAT	Situation Awareness Global Assessment Technique.
SART	Situation Awareness Rating Technique.
SIN	speech-in-noise.

SPAM Situation-Present Assessment Method.

SSA social situation awareness.

SSML speech synthesis markup language.

TTS text-to-speech.

UAV unmanned aerial vehicle.

UCAV unmanned combat aerial vehicle.

UGV unmanned ground vehicle.

List of Publications

This thesis is based in part on the following peer-reviewed publications:

- A. K. Coyne, A. Murtagh, and C. McGinn, "Using the Geneva Emotion Wheel to Measure Perceived Affect in Human-Robot Interaction," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, (New York, NY, USA), pp. 491-498, ACM, Mar. 2020

  Awarded "Best Paper" in its category

- A. K. Coyne, K. Sapkota, and C. McGinn, ""Who said that?" Applying the Situation Awareness Global Assessment Technique to Social Telepresence," *ACM Transactions on Human-Robot Interaction*, pp. 1-22, Apr. 2023

- A. K. Coyne and C. McGinn, "The Effect of Human Prosody on Comprehension of TTS Robot Speech," in *Proceedings of the 32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN*, (Busan, South Korea), IEEE, 2023

as well as the following paper, currently in review:

- A. K. Coyne, A. Murtagh, and C. McGinn, "Circumplex measurement, analysis and visualisation of perceived robot affect," in *the International Journal of Social Robotics*, (manuscript submitted for review September 2023)

Chapter 1

Introduction

1.1 Background

The perception of future technology in popular culture has long involved robots; embodied, automated machines completing tasks that might be difficult or tedious for humans. The term *robot* being derived from Czech terms for work and drudgery [2], the collective mental image of service robots of the future portrays them as multi-purpose, widely accessible tools for physical and social assistance such as performing housework or providing information [3, 4]. This was reflected in predictions made by experts in the field to some degree, who heralded a time in the near future where robots would be commonplace in everyday life [5]. The vast potential benefit of new technologies for social interaction was recently highlighted during the COVID-19 pandemic, when video calls became commonplace, and, as robots cannot contract diseases and are at much less risk of transmitting them, a number of service robot platforms were developed and deployed for disinfection, telepresence and more [6, 7, 8, 9]. In a broader way, the pandemic demonstrated the value of remote work, and that, while imperfect, technology-enabled social interaction could provide much-needed contact and connection with others when lockdown measures prevented in-person meetups.

Advances in robotics are indeed at an all-time high, but their adoption has been primarily within the context of manufacturing. Starting in the mid-20th century, the Digital Revolution and information technology enabled robots to enter production lines in factories, and

today robotic automation is considered a standard practice for many common manufacturing operations, with over 3 million operational industrial robots worldwide at the end of 2021 [10]. The same is not true of social robots, which remain ever absent from everyday life, let alone social interactions. A number of social robot platforms have entered the public eye such as Sophia [11], but they have not demonstrated any meaningful utility in real-world applications. Those that have shown potential in the real world are simple service devices such as automated vacuum cleaners, and seldom have social components. Most well-known platforms such as Asimo [12] and Pepper [13] exist primarily as research platforms, with price tags that render them inaccessible elsewhere. When social robots do enter real-world use, such as in supermarkets and hotel lobbies, their added value is solely as novelty attractions. As they are seldom able to operate without a human supervisor nearby, any potential workforce increase is cancelled out. This is in spite of the fact that on paper, modern social robots are powerful and versatile platforms; they are generally equipped with wide ranges of sensors, and even without any in-depth social interaction algorithms they employ software to strong social benefit, such as gaze following and idle "breathing" animations [13, 14].

A substantial reason for the lack of adoption of social robots is the open-ended nature of social interaction. **Tasks which require structured behaviour in response to predictable environments can much more feasibly be addressed by robots than the complexities of social interaction,** and despite focus on AI being at an all-time high, its capabilities are still too limited for smooth, autonomous operation of robots in social contexts. Recent advances in natural language processing through language models such as ChatGPT [15] give the illusion of social intelligence, but as opposed to what humans are capable of, there are still vast limitations in the the internal ontologies of these agents [16]. Social interaction requires consistent detection and advanced processing of high-level concepts and social cues, which artificial agents still struggle to achieve. As a consequence, there can be a mismatch between the expected and observed abilities of autonomous robots that interact through the social interface, particularly humanoid robots. This results in a degree of distrust, and the robot being relegated to the role of a novelty "toy". When this mismatch is observed with social interaction specifically, it can result in more severe feelings of unease

in observers – the infamous "uncanny valley" phenomenon [17, 18, 19].

With these limitations, the societal value of social robots in theory is much greater than that of recently deployed social robots, which are seldom more than gimmicks. Considering that human oversight is currently essential in any robot automation, particularly when concerning social interaction, what can bridge the gap to adoption caused by the current shortcomings of machine intelligence is to mitigate reliance on AI altogether, and shift to human-controlled **social robot teleoperation**.

This thesis details a body of research which addresses current substantial limitations to this task, enabling the design of social teleoperation interfaces for existing service robots. As shall be explored, the use of social robots in this manner would enable their adoption, and social teleoperation would provide additional societal benefits beyond the hypothetical advantages of automation mentioned above.

1.2 Motivation

Social interaction through human teleoperation of a robot, essentially allowing social robots to become remote surrogates for operators (a form of embodied telepresence), could provide a number of substantial benefits to both individuals and companies.

Enhanced remote working The COVID-19 pandemic proved the value of remote work for both employers and employees. As of yet, working from home is really only suited for office jobs, but an embodied form of social telepresence would allow a great increase of service roles to be completed remotely, in face-to-face roles such as in customer service. This could be done despite any operator distance or mobility constraints, and would potentially broaden a tight workforce (such as in healthcare, which is currently undergoing a severe workforce shortage[1]). Reliance on automation to broaden a workforce is potentially risky in sensitive settings such as hospitals, where operational mistakes can have substantial consequences; **social robot teleoperation would enable this workforce increase without compromising on human oversight and responsibility.**

[1]The World Health Organisation predicted a shortage of 14.5 million healthcare workers in 2030 [20]

Decoupled identity Telepresence robots allow a form of remote social interaction, but there are notable advantages to social interaction in a way that obscures the operator's identity while preserving embodiment. For one, consistency of outward appearance/personality can be advantageous for branding, child psychotherapy, or artistic performance, allowing for seamless changing of operators, or even covert shifting of levels of automation. Another use of this is where the operator's identity may be sensitive information; a social robot can obscure personally-identifiable features. Finally, such a system enables an exploration of identity itself. Decoupling of physical presence from appearance allows a novel opportunity for individuals to engage with others in a fashion similar to online communication – specifically, it allows for the presentation of an identity physically different from that of the operator. This may be useful in alleviating issues such as gender dysphoria or body dysmorphia, while also allowing the opportunity to explore socially without fear of judgement or bigotry [21].

Co-design of autonomous behaviour As mentioned above, artificial intelligence still struggles to achieve the level of social interaction that comes naturally to a human. As such, a substantial boon of human teleoperation of social robots lies in the teleoperation data. Recording how a human responds to social situations while operating the robot can inform the design of social artificial agents, providing insights for engineers and training data for machine learning. With this perspective, **social robot teleoperation would be more than just a stopgap while autonomous social systems improve, but would in fact assist and accellerate their development.**

The overall challenge for social interaction to be possible through robot teleoperation is that **operator social cues must successfully be conveyed using the platform, and remote social cues must be successfully noticed by the operator**. Despite the quality and availability of modern robots and the ever-decreasing cost thereof, there are a number of obstacles and unknowns towards reaching this requirement of social exchange, notably how challenging these qualities of social awareness are to measure. Robot teleoperation in general is very well-studied outside of the social domain, but social teleoperation is substantially less so; human-robot interaction is still an emerging field, that has yet to find its footing with

scientific rigour, particularly by comparison to psychology studies [22, 23]. At present, the small pool of existing research related to social robot teleoperation suffers from a lack of repeatability and result comparability. The majority of research projects use unique specialised robot platforms, along with a wide range of differing analysis tools, the results of which are often only relevant for the robot in question.

In summary, **many of the challenges to the goal of social robot teleoperation stem from the very large possible design space, along with a lack of consensus on best practices in research and design**. For robot designers to begin to narrow down this space, to readily and reliably trial and evaluate social teleoperation interfaces, it is imperative that universally-applicable statistical tools and methodologies are available that can measure performance in specific socially-relevant areas, which are explored in the following section. Without these, research findings cannot truly be compared between different robots or their control interfaces, inhibiting advancement.

1.3 Research aims and objectives

The aim of this research project is to **resolve key issues in the design of social robot teleoperation interfaces**. As mentioned above, many current design publications in HRI are *platform-dependent*—the findings they present heavily depend on the robot platform used, and are seldom applicable elsewhere. While this can be of value to designers when using more common robot platforms such as Nao [24], teleoperated social robots remain practically inexistent in real-world applications, By taking a *platform-agnostic* approach, with methods applicable to any robot platform with social interaction capabilities, this thesis aims to yield findings that can assist in the design process of new robots, as well as to existing robot platforms. A similar constraint applies to the operator interface; novel input devices such as head-mounted displays are not necessarily available or usable for every robot teleoperation scenario - the methodology must be applicable to most teleoperation interfaces regardless of input device. The exception to this can be the standard PC setups (mouse, keyboard and monitor, with headphones and microphone), given their widespread availability and common

choice for robot control stations. This section explores which key issues need to be addressed within these research limitations.

There are many modalities of social communication, the primary two being the *vocal* modality (speech, including the channels of both *content* and *prosody*) and the *visuospatial* modality (body language, which can be broken into channels such as *gesture, gaze, facial expression*, etc) [25]. Many if not all of these can convey *affect* (emotion), speech prosody and facial expression in particular [26, 27].

The first modality of Speech can be considered the primary method of social communication. There are any number of solutions to conveying speech through teleoperation, be it keyboard input or more advanced methods such as speech recognition. A potential disadvantage of keyboard use, while providing more precise input, is the lack of control over speech prosody in the resulting utterance. While prosody itself is a channel of communication, it can also pose disadvantages such as lowered intelligibility. The effect of prosody on the comprehension of TTS robot speech has not been studied.

As for body language, an issue with its study for robots is that of platform-dependency. A gesture expressed by one robot may not be possible to be expressed by another, for example a robot with a fixed head position cannot express head tilts in the way a robot with an articulated neck could. Furthermore, it is possible for robots to convey body language cues through unique non-human ways, such as LED colour. For this research project, a measurable quality is required to evaluate the quality of social cues conveyed through robot body language, without relying on specific robot-dependent metrics. One important feature of body language is *affect expression* as it incorporates facial expressions, another is *attention* as it incorporates gaze and posture. While it is clearly of value to represent the latter for the purposes of conveying intent and joint attention, accurate perception of robot affect is arguably much more important; it is essential that the robot conveys to the viewer the intended emotion that the operator selected to mitigate uncanniness.

The above solely concerns expression of social cues by the operator. As mentioned, the other half of the exchange is reciprocal comprehension of social cues; therefore it is also

essential the operator successfully picks up on remote social cues with the person interacting, in a form of social awareness. With sound/hearing, a common abstraction between in-person and remote sensing is the spatial nature of audio feedback; in-person hearing intuitively conveys audio directionality to the listener, whereas a common mono audio feedback setup does not provide this. Using multi-channel microphone setups and complex signal processing, spatial audio feedback can be replicated remotely, which is known to enhance immersion in a remote environment. However to implement it for robot teleoperation can introduce overhead because of the additional hardware and software resources required. Given that it is currently unknown to what degree spatial audio feedback can assist in the perception of social cues, this is an important design tradeoff which remains unexplored.

Research objectives

The above preliminary analysis has determined a number of key issues which need to be addressed in order to enable social robot teleoperation. Towards this aim, this research project has the following objectives:

- **Develop and validate a methodology for measuring and analysing perceived robot affect**
- **Evaluate the effect of human-informed prosody on comprehension of robot speech**
- **Evaluate the effect of spatial audio feedback on social situation awareness during robot teleoperation**

The research conducted in this thesis shall form a palette of participant-testing tools that any robot designer can use to successfully create a robot teleoperation interface. Towards achieving the research objectives, the following *contributions* shall be made to the field:

- **A suite of complementary statistical methods for circumplex analysis of robot affect data data acquired using the Geneva Emotion Wheel**, for:
  - individual frequentist evaluation of affect expressions

  - overall Bayesian mapping of affect expression range

- **A methodology for applying the Situation Awareness Global Assessment Technique to the domain of social interaction**

1.4 Thesis overview

This thesis is structured as follows: the following chapter is a review of the literature relating to social robot teleoperation, including the theory and current applications thereof, and the modalities of social interaction that it employs. Chapter 3 develops methodological tools for evaluating perceived robot affect expression, through heavy use of an emotion model and input device known as the Geneva Emotion Wheel, and trials them in a participant study. Chapter 4 focuses on robot speech prosody, in a study to evaluate intelligibility of robot speech with and without prosody tuned to match that of a local, drawing from the field of audiology. Chapter 5 details a study evaluating the social benefit of spatial audio feedback during robot teleoperation, introducing the concept of *social situation awareness* (the intersection of social and situation awareness) and proposing a methodology for its measurement. The thesis then concludes with a review of the completed work and research objectives, and identifies potential subsequent topics of research in social robot teleoperation.

Chapter 2

Literature Review

This chapter shall begin by revisiting the fundamental theories in the study of robot teleoperation in 2.1, along with common performance metrics for its evaluation, followed by a review of current real-world uses of robot teleoperation in 2.2. The subsequent section opens with an examination of the specific requirements for social interaction through robot teleoperation, breaking down a number of interaction modalities and key social cues in 2.3.1, and common practices and challenges faced in the literature for each are discussed in the remainder of 2.3. The findings of the literature review are summarised in 2.4.

2.1 Teleoperation theory and concepts

2.1.1 Human control and automation theory

Automation is arguably one of the defining characteristics of robotics, as it allows for complex tasks to be performed without requiring a human. There has long been interest in including humans in the operation of robots however, to ensure oversight and responsibility over the robot's actions, particularly in dynamic environments where there is a larger chance of robot error.

In 1993, Sheridan provided the following definition: "The term teleoperation most commonly refers to direct and continuous control of a [robot]" [28]. Fong and Thorpe considered the term as "encompassing the broader spectrum from manual to supervisory control" [29],

High	10. *The computer decides everything, acts autonomously, ignoring the human.*
	9. *informs the human only if it, the computer, decides to*
	8. *informs the human only if asked, or*
	7. *executes automatically, then necessarily informs the human, and*
	6. *allows the human a restricted time to veto before automatic execution, or*
	5. *executes that suggestion if the human approves, or*
	4. *suggests one alternative*
	3. *narrows the selection down a few, or*
	2. *The computer offers a complete set of decision/action alternatives, or*
Low	1. *The computer offers no assistance: human must take all decisions and actions*

Table 2.1: The Sheridan-Verplank Levels of Automation of Decision and Action Selection [30], from [31, Tab. 1]. In this model, the level of automation is inversely related to the level of human control/teleoperation over the system.

allowing for autonomy in the control loop; robot teleoperation could be at a low level, where individual actuators receive commands from the operator, or more abstracted, where the operator issues high-level orders and the robot calculates how to enact them—"human-in-the-loop" (or supervisory "human-*on*-the-loop") control paradigms. This scale was quantified in the *Sheridan-Verplank Levels of Automation* [30], shown in Table 2.1. Widely cited in the literature, this taxonomy has received criticism in more recent years [32, 33], including from one of the original authors [34], notably for the insufficiency of a single dimension to describe autonomy and the presumption of a strict tradeoff between AI/autonomy and human control of robots. The scale is nonetheless illustrative of the potential connection between the two concepts, and the scope and variety of autonomous and teleoperated systems.

A number of inherent problems with autonomy were reviewed in an article by Shneiderman [33], including the common point that "humans have to spend more effort monitoring autonomous computers, often leading to inferior performance"—historically, with the need for human oversight, unmanned systems do not enable to any reduction of human workforce [35, 36].

2.1.2 Teleoperation and the "dual ecology"

Although it is not always their primary purpose, the ability to be teleoperated is a fundamental feature for many robot platforms. Rudimentary teleoperation interfaces serve as

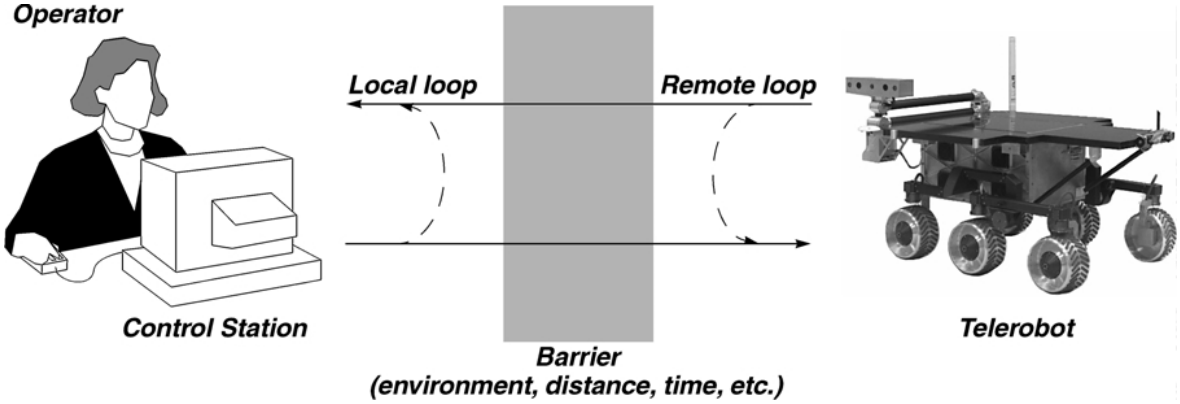


Figure 2.1: Diagram representing the fundamentals of robot (vehicle) teleoperation, by Fong and Thorpe. [29, Fig. 1] Original caption: *Vehicle teleoperation. An operator at a control station generates commands and receives feedback from displays. The remote vehicle executes the commands, often using some level of on-board autonomy.*

development tools (debugging) and more polished systems can supplant the need for automated behaviour entirely. Including teleoperation interfaces can increase robot acceptance, normalising and commodifying the robot in question [37]. As expressed by Fong and Thorpe [29], robot teleoperation allows a robot to be remotely controlled by a human operator in a way that would not be possible otherwise because of environmental, distance or other limitations. This is illustrated in Figure 2.1, which shows the "barrier" between operator and robot—in cognitive interaction terms, this barrier can be considered as separating the two *ecologies* of robot teleoperation.

While autonomous robot operation would theoretically take place in a single given ecology (the environment the robot is present in, along with any other artifacts in the space), remote robot teleoperation leads to a "dual ecology"; the ecology of the remote operator (i.e. the robot teleoperation interface, such as a PC desk) as well as the ecology proximate to the robot.[1] The "dual ecologies problem" was described by Kuzuoka et al. [38], who pointed out the importance of continuity across the dual ecology for teleoperation design; discrepancies in continuity across both ecologies is said to hinder proper social communication, as the remote operator's experience at the robot control station can vary substantially from that of

[1]While the two ecologies can physically overlap in the specific situation where a robot is in the same environment as the operator, "ecology" here is a cognitive interaction concept and not necessarily an understanding of physical location.

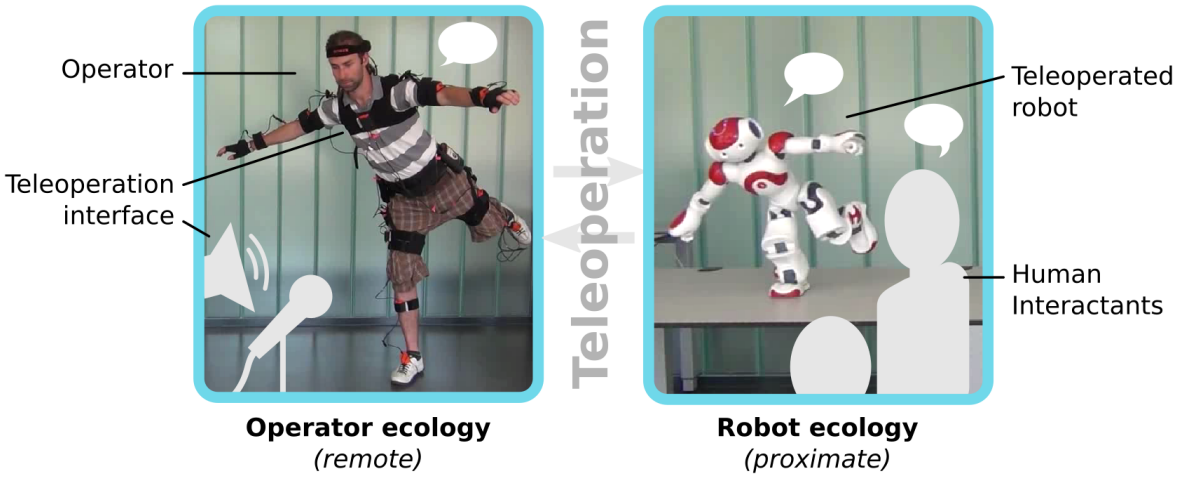


Figure 2.2: The dual ecologies of robot teleoperation. This can lead to the "dual ecologies problem" when inconsistencies between the two ecologies results in communication problems [38], particularly for humanoid robots. Here, an operator in the remote ecology controls the motions of a Nao robot with full body tracking, to interact socially with people in the robot's ecology ("interactants").
Photos from [39, Fig. 1], modified for demonstration purposes

a person interacting locally with the robot. This is the case for discrepancies of awareness (the operator is not aware of things that may be apparent to proximate interactants) [38, 40], embodiment (the operator does not experience the robot's body as their own) [41, 42], and personality/identity (the robot's perceived personality clashes with that expressed by the operator) [41, 43, 44]—a review by Goodrich et al. determined that "dual ecology impacts almost every aspect of designing teleoperation for humanoid robots" [45]. The dual ecology framework has also been used outside of robotics, in the study of augmented reality interactions [46, 47].

As with human-computer interaction, human-robot interaction (HRI) has been considered an intersection of engineering and social or cognitive psychology [22], which has been defined as "a field of study dedicated to understanding, designing, and evaluating robotic systems for use by or with humans" [48]. Therefore, in the dual ecology of social robot teleoperation, both ecologies are human-robot interaction scenarios; the operator interacts with the robot through the teleoperation interface, and the robot interacts socially with people in its ecology (referred to as "interactants"[2]), through what is known as the social interface [18].

[2]Goodrich et al. used the term "interactants" [45] to refer to every human in this scenario (including the

2.1.3 Performance metrics and evaluation

A number of desirable qualities have been identified in the evaluation of robot teleoperation interfaces; while they can sometimes overlap, those relevant to the performance of social teleoperation include *telepresence, embodiment* and *situation awareness*.

Telepresence and embodiment

Where the robot is designed to emulate for the operator the experience of being remotely present in a location, the term "telepresence" is used, originally coined by Minsky in 1980 [49].

What differentiates telepresence from regular robot teleoperation is its focus on the quality known as presence, as experienced by the operator [50]. This was defined as "the degree to which users are involved in and feel part of the task environment" by Witmer and Singer [51], and it was shown to be linked to task performance, situation awareness and simulator sickness. Existing methods to measure one's sense of presence commonly take the form of questionnaires, the first being the Witmer and Singer Presence Questionnaire (PQ) [51] focusing on operator experience. Subsequent questionnaires include the Igroup Presence Questionnaire (IPQ), which purportedly measures presence on a higher level than just through component factors, and the Temple Presence Inventory which allows for multiple "types of presence" [52]. A review by Hein et al. [53] showed that questionnaires are overwhelmingly used for measuring presence in research (as shown in Figure 2.3), with the Witmer and Singer PQ remaining the standard, with any objective measures being either physiological (such as electrodermal activity) or behavioural (through qualitative observation of experiment recordings).

A term with a similar if not identical meaning to telepresence is "telexistence", coined around the same time by Tachi [54]—modern use of this term is similar to that of "tele-embodiment" by Almeida et al. [55], implying a heavy focus on simulated embodiment and immersive feedback for the operator [56, 57, 58]. While Minsky's original meaning for telep-

operator as well as any third parties, such as specialists overseeing the robot's use)—for conciseness, this thesis will use the term "interactant" to refer only to the people in the robot's ecology interacting with it socially, as in Figure 2.2

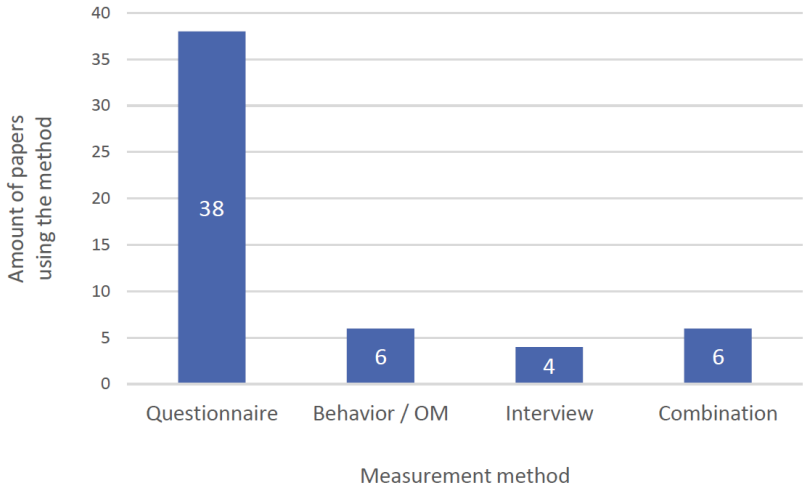


Figure 2.3: Presence measurement methods in 41 presence studies from the years 2016 and 2017 [53, Fig. 2])

resence implied a similarly immersive experience, the modern interpretation of telepresence can be fairly minimal by comparison, only remotely emulating the social aspects of presence.[3] However, studies have shown that increased levels of embodiment, moving towards telexistence, can enhance the operator's sense of presence as well as the perceived depth of social exchange for interactants [42, 59, 60, 44].

Situation awareness

A quality focusing less on operator experience and more on performance is that of situation awareness (SA). Originally designed to evaluate the performance of military air pilots, it was defined by Endsley as "the perception of environmental elements and events with respect to time or space, the comprehension of their meaning, and the projection of their future status" [61]. This was built upon by Mogford [62] concerning the task of aircraft control and coordination; by linking the concept to that of the mental model, situation awareness was defined as "the transient contents of awareness, as structured and supported by an underlying mental model", also described as "the holding area where images and verbal data are maintained for ready use". Modern use has expanded beyond aeronautical applications; Yanco et al. [63]

[3] See 2.1.3 for examples of telepresence robots.

refined Endsley's definition more recently for the purposes of robot teleoperation, defining situation awareness as "the perception of the robots' location, surroundings, and status; the comprehension of their meaning; and the projection of how the robot will behave in the near future". SA is considered an important quality in robot teleoperation, as it is linked to task performance, workload and usability [64].

The connection between situation awareness and the similar but distinct concept of *presence* explored in 2.1.3 has been pointed out in the literature [50, 65, 45], linking conventional robot teleoperation and social robotics. Goodrich et al. [45] explain that while presence and SA do not share the same definition, a high sense of presence implies the remote operator has sufficient awareness of the robot's proximate situation to feel present. Draper et al. [50] considered telepresence to be the maximisation of SA in the robot's environment, and an accompanying loss of SA for the operator's environment.

While seldom if ever applied to social teleoperation, techniques for situation awareness (SA) evaluation are well-established for non-social tasks and contexts. A number of such techniques would be considered standard in human-robot interaction research [63]:

SAGAT Endsley's Situation Awareness Global Assessment Technique (SAGAT) [66] is possibly the most commonly-used SA evaluation technique. It requires a simulated environment for the teleoperated robot. Periodically, the simulation is paused and questions are asked to the operator about the situation. SAGAT questions are divided into three categories, each measuring a particular degree of awareness (immediate, recall and future estimation). This reproducible, empirically validated technique results in an objective score measuring situation awareness, with the caveat of being incompatible with real-world robot teleoperation. While intended for the piloting of aerial vehicles (as shown in Figure 2.4), it has also been applied to medical simulations [67]. SAGAT has also been adapted to use real-time querying of the operator instead of pauses [68]. Overall, SAGAT can be considered the standard technique for objective measurement of situation awareness.

SPAM The Situation Present Assessment Method (SPAM) is another adaptation of SAGAT using real-time querying of the operator [69, 70]. By prompting the operator to answer ques-

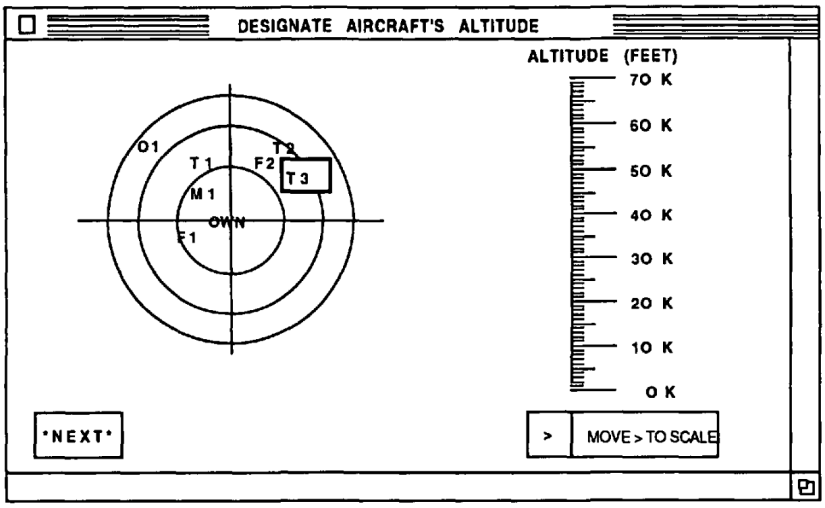


Figure 2.4: A sample aerospace-related SAGAT query from 1995 [1]

tions while robot operation is ongoing, it can be administered during real-world robot operation (outside of simulation). When a query is administered, the operator is first notified, and given the option to open the query, upon which they can read and answer. As the operator is free to seek the required information to answer the question yielding high accuracy rates, response score is not a useful metric compared to SAGAT, rather the time taken to answer the question is considered the indicator of situation awareness, and the time taken for operator to open the query after being notified is considered a measurement of workload.[4] The validity of SPAM as a situation awareness measurement tool is entirely dependent on that of SAGAT however, and it has received little empirical validation in itself.[5]

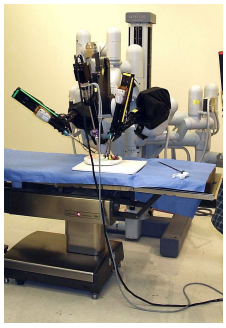
SART Unlike the objective measurement techniques above, Taylor's Situation Awareness Rating Technique (SART) [74] is a commonly-used subjective measurement. A form with three sets of experiential Likert-scale questions is completed by the operator after the tele-

[4]While SAGAT hinges on operator memory of the environment to measure SA, the SPAM proposal argues that a more accurate understanding of SA is how readily available information is to the operator from the environment, rather than what is immediately in the operator's mind when the simulation is paused. This is congruent with the cognitive interaction concept of distributed cognition [71].

[5]A recent meta-analysis by Endsley [72] called the validity of SPAM into question, finding it less sensitive and more intrusive to operators than SAGAT, however this review in itself was critiqued [73] for using spurious methodology, leading to unfalsifiable claims and predetermined results "running contrary to the purpose of meta-analysis".



(a) The USSR *Tele-Tank*, an early example of teleoperation [77, p. 12]



(b) The *DaVinci* laparoscopic robotic surgery machine.[6]



(c) *Robonaut* teleoperated humanoid [78, fig. 1]



(d) *Beam* telepresence robot [79, fig. 1a]

Figure 2.5: Examples of real-world applications of robot teleoperation

operation session, resulting in three scores describing different facets of awareness, which combine into a composite overall score. It is vulnerable to the same disadvantages as other forms of subjective assessment, such as bias from participants' feelings and emotional state [75], and was shown to be less accurate than SAGAT in particular [76], but it has the additional advantage of easy, universal administration. While no longer as broadly used, it can be considered a foundational measurement of situation awareness

Where social interaction is concerned, presence (discussed in 2.1.3) is often the focus of research rather than SA, and no explicit techniques have been developed for evaluating situation awareness in social contexts.

2.2 Real-world applications of teleoperated robotics

2.2.1 Task-oriented teleoperation

Current real-world uses for robot teleoperation are varied, and are predominantly applied to perform concrete, physical tasks (referred to here as "task-oriented"). Among others, these include:

[6] Source: Nimur at the English-language Wikipedia, CC BY-SA 3.0

Space or marine exploration Robots are of clear benefit in environments that are dangerous or inhospitable to humans. Autonomous underwater vehicles (AUVs) are used for marine exploration, allowing researchers to explore areas currently inaccessible by manned craft [80, 81], such as extreme depths or beneath polar ice sheets. In a similar vein, orbital manipulation is a task that is simplified with the use of teleoperated robot platforms; examples used by NASA include the "Canadarm" robotic manipulator on the International Space Station [82] or the humanoids *Robonaut* [78] (Figure 2.5c) and more recent *Valkyrie* [83]. Teleoperated ground rovers are commonly used for lunar [84] or planetary [85] exploration, with onboard sensors for scientific data collection. The high stakes of space exploration also required the creation of NASA-TLX for task load measurement, which has entered widespread use in user experience studies and HRI [86, 87].

Military robots Similarly, the use of teleoperated robots for military purposes can mitigate casualties for the side that deploys them, although the ethics of such use is heavily debated [88, 89]. One of the earliest remotely-controlled ground vehicles was the USSR *Tele-tank* (Figure 2.5a), used in World War II [77, p. 12]. Nowadays, arguably the most widely-recognised form of military robot is the unmanned combat aerial vehicle (UCAV) [90, 29]. To control its navigation and combat modalities, operators use a dedicated control interface with a satellite uplink. Other robots used for military activities include unmanned ground vehicles (UGVs), often for reconnaissance as opposed to combat [91, 92, 93]; maintaining locomotion but foregoing combat modalities. Military-funded robotics research has substantially influenced non-military robot teleoperation studies, such as through the identification and measurement of constructs such as presence [51] and situation awareness [66].

Search-and-rescue robots Navigating rubble or other dangerous environments in emergency situations is a hazardous and can put more humans' lives at risk. Search-and-rescue robots [94, 95] are designed to perform this task, locating injured or trapped people while teleoperated by a human at a safe distance. Research interest in these robots peaked after the 2001 World Trade Center attacks [96, 97] and a recent robot search-and-rescue operation was that of the Titan submersible incident of 2023 [98].

Telemedicine Telesurgery [99] allows specialist surgeons to be available to patients worldwide despite distance constraints, provided the patient's local hospital is equipped with a telesurgery robot and that the surgeon has access to a suitable teleoperation interface. Robotic surgery when performed locally [100] can also help to overcome the limitations of conventional laparoscopic surgery, or keyhole surgery.[7] The high-risk nature of the task requires surgery robots to have qualities such as extra precision or haptic feedback [102], by comparison to other manipulator robots.

By examining current industry applications of robot teleoperation, we can see that *navigation* and *manipulation* are task modalities that commonly recur.

2.2.2 Social teleoperation

While some overlap can exist, social interaction requires an entirely different set of control modalities to the task-driven robots mentioned above. Teleoperated robots currently in use for their social applications are primarily in two categories; telepresence, and robot therapy.

Telepresence Originally named Personal Roving Presence (PRoP) in the 90's [103, 104], telepresence robots allow people to participate remotely in activities as a substitute to being present [105, 106] (as explained in 2.1.3). They often consist of a tablet with teleconferencing software suspended over a wheeled chassis to provide mobility (informally, an "iPad-on-a-stick"), as shown in Figure 2.5d, allowing the remote user to move around and interact with those present at the robot's location; these navigation faculties are the key advantage to current telepresence robots compared to conventional teleconferencing [79, 107]. The current COVID-19 pandemic is leading to an increase in use of telepresence robots [6, 9]—while not humanoid or immersively embodied, they allow social distancing while approximating face-to-face social interaction.

Robot-mediated therapy The controlled social interface that robots can offer presents a unique advantage in the field of psychotherapy. This is useful in particular for children and

[7] The current cost of laparoscopic robots may outweigh the benefits they can provide to hospitals, however [101].

teenagers with autism spectrum disorder (ASD), where social robots have been found to elicit novel engagement and interest [108, 109, 110, 45]. While not yet accepted as standard in the field, some therapists use teleoperated social robots regularly in their practice [111, 112].

Outside of teleoperation, one of the main occurrences of social robots enter real-world use is to exploit their novel nature for advertising and commercial appeal, such as in supermarkets [113], as receptionists [114], or as bartenders [115]. The added value here is primarily as novelty "gimmick" attractions [116, 23].

2.3 Key cues and modalities of interaction for teleoperation

The goal of social teleoperation is to enable social interaction between the teleoperated robot and the interactants to a sufficiently similar level to in-person interaction. The fundamental challenge of social interaction through robot teleoperation can therefore be understood through the bidirectional communication required across the dual ecology, as visible in Figures 2.1 and 2.2—the operator must be able to **successfully convey social cues** (by controlling the robot), while also **picking up on the social cues of interactants** (by reading feedback). The literature surrounding both sides of this exchange shall be explored here, first drawing on multimodal interaction concepts of social psychology to establish the essential modalities and social cues necessary for social teleoperation, then examining the hurdles faced in research for each.

2.3.1 Understanding multimodal interaction

An important trait of face-to-face social interaction, by comparison to more formalised methods of communication such as printed word, is its inherently multimodal nature [117]. Breaking down modalities and channels of social interaction can demonstrate which of them are key focuses in social robotics.

Enfield [25] distinguishes two core modalities, which are each comprised of a number of channels (shown in Table 2.2); the *vocal* modality incorporating the *lexico-syntactic* and *prosodic* channels, and the *visuospatial* modality which includes channels such as *hand gesture*,

Modality	Vocal	Visuospatial
Channels	**Lexico-syntactic**	**Facial expression**
	Textual speech content, utterances etc	**Gaze**
	Prosodic	**Proxemics**
	Pitch, intensity, timbre etc	...

Table 2.2: Partial breakdown of social interaction modalities in Enfield's model [25]. Noteworthy modalities and channels are listed here; many other channels can exist in or out of these modalities.

facial expression, *gaze* and more. These modalities and their channels act in tandem to enable complex social exchange, beyond what a single channel can provide. Affect is increasingly understood as being inherently social [118, 119], and it is one key piece of social information that these modalities can collectively communicate, notably through facial expression [120] and speech prosody [121].

2.3.2 Task-oriented modalities: Navigation, manipulation

Although not social modalities or cues, as seen in 2.2.1 navigation and manipulation are fundamental to the literature around robot teleoperation.

Navigation For remotely controlled robots considered "vehicles", navigation can often be the only controlled modality. This applies to many of the vehicles reviewed by Fong and Thorpe in 2001 [29, 122]. Being a non-social modality of interaction with the environment, little research exists focusing on navigation during robot teleoperation in explicitly social contexts. Initial research on teleoperated robot navigation emerged from military applications, but more recent research concern navigation of smaller UAVs and UGVs, as well as humanoid robots.

Many examples of robot navigation control in research have used conventional input devices, such as joysticks [29, 123, 40, 124], gamepads [125, 126, 127, 128] and the mouse and keyboard [129, 104, 130]. Slightly less conventional are haptic devices, which have been explored to provide additional feedback during navigation of a UGV [131, 132]. Natural user interfaces (NUIs) have also been used for navigation control in research, tracking an operator's

body motions with depth cameras [133, 134, 126, 135] or wearable IMUs [136, 55, 125], mapping head and arm gestures to velocity or positional commands.

Manipulation Robotic manipulation frequently finds use in industrial or research applications, the former field in which it is seldom teleoperated. It is common knowledge that, as opposed to humans, machines when properly designed can do a single task very well, but often struggle with variation. This is particularly the case with manipulation tasks; industrial robots can perform the same precise movement with very high accuracy, but are rarely designed to adapt to unusual conditions.

Teleoperation interfaces for manipulation were compared in a paper by Kent et al. [137], for supervisory control. The studied control systems used mouse and keyboard input for supervisory control of the robot manipulator, defining a grasp points in three dimensions using regular two-dimensional input. Manipulation has also been handled on a high level through a touchscreen interface [138]. For more lower-level control, the use of NUIs found interest for manipulation, through the use of a Microsoft Kinect depth camera for input [139, 140, 141], as well as through wearable IMUs [142, 143].

Given that teleoperation research began for practical, task-driven robots such as search-and-rescue or military robots, the modalities these robots act through in order to complete these tasks are well-researched. While they do not focus on social interaction, the design approaches for some may yield benefits for the design of social robots, particularly for conveying gesture through body tracking.

2.3.3 The vocal modality

Definitions

In the field of semiotics, the vocal modality is be considered the more symbolic method of social communication [144, 25, 117]—the symbols it communicates (e.g. language) have tangible meanings they are associated with. In Enfield's model, the vocal modality is composed of both the *lexico-syntactic* channel (the content of speech, "what is being said") and the

nonverbal *prosodic* channel, which "in perceptual terms, accounts for *how* something is said" [145][8] (accent, tone of voice etc). In this regard, speech is arguably the primary form of social communication, particularly the lexico-syntactic channel.

While less symbolic, prosody is a channel that nonetheless communicates information, discretely from the lexico-syntactic channel [117]. It provides syntactic information pertaining to a sentence [121], and can convey implicit social cues such as affect ("emotion") and "personality" [146]. There is little consensus on the amount of components of prosody. Some of the principal components of include the pitch (fundamental frequency), loudness (intensity), length (duration) and timbre (spectral characteristics) of speech segments [147]. Munhall et al. [148] demonstrated that the prosody construct as it is understood may also include a visual component, somewhat breaking out of Enfield's bimodal model of vocal versus visuospatial. In 1969, Crystal proposed that *pitch*, *loudness* and *length* of speech segments were the three primary components [149], and this has been recurrently used in more recent engineering research on the topic [150, 151, 152]. It has been shown to contribute to speech intelligibility by disambiguating sentences [153]; prosody is intuitively understood to map to certain syntactic structures, and if a sentence itself was not fully heard, the listener can often use the prosody of the sentence to infer the missing content. This is only possible if the listener is familiar with local prosodic patterns however; prosody is language- and culture-dependent [121, 147, 154], and is a contributor to accent, both foreign [155, 156] and native [157, 158, 159]. Indeed, in a 2000 experiment, it was shown that for native Cantonese speakers listening to English, some of the disambiguating benefits of prosody were diminished as "*they did not seem to have the relevant knowledge of how [prosody] resolves ambiguity*" [154, p. 187]. It was also noted that the prosody of casual speech was less intelligible than that of certain deliberate styles of speech [160].

Both through prosody and in general, speech can convey a number of indicators to the listener on the speaker's identity, social and cultural background. In the context of speech interfaces, it is said that speech informs the *partner model* that the listener forms of the speaker—in human-machine interactions, partner models have been shown to greatly impact

[8] quoted from [146, p. 1063]

listener assumptions of knowledge and capability [161] as well as perceived humanness [162, 163]. In a phenomenon known as the "similarity-attraction effect", partner models that are more similar to the listener's own identity (e.g. informed by a voice with a similar accent to the listener) will be perceived more favourably, assumed to have more knowledge and greater ability than speakers identified as different [164, 165].

Speech teleoperation

It is widely known that language is the most important medium of communication between humans. Robots without speech or other language faculties will always be limited in their communication with humans, and thereby in their applications.

As pre-recorded audio is very simple to play back through a robot's speakers, it has been used to play back utterances in several research projects [166, 167, 168]. This is likely worth the added functionality compared to no speech at all [169], at the cost of providing much less flexibility for the robot than full speech generation. For telepresence robots, playing back the operator's voice is the natural option [105] as their purpose is to represent the operator at a remote location. It has been used for social robots in several studies [170, 171], with the caveat that the robot's identity is inconsistent.

Speech capabilities are not rare in social robots, with many social robots being easily equipped with TTS engines [12, 172, 173, 174, 13, 175, 14], and modern engines can purportedly be indistinguishable to recorded human speech [176]. Despite this, few solutions have been considered for *teleoperated* control of speech. As a consequence, existing teleoperation interfaces nearly all consist of keyboard input to the robot's TTS engine, without examination of validity or performance—a potential disadvantage of this basic keyboard input is the lack of control over speech prosody. A alternative proof-of-concept of "voice puppetry" was showcased by Aylett et al. [177], combining operator voice recognition and signal analysis to supply the text-to-speech engine, allowing minimal prosody control. Voice modulation is an non-TTS solution that has been used in Wizard-of-Oz experiments [178, 179] and although it has not been used for robot teleoperation, the recent rise of neural voice modulation or "voice filtering" to the public eye, using machine learning models to alter speech audio, may allow

smooth prosodic voice teleoperation in real time bypassing a TTS engine [180, 181, 182].

A recognised problem with TTS engines was that the speech generated would lack expressivity [169, 183, 184], and mechanical-sounding voices have been shown to decrease robot acceptance [185]. As argued by Yilmazyildiz et al. [179], this is a major disadvantage of TTS-generated audio in teleoperation as opposed to playing back human voice audio (as is done for robot speech in their paper). Overall, despite speech synthesis being essential for speech-capable robots, recent research has indicated that researchers often do not give much consideration to the voice the robot is given during HRI experiments, which has been shown to affect the partner models that they subsequently form of the robot [184, 186].

Prosody control

The ability to produce prosodic speech to differing extents is a feature in many speech synthesis engines [187], often through the use of the World Wide Web Consortium standard Speech Synthesis Markup Language (SSML) [188, 189], although Ezzerg et al. observed that producing more prosodically expressive speech tends to result in inferior audio quality [180].

Existing research into robot speech prosody has primarily been towards the goal of conveying emotion [190, 191]. As reviewed in 2016 by Crumpton and Bethel [192], conveying emotion through robot speech has been done both by fine-tuning prosody in speech sections (often using SSML) or by applying more global prosody settings to entire statements (with other markup tools such as EmotionML [193]). Other studies explore the impact of prosody on naturalness, robot acceptance and expressivity [183, 194], notably through the prosodic feature of intonation. In a study by van Straten et al. [110] robot intonation was shown to impact the affective state of children with autism spectrum conditions, and more recently, Velner, Boersma, and de Graaf [195] studied how humanlike intonation for robot speech could affect how it was perceived. The less-explored visual component of prosody has also seen initial research in the context of human-robot interaction by Hill and Vaikiotis-Bateson, through a virtual talking head “adaptable to robots" [196].

The use of prosody to increase robot speech intelligibility was explored in a preliminary study by Lee [152]. In a within-subjects study, comprehension was measured for 21 partici-

pants through recall of weather forecast information expressed by the robot, and differences were compared between three levels of overall pitch in the robot's voice (presented in fixed order to participants).

The study of robot speech being inherently multidisciplinary (drawing from robot engineering as well as linguistics and audiology), robust methods for assessing intelligibility are seldom encountered, despite being common in the study of speech synthesis [197, 198], including when pertaining to prosody [199, 200, 201]. For assessing intelligibility of speech in the fields of audiology and linguistics, speech-in-noise (SIN) methods are often employed [202, 203, 204], in which speech is overlaid with noise to increase the difficulty of the listening task to distinguish between control conditions, avoiding "ceiling performance" [204, p. 2]. For this purpose, standard sentence lists such as the Bamford-Kowal-Bench (BKB) list [205] are used. The BKB sentence lists were used in the development of several speech-in-noise tests [206] such as HINT [207] and BKB-SIN [208]. BKB sentence lists and similar lists have also used in context of speech prosody [209, 210].

2.3.4 The visuospatial modality

Definitions

Returning to Enfield's model, "body language", or the visuospatial modality, encompasses all the physical cues expressed in a social interaction—interacting through perceived physical states. This includes channels such as manual gestures, facial expressions, body posture and gaze. Semiotically, these physical cues are understood as being iconic [117] rather than symbolic—they do not concretely map to information as symbols, but are physically representative portrayals, such as a swiping hand wave to indicate dismissal.[9]

[9]This primarily applies to spoken language interactions; interactions involving sign language differ to a degree, as gesture becomes more symbolic [211]. Sign language is a complex method of social communication which is outside the scope of this thesis.

Gesture teleoperation

As body language relies on physicality, findings of publications focusing on the visuospatial modality are rarely cross-comparable. Commonalities can nonetheless be found however for robots which share humanoid features.

For those with articulated legs such as the Nao robot platform [175], full-body tracking of the operator and subsequent imitation by the robot have been developed and studied for the purposes of navigation and body pose [212, 213, 214, 39, 215], focusing on accuracy and fidelity of motion rather than social benefit.[10]

2.3.5 Gaze teleoperation

Gaze is an important social tool; as expressed in a review by Ruhland et al. [216] aside from gathering information it signals interest and affect, and regulates conversation. Research on gaze control in social robotics generally takes one of two avenues; a focus on vergence and camera focus, or a focus on the more social aspects of gaze.

Research on robot gaze is primarily for stereo vision purposes when the eyes are equipped with cameras [217, 218], and social gaze has seen minimal focus. It was found that for telepresence robots, mutual gaze resulted in better social impressions of the robots by interactants, but most robot platforms, particularly those with screen-based eyes that do not convey depth, can rely on the "Mona Lisa effect" to passively enable simulated mutual gaze without complex gaze targeting systems [219]. Ruhland et al. mention a similar effect [216] drawing from the study of animation and virtual avatars; for social purposes, gaze does not need to be accurately directed at the point of focus—the social aspects of gaze can be conveyed in a symbolic manner and are not strictly tied to accurate eye positions.

[10] Shown in Figure 2.2 (page 12)

2.3.6 Affect expression

Definitions

Affect (or emotion) is a quintessential ingredient in social interaction. As expressed in the review by Goodrich et al. [45], the value of "socioemotional" aspects in robots is increasingly recognised in the literature; allowing robots to communicate emotions during social interaction can convey information about the robot's state [220], allow it to be more persuasive [221], and increase its acceptance [222]. In particular, it has been found that robots that can convey emotions are more easily anthropomorphised, and more closely inform the mental models that observers form through interaction with the robot [223, 224].

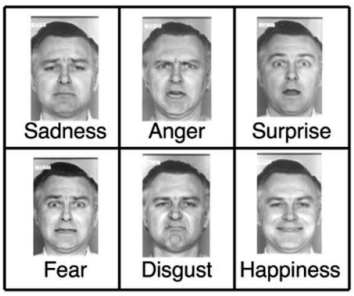
While it is primarily communicated through facial expression [120], it can be conveyed through both visuospatial (pose, movement, etc.) and vocal modalities (notably prosody) [225, 26, 226, 27].

Quantifying perceived affect

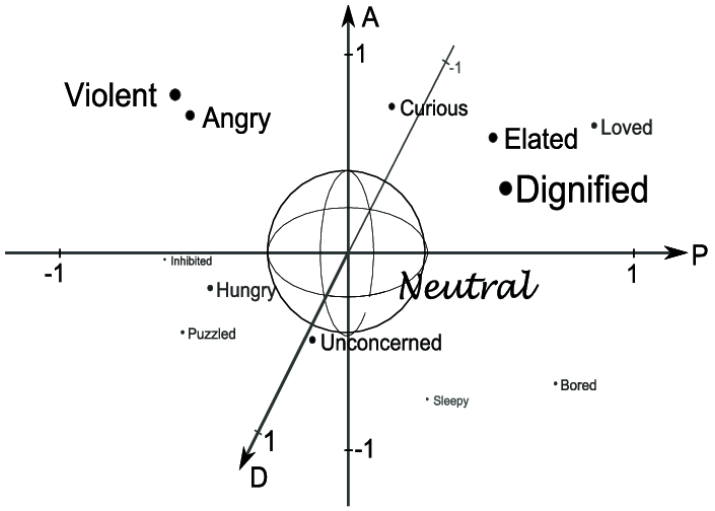
The relationship between emotions and how they are expressed has been studied for decades. As they are by definition subjective, efforts have been made to establish *models of emotion* in order to measure and understand them internally, which in turn can inform *emotion coding schemes* that link internal emotions to outward expressions of affect, such as facial movements or changes in prosody.

Ekman [232], proposed one-to-one coupling of emotions and facial expressions, and this has led to widespread adoption of Ekman's basic emotions [233, 227] (Figure 2.6a) and the Facial Action Coding System (FACS) [234] for the modeling of emotional states in human-robot interaction. FACS is a categorical emotional model; like others of the same type [235, 236] it assumes that emotions are independent of one another.

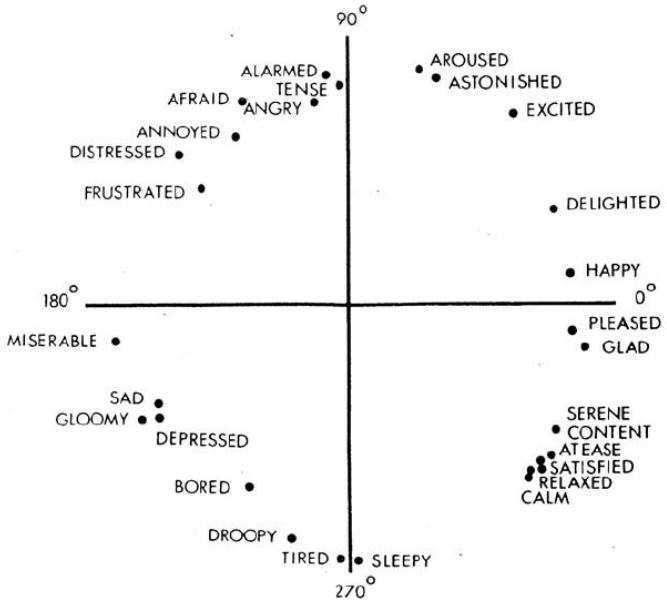
By comparison to categorical models, dimensional models where emotions are represented on a continuum can provide a richer way to represent emotional states [237] (although these states can sometimes be ambiguous and ill-defined). Commonly used in human-robot interaction is Russel's circumplex model of emotion [238], with axes "valence" and "arousal"



(a) Six of Ekman's basic emotions [227] in a categorical model. Source: [228]



(b) A representation of Mehrabian's three-dimensional PAD model [229]. Source: [230, fig. 2]



(c) Russel's circumplex model of affect, a dimensional model. [231]

Figure 2.6: Examples of emotion models

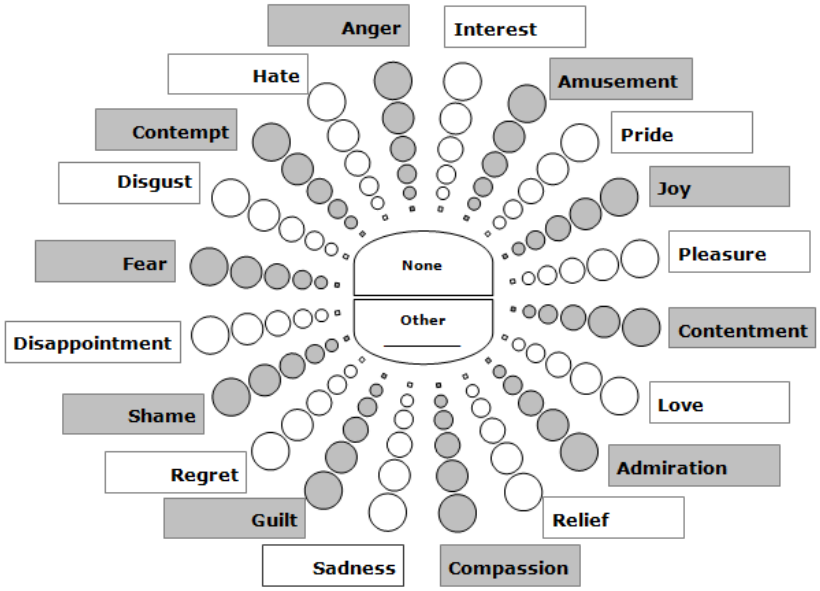


Figure 2.7: The Geneva Emotion Wheel [248]

(Figure 2.6c). An extended form of this, the PAD model (adding an extra axis of "dominance", creating a three-dimensional space, shown in Figure 2.6b) has also been used in HRI [239, 240].

In some HRI studies, dimensional models have been implemented in conjunction with categorical models. For example, it is common for researchers to first develop a range of expressions corresponding to 'basic emotions' using a coding system like FACS, and then map these expressions to defined locations in an affective space (2D/3D) which are deemed facilitate better control and more continuous transitions [241, 242, 243, 244, 245, 246, 247]. One model that combines the benefits of both categorical and dimensional emotion models, overcoming their respective drawbacks, is the Geneva Emotion Wheel (GEW) [249]. Shown in Figure 2.7, it is a dimensional model with two cartesian axes, but also with discrete categorised emotions arranged in segments of a circle (also making it a circumplex model). Its performance as an emotional model has been empirically validated [248]. It has received tentative use in human-robot interaction, generally as a self-assessment tool for participants interacting with a robot locally [250, 251], in seldom cases relating to the robot's portrayal of emotions [252].

Despite recent progress to adapt the GEW model for use in HRI applications [252, 253], it has been most commonly adopted (or recommended) as a means of capturing the individual experience of study participants as they engage with a robot [250, 254, 251], and a major limitation of the GEW and other dimensional models is a lack of statistical methods suited to analyse the resulting data [255].

Affect teleoperation

In practice in HRI research, robot emotion has generally been portrayed by changing facial expressions; sometimes other cues or body language are used. Examples of existing emotion control systems in the literature often use categorical models of emotion. The face of the *Roboceptionist* by Gockley et al. [235] as a 3D model displayed on a screen used Ekman's basic emotions, interpolating them to produce different intensities and intermediaries, and using Delsarte's code of facial expressions to translate these visually (Figure 2.8a). A similar approach was done for the actuated robot head *ROMAN* by Berns et al. [244] (Figure 2.8b), capable of portraying six basic emotions, but these emotions being based on the dimensional AVS model (arousal, valence, stance) in the same way as the emotive *Kismet* robot [220]. Out of these three research projects, none concern the control framework or interface from the operator's perspective, the focus being on the remote robot's interaction with other humans. One approach encountered for control of robot emotions is to recognise operator expressions from a camera feed using modern image processing software, with the robot imitating the expression. Liu et al. [256] developed a system for this purpose using a machine learning engine, allowing a robot to imitate one of eight recognised emotions (Figure 2.8c).

The control of robot emotions is not a well-researched field, but further insight into emotion teleoperation may be gained by looking towards the animation industry. Modern animation techniques involve full markerless facial tracking for CGI characters, and amateur and professional online content creators often employ face tracking for virtual online avatars ("VTubers") [257]. A popular, simple method for face tracking is the use of an iPhone with its integrated depth camera and facial mapping technology, however other methods may be more efficient and with higher performance [258, 259].



(a) *Roboceptionist* [235] (b) *ROMAN* [244] (c) Unnamed robot [256] (d) *Stevie* [252]

Figure 2.8: Examples of emotion portrayal (here, the Sad emotion) by robots in literature.

2.3.7 Awareness of interactant social cues

Other potential modalities for human social interaction can exist outside of the above model, involving using the "olfactory, thermal and tactile channels" [260]—depending on the approach, countless smaller channels can be identified. Concerning robots, in *Anthropomorphism and the Social Robot* [223, p. 183], Duffy states:

> "For artificial sociability the social robot does not necessarily need to be able to communicate information in as complex a manner as people but just sufficient enough for communication with people. [...] [T]he social robot needs to be able to communicate enough to produce and perceive expressiveness (as realised in speech, emotional expressions, and other gestures)".

One of the core tasks of designing robots that can socially interact with humans is therefore the choice of modalities to implement to this effect [260]; for social robot *teleoperation*, the consideration is which of these should be prioritised that can be simultaneously teleoperated **without compromising on the operator's perception of social cues of interactants**. Examples in the literature have taken different approaches to this, particularly in the chosen range of controlled modalities and also in the level of automation of their control, however this also highlights the importance of the other side of the exchange: ensuring the operator has sufficient situation awareness to successfully identify social cues.

As detailed in 2.1, situation awareness has not been studied in the context of social interaction so it is currently unknown what measures specifically can be taken to maximise remote social awareness.[11] However, studies on the dual ecology dictate that immersive feedback can mitigate miscommunication. The majority of social robot teleoperation interfaces in the literature are common PC setups, although head-mounted displays have been investigated [261, 262]—while these represent an immersive alternative to a regular PC monitor, they introduce additional costs and may not be accessible to all, as well as notoriously causing "simulator sickness" (similar to motion sickness) [141, 263, 264]. A more accessible method of increasing immersion, focusing on audio as opposed to video, is spatial audio feedback.

Spatial audio

By default, audio feedback from a single microphone consists of a single mono channel. This is opposed to the binaural nature of human hearing, which allows the human brain to instinctively localize where sounds are coming from, primarily through interaural time difference or ITD (the difference in sound arrival times between the ears) [265]. By mathematically modelling the human head in a head-related transfer function (HRTF), and using a pair of headphones, this effect can be emulated and sounds can be made to be perceived from different locations—this is often used in interactive media such as video games [266] to increase operator immersion in a virtual environment. The binaural nature of hearing is also a major contributor to the "cocktail party effect", [267] our ability to distinguish and focus on an auditory source in noisy environments, including speech when multiple people are talking simultaneously, however this can occupy substantial mental resources (the "cocktail party problem") [268].

The potential benefits of spatial audio during conference calls was investigated by Inkpen et al. [269] and Ahrens et al. [270], and this approach to conference calls was extended to that of a shared virtual room in recent years by Wong et al [271]. Concerning robotics, proposals of spatial audio feedback have predominantly been to enhance telepresence. This includes

[11] The academic use of the term "social awarenes" often concerns high-level concepts such as social role and socioeconomic background; here, it is used to refer to awareness of low-level social cues.

the technical work of Keyrouz and Diepold [265] and Saraiji et al. [58]; both concerned the teleoperation of a humanoid robot, using spatial audio in order to enhance the operator's sense of presence. Combining audio directionality with telepresence has also seen applications outside of social robotics; while not involving an embodied teleoperated robot, a telepresence simulation concept by Torrejon et al. was enabled in part by spatial audio [272]. Telepresence was also employed for industrial machinery operation by Ahn et al. [273], using spatial audio feedback on the predicate that it would enhance immersion and situation awareness. A lack of spatial audio [274, 79] or otherwise unnatural audio behaviour [107, 41] has been expressed as an issue during teleoperation for some robot telepresence platforms.

2.4 Summary

A review of the literature was conducted over 269 publications concerning the teleoperation of social robots. Overall, robot teleoperation has seen substantial research, but seldom in the social domain. Non-social robot teleoperation studies predominantly see focus on navigation and manipulation, whereas research on teleoperation of social modalities is slim. The review identified unique design challenges to social robot teleoperation in three key areas: affect expression, speech expression, and operator situation awareness.

Affect expression is core to social interaction—it exists purely as a social cue; it could be considered in some ways what makes interactions social. Knowledge and theory on emotions in robots is a more substantially explored area, but this is not the case for its teleoperation. 2.3.6 showed that emotion teleoperation in robots has seen vanishingly little research focus, and the few examples only allow the display of discrete, categorical emotions. With low sample sizes and methods that do not evaluate social benefit, their contributions to the field of HRI are mainly as demonstrations of a developed system, more so than developing theory. Echoing what was said on speech control, there is the need for the theory of emotions to be incorporated in both the design and evaluation of emotion control interfaces to enable deliberate emotion teleoperation, as currently there is no way to ensure affect is successfully conveyed by a robot; this could be addressed primarily through the development of universally-applicable repeatable statistical methods for evaluation.

As seen in 2.3.3, most speech-capable robots have TTS engines, meaning deliberate prosody (an inherently social facet of speech that can communicate further information) can often be conveyed. However, robot speech teleoperation is an emergent area of research. Theory is being established about how voice qualities can changes an onlooker's perception of a robot, but the cutting edge of robot speech teleoperation remains at a level of proof-of-concept, and prosody control during teleoperation is practically inexistent. Additionally, the validation of research results is frequently seen in the context of specific robot platforms—speech quality is evaluated through overall robot quality, introducing confounding factors, preventing comparisons with other papers and decreasing repeatability. Better design of

speech control interfaces that include prosody control is possible, however because of the strong cultural dependency of prosody it has been found to decrease intelligibility in some circumstances; this potentially confounding effect of prosody during robot teleoperation is yet unknown. The field of linguistics can contribute strong methods for the evaluating of speech teleoperation, potentially overcoming the aforementioned methodological issues.

With the dual ecology model for teleoperation explained in 2.1.2, ensuring a more seamless, immersive experience connecting both operator and robot can improve social performance, as discontinuities between both ecologies can result in miscommunication of social cues. A relatively accessible method for designers to increase immersion and better convey social cues may be through the use of spatial audio feedback, the potential benefits of which are detailed in 2.3.7. However, measuring perception of social cues currently uses qualitative or subjective measures, and has not yet been objectively, quantitatively measured, despite the fact that Situation Awareness is very similar (albeit nonsocial), with the gold standard of objective SA measurement being SAGAT.

Chapter 3

Affect expression in social robot teleoperation

This chapter is based on the following peer-reviewed publication (awarded "Best Paper" in its category):

- A. K. Coyne, A. Murtagh, and C. McGinn, "Using the Geneva Emotion Wheel to Measure Perceived Affect in Human-Robot Interaction," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, (New York, NY, USA), pp. 491-498, ACM, Mar. 2020

as well as the following paper, currently in review:

- A. K. Coyne, A. Murtagh, and C. McGinn, "Circumplex measurement, analysis and visualisation of perceived robot affect," in *the International Journal of Social Robotics*, (manuscript submitted for review September 2023)

3.1 Introduction

For robots to communicate and collaborate with people in a natural way, it is important that they possess the capacity for social interaction. Emotion is a vital ingredient in social interaction, and while it is conveyed in part through visual cues (facial expression, pose, movement, etc.) and tone of voice. It is primarily communicated through the face, and many researchers have advocated the need for robots to be capable of providing emotional feedback

through facial expressions [275, 243, 276, 277, 184] and/or through non-anthropomorphic cues [278, 279]. While there remains significant disagreement among scholars regarding the true nature of emotion, it is widely accepted that certain emotions (and their physical manifestation) are universal among humans. This basic assumption provides a common ground for understanding, and facilitates the reliable inference of emotional states between humans, and possibly between humans and social robots.

Conveying emotion accurately is key to successful human-robot social interaction, but as explored in 2.3.6, there are currently no established methods for ensuring that robot affect is correctly perceived. Affect expression is seldom a conscious decision in humans, and is expressed spontaneously, with little delay - this can be complicated for robot teleoperation, as an operator may need to select an emotion manually based on objective judgement, while trusting that the expression conveys the affect they require. The problem is compounded by the potentially limited range of expressions available to select from on a given robot platform. To evaluate and address these limitations for a specific robot requires a statistical understanding of perceived affect, which can be found by drawing from models of emotion.

Different frameworks exist for modelling emotions, the relationship between emotions, and how emotional states may be expressed externally (positioning of facial features, tone of voice, etc.). These frameworks have been traditionally categorised as being either 'categorical/discrete' or 'dimensional'. 'Categorical' emotional models separate emotions (normally based on Ekman's 'basic emotions') into semantically distinct states, whereas dimensional models map emotional states to a Cartesian space (normally 2D or 3D).

In adapting for HRI research, there are problems with using existing emotional models. While categorical models benefit from their simplicity and ease of drawing statistical measures, they provide a very crude and limited framework for representing the true complexity of emotions. On the other hand, dimensional models better represent the complexity and range of possible emotions, but they generally provide a weak basis for generating statistical measures and positions in dimensional space can have little intuitive meaning, which can make it challenging for both synthetic emotion generation and self-report measurement. As such, there is a need for new frameworks that combine the benefits of categorical and

dimensional systems, namely that they can represent a wide range of emotions, are intuitive to understand, can aid in the design of synthetic emotions and can produce quantifiable statistical measures.

Existing emotion models are tailored to measuring the emotional expression of a person, either through self-report or observation. As the portrayal of emotional states is a core part of social robotics, both teleoperated and automated, there is a need within the field for new design frameworks for better supporting synthetic emotion design and evaluation.

We propose that the Geneva Emotion Wheel (GEW), an appraisal-based measurement tool that has gained popularity in affective computing, but has been seldom used in robotics, offers compelling advantages over standard categorical and dimensional approaches to modelling emotions on robots and/or measuring perceived affect. This chapter details the development of frequentist statistical methods for the GEW, and introduces complementary novel analyses allowing continuous mapping across the Wheel, yielding rich inference and design insights for affect teleoperation. In the proceeding section, we first present a more detailed description of the GEW, and the proposal of frequentist and Bayesian statistical measures for the GEW in HRI, followed by a case study of using the GEW in an application to measure the perceived affect of a social robots. Results of the experiment are then presented, followed by a discussion of key insights, the initial impact of the above research, and future work.

3.2 Development of statistics for the Geneva Emotion Wheel

The GEW is a graphical self-report measuring tool; the model is enclosed by a circular shape, evenly divided into into 20 slices (thus resembling a 'Wheel'), with each slice mapping to a distinct emotion family. Each slice is then subdivided into several segments according to the intensity of the emotion represented; segments close to the centre represent low emotional intensity, segments towards the outside represent high intensity. The Wheel has two dimensions corresponding to valence (horizontal axis) and control (vertical axis). A circular section in the centre of the Wheel is dedicated to neutral (i.e. no emotion) states, and the model also permits participants to register 'other' emotions, if they do not identify the experienced

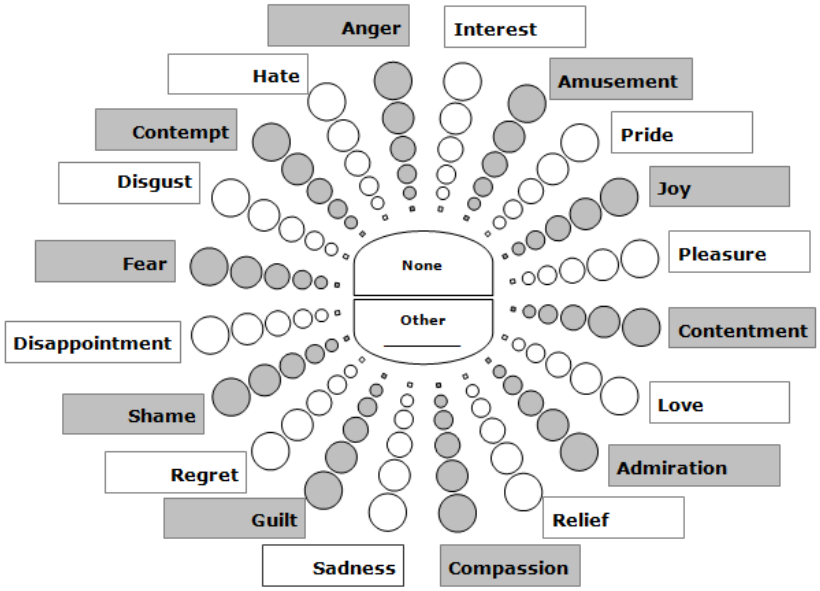


Figure 3.1: The Geneva Emotion Wheel (ver. 3) [248]

emotion with any of the emotion families listed. Since its formulation, the GEW has undergone several revisions [280]; the most recent version of the Wheel (ver. 3), and the one used for this analysis is given in Figure 3.1.

In the proceeding subsections, we make four significant contributions. First (3.2.1), we develop statistics which enable better quantification of findings from the Geneva Emotion Wheel. What makes these metrics most compelling is that they account for the similarity of neighbouring emotions on the Wheel, thus overcoming limitations of confusion matrices and other approaches that have tended to treat each emotion as being categorically distinct. These bottom-up measures are especially relevant in the measurement of perceived robot affect, since researchers often want to know how clear (or unclear) emotional expressions are to observers, comparing the clarity of emotions. Second (3.2.2), we propose a Bayesian statistical model through which these statistics can be analysed. The adoption of a Bayesian paradigm generalises collected data across the Geneva Emotion Wheel in continuous distributions, which allows for top-down analysis such as the identification of gaps in a robot's range of emotional expression. Following this, we introduce the use of new graphing methods that visualise findings from using the Wheel, thus providing an accessible and informative resource

to HRI researchers and designers. Finally, we propose a new design process based on the GEW. This process is cyclic, and emphasises the need for designers to focus on emotions that are clear and unambiguous when mapped to the Wheel, as opposed to creating limited number of emotional states based on vague heuristics or coding systems designed primarily for human use.

3.2.1 Quantifying Findings from the GEW

Prior papers employing the GEW generally analysed emotion data with minor statistical depth; often studying emotion slices discretely [281, 282, 250, 283] or studying emotion intensity in isolation from emotion distribution [281]. For the purposes of our study, we need a statistical measure of the *goodness of fit* of data collected with the GEW; finding variances of perceived emotions on the Wheel will allow us to compare emotion distinctness.

The Geneva Emotion Wheel is a circumplex model of affect, therefore circular statistics[1] can be used to interpret its data. For our analysis, we will apply Mardia's vector method [285] for profile averaging.

For each emotion (e) tested, vector arithmetic is performed to sum all datapoints together. For n_e datapoints, with θ_i the angular location of a datapoint and I_i its intensity, we can first calculate the resulting coordinates (X_e, Y_e) of the averaged emotion profile[2].

$$X_e = \frac{1}{n_e} \sum (I_i \cdot \cos \theta_i) \tag{3.1}$$

$$Y_e = \frac{1}{n_e} \sum (I_i \cdot \sin \theta_i) \tag{3.2}$$

The resulting dimensional coordinates are an expression of the average valence (X_e) and control (Y_e) of the emotion according to the sample data. Expressing this vector in its polar coordinates yields the angular direction of the emotion profile (θ_e) and its intensity (I_e).

[1] A good paper reviewing methods for circumplex analysis is provided in [284].

[2] As advised by Gurtman et al. [284], care is taken to correct for angle values, given the signs of X_e and Y_e and the periodicity of trigonometric functions used.

$$\theta_e = \tan^{-1}\left(\frac{Y_e}{X_e}\right) \tag{3.3}$$

$$I_e = \sqrt{X_e^2 + Y_e^2} \tag{3.4}$$

This vector is considered the centre of gravity of the data [286, 284]; θ_e is the circular mean of the emotion profile, pointing to the predominantly perceived emotion, and I_e the mean intensity, also indicating emotion distinctness.

To find the variance of the distribution, the *unweighted* versions of the above equations apply to calculate circular mean and its corresponding variance; this is done by removing the intensity factor (I_i) and scaling ($1/n$) from equations (3.1) and (3.2) and continuing on to find the unweighted circular mean (θ'_e); the corresponding unweighted circular variance (V'_e) would then be expressed as:

$$V'_e = 1 - \frac{\sum \cos(\theta'_e - \theta_i)}{n_e} \tag{3.5}$$

Considering that emotion intensity is the defining feature of the Geneva Emotion Wheel compared to other circumplex emotion models [249], this formula is adapted to express the variance weighted by intensity, or *weighted circular variance* (V_e):

$$V_e = 1 - \frac{\sum(I_i \cos(\theta_e - \theta_i))}{\sum(I_i)} \tag{3.6}$$

The variance in degrees is found using the arccosine of the second term in equation (3.6):

$$V_e^\circ = \cos^{-1}\left(\frac{\sum(I_i \cos(\theta'_e - \theta_i))}{\sum(I_i)}\right) \tag{3.7}$$

To make the results more intuitive, this can be expressed as *variance of emotions*, by scaling to the total number emotion slices on the Wheel:

$$V_e^e = V_e^\circ\left(\frac{N}{360}\right) \tag{3.8}$$

where N is the total number of slices; 20 for the standard GEW. The intensity of the emotion profile (I_e) and the weighted and unweighted circular variances (V_e) and (V_e') respectively all provide an indication of emotion distinctness; the lower the variances or the higher the intensity, the more distinct an emotion appears. Scaling variance by emotion slices allows an intuitive understanding of its value; a variance of 1 or less indicates that the emotion is distinctly defined on the Wheel, insofar that the closest Wheel slice to the circular mean (θ_e) can be considered an accurate descriptor of the emotion (for example, an emotion with variance of 0.7 with circular mean of 0° can be said to be “Pleasure" – the label of the slice at 0°).

The above statistics can be used in inferential analysis, the simplest approach being that of significance testing. To find if variances are significantly different between two emotions, e_1 and e_2, a standard F-test (assumed $\alpha = 0.05$) with the variances can be performed. The F-statistic is calculated with both weighted and unweighted variances, F and F' respectively. Different degrees of freedom apply to the two types of variances, therefore there will be different critical F-values for each; as seen in equation (3.5) the unweighted circular variance (V_e') has $n_e - 1$ degrees of freedom (from the θ_i variables and 1 mean θ_e), whereas the weighted circular variance (V_e) has $(2n_e - 1)$ degrees of freedom as seen in equation (3.6) (because of the additional I_i variables in the denominator).

3.2.2 Bayesian modelling of the GEW

This subsection details a basic Bayesian approach to analysis using the statistics of 3.2.1, to derive continuous maps of data from the collected discrete datapoints. For each evaluated affect expression, an informative prior distribution is selected, a likelihood distribution is generated by fitting the participant data to two-dimensional normal curves across the GEW, and the posterior distributions of all affect expressions are condensed together into easily interpretable profiles indicative of the robot’s overall emotive range.

Model and parameters A simple statistical model is used, based on the Geneva Emotion Wheel. It is comprised of two parameters, which correspond to the polar coordinates of the

Wheel;

- I, the radial coordinate (which corresponds to *emotion intensity*),
- θ, the angular coordinate or slice of the Wheel (which corresponts to *emotion type*).

I can take values between 0 (no emotion) and 6 (very intense emotion), and θ can take any value of angle between 0° and 360°.

With the discrete, grid nature of the Wheel, grid approximation is considered apt to conduct probability calculations - each discrete section (I, θ) of the Wheel can be represented by a scalar probability $P(I, \theta)$. Given I can represent one of 6 intensities, θ one of 20 emotions on version 3.0 of the Geneva Emotion Wheel, and including the central "None" emotion, this results in a two-dimensional probability density function composed of 121 scalar probabilities over the entire Wheel.

Selecting the prior As with all Bayesian analysis, a prior probability density distribution (a "prior") is selected for each affect expression to be evaluated. This subsection details a proposed method for choosing priors on the GEW.

Normal distributions will be used to represent this prior; for each *non-neutral* affect expression e to be evaluated;

$$p_e(I, \theta) = \mathcal{N}(I|\mu_{I,e}, \sigma_{I,e}) \times \mathcal{N}(\theta|\mu_{\theta,e}, \sigma_{\theta,e}) \tag{3.9}$$

with

- $p_e(I, \theta)$ the prior probability at a point (I, θ) for affect expression e
- $\mathcal{N}$ the normal distribution function
- variants of μ and σ the mean and standard deviation parameters for $\mathcal{N}$ respectively, along both the radial $(\mu_{I,e}, \sigma_{I,e})$ and angular $(\mu_{\theta,e}, \sigma_{\theta,e})$ coordinates.

To determine the values of these parameters, we propose employing a form of *subjective Bayesianism* [287, 288]; as each affect expression evaluated was likely designed with the intent

of portraying a given emotion, it is proposed that the closest matching label on the Wheel be selected by the researcher to inform their selection, to inform $\mu_{\theta,e}$. For instance, if one of the affect expressions is designed to portray Anger, the "Anger" slice on the Wheel would be ideal; whereas if an expression is designed to portray Happiness, the researcher can pick a high-valence emotion from the Wheel that they deem most appropriate. If affect expressions are unlabeled or ambiguous, or if biasing the results is a substantial risk, an uninformative prior such as a *flat prior* may be used.

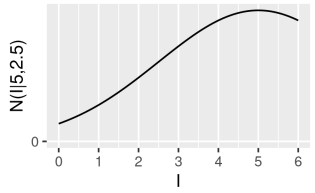
This approach for determining prior probability distributions has caveats. For one, the labels ascribed to the Geneva Emotion Wheel are considered secondary and less reliable than the two dimensions of valence and control for describing emotion [249], and the reliance on such linguistic terms for determining priors is a weak link in the rigour of the method. The use of a flat prior can circumvent this issue, but a well-selected informative prior is considered superior to an uninformative one [289]. Hesitancy to employ informative priors and defend the choice thereof in Bayesian analysis has furthermore been expressed as a shortcoming by some [288]. Therefore, to ensure adequate priors that are robust to any poor association of labels, we propose the use of high-variance, weakly informative priors. The angle of the selected emotion, the mode of the prior, is assigned to $\mu_{\theta,e}$. In order to maintain a broad prior that does not exclude unexpected values, we assign $\sigma_{\theta,e} = 90°$; this ensures ample coverage of emotions within the same quadrant, while giving minimal probability to emotions on the opposite side of the quadrant. This angular factor of the prior is shown in Figure 3.2b. Finally, for higher credibility at stronger intensities of emotion, the parameters of the radial factor are assigned $\mu_{I,e} = 5$ and $\sigma_{I,e} = 2.5$, resulting in the curve shown in Figure 3.2a.

Equation (3.9) can therefore be expressed:

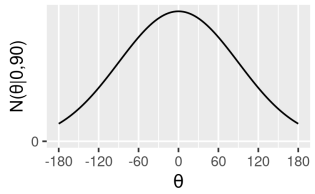
$$p_e(I, \theta) = \mathcal{N}(I|5, 2.5) \times \mathcal{N}(\theta|\mu_{\theta,e}, 90) \tag{3.10}$$

with $\mu_{\theta,e}$ the selected angle in degrees of the closest matching emotion on the Geneva Emotion Wheel. The resulting prior distribution for a non-neutral emotion is shown in Figure 3.2.

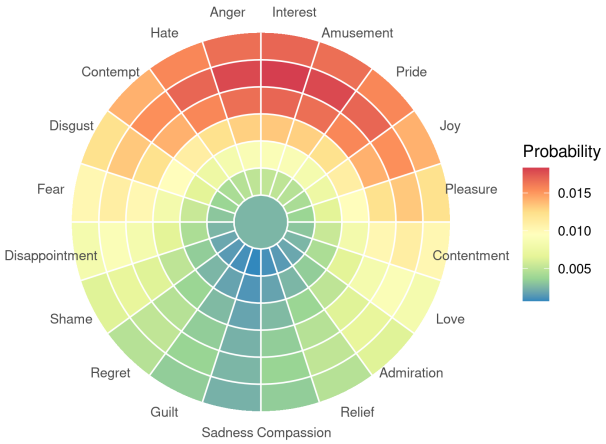
For affect expressions considered *neutral* or *emotionless*, a different prior is generated,



(a) Radial component



(b) Angular component (μ_θ=0)



(c) Prior probability distribution (the product of both components as shown in Equation 3.10)

Figure 3.2: The proposed components and calculated prior distribution for a non-neutral expression on the Geneva Emotion Wheel (Equation 3.10 after normalization). Here, the angular component was ascribed a mean value (μ_θ) of 0, corresponding to the "Interest" label; this is set to the angle of the closest matching emotion label for the expression.

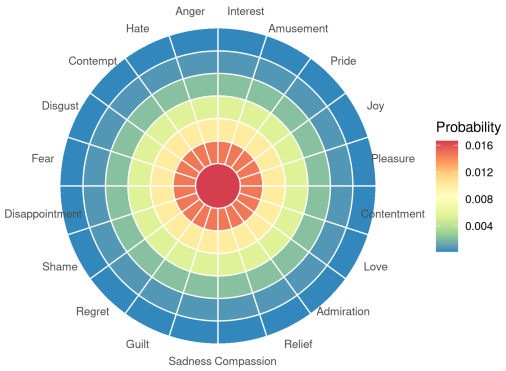


Figure 3.3: The proposed prior distribution for a neutral expression on the Geneva Emotion Wheel (Equation 3.11 after normalization).

whereby the angular factor is set to a nonzero constant, and the radial factor to a normal function centred on the origin, with a standard distribution σ set to 2:

$$p_{neutral}(I, \theta) = C \times \mathcal{N}(\theta|0, 2) \tag{3.11}$$

As the resulting function represents a two-dimensional probability density, the sum of all probabilities over the Wheel should total to 1. Normalization is simple with the grid nature of the Wheel; it can be done by dividing each probability by the sum of all probabilities. As long as the function is normalized afterwards, the value of C is irrelevant and can be set to 1. The resulting prior is shown in Figure 3.3.

The advantage of the above methods of prior selection lie in their accessibility. They allow robot researchers and designers to incorporate information about a robot into analyses even if no empirical studies on its expressive abilities have been performed in the past. Should more robustly informative data be available however, the careful and justified use of novel priors that incorporate it is encouraged.

Likelihood function To incorporate collected data into our statistical model, a probability is required for each of the 121 sections of the GEW, but it is unlikely that the data is continuous across the Wheel (e.g. no participants would report "Sadness" for a clearly positive emotion, leaving that section without data). To generate a continuous probability distribution for each emotion, the data are fitted to normal distribution curves using the sample statistics of 3.2.1. The likelihood function is defined as follows:

$$p_e(D_e|I, \theta) = \mathcal{N}(I|I_e, \sigma_{I_e}) \times \mathcal{N}(\theta|\theta_e, V_e) \tag{3.12}$$

with

- $p_e(D_e|I, \theta)$ the likelihood probability of the data (D_e) at point (I, θ) for affect expression e
- I_e the *intensity of the emotion profile* (equation (3.4))

- σ_{I_e} the standard deviation of intensity across each datapoint i:

$$\sigma_{I_e} = \sqrt{\frac{\sum (I_i - I_e)^2}{N}}$$

- θ_e the *angular direction of the emotion profile* (equation (3.3))

- V_e° the *circular variance in degrees* (equation (3.7))

Calculating the posterior Bayes' theorem states that:

$$p_e(I, \theta | D_e) = \frac{p_e(D_e | I, \theta) p_e(I, \theta)}{p_e(D_e)} \tag{3.13}$$

with

- $p_e(I, \theta | D_e)$ the "posterior" probability distribution, the output data expressing the probability that expression e is perceived as the emotion at point (I, θ)

- $p_e(D_e)$ the "evidence", the probability of the data according to the model.

The evidence term can also be expressed in terms of model parameters:

$$p_e(I, \theta | D_e) = \frac{p(D_e | I, \theta) p(I, \theta)}{\iint p(D_e | I, \theta) p(I, \theta) dI d\theta} \tag{3.14}$$

The discrete but contiguous grid nature of the Geneva Emotion Wheel enables the use of grid approximation for calculating the posterior, which allows the equation to be expressed using sums instead of integrals:

$$p_e(I, \theta | D_e) = \frac{p_e(D_e | I, \theta) p_e(I, \theta)}{\sum_I \sum_\theta p_e(D_e | I, \theta) p_e(I, \theta)} \tag{3.15}$$

The denominator is now simple to calculate, as a sum of the values of the numerator across all sections of the Wheel, normalizing the posterior.

3.2.3 Visualising Findings from the GEW

The most common way to present findings from emotion classification tests conducted using categorical models has been through confusion matrices. These can be limiting in two ways; first, they can become unwieldy as the number of emotional states increases, and second, they assume that each expression evaluates to a unique absolute state that can be defined from the outset. Visualising findings from dimensional models can also be problematic and unintuitive. Typically, the dimensions are reported using bar plots and line graphs with points/bars corresponding to the dimensions in the model (i.e. arousal, valence).

Discrete, "raw" data To visualise discretely reported emotions using the GEW, we propose using circular histograms (or rose plots) and heatmaps. Both approaches exploit the ability to discretize the GEW. Using circular histograms, we can view which emotion families, and which quadrants of the Wheel, were most frequently connected with a corresponding expression. From this graph, we can easily identify emotions that are distinct (i.e. Fig. 3.4a) from ones that are less distinct (i.e. Fig. 3.4c). Using heatmaps, we can observe the full impact on the Wheel of each perceived emotion. Here, a distinct emotion would be one where few segments, closely spaced together, are selected most frequently (Fig. 3.4b), whereas an indistinct emotion would be one where there are multiple segments, widely distributed across the Wheel, are commonly selected (Fig. 3.4d).

Continuous mapping To assess the overall emotion expression capabilities of a robot outside of singular emotions, we propose continuous visualisations across the Wheel using the Bayesian methods of 3.2.2.

These methods yield posterior probability distributions for each tested affect expression. By taking the maximum probability value on each section of the Geneva Emotion Wheel for all expressions, a *certainty profile* can be obtained for the robot:

$$\mathcal{C}(I,\theta) = \max(p_1(I,\theta|D_1), p_2(I,\theta|D_2), ..., p_n(I,\theta|D_n)) \tag{3.16}$$

with

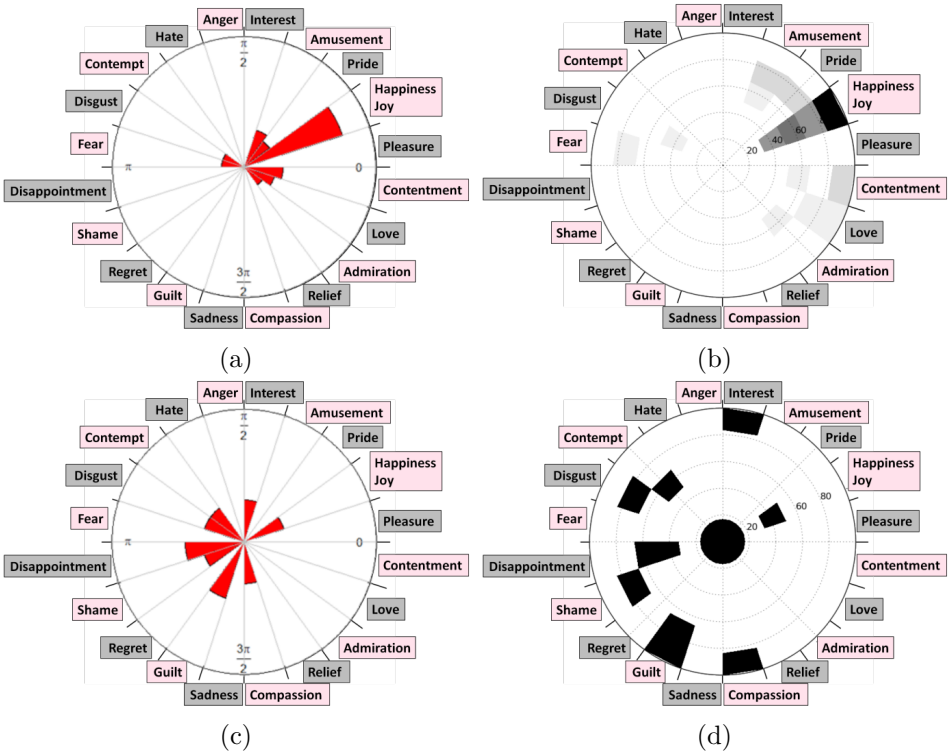


Figure 3.4: Raw data visualisations for two perceived emotions; rose plots (left) and heat maps (right). Emotion 1 (a, b) is observed to quite distinct. Emotion 2 (c, d) is visibly less distinct.

- $\mathcal{C}(I, \theta)$ the certainty profile function for the robot, at point (I, θ) of the GEW,
- $p_e(I, \theta|D_e)$ the posteriors calculated for each affect expression e as in eq. (3.15), for n total expressions tested.

The certainty profile expresses on the GEW how successfully the robot can portray each emotion; visualization of this profile can be used to inform robot design improvements. Peaks in the profile can show emotions with strong certainty of being conveyed, emotions the robot can reliably be made to express. Valleys can identify parts of the robot's emotional palette that it struggles to successfully convey, that may benefit from further design attention (e.g. the implementation of additional affect expressions).

Another profile that can yield insights into the data is the categorical *closest-expression map*, obtained by selecting the expression with the highest certainty for each point of the GEW:

$$\begin{aligned} E_{\max}(I, \theta) = e \in \{e_1, e_2, ..., e_n\} \\ \text{where } p_e(D_e|I, \theta) = \mathcal{C}(I, \theta) \end{aligned} \tag{3.17}$$

with $E_{\max}(I, \theta)$ the affect expression with the maximum probability at point (I, θ). This splits the Geneva Emotion Wheel into at most n regions, one for each evaluated expression, showing which expression can most reliably convey each emotion on the Wheel. This profile can be used as a tool for teleoperation or automation design, as it can inform, in tandem with the above certainty profile, which expression to select based on the required emotion.

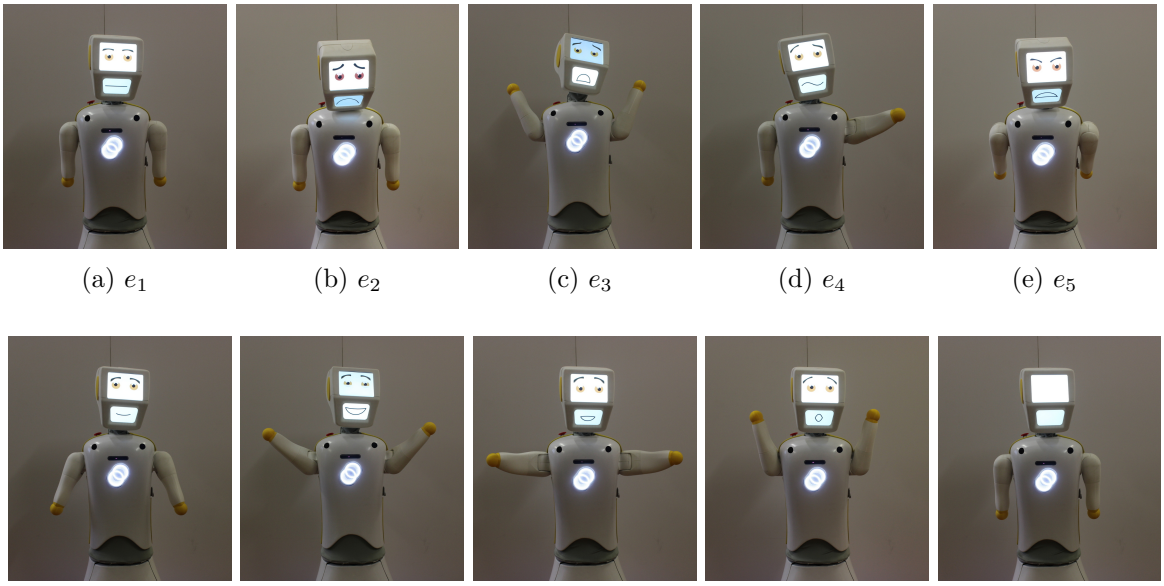
3.2.4 Adapting the GEW as a HRI design tool

Systematic approaches exist for developing and validating synthetic emotional expressions on robots. Normally they begin with the selection of 6-8 discrete emotion labels. Then, a process is undertaken whereby the emotion is synthesised on the robot, frequently through the use of a formal coding scheme, such as FACS, or using heuristics from animation. Next, an evaluation process takes place that usually involving an experiment where an observer is asked to estimate the perceived emotion on the robot. The success criteria is if the perceived

emotion matches with the pre-defined emotion label.

As mentioned in 3.2.2, an underlying flaw of this approach is that it assumes all emotions can be pre-labelled in advance, and the ultimate design objective is to create expressions that evaluate to a defined absolute state when perceived by a human observer. We argue that designing affective expressions to meet the dictionary definitions of emotion labels limits the size of the emotional space that can be represented by the robot, and according to Scherer [249], “is neither sufficiently comprehensive nor consensual enough to be appropriate for scientific profiling of emotion terms". Instead, we present a design approach that uses the dimensional labels of the GEW to overcome this limitation. In this new paradigm, designers focus on developing expressions that map to positions in two dimensional space on the GEW, and which then resolves to a point that can be intuitively understood from observation. It is important to note that while the GEW possesses emotional labels, these are intended primarily for guidance, and to inform participants of its underlying structure.

Designers following this new approach begin by creating expressions corresponding to the dimensions of valence (how positive/negative the expression is?) and control/power (do I have sufficient power to exert control if possible?). This process can be informed by heuristics, emotion coding manuals, prior research or intuition as appropriate. The success of each expression is measured empirically, through use of the GEW for self-report and post-test application of the statistical analyses outlined in 3.2.1 and 3.2.2, and the visualisations in 3.2.3. The performance of each expression is measured by how distinct it is, namely how well it clusters to a local region on the Wheel. The observed variance provides important insight on the clarity of the perceived emotion; high variance indicates that not all observers agreed on the perceived valence and/or control of the affective state, indicating that it could be ambiguous if used on the robot, while low variance indicates that the expression was perceived similarly by most observers. These insights allow designers to evaluate individual affect expression performance, whereas continuous Bayesian visualisations allow them to identify how well the robot performs overall in each region of the Geneva Emotion Wheel, notably what available expression maps best to a point on the GEW, and what parts of the GEW show low performance and might benefit from further expression generation.



(a) e_1 (b) e_2 (c) e_3 (d) e_4 (e) e_5

(f) e_6 (g) e_7 (h) e_8 (i) e_9 (j) Blank

Figure 3.5: Emotion expressions e_1-e_9 for the robot. (j) Transient "blank" state shown between emotions.

3.3 Methodology

A trial study was performed where the GEW was used to evaluate the affective expression states of the Stevie (version II) socially assistive robot. All experiments took place over two days and participants for the study were recruited from an innovation and collaboration space located within the University. Informed consent was provided by each participant and and the study received approval from the appropriate ethical review committee.

Participants

In total, 34 participants took part in the experiment. The average age was 22.3 $\pm$ 4.2, with the youngest being 17 and oldest being 33. The gender breakdown was 20:13 male-to-female with 3% preferring not to share their gender information. The experience of interacting with a robot was new to most participants, with 88.2% admitting that they had never interacted with a social robot before. The nationality of participants was diverse, with people coming from USA, Germany, Ireland, India, Russia, Trinidad and Tobago, China and Italy. All

participants spoke English at a fluent or native level. One potential participant was required to leave the experiment prior to completion, so their data was discarded.

Procedure

To ensure similar conditions to real life interactions without additional confounding factors, a real hardware robot was chosen over video clips to portray the emotions to participants. As the Stevie II robot platform was readily available to the researchers, it was chosen for this case study. The affective expressions portrayed by the robot were adapted from those of its predecesor, Stevie I [14].

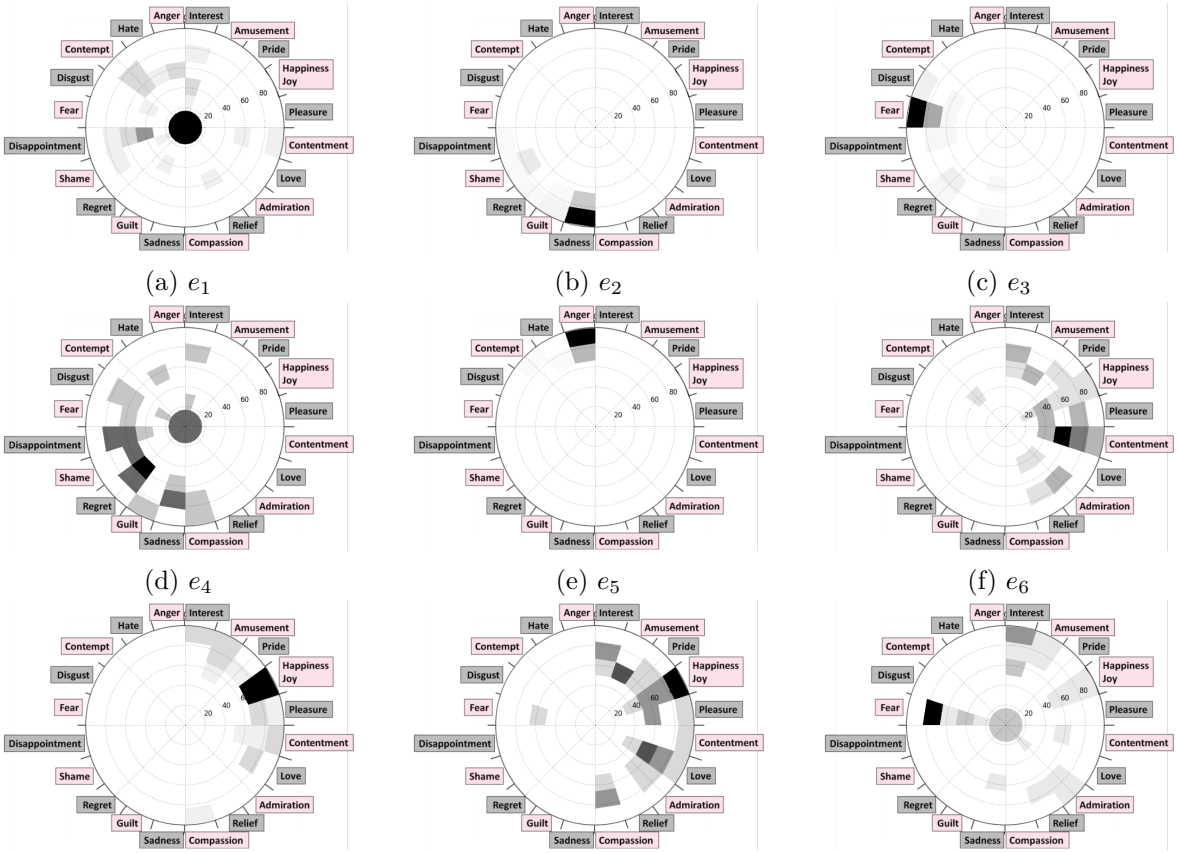
After providing informed consent, a brief training period was given on how to rate emotions using the most recent version of the GEW (version 3). To further familiarise participants with the process of using the Wheel, participants were required to rate seven videos where a human actor expressed the emotions corresponding to surprise, happiness, anger, disgust, contempt, sadness and fear [3].

When this training exercise was completed, the robot was brought out and revealed to the participants. Participants initially rated 9 emotions; these are indicated in Figure 3.5. The sequence that the expressions were shown to participants was randomised in all cases, and between expressions the robot was returned to a blank state (Fig. 3.5j). After the tests were completed, participants were asked to provide some basic demographic information.

3.4 Results

Findings from the emotion state recognition tests are presented graphically in Figure 3.6 and the associated statistics expressions provided in Table 3.1. Posterior probability distributions for each expression were condensed into a certainty profile shown in Figure 3.7a, and a closest-expression map shown in Figure 3.7b. These individual posterior distributions (along with chosen priors and heatmaps) are provided in Appendix A.

[3]The videos used were taken from a widely viewed YouTube video on basic emotions, which can be found at `https://www.youtube.com/watch?v=embYkODkzcs`



(a) e_1 (b) e_2 (c) e_3

(d) e_4 (e) e_5 (f) e_6

(g) e_7 (h) e_8 (i) e_9

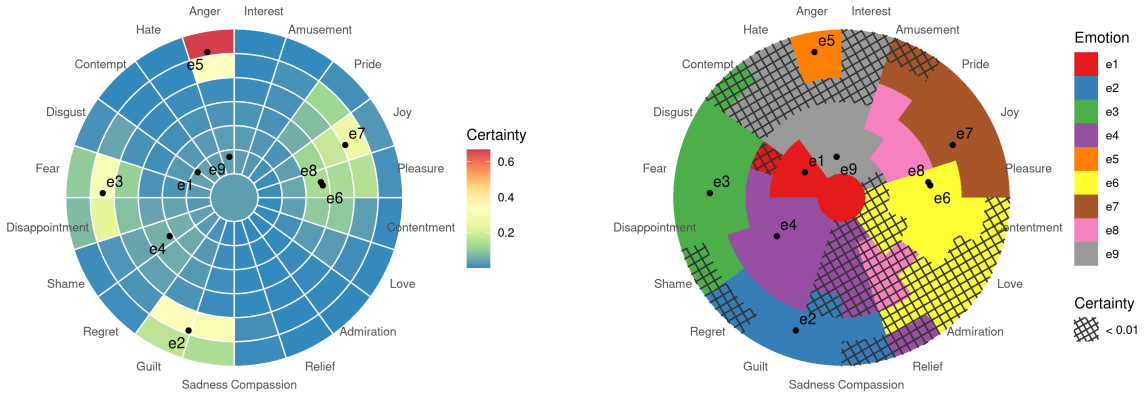
Figure 3.6: GEW heatmaps for each tested emotion.

	Original label	**Assigned prior**	$\mu_{\theta,e}$	I_e	σ_{I_e}	θ_e	V_e°	V_e^e	**Closest perceived**
e_1	Neutral	*neutral prior*		1.35	1.78	295.76	62.56°	3.48	Disgust
e_2	Sad	Sadness	180	5.35	0.46	189.96	20.29°	1.13	Guilt
e_3	Scared	Fear	270	4.98	0.71	262.96	24.35°	1.35	Fear
e_4	Confused	*flat prior*		2.63	1.49	230.20	48.81°	2.71	Amusement
e_5	Angry	Anger	342	5.67	0.46	340.49	6.70°	0.37	Anger
e_6	Happy	Contentment	90	3.23	1.04	73.17	39.89°	2.22	Pleasure
e_7	Joy	Joy	54	4.62	0.61	55.66	31.35°	1.74	Joy
e_8	Interested	Interest	0	3.17	0.88	70.88	47.00°	2.61	Pleasure
e_9	Surprised	*flat prior*		1.21	1.54	344.14	73.82°	4.10	Anger

Table 3.1: Prior selection and likelihood statistics for each evaluated affect expression. "Original label" denotes the name given by the robot designer to the expression, "Assigned prior" shows the closest matching label on the GEW, or the nature of the prior used if ambiguous. A **flat prior** is a distribution with equal probabilities across the entire GEW, and a **neutral prior** is shown in Figure 3.3.

Emotion	**Other name(s), times specified**
e_1	"Boredom", 1
e_4	"Confusion", 6 ; "Unsure", 1
e_9	"Surprise", 4 ; "Shock", 1

Table 3.2: Other emotion names specified



(a) Certainty profile created from participant data

(b) Closest-expression map created from participant data, with lowest certainty regions highlighted

Figure 3.7: Continuous map visualisations created from participant data. Black points indicate centres of gravity of likelihood data for each evaluated expression (I_e, θ_e).

3.5 Discussion

Individual expressions

The results show a wide range of variances for each emotion, from 0.37 to 4.35 in the case of the weighted variance (V_e^e). Variance appears to be inversely correlated to intensity for this dataset - high variance and low intensity show an indistinct emotion (e.g. e_1 and e_9). Conversely, low variance indicates a distinct emotion (e.g. e_2 and e_5). In this case, the intensity is representative of the average intensity chosen for the emotion. This is mirrored in the heatmaps in Figure 3.6; emotions e_2, e_3 and e_5 are visually very focused in certain regions, whereas most of the other emotions' heatmaps show chosen emotions more dispersed over the GEW.

The highest variances are reported for e_9. The closest label on the GEW, indicated by the circular mean θ_e, is "Anger", which the researchers consider an unintuitive label for the corresponding expression shown in Figure 3.5i. The heatmap for e_9 (Fig. 3.6i) shows that participants did not once assign a segment from the "Anger" slice of the GEW to this emotion. This observation indicates that when variance levels are high, the circular mean (θ_e) is a poor indicator of trend and may not correctly identify the dominant label selected on the Wheel.

The variances of e_5 was the lowest of the emotions tested. This indicates that, according to the population sample, it was the most unambiguous emotion out of those shown; the emotion heatmap 3.6e, showing a single focused region, confirms that the emotion is highly defined. V_e^e is the weighted variance scaled to the 20 emotion slices; as the variance is less than 1 emotion, θ_e fits the data strongly, and the closest label on the GEW (here, "Anger") can be considered an accurate, unambiguous descriptor of emotion e_5.

F-testing can be used to compare variances for significant differences, as explained in subsection 3.2.1; for example, using the weighted variances obtained in the case study, we can say that the variances of e_4 and e_8 are not significantly different [$F(54, 68) = 1.20$, $p > 0.05$] whereas e_4 has a significantly higher variance than e_7 [$F(54, 68) = 1.86$, $p < 0.05$], i.e. results show that e_7 is a significantly less ambiguous emotion than e_4.

It is important to note that these statistics consider that *angle* defines emotion; this is

normally accurate except when significant numbers of participants label the state as being “Other" (shown in table 3.2) or "None", the latter having zero associated angle and intensity. Datapoints indicating “None" therefore have no impact on the calculation of the weighted circular variance (V_e), and only decrease the final vector intensity (I_e). This case was observed in the data for state e_1; like e_9, the variance is high and the intensity low compared to other emotions, and while it would be intuitive to dismiss it as an indistinct emotion, its heatmap (Fig. 3.6a) shows the strong prevalence of “None" in the data unlike e_9 (Fig. 3.6i). This shows how important graphical tools are for visually understanding the data and overcoming the limitations of the statistics.

Continuous models

The Bayesian visualisations of Figure 3.7 reveal at a glance to what degree the regions of a robot's emotional palette are successfully conveyed. It has also provided a mapping of each emotion on the GEW to one of the robot's affect expressions, which can be used for robot teleoperation or automation, in tandem with an internal emotion engine. In short, these visualisations show how best to utilise the current range of affect expression of a robot, and what emotions require further design attention.

It can be observed in Figure 3.7a that for this dataset, emotions such as moderate Joy or strong Sadness are of high certainty with strong Anger the highest, while emotions such as strong Interest, Hate or Relief, or mild Pleasure, are of very low certainty. This indicates that in isolation from context, the affect expressions of the robot would struggle to express the latter emotions. This is not to say that the robot is incapable of expressing these emotions when accompanied with contextual cues—for example, while the emotion of intense Hate has low certainty, it is likely that this can be conveyed through expression e_5 (Fig. 3.5e) in the right context. However this choice of which expression to select is subjective, there is no systematic method that can be relied on in low certainty regions for automation or teleoperation purposes—the data indicates that the highest certainty for intense Hate is for expression e_9 (Fig. 3.5i), which the researchers judge to be an inappropriate expression for the sentiment. Low certainty regions can therefore be considered areas requiring further

design attention to complete the robot's emotive palette, either by extending regions of high certainty (for example considering e_8 to be capable of expressing intense Love) or by implementing new affect expressions in the centre of low certainty regions (generating an expression of Admiration that might also be suitable for Love and Relief).

Important insights can also be gained by analysing where expressions tend to fall on the Wheel. One design insight of interest is shown for expression e_1; this had been conceived as a "Neutral" expression by the designer, but the data in Figure 3.7b shows a drift away from the centre of the GEW, towards negative valence and emotions of Disgust, Fear, Contempt. These can be potentially very undesirable emotions to mistakenly convey, which indicates that expression e_1 may also require design attention—for example, generating an alternative neutral expression that tends more towards positive valence. This expression may still be of value for what it is perceived as however; as informed by the data, it could be used to express these emotions of mild discontent.

Limitations and future work

While accessible to calculate, the technique used to fit collected data to the continuous model (the product of normal distributions across both polar coordinates) considers that expression data is unimodal. This can appropriately represent data that would otherwise have a large number of peaks (expressions that are perceived as a large number of different emotions are unreliable), however data with a small number of peaks (e.g. bimodal) may be of interest to robot designers. Regardless of utility, future work might improve fitting accuracy by employing techniques such as cluster analysis, preserving more information from the collected data.

Grid approximation for Bayesian calculation is not only simple to calculate, but is also well-suited for the discrete grid of the Geneva Emotion Wheel. However there are advantages to using *hierarchical* Bayesian models [289], in which the parameters of the curves themselves (I_e, σ_{I_e} etc) are given a probability distribution over a range of potential values, which requires more complex calculation methods such as Monte Carlo Markov Chain. This also requires more strict specifications of prior distributions, but the resulting probability distributions for

higher-level parameters would enable richer statistical comparisons, including an analogue to the F-test demonstrated in 3.5 by calculating the Bayes factor over V_e° or σ_{I_e}.

Future work may devise a more complex visualisation of the closest-expression map (Figure 3.7b) incorporating Bayes factors across every expression pair, enhance the richness of the analysis and providing numerical indicators of evidence (different but loosely analogous to the statistical significance of p-values).

Finally, a potential limitation in any study involving Bayesian inference is that of prior selection, as mentioned in 3.2.2. A methodology was proposed in this paper for selecting a prior based on design choices that went into the robot; future studies might propose alternative methods for selecting a prior, ideally using data from previous Bayesian studies.

Chapter 4

Speech expression in social robot teleoperation

This chapter is based on the following peer-reviewed publication:

- A. K. Coyne and C. McGinn, "The Effect of Human Prosody on Comprehension of TTS Robot Speech," in *Proceedings of the 32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN*, (Busan, South Korea), IEEE, 2023

4.1 Introduction

Language can be considered be the primary channel of communication between humans, and as such can be of critical value for social robot teleoperation. The open-ended nature of language makes it a universal, versatile tool for issuing commands and receiving feedback from robots, but it can also pave the way for less utilitarian, more social paradigms of interaction.

As well as explict semantic content that enables people to exchange information rapidly and effectively, speech also contains implicit information that informs the listener's *partner model* of a speaker, providing insight into their possible background (i.e. age, nationality, gender), their affective state, and identity [290, 161, 163]. Much of this implicit information is captured in the prosody of the voice. Prosody is understood to be a vocal nonverbal signal,

which *"in perceptual terms, accounts for* how *something is said"*[4] [145]. It refers to the rhythm, stress, intonation, etc. of speech, and *"plays a role in the comprehension of spoken language by human listeners"* [291]. It is language- and culture-dependent [121, 147, 154], and is a contributor to accent, both foreign [155, 156] and native [157, 158, 159].

While some HRI studies have utilized human voices, namely through tele-operation or using pre-recorded soundbites, most tend to use text-to-speech (TTS) engines that generate synthetic speech in accordance with the semantic text content provided and the markup language supported by the software. These TTS programs provide robots with different voices, accents and prosody, offering the potential for robots to appear across a spectrum of gender, age, and expressiveness.

Teleoperation of robot speech is still an emerging area of study (explored in 2.3.3), and it is very rare that teleoperated speech systems leverage the features of their TTS engines to incorporate operator prosody. This is despite the fact that an imitation-based speech teleoperation pipeline, in which a teleoperated robot replicates the speech spoken by an operator in its own voice, provides ample prosodic information that can be extracted from voice audio to inform the accompanying robot prosody (as shown in [177]).

Whether or not the implementation of such a system should be a design priority however is unsure, as while it may seem intuitively advantageous for a robots' voice to exhibit high levels of prosody, there are several factors that could negatively affect the experience. On one hand, if the TTS system approaches human speech in terms of naturalness, it is possible an uncanny effect is observed, which may in turn cause a mismatch between how the robot sounds, and how it appears and behaves. Secondly, while changing prosody may enhance the robots ability to convey affective expression, artifacts introduced through modification of the intonation and pitch of the voice may affect intelligibility of the speech signal. It has been shown that prosody can increase intelligibility, but only if the listener is culturally familiar with the prosody source, [154] potentially making the robot's speech harder to comprehend if not.

Recent HRI research has found that changes to the prosody of a robot's voice can enhance

[4] quoted from [146, p. 1063]

its ability to express affective states, however the effect of prosody on intelligibility, while understood for human speech, remains almost entirely unexplored for robot speech, leaving robot designers with little insight into how best to implement such a feature.

In this chapter, to inform the design of speech teleoperation interfaces, we aim to explore how people comprehend robot speech that is tuned to imitate human prosody. Using a mainstream TTS engine and experimental methods drawn from the field of linguistics research, the intelligibility of unaltered, 'default' sentences was compared to 'prosodic' sentences manually tuned to match the prosody of human actors. We examined factors that we postulated might affect intelligibility of prosodic speech; the length of the sentence uttered (long/short), and whether or not the listener was a native English speaker that learned English in the same region as the human actor used to model the prosody of the synthetic speech. We anticipate that the findings of this research will be of broad interest to the HRI community, especially to designers and researchers engaged in experiments involving speech interaction.

4.2 Methodology

A within-subjects speech-in-noise experiment was conducted to investigate the effects of human-like prosody on robot speech comprehension. Participants viewed short videos involving synthesized audio sentences being expressed by a robot, and were then asked to transcribe them as accurately as possible. Sentences were generated either with tuned expressive prosody, or using the default prosody of the speech synthesis engine as a control (referred to as "human' and 'default' respectively). The accuracy of transcribed sentences to the originals was the studied dependent variable.

4.2.1 Stimuli creation and administration

Stimuli for this experiment took the form of videos. In each video, a social robot expressed a sentence in one of the four synthetic voice conditions (either male or female, and either 'default' or 'human'). To increase the difficulty of perception of the sentence, the audio in each video was overlaid with speech-shaped noise. This approach, known as speech-in-

noise testing is widely used in audiology testing as playing clear audio of each sentence to participants may lead to near-100% comprehension rates, obscuring any differences in test conditions–adding noise avoids “ceiling performance" [204, p. 2]. For this task, speech-shaped noise provided by Harvard Speech Corpus [292] was added to the videos at a signal-to-noise ratio of 4dB. To avoid a harsh start and stop, the noise was set to fade in before the sentence was heard and fade out after.

Synthetic speech generation

Synthetic speech was generated using the Google Speech Synthesis API (June 17 2021 release). This was selected because it was among the most natural-sounding commercial TTS system available, it appeared to be widely used in the HRI community [293, 294, 295], and it supported speech synthesis markup language (SSML), the markup language that allows relatively fine grain control of speech prosody.

First, synthetic speech was created by passing the sentences in the sentence corpus to the TTS system without modification. This produced utterances with relatively generic prosody, that generally lacked contextual expressiveness.

To generate speech with human-like prosody, recordings were first made of 5 voice actors who were instructed to speak each of the sentences in the sentence corpus in an expressive way. All actors spoke with a dialect that was native to the region where the experiment was performed. An informal survey was then conducted in which 5 college students were asked to rank each voice recording by “most-to-least natural"; the top-ranking recording of each sentences was retained and subsequently used as the model to construct the synthetic humanly-prosodic voice. Using Audacity, the speech signal of each of these model sentences was analysed for pitch, timing (i.e. duration of pauses between words) and rate (i.e. speed that words were pronounced). This information was encoded in SSML and was subsequently used by the TTS engine to generate the synthetic prosodic speech. Pitch was controlled for each individual word using the `pitch` attribute of the `prosody` element, timing was controlled using the `break` element, and the `rate` attribute of the `prosody` element was used to slow down or speed up words. This is illustrated in the following example SSML snippet:

```
<speak>
    <s xml:lang="en-UK">
        they <prosody pitch="-2st">waited</prosody> for
        <prosody pitch="+2st" rate="slow">one</prosody>
        <prosody pitch="+1st">hour</prosody>
        <break time="0.1s"/>
        <prosody pitch="+1st">while the</prosody>
        <prosody pitch="-3st">train</prosody>
        <prosody pitch="+3st">stopped</prosody>
        <prosody pitch="+1st">at the</prosody>
        <prosody pitch="-1st">station.</prosody>
    </s>
</speak>
```

Recordings using both 'default' and 'human' configurations were made in two voice conditions: one male ("en-US-Wavenet-D") and one female ("en-US-Wavenet-G").

Visual presentation

Videos were created showing a social robot speaking each sentence. The Stevie robot [296] was selected for the experiment due to its availability to the researchers as well as its ability to move its mouth while speaking, which was deemed important for the purpose of a video. For this experiment, videos of Stevie talking were recorded on a chroma key backdrop which was replaced with a white background, and overlayed with the generated audio sentences to produce videos for each sentence and test condition.

Sentence corpus and measures

Sentences used in the experiment were taken from the Bamford-Kowal-Bench (BKB) sentence list [205]. Sixteen BKB sentences in total were used. Sentences were only chosen if they were not deemed to be age or culturally dependent. Sentences 1-8 were extracted verbatim from the BKB sentence list, while sentences 9-16 were constructed by combining two standard BKB sentences using connective words. This allowed us to explore how prosody affected the comprehension of relatively larger and shorter sentences. The 16 sentences used in the study are given in Table 4.1.

The BKB lexicon was designed to measure speech comprehension. Each sentence contains 3 or 4 "keywords", and each correctly-transcribed keyword counts as a single point, providing a score out of 3 or 4 for each sentence. For the dual-clause sentences, being composed of two

basic sentences, the score was out of 6, 7 or 8. For each sentence, scores were normalized and represented as a *comprehension rate* between 0 and 1, representing the proportion of correctly transcribed keywords.

4.2.2 Experimental design

Participants

Participant recruitment was conducted in two locations; a university campus and a gift shop in a department store. For each participant, country of origin, country where English was learned, and English language fluency were recorded, as well as whether the participant had any conditions that might impair hearing or speech comprehension.

Procedure

Participants were first asked to read an information sheet and provide written informed consent in accordance with ethics requirements. Afterwards, participants were seated at a PC and asked to put on high quality over-ear headphones.[5] The experiment was implemented in Python using PsychoPy libraries [297]. The experiment was programmed to play sixteen videos of the robot, one for each expressed sentence. Following each video, the participant was required to type the perceived sentence, transcribing as completely and accurately as possible. Once completed, the participant clicked a button to advance to the next video. Screenshots showing different stages of the testing procedure are given in Figure 4.1, and a video demonstrating the simulator and stimuli is available as supplemental material.[6]

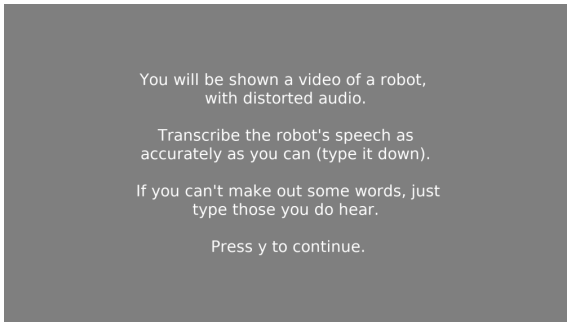
Videos were presented in random order to each participant, with no repeated sentences and an even distribution of factor permutations (prosody type,). Subsequent participant stimuli were counterbalanced, ensuring each permutation was evenly represented across the sample.

[5] Precautions were taken in regards to the ongoing COVID-19 pandemic; all equipment was thoroughly wiped using antibacterial wipes before participant contact, and personal protective equipment (PPE) was provided.

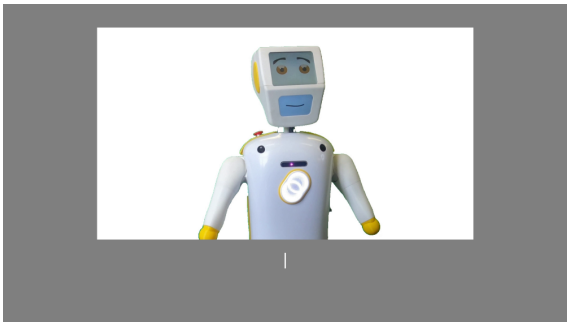
[6] Accessible at the following link: `https://adm.pizza/thesis/supplemental/`

Table 4.1: List of the shorter (1-8) and longer (9-16) sentences used in the experiment. In accordance with the BKB lexicon, one point was awarded for each word in capital letters that was correctly identified.

	Single-clause (extracted from BKB sentence corpus)
1	"YELLOW PEARS were LOVELY"
2	"Some MEN SHAVE in the MORNING"
3	"The BIG FISH GOT AWAY"
4	"SHE'S CALLING her DAUGHTER"
5	"THEY KNOCKED on the WINDOW"
6	"THEY WALKED ACROSS the GRASS"
7	"HE DROPPED his MONEY"
8	"The DOGS GO for a WALK"
	Dual-clause (composite of 2 sentences from BKB sentence corpus)
9	"THEY WAITED for ONE HOUR while the TRAIN STOPPED at the STATION"
10	"The FIRE is VERY HOT but the ROOM'S GETTING COLD"
11	"The LADY WASHED the SHIRT since the WASHING MACHINE BROKE"
12	"The OLD MAN WORRIES that a TREE FELL ON the HOUSE"
13	"The FOOTBALL GAME' S OVER and PEOPLE are GOING HOME"
14	"The TWO FARMERS are TALKING as THEY SIT on a WOODEN BENCH"
15	"The DAUGHTER LAID the TABLE, and the CHILDREN are ALL EATING"
16	"While the TRUCK CLIMBS the HILL, the DRIVER WAITS by the CORNER"



(a) Test guidelines shared with participants at the beginning of the experiment



(b) A video of the robot taking was shown to the participant with randomly selected and counterbalanced speech conditions, overlaid with noise



(c) Participants could begin to transcribe what they heard before the end of the video



(d) When the video ended, participants could finish transcribing what they heard and move on to the next video

Figure 4.1: Screenshots of the testing software.

Videos were presented in random order, with randomly selected prosody type for each ('default' vs 'human'), selected in advance to ensure even distribution of test factors (*prosody type* and *sentence length*). Test factors were also counterbalanced for subsequent participants—as each participant would only hear each sentence once , this ensured each sentence was expressed in all conditions the same number of times over the course of the experiment.

Upon completion of the experiment, the participant was asked to complete a demographics form. The participant was supervised by a researcher during the experiment and was thanked for their participation upon completion. When the experiment was finished, data was logged to CSV files on the host PC.

Analysis

In order to analyse differences in means for comprehension rates for both prosody test conditions over the two studied factors of *sentence length* and *familiarity with local English*, a mixed-measures 3-way analysis of variance (ANOVA) was conducted. *Sentence length* and *prosody type* were within-subjects factors, and *familiarity with local English* was a between-subjects factor. This form of analysis was selected so that both first-order and interaction effects between factors could be studied, and was calculated in R [298].

4.3 Results

In total, 53 participants volunteered for the experiment, of which one that did not complete the experiment, resulting in 52 total participant data points. 11 participants were recruited in the gift shop (mean age of 26, standard deviation of 7.1, 4 female, 7 male), and 41 from a college campus (mean age of 22, standard deviation of 3.5, 10 female, 14 male, 3 other). 31 participants reported being native English speakers, and 26 participants were fluent English speakers that learned English in the local region.

Analysis of first-order effects, with both sentence lengths and both voice types aggregated for each participant, shows a significant difference in the comprehension of 'default' speech ($\mu = 0.742$) and 'human' speech ($\mu = 0.679$); $F(50) = 10.329$, $p = 0.002$. Two second-order interactions with prosody levels were shown; with the factor of sentence length; $F(50) = 5.2935$, $p = 0.026$, and listener locality; $F(50) = 5.1114$, $p = 0.028$. No significant third order interaction was identified between all three factors of prosody level, listener locality, and sentence length; $F(50) = 0.51082$, $p = 0.48$.

Visualisations of results are shown in Figures 4.2 and 4.3.

4.4 Discussion

In this study, we did not find that robot speech exhibiting prosody was easier to comprehend. Rather, the results indicated that the presence of human-like prosody made the speech less

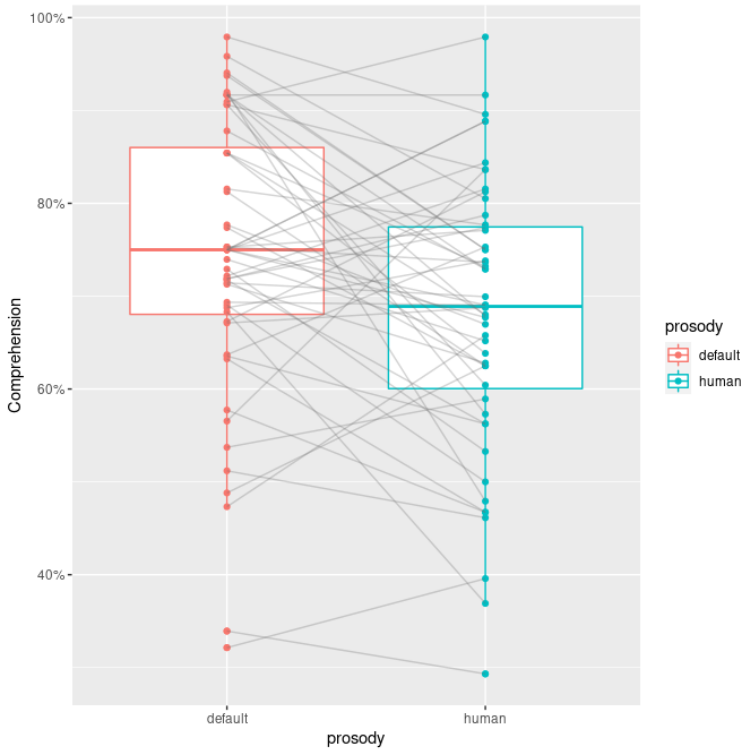


Figure 4.2: Overall paired comprehension ratios for 'default' and 'human' prosody conditions. Each grey line connecting points overlaid on the box plots represents the paired comprehension ratios of a single participant.

intelligible.

One plausible explanation for this effect is that, while prosody might enhance second order effects in human-robot interaction (boosting expressiveness, conveying affect, etc.), possibly increasing robot acceptance, changes it introduces to the pitch, rate and timing of the spoken sentence may come with the trade-off of making *comprehension* harder, not easier. This is correlated with the prior finding that casual human speech was shown to yield less intelligible prosody than other, more deliberate ways of speaking [160]. Furthermore, as prosody can provide contextual cues, sentences uttered with little context may have been harder to understand. This is supported by the finding that shorter sentences were significantly harder to comprehend when spoken in the 'human' voice condition. As the two clauses of the longer sentences were semantically linked, it may have been easier to reconstruct misheard words, whereas the above issue is compounded for shorter sentences.

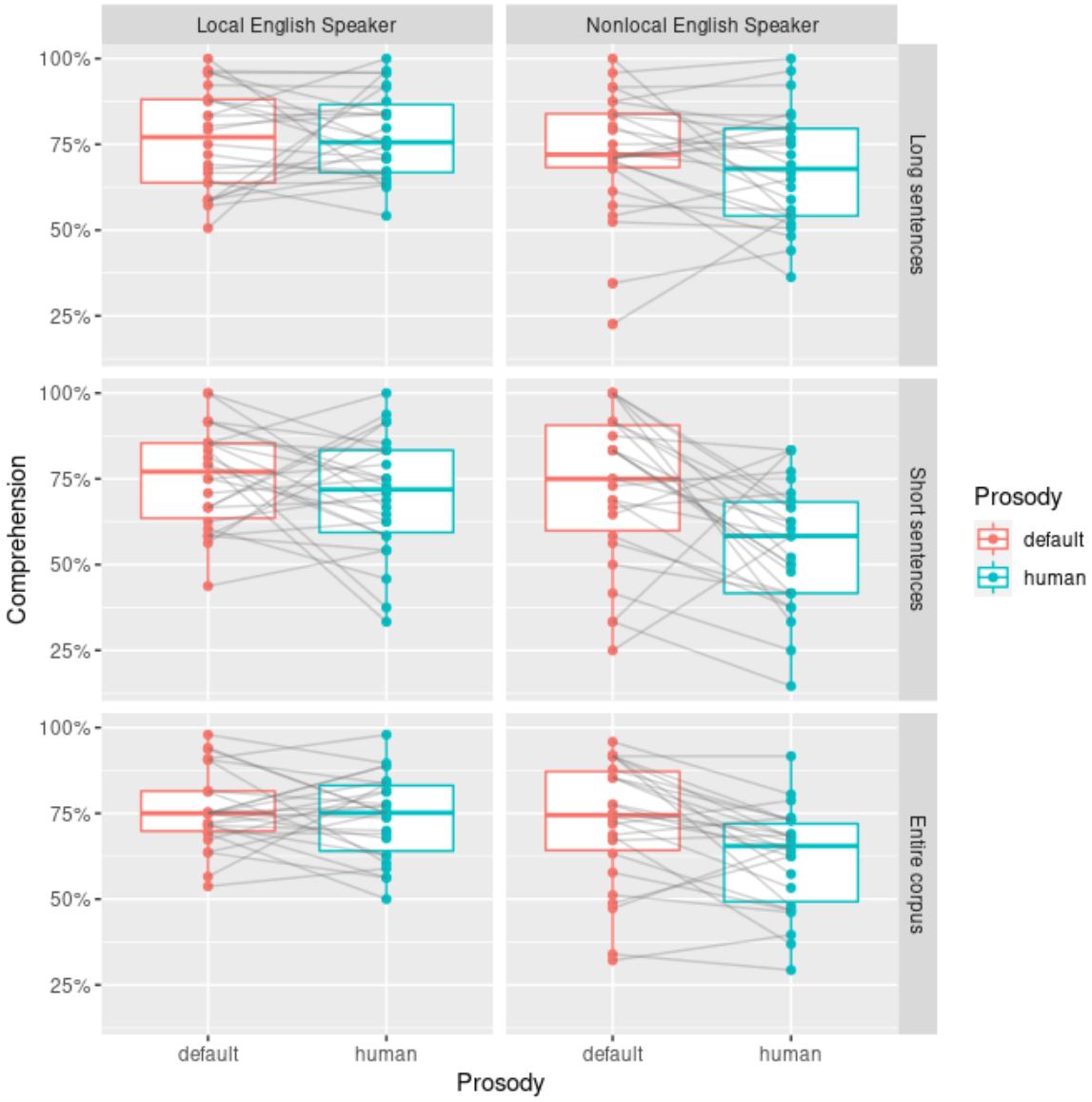


Figure 4.3: Paired comprehension ratios along both levels of participant English dialect locality (local and nonlocal) and sentence length (short, long and overall)

The interaction effect identified between prosody levels and locality on comprehension can be visually observed in Figure 4.3; the degree to which 'human' prosody negatively affected comprehension was substantially more severe for non-locals (graphs on the right), with substantially smaller differences in means for locals (graphs on the left). This also supports the above explanation; participants who did not learn English in the local region found it more challenging to comprehend prosodic speech modeled on an actor with a dialect from that region. An intuitive reason for this is the region-dependent nature of prosody; prosody that matches an accent the listener is closely familiar with is less likely to cause comprehension issues, but contextual cues provided by prosody from a foreign dialect may be lost, and the resulting speech less intelligible to the listener. This explanation is consistent with findings in the literature regarding foreign prosody comprehension reviewed in 2.3.3.

It is also important to put these findings in the context of partner models. The partner models that participants formed of the robot can be impacted by cultural biases, and can influence the perceived competency of the robot—if a participant judged a particular cultural style of prosody higher, even subconsciously, they may have seen the robot as more capable and understood more as a result. Considering the similarity-attraction effect observed in the literature [164, 165], it is possible this bias occurred based on how similar the robot prosody was to one's own, so non-locals may have formed a less favourable partner model of the robot with unfamiliar prosody and vice versa.

While not accounting for the observed interaction effects, an alternative explanation for our findings may be connected with limitations in the TTS engine's ability to produce prosodic speech. While efforts were made to model the synthetic speech as closely as possible on natural speech, commercial TTS remain limited in their ability to reproduce the richness of human speech. Although the pitch of whole words could be modified, it was not possible to control the pitch contour, leading to the loss of, for example, any rising inflection on individual words. As a result, it is possible that comprehension was decreased was because the prosody was "not sufficiently human-like".

We know from previous research [121, 146, 183, 194, 195] that prosody has an impact on intelligibility as well as naturalness and acceptance by humans; these results provide

analogous evidence that prosody also affects robot speech comprehension, and suggests that design efforts to increase the latter may negatively affect the former. These findings provide new insights into the use of prosody in robot speech, and make clearer its potential benefits and drawbacks when used in robot design and human-robot interaction.

As prosody is context-dependent, robot developers should consider how much context will apply to a robot's utterances, and envisage discarding human-like prosody if there is little context to utterances in its use case. Perhaps more importantly, developers should consider their target audience and culture; depending on the manner of speaking that an audience is accustomed to, a mismatch may lead in a decrease in overall comprehension, possibly because of the resulting partner model.[7] What these results suggest in this regard is that tailoring robot prosody to match the cultural expectations of the audience may increase acceptance and anthropomorphism without sacrificing intelligibility, but developers should remain aware of the broader social and ethical implications in assigning ethnic traits to a robot, notably if this would contribute to existing racial biases.

Limitations and future work

Several limitations were identified in this experiment, the first being those of the text-to-speech engine. Although mainstream in use and availability, the TTS engine we used did not allow absolute pitch specification, and instead pitch must be determined iteratively relative to the default output for each sentence. Also, as mentioned earlier in the section, the software did not allow us to implement pitch contour control. Although unlikely to be the only reason for the observed difference, especially considering the difference in results between the studied test conditions, the decrease in comprehension when prosody was modified may have been influenced by this shortcoming–more control over prosodic alterations are encouraged in future studies. In the same regard, future experiments building on this might explore the mapping of other human prosody elements aside from pitch and duration, such as loudness and timbre, experimenting with different TTS systems that might offer more varied (or

[7] A deeper factor yet to consider is that, depending on the context, being perceived as a certain ethnicity may trigger discriminatory, even xenophobic biases in the human observer [299, 300]. This has been a topic of conversation in robo-ethics in recent years [301, 302].

improved) parameter control, or in the case where source prosody is immediately available, exploring non-TTS methods of speech generation such as voice filtering.

A single type of noise (speech-shaped noise from the Harvard speech corpus [292]) was used at a fixed signal-to-noise ratio in this experiment. Future work may explore a variety of noise types, either contextual (such as environmental background noises) or generated to match the audio characteristics of each individual synthetic voice. Also, although not necessary for speech-in-noise testing, a range of different signal-to-noise ratios would provide more thorough results.

We explored the impact of cultural background in regards to robot speech comprehension, based on the prosody of local English, showing preliminary indications that a foreign prosody can have an adverse effect on comprehension. The relatively small sample size in this study made it difficult to study distinctions between groups any further; future research may investigate the effects of local prosody for both local and foreign participants, restricted to only native English speakers. The impact of cultural familiarity and bias in this study would also benefit from further investigation, through the use of partner models and other interaction design concepts. Pushing this approach further, as accents have been shown to have an effect on robot perception [164, 303, 165], of interest would be a larger study involving multiple prosodies from different world origins, and their intelligibility by a more diverse pool of participants.

Finally, the robot used in this context was virtual, presented to participants in the form of a video on a PC screen. Employing a physical embodied robot would more accurately simulate the experience of interacting with a robot in a real-world use-case. However, we also anticipate drawbacks of this embodied approach, namely in terms of overall repeatably since ensuring the same level of control over acoustics parameters during the experiment would be notably weaker.

Chapter 5

Spatial audio in social robot teleoperation

This chapter is based on the following peer-reviewed publication:

- A. K. Coyne, K. Sapkota, and C. McGinn, “"Who said that?" Applying the Situation Awareness Global Assessment Technique to Social Telepresence,” *ACM Transactions on Human-Robot Interaction*, pp. 1-22, Apr. 2023

5.1 Introduction

With the ongoing COVID-19 pandemic, remote social interaction has shown itself to be an essential part of modern life, with remote work and videoconferencing becoming widely accepted alternatives to their in-person equivalents. With this, the phenomenon of “Zoom fatigue” has also emerged [304], where the excessive use of virtual communication can result in decreased energy and motivation [305]. This has been attributed in part to the increased difficulty perceiving and conveying non-verbal information (social cues, such as turn-taking during conversations) through a conventional monitor-webcam setup [306]. In response to the limitations of existing solutions, new technologies are emerging to facilitate remote interaction such as robots or virtual reality applications. Their informed use may help mitigate the above communication issue through changes in interaction paradigms, such as allowing remote social

interaction to more closely mirror in-person interaction. Self-effacing natural user interfaces such as this can allow for higher exchange of non-verbal information without the added cognitive load.

Teleoperated social robots are suitable platforms to address this in practice, notably telepresence robots or social humanoid robots. The former are already in use commercially [307], providing the operator the experience of being socially present at a remote location. The latter, while still mostly restricted to research applications, could also act as a natural user interface for remote interactants on behalf of the operator, providing more human social cues (body language etc) through humanoid embodiment.

The previous two chapters have explored expression of social cues during robot teleoperation (affect and speech), but an equally important part of social interaction is remote perception of the social cues of others - for a robot operator, this can be understood through the lens of Situation Awareness (SA). As with any any teleoperated robot, operator SA is important for adequate performance and low cognitive load, so the subtleties of non-verbal information during social interaction makes SA all the more important for telepresence robots. However, where situation awareness has been improved for non-social robots through representative instrumentation (e.g. minimaps, haptic feedback, etc), social cues are less feasibly abstracted—they require the *human* information feedback of an NUI to be easily perceived.

One potential avenue for enhancing situation awareness through the paradigm of an NUI is via spatial audio feedback. As reviewed in 2, the human brain can instinctively localize where sounds are coming from by combining several different processes, one of the most significant being difference in sound arrival times between the ears [265]. By mathematically modelling the human head in a head-related transfer function (HRTF), and using a pair of headphones, sounds can be made to be perceived from different locations - this is often used in interactive media such as video games [266] to increase operator immersion in a virtual environment. The binaural nature of hearing is also a major contributor to the "cocktail party effect", [267] our ability to distinguish and focus on an auditory source in noisy environments, including speech when multiple people are talking simultaneously. As spatial audio feedback increases the quantity of information a user receives through the NUI paradigm, it would

seem logical that its inclusion over mono audio feedback would increase situation awareness of the remote location. The positive effect of spatial audio on situation awareness has largely been proven for non-social teleoperation tasks, such as search tasks or UAV piloting; its effect when applied to social contexts has yet to be explored.

From the practical perspective of the robot designer, the casual inclusion of spatial audio feedback features is not always possible. Its most limited form requires at least two microphones mounted on the robot platform, along with complex audio processing techniques. The time and effort required to implement it can be substantial, and while commercial solutions exist, software implementations of audio spatializers can be taxing on computing resources. Dedicated hardware for spatial audio capture and processing is an alternative, but its viability is limited by the high cost of such equipment.

With this in mind, the benefits of spatial audio feedback for social robot teleoperation require proper examination, to inform important decisions of robot designers and manufacturers. In this chapter, we aim to evaluate the effect of spatial audio feedback during social robot teleoperation, both on situation awareness metrics, as well as through qualitative analysis. Towards this goal, we attempt to develop a repeatable methodology for measuring social situation awareness by adapting existing tools.

The following section details the development of a technique for measuring situation awareness in social contexts, by adapting the widely-used SAGAT. Following this we present the methodology used for participant testing, employing this technique. A summary of experimental results are then presented in the subsequent section, followed by their discussion and a conclusion on the findings.

5.2 Developing an assessment technique of social situation awareness

An obstacle to the aim of this paper is the lack of methods available for evaluating situation awareness (SA) when applied to social contexts, which we will name social situation awareness (SSA). With caution taken to ensure reliability, this section details the development of a

measurement of social situation awareness.

5.2.1 Social interaction and awareness traits of interest

A defining feature of social robotics, and by extension social telepresence, is the use of a *social interface* for human-robot interaction [308, 18]. The social interface can be seen as the communication medium used for everyday human-human social interaction, which includes channels such as speech, gesture and affect expression.

According to Endsley's model of situation awareness [61], SA is defined strictly relative to performance in a task, towards one or multiple goals and objectives. Social interaction on the other hand is seldom goal- or task-driven, however the design exercise of social telepresence has goals, one of which is *to simulate the social interface as closely as possible to the experience of in-person interaction.* SA in social telepresence can be therefore be informed by this goal of successful social interaction.

Social awareness it is not a concretely-defined concept, but its mention in psychosocial literature [309, 310] as well as technology-related publications [311] often concerns high-level awareness during social interactions, such as awareness of emotional states (as provided by empathy) or of socioeconomic factors. While it could be argued that these high-level concepts are more important for social interaction, it is difficult to conceptualise a consistent measurement of awareness thereof, as it possibly depends much more on the operator than the teleoperation interface.[1] Furthermore, if an operator fails to pick up on more basic cues (e.g. who is saying a phrase), then their ability to pick up on more complex ones (e.g. who is surprised to hear the phrase) is likely impacted nonetheless.

Because of this, we will target a low-level form of social situation awareness: knowledge of basic non-emotional information about a social situation that can be obtained only through the social interface. This can include for example the names or explicit social roles of people interacted with, but not information about their appearance (which can be acquired non-socially). Targeting this low level of awareness will minimise the complexity of the problem

[1] Measurement of the awareness of these higher-level concepts can take entirely different approaches that do not involve the teleoperation interface, for example through Baron-Cohen's Empathy Quotient [312].

at hand, potentially increasing the reliability of the technique, while also allowing for shorter, more practical procedures for participant testing.

5.2.2 Existing technique selection

To maximise reliability, an existing situation awareness assessment technique will be adapted to measure awareness of the above traits. Despite being one of the most commonly used subjective techniques, it is judged that the abstract experiential questions of Taylor's Situation Awareness Rating Technique (SART) may be confusing to non-expert participants[2] and are overall unsuited for measuring situation awareness for social interaction. Although the questions of Witmer and Singer's Presence Questionnaire may be of value for its relevance to telepresence studies, an objective technique is preferred for consistency.

While real-time variant techniques of SAGAT may provide additional information of interest (workload measurements in the case of SPAM [70]), it is unknown whether the potentially intrusive nature of the questions might introduce a confounding factor during social interaction. This risk, combined with the widespread acceptance and stronger empirical validation of SAGAT, makes the latter the most reliable choice.

5.2.3 Adapting SAGAT for social contexts

The body of literature on SAGAT details requirements and guidelines for its administration. However, as its original intended use concerned the piloting of unmanned aerial vehicles [66] there are complications adapting it to social interaction. As many of SAGAT recommendations as possible will be replicated, with some changes for it to pertain to social situation awareness.

For one, it is instructed that SAGAT queries be created through a goal-directed task analysis [313] i.e. based on the operator's goal, but as mentioned in 5.2.1 the average real-world social situation is seldom intentionally goal-driven, and the telepresence goal of *engaging in social interaction comparable to face-to-face interaction* is too broad to analyse in this

[2] A SART questionnaire includes queries such as "How many variables are changing within the situation?" and "How much information have you gained about the situation?"

manner. We propose instead that analysis of the *context* of the social interaction determine the selection of important information to query.

SAGAT divides awareness into three levels:

1. **Perception of Data**
2. **Comprehension of meaning**
3. **Future projection**

Each SAGAT query concerns awareness of one of the three levels.

The first level, *Perception of data*, translates well to the social domain. Originally, this would consist of awareness about the position and speed of remote aircrafts. Adapted to social interaction, this data can incorporate most of what was targeted in subsection 5.2.1—data explicitly expressed through the social interface, such as names.

Comprehension of meaning, while slightly more complex to translate, can still be adapted to social interaction. One potential consideration is the awareness of emotional states based on secondary social cues such as speech prosody [314] but expressing, perceiving and even defining emotional states in this manner is potentially difficult [315]. We will instead consider Level 2 awareness to concern implicit contextual *objective* data, requiring understanding of the social situation to identify. For instance, while awareness of the information "Brian wishes to leave" is a deeper comprehension of implicit social meaning, it is subjective and may be difficult to pick up on, so is considered unreliable for participant testing. Contextual information such as "Brian is chairing the meeting", however, is more suitable; it is objective while remaining implicitly expressed and socially relevant, and it requires awareness and synthesis of lower-level social signals to comprehend.

The third level, *Prediction of future events*, is less applicable to social contexts. Social interaction can be inherently unpredictable, and while predictions during a conversation might occur intuitively, it is assumed that conscious attempts to anticipate the outcome of an interaction to be unneccesary for the majority of cases (akin to considering social interaction goal-driven). This level shall therefore not be considered during generation of SAGAT queries.

Beyond the selection of queries, a substantial issue that may exist when applying SAGAT socially may be that of *demand characteristics* [316]. While awareness of socially-relevant information is maintained passively, participants that are explicitly aware that they are being queried on that awareness may attempt to overcompensate, deliberately seeking pieces of information they would not otherwise. This could drastically bias results in a positive direction, particularly in the case of a repeated-measures trial, where participants may overcompensate in this manner during the second test condition. This foregrounds the need for measures mitigating demand characteristics and order effects; it is recommended that the study be conducted in between-subjects design, and/or that the intent to measure situation awareness is obscured or distracted from.

5.2.4 Practical considerations

SAGAT requires the situation be simulated, so that it can be paused to administer queries. The simulation used for this must be a sufficiently valid representation of reality, in this case a valid representation of social interaction. The ubiquitous nature of remote video communication in modern life (and the acceptance of video chats as a valid form of social interaction) makes this easier to accomplish; the simulator does not need to have a similar level of immersion to in-person interaction, as long as it is at least as immersive as a video chat.

Communication through social interaction can nonetheless include dozens of channels [260], which cannot feasibly all be accounted for, therefore to construct a simulation of social interaction and faithfully convey as many natural social signals as possible, live video recordings of actors must be used rather than computer-generated images. Consequentially, conditions cannot be easily randomized, limiting experimental design. While the conditions of a flight simulator can be changed in software (aircraft positions, velocities etc), with potentially no two identical situations across all participants, for a social simulator every video used must be captured in advance. Although variations can be generated by editing videos together programmatically, two issues prevent them from being viable for use. For one, to maintain immersion and realism, individual phrases should not be interrupted by a cut in

the video, which means that the information provided in each utterance cannot be altered, drastically decreasing the random sample space, and with it the utility of programmatic editing. But even without splitting sentences, cuts in the video may interrupt the natural flow of the overall interaction portrayed by actors, possibly destroying subtler social signals.[3] This limits the available test conditions to the amount of videos the researcher can feasibly record. While requiring more resources, having a larger pool of different videos to draw from for each participant will mitigate confounding factors, however it may also decrease the sensitivity of the trial. Regardless of the quantity of videos used, they should all concern the same type of social encounter and the same SAGAT queries should apply across all, but with different ground-truths determining the correct answer.

5.2.5 Summary

The final technique developed to assess social situation awareness (SSA) is as follows:

- A simple, common social context for the study is chosen, depending on the research focus
- The information to query participants about is chosen through analysis of this context, along the first two of the three SAGAT levels of awareness:
  - **Perception of data**, information explicitly available via the social interface
  - **Comprehension of meaning**, implicit information deduced using first-level information and through context
- This information must be:
  - socially relevant
  - objective in nature

[3]The use of *live acting* for participant testing is a hypothetical alternative allowing for shuffled test conditions, in which actors engage with each participant in real-time during the experiment as opposed to through pre-recorded videos. While this approach would allow for a near-perfect simulation of natural social interaction, it is would require rigorous caution to ensure consistency across participants, and its replicability would presumably be unsound.

  - non-emotional

- A simulation is created using recorded footage of actors, simulating social interaction in the chosen context

- SAGAT is administered as appropriate during testing, using queries on the selected information

Section 5.3 below details the practical use of this technique to assess the effect of spatial audio feedback on social situation awareness.

5.3 Methodology

The goal of the experiment was to evaluate the effect on social situation awareness (SSA) of spatial audio feedback, by comparison to mono audio feedback, during social robot teleoperation. To do so, we followed the technique devised in the previous section and created a simulation of social robot teleoperation, with the robot at the centre of a social scene; participants were asked to follow along as though teleoperating the robot remotely. One group of participants received mono audio feedback as a baseline, the other received spatial audio, and were therefore capable of localising sounds in the virtual environment.

5.3.1 Stimuli creation and administration

To simulate a social interaction as closely as possible while relying on pre-recorded video, omnidirectional footage was recorded of the portrayed scene.

Social context The form of social interaction chosen for this study is a classroom quiz scenario. A scene like this is a structured, systemic way to represent a social interaction, as other typical social interactions can be more complex. This situation is also a potential use-case for social robot teleoperation, in the form of robot-mediated remote learning or teaching.



Figure 5.1: Side view of the set used for recording stimuli. Omnidirectional camera and ambisonic microphones can be seen in the centre.

Design Actors were recruited (two men and four women), to portray one quiz master and five contestants. So that the simulation would adequately represent robot teleoperation by contrast to a conventional video call, actors were seated at desks in a circle around the camera (as shown in Figure 5.1)—this was to encourage the operator to look around the virtual environment rather than simply view it like a static camera feed, so that the workload would incorporate the control input component.[4] Quiz questions that the quiz master asked of the contestants were explicitly chosen to be extremely difficult or very vaguely worded, so although the questions and answers seemed conventional in a quiz, it was unlikely that contestants would be able to answer using prior knowledge. Two videos were recorded using two distinct scripts following the same format. Actor positions were shuffled, and a different set of quiz questions, character roles and names were used between the two. Quiz questions and answers for both videos are detailed in Appendix B.

Duration One recommendation for the administration of SAGAT [313] is that a minimum of 3 minutes must have elapsed from the start of the simulation before the first pause, and

[4] A secondary reason for this robot-centric layout (as opposed to including the robot in the social circle with the actors) was motivated by an informal observation made some years prior by the researchers. It was observed that when a telepresence robot brings a substantial degree of novelty to a social gathering, it can draw the attention of groups of people, who observe and interact with it by surrounding it in a circle in this manner, requiring a mentally-taxing 360 degrees of social awareness for the operator.

that further pauses must be spaced apart by a minimum of 1 minute. In order to keep simulation time relatively short, both for ease of testing with participants and to facilitate video recording, it was decided that SAGAT pauses be administered twice per video; first at a random time between 2 and 4 minutes, and then at a random time at least one minute after the first. Each video would therefore last 5 minutes total.

Script The quiz master began by introducing the experiment to the operator, initiating a round of introductions where every contestant spoke their name. This provided the operator a chance to hear every contestant name, and also encouraged them to visually explore the room. Then the main loop of the scene began: every 15 seconds, the quiz master would ask a question to a specific contestant, addressing them by name (e.g. “Which European city hosted the 1936 Summer Olympics? Annie?”). The chosen contestant would begin by saying “I think the answer is-” to prime the operator for the answer. Then both the contestant and another "distractor" contestant would speak different answers simultaneously, of which one at random was correct, talking over one another (Annie: “I think the answer is...London”; Maria: “Berlin”). The quiz master would then congratulate the contestant who answered correctly (“That’s right, Maria"), pause for the remainder of the 15 seconds, and continue with the next question. This main loop of the script would continue until the video time reached 5 minutes. A complete list of quiz questions and answers is provided in Appendix B.

Recording Video was recorded using a Ricoh Theta Z1 omnidirectional camera. The camera captures videos in a spherical format using its dual fish-eye lens, which are converted to a conventional 360-degree format (MPEG-4) using Ricoh Theta proprietary software. Audio was captured using a Sennheiser AMBEO VR Mic, a ambisonic microphone array consisting of four high-fidelity microphones in tetrahedral arrangement, via the Zoom H6 Audio Recorder. Sennheiser’s AMBEO A-B proprietary software was used to convert the recording to standard Ambisonics-B format (WAV).

Implementation

Simulator For its support in the literature for use in human-robot interaction simulators [317, 318] as well as its ease in handling multimedia content, Unity was chosen as a framework to create the simulator. The 360-degree MPEG-4 videos were projected on the interior of a Sphere object in Unity, with a Camera object at its centre to represent the teleoperated robot. The open-source plugin Resonance Audio [319] was used to generate spatial audio in real time based on the ambisonic recordings, conveying both interaural level and time differences using head-related transfer functions.

Teleoperation interface A conventional desktop PC setup was used as a teleoperation interface, for its widespread familiarity and ease of access. Audio feedback was provided through a pair of high-fidelity over-ear headphones. Pressing the left and right arrow keys of the keyboard would rotate the camera object, as if rotating the camera feed of a robot—the spatial audio feed would rotate accordingly.

5.3.2 Experimental design

Task

In order to mitigate SAGAT query demand characteristics as explained in 5.2.3, as well as to engage and maintain focus on the scene, a simple task was created; participants were asked to follow along with the video shown to them, and to identify for each question the correct answer. After a question was answered, both the correct and incorrect answers that were uttered were displayed at the bottom of the screen, and the participant chose which one they believed was correct by pressing a key on the keyboard. In this manner, the task pertained to the situation and awareness thereof, without any direct overlap with the content of the SAGAT queries. A video demonstrating the simulator and experimental process is provided as supplemental material.[5]

[5] Accessible at the following link: `https://adm.pizza/thesis/supplemental/`

Manipulations

Two videos were recorded, and two levels of the independent variable (audio feedback) were to be evaluated, resulting in four test condition permutations. Each participant would view one of the four permutations in between-subjects experimental design, resulting in two groups; one having received Mono audio feedback, the other Spatial audio feedback.

The increased sensitivity of conducting a within-subjects trial a.k.a. repeated measures would have been advantageous (presenting one of each video in random order with one for each audio feedback condition). However it was judged based on informal testing of the simulator that ordering effects could be significant between the two audio feedback conditions, potentially introducing confounding factors, so the decision was made to use independent measures.

Measures

While the SAGAT score of each participant is the primary metric of interest to this study, measuring situation awareness, secondary metrics were also recorded and analysed to gain further understanding of test results.

SAGAT score This score of social situation awareness is measured using the technique devised in Section 5.2.1. Queries were created that pertained socially to the quiz scenario, such as "Which contestant last answered correctly?". No queries required the participant to identify the correct answer to a quiz question, as this task was already asked of participants (see 5.3.2). The full pool of questions, randomized between both SAGAT pauses, is shown in Table 5.1.

Each query allowed the participant to select from multiple response options—six options available, with a single correct answer, as well as an additional "I don't know" option. Two SAGAT pauses were administered per participant of five queries each, resulting in a final SAGAT score out of 10 (a score of 10 showing that every question was answered correctly, indicating high situation awareness). A sample SAGAT query screen displayed during the simulation is shown in Figure 5.2

Level of awareness	Queries
1. Perception of Data	*"What was the most recent question?"* *"Who is sitting left/right of the quiz master?"* *"Who is sitting two seats to the left/right of the quiz master?"* *"What is the colour of the quiz master's shirt?"* *"Identify one person who was in your field of view just before the simulation paused."*
2. Comprehension of Meaning	*"Who was asked to answer a question most recently?"* *"Who is the quiz master?"* *"Who last answered correctly?"* *"Who last answered incorrectly?"* *"What kind of social event is taking place here?"*

Table 5.1: Pool of all potential queries delivered during SAGAT pauses.

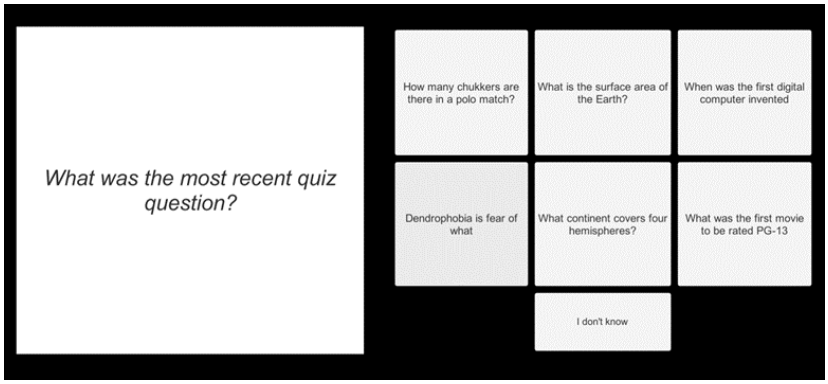


Figure 5.2: A sample view of the teleoperation screen during a SAGAT pause, during which the participant was queried about their awareness of aspects of the scene.

Presence Questionnaire With the value of *presence* in relation to situation awareness, particularly where telepresence is concerned, of a subset of questions from Witmer and Singer's Presence Questionnaire [51] was administered to participants. Each response was provided through a 7-point Likert scale. Question labels were slightly adjusted and extended for clarity, as shown in Table 5.2. The sum of scores for all component questions constitutes an overall presence score—for 6 questions, this will be a score out of 42.

Positional metrics Spatial audio can enable an operator to localise sounds without needing to see the source. It is therefore predicted that mono audio feedback will incite participants to visually pan around the simulation more by comparison. To investigate this, the following metrics were derived using positional data and user input logs from the teleoperation interface:

- *Mean answer time* - the mean time in seconds for each participant to select which of the two answers to a quiz question they deemed correct.
- *Ratio of time in motion* - the ratio of experiment time spent turning the camera to total experiment time.
- *Mean viewing angle* - the mean angle in degrees (where an angle of 0 is facing the quiz master).
- *Heading angle variance* - we wish to evaluate the angular range of motion employed for each participant, but it is assumed participants will pan to view the entire scene (360 degrees) at least once during the experiment. The variance of the heading angle over time for each particpant can represent the variability of angles in the given timespan; a low variance indicates an generally smaller angular range was viewed, while a high variance indicates the participant was more willing to cover wider ranges over the course of the experiment.

Task performance A lesser metric is that of task performance; the total amount of correct quiz answers identified. Every time a question is asked by the quiz master to a contestant,

Label	Original number [51]	Original question [51]	Adjusted for clarity
P1	5	*"How much did the visual aspects of the environment involve you?"*	*"How much did the visual aspects of the environment involve you? In other words, how much did the visual component of the experience contribute to the awareness of the situation?"*
P2	6	*"How much did the auditory aspects of the environment involve you?"*	*"How much did the auditory aspects of the environment involve you? In other words, how much did the audio component of the experience contribute to the awareness of the situation?"*
P3 *	15	*"How well could you identify sounds?"*	*"How well could you identify sounds?"*
P4	16	*"How well could you localize sounds?"*	*"How well could you tell where sounds were coming from? In other words, how well could you localize the direction of sound?"*
P5 *	12	*"How much did your experiences in the virtual environment seem consistent with your real-world experiences?"*	*"How much did your experiences in the virtual environment seem consistent with your real-world experiences?"*
P6 *	23	*"How involved were you in the virtual environment experience?"*	*"How involved were you in the virtual environment experience?"*

* Question unchanged from original questionnaire

Table 5.2: List of questions used from Witmer and Signer's Presence Questionnaire [51], adjusted for clarity and administered post-experiment.

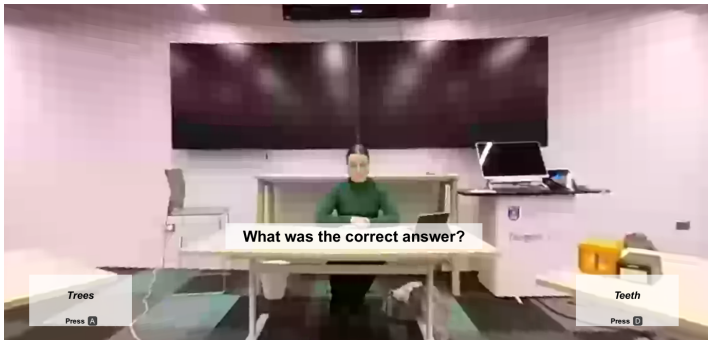


Figure 5.3: A view of the simulator screen shortly after a quiz question was asked, illustrating the task asked of the participant to identify the correct answer.
Screenshot captured after testing was completed, with inferior video resolution than what was used during data collection.

the participant is prompted on-screen to identify which answer was correct, as shown in Figure 5.3. The task was not created with the intent to measure performance but rather to mitigate demand characteristics (drawing attention away from the SAGAT queries), and with no particular precendent in the literature for the use of such a metric, it was considered auxiliary to the others.

Qualitative feedback The post-experiment questionnaire includes the following optional fields for positive and negative open feedback from participants:

- *"Is there anything in particular that you liked about the experience?"*
- *"Is there anything in particular that you disliked about the experience?"*

Participants

The goal was to recruit a representative sample of the lay population, controlled for English language fluency, hearing or spatial awareness issues, and colourblindness. This was done through random recruitment of the footfall in semi-public locations.

Procedure

Participants were recruited one by one. After providing informed consent, the participant would begin a 2-minute trial of the experiment during which no data was recorded, which

included one SAGAT pause. The participant was encouraged to ask the researcher any necessary questions during the trial. Once complete, the screen would fade to black, and the participant informed that they could begin the experiment proper when ready, which began once they accepted through the simulator interface. After the simulation, participants completed a questionnaire based on their experience, and were thanked for their participation.

Analysis

Data was pre-processed using Python scripts. Statistical analysis was performed using R. [298]

We wish to evaluate whether the difference in means for SAGAT scores between the two groups is significant. To first determine parametricity, a Shapiro-Wilk normality test was conducted. A two-tailed independent-measures t-test would be conducted for normally distributed data, a Wilcoxon rank-sum test if not normally distributed. The same process was followed to determine a difference in means between the answers given in the presence questionnaire. The chosen significance level in all cases was $\alpha = 0.05$.

A Pearson correlation matrix was also calculated across all quantitative variables, to investigate and evaluate their research value. Finally, qualitative analysis was conducted to identify any overarching themes in participant feedback.

5.4 Results

A total of 56 participants were recruited from two separate locations, a college building (28) and an office block (28). Participant age ranged from 18 to 60 ($\mu = 28.9$, $\sigma = 11.5$), of which 55% were male. All participants reported to be fluent in the English language, and none reported any form of colourblindness, hearing issues or spatial awareness issues.

Quantitative results

SAGAT score data across all participants was not normally distributed ($W = 0.926, p = 0.002$). A Wilcoxon rank-sum test showed that the group that received spatial audio feedback

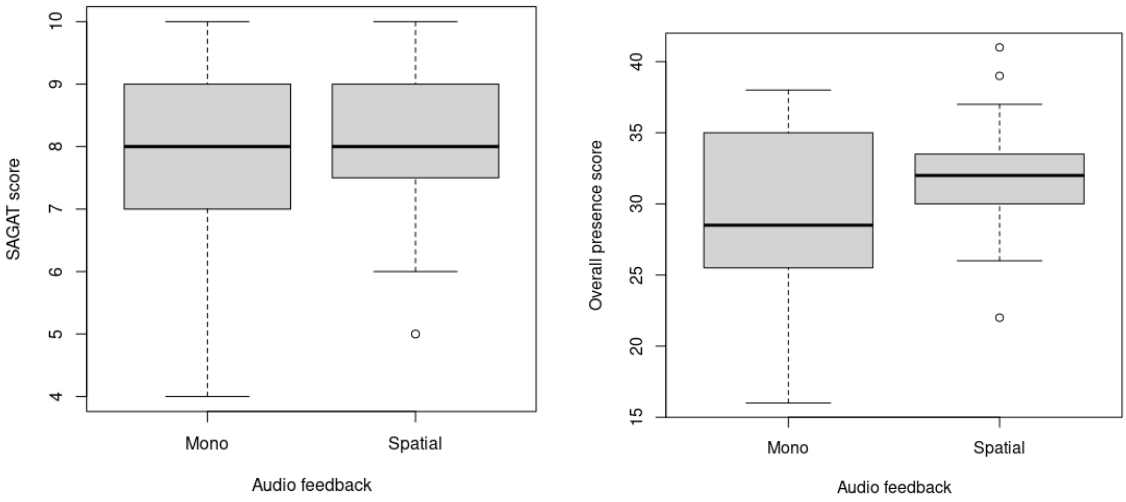


Figure 5.4: Box plot summaries of SAGAT scores and overall presence scores for both Mono ($n_1 = 28$) and Spatial ($n_2 = 28$) audio feedback groups.



Figure 5.5: Density of SAGAT scores observed for both Mono ($n_1 = 28$) and Spatial ($n_2 = 28$) audio feedback groups.

SAGAT Query	Mono audio			Spatial audio			Overall			
	n	C	r	n	C	r	n	C	r	Δr
"What was the most recent quiz question?"	27	23	0.85	33	28	0.85	60	51	0.85	0
"Who is the quiz master?"	21	20	0.95	22	22	1.00	43	42	0.98	+ .05
"Who is sitting two seats to the [l/r] of the quiz master (your [l/r])?"	29	20	0.69	33	22	0.67	62	42	0.68	− .02
"What is the colour of the quiz master's shirt?"	33	23	0.70	28	24	0.86	61	47	0.77	+ .16
"Who was asked to answer a question most recently?"	29	24	0.83	32	27	0.84	61	51	0.84	+ .01
"Who last answered incorrectly?"	28	14	0.50	32	20	0.62	60	34	0.57	+ .12
"Who last answered correctly?"	27	22	0.81	26	22	0.85	53	44	0.83	+ .04
"What kind of social event is taking place here?"	21	21	1.00	19	16	0.84	40	37	0.93	− .16
"Who is sitting [l/r] of the quiz master (your [l/r])?"	28	19	0.68	27	23	0.85	55	42	0.76	+ .17
"Identify one person who was in your field of view just before the simulation paused."	37	30	0.81	28	23	0.82	65	53	0.82	+ .01

Table 5.3: SAGAT scores detailed by each individual SAGAT query. As query selection was random, not every query was equally represented. With n the number of times the query appeared, C the number of times the query was answered correctly, r the ratio of correct answers, Δr the difference between the correct answer ratios of spatial and mono audio conditions.

did not yield significantly higher situation awareness scores compared to the group with only mono audio ($W = 338.5, p = 0.373$). Indeed, the median score was the same for both groups (8). No instance of a participant selecting the "I don't know" answer was recorded for any of the queries. SAGAT score data is summarised graphically in Figure 5.4 and numerically in Table 5.4, with detailed information on responses to each query shown in 5.3.

Although composite overall responses to the Questionnaire formed a normal distribution ($W = 0.976, p = 0.331$), individual responses to Presence Questionnaire were not normally distributed ($p \leq 0.004$). Cronbach's alpha showed that the six items of the Presence Questionnaire were poorly internally consistent ($\alpha = 0.577$). Wilcoxon rank-sum tests for each individually did not show any significantly different means between the spatial audio and mono audio groups ($p \geq 0.104$). These results are summarised in Table 5.5.

The other quantitative measures (task performance, time to answer etc) are summarised

	Mono audio		Spatial audio			
	μ	σ	μ	σ	W	p
SAGAT scores	7.71	1.61	8.11	1.31	338.5	0.373
Task performance	9.11	1.17	9.25	0.89	377.5	0.804
Mean answer time (s)	2.99	0.70	3.19	1.33	385	0.916
Ratio of time in motion	0.14	0.09	0.15	0.07	371	0.739
Mean viewing angle (°)	2.81	13.7	-0.43	10.6	475	0.178
Variance of viewing angle	69.0	20.4	79.0	20.1	296	0.118

Table 5.4: Statistical summary of objective results, with results of Wilcoxon rank-sum tests between both Mono ($n_1 = 28$) and Spatial ($n_2 = 28$) audio feedback groups

		Mono group		Spatial audio			
Question	Content	μ	σ	μ	σ	W	p
P1	"Visual involvement"	5.04	1.48	5.19	1.30	358.5	0.742
P2	"Auditory involvement"	5.79	1.37	5.85	1.32	364	0.808
P3	"Sound identification"	5.21	1.34	5.56	1.01	334	0.445
P4	"Sound localisation"	4.21	1.89	5.00	1.62	283	0.104
P5	"Real-world consistency"	4.64	1.31	4.74	1.70	337.5	0.492
P6	"General involvement"	4.89	1.71	5.56	1.25	296.5	0.163
	Overall presence score	29.8	5.51	31.9	4.22	306.5	0.230

Table 5.5: Statistical summary of results of Presence Questionnaire questions, with results of Wilcoxon rank-sum tests between both Mono ($n_1 = 28$) and Spatial ($n_2 = 28$) audio feedback groups. Shown here with reminder phrases for question content (full questions as seen by participants shown in Table 5.2)

in Table 5.4. Pearson correlations between all measures are shown in Table 5.6, identifying a number of significant correlations.

Qualitative results

32 items of positive feedback and 20 items of negative feedback were provided through the post-experiment questionnaire. Thematic analysis identified several themes of interest, shown in Table 5.7. Individual answers to positive and negative open feedback questions are detailed in Appendix B along with their thematic relevance.

"Immersion" and "Entertainment" in overall positive feedback The majority of feedback given across both groups (14 items total) was of appreciation for the immersion of the interface, in particular the ability to look around. A smaller number of items (8) across both groups reported appreciation for the entertainment of the experience.

"Sound localisation" in Spatial group positive feedback Participants of the Spatial audio group reported appreciation for the spatial nature of audio feedback and the ability to localise sounds (6 items).

"Difficulty, Frustration" in Mono group negative feedback Negative feedback provided by the group having received mono audio feedback describes experiencing *difficulty and overall frustration* with the task and interface (8 items, 6 more than in the Spatial audio group).

"Boredom, Impatience" in Spatial group negative feedback Negative feedback provided from the spatial audio feedback group reports impatience with wait times between questions or a desire for a faster, more dynamic interface (6 items, 5 more than in the Mono group).

	(1)	(2)	(3)	(4)	(5)	(6)	(P1)	(P2)	(P3)	(P4)	(P5)	(P6)
(1) SAGAT score	1											
(2) Task performance	0.28*	1										
(3) Mean answer time	-0.21	-0.49***	1									
(4) Ratio of time in motion	0.23	0.1	0.02	1								
(5) Mean viewing angle	0.11	0.26	-0.13	-0.05	1							
(6) Variance of viewing angle	-0.02	-0.12	0.15	0.19	-0.2	1						
(P1) "Visual involvement"	-0.02	0.04	0.11	0.09	-0.06	0.23	1					
(P2) "Auditory involvement"	0.14	0.02	-0.14	0.1	0	-0.03	0.04	1				
(P3) "Sound identification"	0.06	0.35**	-0.12	0.08	-0.09	-0.17	0.25	0.36**	1			
(P4) "Sound localisation"	-0.04	0.25	-0.06	-0.06	0.17	-0.29*	0.12	0.01	0.33*	1		
(P5) "Real-world consistency"	-0.11	0.33*	-0.17	-0.12	0.05	0.13	0.21	0.17	0.38**	0.33*	1	
(P6) "General involvement"	0.04	0.05	-0.02	0.06	-0.13	0	0.18	-0.01	0.39**	-0.01	0.21	1

*$p < .05$
**$p < .01$
***$p < .001$

Table 5.6: Pearson correlation matrix of all quantitative measures. (P1) through (P6) represent Presence Questionnaire answers, shown here with reminder phrase (full questions as seen by participants shown in Table 5.2)

97

Themes identified	Feedback categories	Mono	Spatial
	Positive feedback		
Immersion	Immersion / Ability to look around	7	7
	Spatial audio feedback	0	6
Entertainment	Challenging / Gamelike	3	2
	Novel experience	1	2
	Other	3	1
	Total	**14**	**18**
	Negative feedback		
Boredom, impatience	No vertical camera control	0	2
	Slow camera panning speed	1	2
	It was repetitive / boring	0	2
Difficulty, frustration	Audio was poor (increasing difficulty)	6	2
	Field of view too small	2	0
	Other	2	1
	Total	**11**	**9**

Table 5.7: Categorised summary of open feedback; with number of items from respective groups, and identified themes. *Positive feedback* consists of responses to the question "Is there anything in particular that you liked about the experience?" and *negative feedback* responses to the question "Is there anything in particular that you disliked about the experience?".

5.5 Discussion

Quantitative findings

While researchers predicted quantitative differences in favour of spatial audio, no statistically significant differences were found between both groups (for a global significance threshold of $\alpha = 0.05$) for situation awareness measurements, presence questionnaire results, or any secondary metrics. Although it is possible that spatial audio feedback has no effect on any of these factors by comparison to mono audio, the sensitivity of the study may have been poor for a number of reasons. For one, as shown in Figure 5.5, greater SAGAT scores were observed with higher densities, with 75% of SAGAT scores between 7 and 10. With data such as this grouped around the maximum, the *ceiling effect* may have occurred, whereby the upper limit placed on the measure reduces the meaningfulness of the data, potentially obscuring an effect that would otherwise be observable. Another reason is the low statistical power of the study (estimated at 0.45).

Table 5.6 shows several correlations of note. While not pertaining directly to the research hypothesis, a number may be of interest for improving the SSA assessment technique, or for the experimental design of subsequent studies. Several were identified among component questions of the Presence Questionnaire—with how the Questionnaire was designed, along with the similarity of the component questions and the method they are administered, any significant correlations between them are to be expected. A strongly significant negative correlation is that between *Task performance* and *Mean answer time*, which is also to no surprise—participants that are more confident of their answers (or simply more focused on the simulator) would be more likely to input the answers quickly. The positive correlation between *Task performance* and the Presence Questionnaire component on *Sound identification* can be similarly explained; the task itself required identifying sounds.

A weaker positive correlation is that of *SAGAT score* and *Task performance*. This reflects the similar, albeit non-overlapping nature of both SAGAT queries and the quiz task; both require information obtained through social signals, and the success rate of identifying the correct answer to a question based on events in the scene could be considered a very focused measurement of situation awareness. Finally, the positive correlation of *Variance of viewing angle* with *Sound localisation* shows how participants would look around the room more to compensate for difficulty localising sounds. This also shows initial promise for the use of the viewing angle variance as an objective measure, either of ease of sound localisation directly, or of workload related to difficulty localising sounds.

Qualitative findings

Examining Table 5.7, negative feedback provided by the Mono audio group predominantly pertained to the *difficulty and frustration* theme identified, while negative feedback from the Spatial audio group pertained to the theme of *boredom and impatience*. This illustrates a substantial difference in how the trial was experienced between both groups. What participants disliked about the Mono audio experience was the workload, particularly citing the audio feedback as a cause.[6] By comparison, what participants expressed they disliked about

[6]This can be seen in the full list of feedback items in Appendix B.

the Spatial audio experience was tedium, a desire for it to be more dynamic, and to engage more in the interface. In tandem with the fact that a substantial amount of positive feedback for the Spatial group was on the spatial nature of the audio, it can be induced that introducing spatial audio over mono likely *decreased* the workload for the task at hand. This is in accordance with the neurological basis of the "cocktail party problem" [268]; attempting to distinguish one thread of speech spoken simultaneously with others can heavily engage the brain, to the point of decreasing performance at concurrent tasks. The binaural hearing of spatial audio is known to be a major contributor to our ability to distinguish speech in this manner [267], so it can be intuited this would decrease the associated workload.

For the above reasons, along with the distribution of feedback (the spatial audio group having provided 29% more positive feedback items and 10% less negative feedback items than the mono audio group), it can be concluded that spatial audio was a qualitatively superior experience to mono audio, likely due in most part to the potential workload alleviation—the findings of this chapter indicate that spatial audio feedback may improve overall experience for a social robot operator. This is consistent with observations made in the literature around spatial audio; as reviewed in 2.3.7, participants often report appreciation for being able to localize sounds in prior studies. Finally, the praise given by both groups to the ability to look around using the interface underlines the potential value of embodied telepresence over conventional videoconferencing, as noted in 2.1.3.

Limitations and recommendations

The most substantial limitation of this evaluation was its statistical power, estimated at 0.45 for an assumed effect size of 0.5.[7] An increase in sample size could overcome this, as well as a redesign of the experiment to allow for repeated measures over both test conditions for each participant, although as explained in Section 5.2.1, it may be challenging to adapt to within-subjects design because of ordering effects and demand characteristics.

While the robot being placed at the centre of the group of actors was relevant to some teleoperation situations, it may be less representative of the majority of real-life interactions.

[7] as informed by [320].

Drawing on the domain of proxemics can be used to improve on this in future studies, such as through the use of Kendon's F-formations [321, 322].

Although care was taken to preserve sound directionality in the audio pipeline, a complementary study to validate the spatial audio of the setup could ensure with full confidence that spatial audio was properly conveyed, such as that performed by Kiselev et al. [323]. The setup of this experiment used first-order ambisonics—improvements can be made to the fidelity of spatial audio by increasing the ambisonic order of the microphone (using a larger microphone array).

The potential observed ceiling effect on SAGAT scores is another limitation of note. Adjusting for this in future studies can be done based on the data in Table 5.3, which breaks down the individual scores for each SAGAT query type. This can be used to assess the relative difficulty of future queries, and if a similar "quiz" scenario is repeated in a future study, can be prioritized for re-use.

Chapter 6

Conclusion

This research project aimed to enable the design and evaluation of social teleoperation interfaces for existing service robots, by addressing issues currently present in the design process.

Through analysis of the modalities of social communication, three key areas of focus were identified, aligned with the thesis research objectives. Towards accomplishing these objectives, several contributions were made to the field, as expressed in Chapter 1.

6.1 Reflection on research objectives

This section reiterates the research objectives and their motivation, along with how this thesis has addressed them.

Develop and validate a methodology for measuring and analysing perceived robot affect

Expression of emotions is a key part of social interaction. It had been shown in the literature that affect expression is a powerful tool for social robots and their teleoperation, as it can greatly enhance social interactions but also compromise on social communication if emotions are wrongly interpreted. Despite this, very few studies had been done to evaluate the perceived affect of robot platforms. Those that attempted to do so often focused on models of emotion with fundamental limitations such as categorical models, and overall suffered from

a lack of repeatability and very little methodological consensus.

In Chapter 3, a suite of statistical tools was created for quantitative evaluation of perceived robot affect, relying on the empirically validated Geneva Emotion Wheel. By adapting existing circumplex analyses, these tools allowed both frequentist analysis (such as null hypothesis significance testing) and probabilistic mapping of a robot's emotional palette through Bayesian inference. The former allows for evaluation and comparison of individual emotion expressions (such as through Student t-testing or F-testing) to ensure they are precise in how they are perceived, while the latter yields comprehensive maps that can assist in evaluating a robot's affect expression abilities overall, as well as connecting a desired perceived affect to the currently available range of expressions. These maps can be integrated into teleoperation interfaces as they can determine what the closest matching affect expression available is for a given emotion on the Geneva Emotion Wheel. Furthermore, the methods can also be extended for use to assess the perceived emotion of any robot, even when not equipped with deliberate affect expression. It could be argued that affect is universally relevant to human-robot interaction studies, and most robots are instinctively anthropomorphised by interactants, so using these tools on non-affective robots could determine the emotive qualia of their anthropomorphisation—they could determine how the robot appears to "feel" about them, which can inform the following iteration of design.

A trial of these tools demonstrated how they could reveal such insights into the strengths and flaws of a robot's affect expression palette, notably revealing which affect expressions of the Stevie II robot platform were misinterpreted—in particular, the "neutral" expression for Stevie was mildly perceived as "Disgust", which allows for the following iteration of design to remediate this issue. In a broader sense, Chapter 2 showed that one of the only current applications of robot teleoperation was for therapy for children with autism, in which a simple expression of emotion can have substantial impact—this research can ensure that affect is perceived as it is intended. Finally, the ability of these tools to map on the Wheel which of the robot's affect expressions most closely matched a given emotion also showed their potential use for the creation of empirically validated affect expression control schemes, enabling accurate emotion teleoperation for potentially any affective robot, successfully accomplishing

this research aim.

Evaluate the effect of human-informed prosody on comprehension of robot speech

Speech is arguably the primary channels of communication in social interaction. Many researchers have stressed the importance of expressive robot speech, and while a number of prior HRI studies involved speech, notably through text-to-speech engines, how best to deliver or teleoperate that robot speech in the context of social interaction had not been scrutinised, particularly in regards to speech prosody. This was despite the fact that a robot's voice can greatly impact its perceived partner model for interactants, potentially influencing humanness and acceptance. As was seen in Chapter 4, implementation of teleoperated speech prosody is not always beneficial for comprehension.

Natural human prosody was carried over to TTS robot speech and the resulting effect on speech comprehension was evaluated in Chapter 4, using common audiological methods. It was shown that, while the human prosody had little if any effect on comprehension for those familiar with it, it resulted in a significant decrease in comprehension for those unfamiliar with the regional prosody, regardless of whether or not they were native English speakers. This provides important information for the design of speech teleoperation, potentially reducing the priority that might be put on relaying human prosody. Another aspect of these findings is that prosody and expressive speech can render a robot more human-like, and the potential downsides of its implementation as shown here can be seen as a questioning of human-inspired design. This is in alignment with the concept of the "uncanny valley" mentioned in 1—while it is not yet fully understood, very human-like features in robots can be unexpectedly off-putting to observers. The findings of this study might encourage designers to embrace more symbolic representations of speech, drawing from fields of animation and online interactions (such as sound effects or emojis) to engage in a less-human but more modern, readily understood manner. While further research is required on the topic, given the scope of the thesis this study has answered this research goal by yielding important insights for the development of robot speech teleoperation interfaces.

Evaluate the effect of spatial audio feedback on social situation awareness during robot teleoperation

As established in the literature, the perception of social cues is one half of the bidirectional nature of social communication. The paradigm of an immersive natural user interface may enhance perception of social cues during robot teleoperation, in particular through spatial audio feedback, an accessible method for increasing operator immersion. However, while social awareness can be considered a social form of situation awareness (SA), linked to qualities such as presence in robot teleoperation, no methodologies existed for measuring it that exploit this connection, and any measure of perception of social cues relied on qualitative or subjective measures.

After laying out the process of creating an objective measurement for social situation awareness, by applying tried-and-tested SA measuerment techniques to the domain of social robot teleoperation, Chapter 5 detailed a participant study comparing social situation awareness between mono and spatial audio feedback. While the created methodology did not yield any significant differences in SA means, differences in qualitative and subjective metrics observed between the two groups indicated that spatial audio feedback resulted in a superior operator experience overall. Therefore, while not always possible because of robot platforms with limited audio capture abilities when it comes to robot teleoperation design, it would be an obvious feature to implement. This is made easier by the many freely-available audio processing software tools that have been popularised by the study of virtual avatars and the video game industry.

As such, while the SA assessment methodology requires further tuning and validation in order for it to be a viable option in future research and this research objective remains to be achieved, the broad benefits of spatial audio feedback during robot teleoperation were nonetheless clear, and foundations have been laid to apply robust measurements of situation awareness to the social domain.

The initial two chapters of this thesis identified key areas of focus for social robot teleoperation, each of which presented a unique unexplored design challenge. The research findings of

this thesis addressed these challenges in a platform-agnostic manner applicable to the majority of social robots, answering important questions and unknowns in the robot teleoperation design process, while offering methodological contributions that can form a toolkit for future research.

With its open-ended nature, the design space for social robot teleoperation is still enormous. But the work of this thesis reduces that space; these findings enable robot designers to move forward in the areas of greatest import for social teleoperation. Through this thesis, we hope to have revealed new ways for social robots to develop in academia, industry and everyday life.

6.2 Academic impact

The research conducted in this project has resulted in three peer-reviewed publications in high-ranking journals and conferences; the ACM/IEEE International Conference on Human-Robot Interaction with low acceptance [252], the leading journal ACM Transactions in Human-Robot Interaction [324], and the established IEEE International Symposium in Robot and Human Interactive Communication [325]. The research detailed in Chapter 3 on measuring perceived robot affect [252] was awarded Best Paper in its category at the Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI 2020). The circumplex statistical analysis of Geneva Emotion Wheel data, along with the open-source software tools that have been made available, has been used in subsequent research projects [253, 326].[1]

6.3 Limitations and future work

While strong scientific rigour was prioritised in the experimental design of these research projects, their primary limitations lie in their small sample sizes, which is a common obstacle in human-robot interaction research. As such, many of the methodologies presented above would benefit from further empirical validation.

[1] This publication has been cited by 11 others indexed by Scopus as of July 2023.

This work used the Stevie II robot platform for its research [14]. A full teleoperation system for Stevie could be made that combines the insights gained for the system, and evaluated for performance in real-world social interaction scenarios, with complementary measurements such as workload, presence and qualitative analysis.

Concerning the latter, in the same manner as other recent HRI publications [318], Chapter 5 demonstrated the oft-dismissed value of qualitative analysis in human-robot interaction research. This thesis laid quantitative foundations for the topics it explores, but future work may investigate complementary qualitative methods.

Another limitation of this research, one which still applies to many studies in the field of human-robot interaction, is the controlled nature of the experiments. The findings of this thesis are assumed to be applicable to real-world interaction, however field studies that rely on more qualitative analysis can sometimes yield very different results, particulary with the unpredictability of the social domain [327]. This research would benefit from analogous experiments in real-world contexts to explore the applicability of the findings.

A different approach to methodology that might be explored is to rely further on interaction design concepts, drawing from broader human-machine literature. The study of partner models (how a dialogue partner is mentally modeled by an interactant) are one example of this [163]; a partner model is built up in an interactant's mind over longer back-and-forth exchanges of dialogue, so this construct might more accurately represent robot perception in real-world scenarios [162], as well as a method to factor in cultural bias.

The visuospatial modality of social interaction is difficult to evaluate in a system-agnostic method, which in part motivates the focus on affect—aside from its importance in social interaction, the visuospatial modality is closely tied to it. Another high-level social cue that may be of value to evaluate is that of attention/focus. Studies on joint attention or conveying attention would further allow robot teleoperation systems to be improved, with a similar level of statistical rigor.

The methods developed and demonstrated in this thesis aimed to be accessible to the majority of research groups, however additional equipment for participant testing may yield further insights. In particular, non-invasive physiological measurement methods are increas-

ingly available to researchers, such as pupillometry and skin conductivity measurements, and future work can leverage the objectivity of their measurement to yield more robust results. As such tools often measure emotional arousal, they may be particularly relevant for studies on robot affect evaluation, or for assessing perceived uncanniness in general.

A more distant goal for potential future work was mentioned in Chapter 1; enabling robot teleoperation systems will allow for collection of valuable operator data, which can be used to inform the behaviour of artificial agents.

Finally, the research presented in this thesis was centred around design for human-robot interaction. While further investigation and broader researcher skillsets would be required, future work in a collaborative multidisciplinary environment may adapt some of the methods developed here for use in non-design HRI studies, or even human-human studies, in particular those on affect.

Appendix A

Geneva Emotion Wheel data visualisations

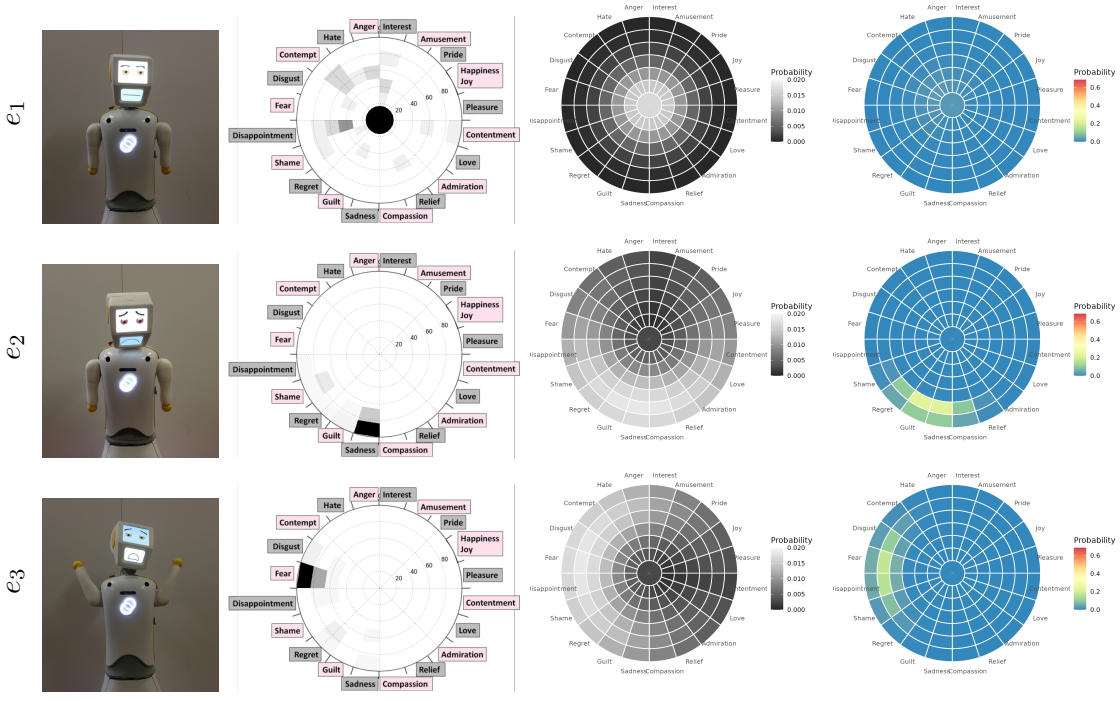


Figure A.1: Visualisations for all evaluated affect expressions e_1-e_9 (continued overleaf)

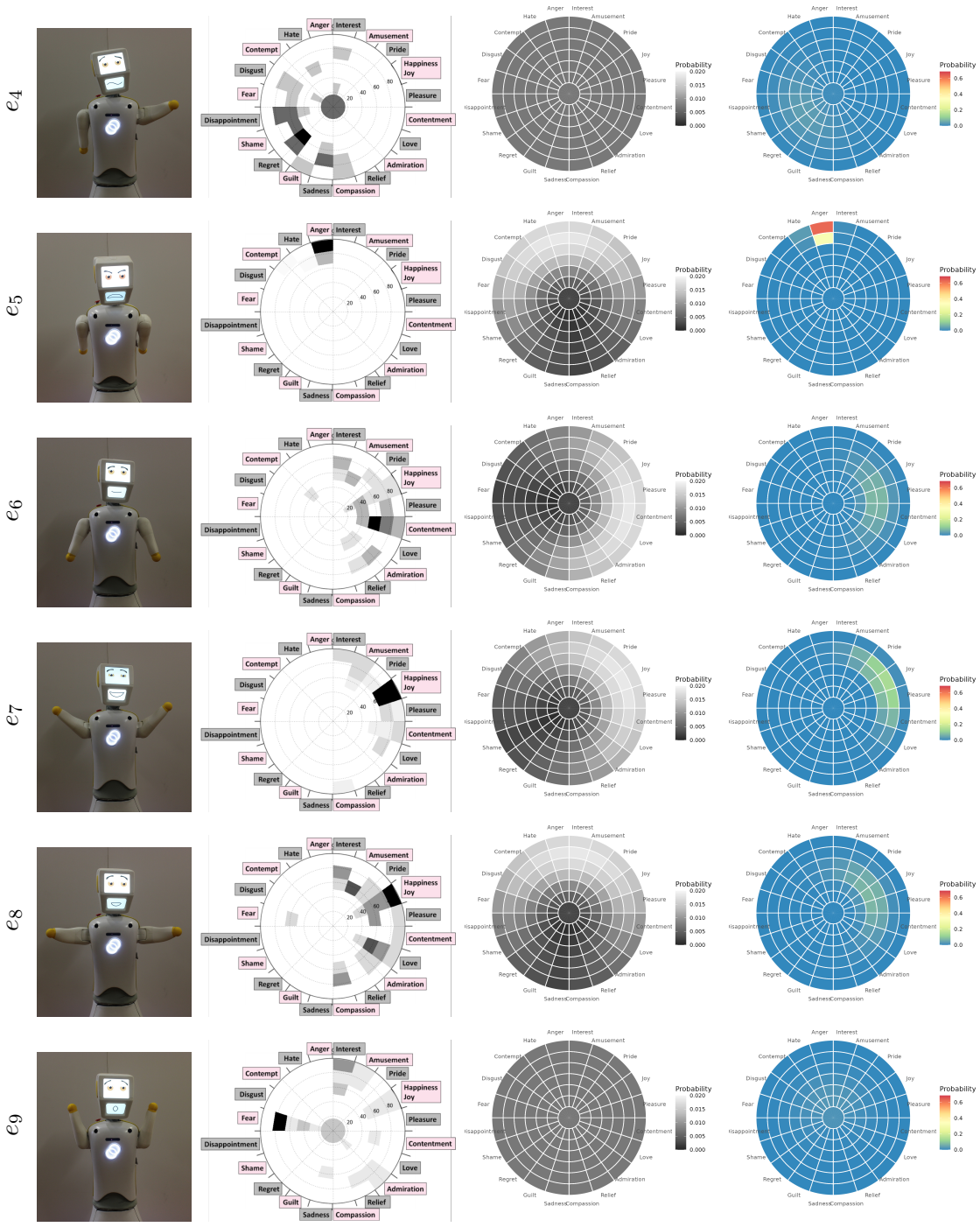


Figure A.1: (cont.)

Appendix B

Situation awareness testing data

Quiz questions and answers

Table B.1 is the complete set of quiz questions asked by the quiz master in simulation for each of the two videos used, along with the two answers spoken simultaneously; that of the contestant targeted by the question, and of the "distractor" (see 5.3).

Open participant feedback

The feedback in the following tables was provided via post-experiment questionnaire; positive feedback in Table B.2, negative in Table B.3. This feedback was used for quantitative analysis in 5.4.

Video 1

Quiz question	Target	Target's answer	Distractor	Distractor answer
Which European city hosted the 1936 Summer Olympics?	Annie	London	**Maria**	**Berlin**
How far is the target in Olympic Archery?	John	70 metres	**Annie**	**80 metres**
When is World Literacy Day celebrated?	John	October 7	**Susan**	**September 8**
What was invented by James Dewer in 1872?	John	Steam pump	**Susan**	**Thermos flask**
Where is the largest bowling alley?	**Elizabeth**	**Japan**	Maria	Minnesota
Who was the first Tsar of Russia?	**Susan**	**Ivan the Terrible**	John	Catherine the Great
Where is the oldest tree in the world?	Elizabeth	Colorado	**Maria**	**California**
What is the tallest mountain in Canada?	Annie	Mount Columbia	**John**	**Mount Logan**
Who wrote Flowers for Algernon?	**Elizabeth**	**Daniel Keyes**	Maria	Isaac Asimov
What colour is the Mallow flower?	Susan	Pink	**John**	**Purple**

Video 2

Quiz question	Target	Target's answer	Distractor	Distractor answer
What country was previously called Formosa?	**Amy**	**Taiwan**	Paul	Indonesia
How many chukkers are there in a polo match?	Amy	4	**Michael**	**6**
Dendrophobia is the fear of what?	**Carol**	**Trees**	Michael	Teeth
What was the first movie to be rated PG-13?	Michael	Gone with the Wind	**Carol**	**Red Dawn**
What is the surface area of the Earth?	**Paul**	**510 million km**	Emma	197 million km
When was the first digital computer invented?	Amy	1960	**Michael**	**1946**
What species of land animal has the largest eyes?	Paul	Elephant	**Emma**	**Ostrich**
Who was the first performer at Woodstock?	Emma	Fleetwood Mac	**Paul**	**Richie Havens**
Where are the most Trappist breweries?	Michael	Denmark	**Carol**	**Belgium**
How many countries are in the Southern Hemisphere?	Emma	35	**Paul**	**32**

Table B.1: Quiz questions asked during the experiment for videos 1 and 2. Correct contestant and answer for each question shown in **bold**.
Video 1 quiz master: "Carlos", Contestants: "John", "Elizabeth", "Annie", "Susan", "Maria"
Video 2 quiz master: "Megan", Contestants: "Michael", "Emma", "Amy", "Carol", "Paul"

Feedback item	*Category*
Mono group	
The ability to replicate a real world environment by aiding the auditory senses with visual to create a more complete picture	Immersion / Ability to look around
Fun to try and work your way around the posed challenge in creative ways	Challenging/Gamelike
Questions and answers very easy to follow, visuals clear, instructions and questions easily understood - no confusion throughout the experiment as to what I had to do	Other
Sound very clear, felt that I had good awareness of what was happening	Other
Being able to see whos talking	Immersion / Ability to look around
the instructions were very clear	Other
I liked being able to look around the room.	Immersion / Ability to look around
a bit like a game for me	Challenging/Gamelike
it was cool being able to move it around like you are in the room	Immersion / Ability to look around
being able to look around	Immersion / Ability to look around
I liked being able to look around at who was talking, having control over that felt immersive.	Immersion / Ability to look around
ability to see who is speaking	Immersion / Ability to look around
wanted to concentrate more to get answers right	Challenging/Gamelike
I learnt things!	Novel
Spatial group	
A new concept I've never experienced	Novel
The directional audio	Spatial audio feedback
The controls, being able to witness the surroundings	Immersion / Ability to look around
I really wanted to try to get the answer correct and that made me look around the room a lot.	Challenging/Gamelike
I liked the way you could move around, although I don't think it was necessary to answer the questions.	Immersion / Ability to look around
Was interesting that it was possible to tell from which direction the sound was coming.	Spatial audio feedback
In was an interesting idea	Novel
the audio came from different directions	Spatial audio feedback
The quiz questions	Challenging/Gamelike
The clarity of direction where the sound come from	Spatial audio feedback
real time 360 field of view, very useful for quiz/ conference viewing	Immersion / Ability to look around
the way I could identify where the voice was coming from even though the camera was pointing somewhere else	Spatial audio feedback
easy to navigate	Other
immersion	Immersion / Ability to look around
Questions were unpredictable which made this more engaging	Challenging/Gamelike
Spatial audio let me locate the speaker much easier.	Spatial audio feedback
Getting to see the whole room and everyone involved	Immersion / Ability to look around
Got more used to looking around the room as the experiment continued	Immersion / Ability to look around

Table B.2: Positive feedback provided in answer to the question *"Is there anything in particular that you liked about the experience?"*

Feedback item	*Category*
Mono group	
Not having a wider field of view - not being able to see more than one person at a time	Field of view too small
Not having a 360 degree view and taking the time to shift between people.	Field of view too small
Was a bit strange being in the centre of a group, almost like I was being looked at from every direction	Other
difficult to differentiate the girl voices, I would need to see who speaks	Audio was poor (increasing difficulty)
sound quality!	Audio was poor (increasing difficulty)
sound had a bit of an echo to it	Audio was poor (increasing difficulty)
Sometimes I tried to lipread the participant that was asked to answer and the video lagged slightly making that difficult.	Audio was poor (increasing difficulty)
using the arrows to move left and right. in a real environment I would just position my body, move my head or just my eyes	Other
when answered together, I could only hear one answer rather than both	Audio was poor (increasing difficulty)
frustrating when cant make out answers as someones talking over	Audio was poor (increasing difficulty)
Found it rather annoying cause I like to see who is answering while they answer, but it takes a while to shift focus to a person and figure out who answered till the teachers says it out loud.	Slow camera panning speed
Spatial group	
slow panning speed of the visuals	Slow camera panning speed
The wait time between the questions was quite long. If the video paused until the questions was answered, and then resumed straight away. I think it would've been slightly better.	It was repetitive / boring
No vertical control	No vertical camera control
you cant see up and down	No vertical camera control
the overlay of two answers required a lot of concentration and often the choices were not clear until the graphic with the answer selection appeared on screen.	Audio was poor (increasing difficulty)
the talking over each other so it was hard to concentrate on correct answers	Audio was poor (increasing difficulty)
wish the arrows would move quicker	Slow camera panning speed
It was somewhat repetitive	It was repetitive / boring
How I experienced volume differed to how I experience volume normally. I expected the people behind me to be quieter.	Other

Table B.3: Negative feedback provided in answer to the question *"Is there anything in particular that you disliked about the experience?"*

Bibliography

[1] M. R. Endsley, "Measurement of Situation Awareness in Dynamic Systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, no. 1, pp. 65–84, 1995. ISBN: 0018-7208.

[2] D. Zunt, "Who did actually invent the word "robot" and what does it mean," 2002.

[3] M. Scopelliti, M. V. Giuliani, A. M. D' Amico, and F. Fornara, "If I had a Robot at Home... Peoples' Representation of Domestic Robots," in *Designing a More Inclusive World* (S. Keates, J. Clarkson, P. Langdon, and P. Robinson, eds.), (London), pp. 257–266, Springer, 2004.

[4] C. Ray, F. Mondada, and R. Siegwart, "What do people expect from robots?," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3816–3821, Sept. 2008. ISSN: 2153-0866.

[5] B. Gates, "A robot in every home," *Scientific American*, vol. 296, no. 1, pp. 58–65, 2007. arXiv: http://www.jstor.org/stable/26016002 ISBN: 1111111111.

[6] E. Ackerman, "Telepresence Robots Are Helping Take Pressure Off Hospital Staff," *IEEE Spectrum*, Apr. 2020.

[7] A. Dickson, "You Can't Visit the Museum. But Your Robot Can," 2020.

[8] C. Purtill, "Meet Violet, the Robot That Can Kill the COVID-19 Virus," Apr. 2020.

[9] S. Rao, "Covid-19: Short of doctors, Karnataka set to deploy robots at care units," *Times of India*, July 2020.

[10] C. Müller, *World Robotics 2022 – Industrial Robots*. Frankfurt am Main, Germany: IFR Statistical Department, VDMA Services GmbH, 2022.

[11] M. Greshko and G. Di Sturco, "Meet Sophia, the Robot That Looks Almost Human," May 2018. Section: Photography.

[12] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, and K. Fujimura, "The intelligent ASIMO: system overview and integration," in *IEEE/RSJ International Conference on Intelligent Robots and System*, vol. 3, pp. 2478–2483, IEEE, 2002.

[13] E. Guizzo, "Meet Pepper, Aldebaran' s new personal robot with an "Emotion Engine" ," *IEEE Spectrum*, vol. 5, 2014.

[14] C. McGinn, E. Bourke, A. Murtagh, C. Donovan, P. Lynch, M. F. Cullinan, and K. Kelly, "Meet Stevie: a Socially Assistive Robot Developed Through Application of a 'Design-Thinking' Approach," *Journal of Intelligent and Robotic Systems: Theory and Applications*, pp. 1–20, July 2019. Publisher: Springer Netherlands.

[15] OpenAI, "ChatGPT —Release Notes."

[16] I. Bogost, "ChatGPT is dumber than you think," Dec. 2022.

[17] M. Mori, K. F. MacDorman, and N. Kageki, "The Uncanny Valley [From the Field]," *IEEE Robotics and Automation Magazine*, vol. 19, pp. 98–100, June 2012. Publisher: IEEE ISBN: 1070-9932 VO - 19.

[18] C. McGinn, "Why Do Robots Need a Head? The Role of Social Interfaces on Service Robots," *International Journal of Social Robotics*, pp. 1–15, June 2019. Publisher: Springer Netherlands.

[19] A. Reuten, M. Van Dam, and M. Naber, "Pupillary responses to robotic and human emotions: the uncanny valley and media equation confirmed," *Frontiers in psychology*, vol. 9, p. 774, 2018. Publisher: Frontiers Media SA.

[20] World Health Organisation and others, "Health workforce requirements for universal health coverage and the Sustainable Development Goals," *Human Resource for Health Observer*, no. 17, 2016. ISBN: 978 92 4 151140 7.

[21] O. L. Haimson, A. E. Bowser, E. F. Melcer, and E. F. Churchill, "Online Inspiration and Exploration for Identity Reinvention," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, (Seoul Republic of Korea), pp. 3809–3818, ACM, Apr. 2015.

[22] B. Irfan, J. Kennedy, S. Lemaignan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Social Psychology and Human-Robot Interaction: An Uneasy Marriage," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, (Chicago IL USA), pp. 13–20, ACM, Mar. 2018.

[23] A. Henschel, G. Laban, and E. S. Cross, "What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You," *Current Robotics Reports*, vol. 2, pp. 9–19, Mar. 2021.

[24] A. Robaczewski, J. Bouchard, K. Bouchard, and S. Gaboury, "Socially Assistive Robots: The Specific Case of the NAO," *International Journal of Social Robotics*, vol. 13, pp. 795–831, July 2021.

[25] N. J. Enfield, "The body as a cognitive artifact in kinship representations: Hand gesture diagrams by speakers of Lao," *Current anthropology*, vol. 46, no. 1, pp. 51–81, 2005. Publisher: The University of Chicago Press.

[26] A. Damásio, *Descartes' error: emotion, reason, and the human brain*. 1994. Publisher: Putnam.

[27] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of nonverbal behavior*, vol. 28, no. 2, pp. 117–139, 2004. Publisher: Springer.

[28] T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, 1993. ISSN: 02780097.

[29] T. Fong and C. Thorpe, "Vehicle teleoperation interfaces," *Autonomous Robots*, vol. 11, no. 1, pp. 9–18, 2001. Publisher: Kluwer Academic Publishers ISBN: 09295593 (ISSN).

[30] T. B. Sheridan and W. Verplank, "Human and Computer Control of Undersea Teleoperators (Man-Machine Systems Lab Report)," tech. rep., Massachusetts Institute of Technology, Cambridge, Massachusetts, 1978.

[31] R. Parasuraman, T. Sheridan, and C. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, pp. 286–297, May 2000. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans.

[32] M. Hoffman and R. R. Johnson, "The Quest for Alternatives to "Levels of Automation" and "Task Allocation"," in *Human Performance in Automated and Autonomous Systems*, CRC Press, 2019. Num Pages: 26.

[33] B. Shneiderman, "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy," *International Journal of Human-Computer Interaction*, vol. 36, pp. 495–504, Apr. 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2020.1741118.

[34] T. B. Sheridan, "Function allocation: algorithm, alchemy or apostasy?," *International Journal of Human-Computer Studies*, vol. 52, pp. 203–216, Feb. 2000.

[35] J. Blackhurst, J. Gresham, and M. O. Stone, "The autonomy paradox," *Armed Forces Journal*, 2011.

[36] B. Strauch, "Ironies of Automation: Still Unresolved After All These Years," *IEEE Transactions on Human-Machine Systems*, vol. 48, pp. 419–433, Oct. 2018. Conference Name: IEEE Transactions on Human-Machine Systems.

[37] F. Amirabdollahian, R. Op Den Akker, S. Bedaf, R. Bormann, H. Draper, V. Evers, G. J. Gelderblom, C. G. Ruiz, D. Hewson, N. Hu, I. Iacono, K. L. Koay, B. Krose, P. Marti, H. Michel, H. Prevot-Huille, U. Reiser, J. Saunders, T. Sorell, and K. Dautenhahn, "Accompany: Acceptable robotiCs COMPanions for AgeiNG Years - Multidimensional aspects of human-system interactions," *2013 6th International Conference on Human System Interactions, HSI 2013*, pp. 570–577, 2013. Publisher: IEEE ISBN: 9781467356374.

[38] H. Kuzuoka, K. Yamazaki, A. Yamazaki, J. Kosaka, Y. Suga, and C. Heath, "Dual Ecologies of Robot as Communication Media: Thoughts on Coordinating Orientations and Projectability," in *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, (New York, New York, USA), pp. 183–190, ACM Press, 2004.

[39] J. Koenemann and M. Bennewitz, "Whole-body imitation of human motions with a nao humanoid," *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12*, p. 425, 2012. ISBN: 9781450310635.

[40] C. W. Nielsen, M. A. Goodrich, and R. W. Ricks, "Ecological interfaces for improving mobile robot teleoperation," *IEEE Transactions on Robotics*, vol. 23, pp. 927–941, Oct. 2007. ISBN: 9783642212673.

[41] M. K. Lee and L. Takayama, ""Now, i have a body": uses and social norms for mobile remote presence in the workplace," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, (New York, NY, USA), pp. 33–42, Association for Computing Machinery, May 2011.

[42] D. Sirkin and W. Ju, "Consistency in Physical and On-Screen Action Improves Perceptions of Telepresence Robots," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, (New York, NY, USA), pp. 57–64, Association for Computing Machinery, 2012. event-place: Boston, Massachusetts, USA.

[43] K. Kuwamura, T. Minato, S. Nishio, and H. Ishiguro, “Personality distortion in communication through teleoperated robots,” in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pp. 49–54, Sept. 2012. ISSN: 1944-9437.

[44] J. J. Choi and S. S. Kwak, “Can you feel me?: How embodiment levels of telepresence systems affect presence,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 606–611, Aug. 2016. ISSN: 1944-9437.

[45] M. A. Goodrich, J. W. Crandall, and E. Barakova, “Teleoperation and Beyond for Assistive Humanoid Robots,” *Reviews of Human Factors and Ergonomics*, vol. 9, no. 1, pp. 175–226, 2013. ISBN: 1557-234X.

[46] C. Schnier, K. Pitsch, A. Dierker, and T. Hermann, “Collaboration in Augmented Reality: How to establish coordination and joint attention?,” in *ECSCW 2011: Proceedings of the 12th European Conference on Computer Supported Cooperative Work, 24-28 September 2011, Aarhus Denmark* (S. Bødker, N. O. Bouvin, V. Wulf, L. Ciolfi, and W. Lutters, eds.), (London), pp. 405–416, Springer London, 2011.

[47] M. Prilla, “"I simply watched where she was looking at": Coordination in Short-term Synchronous Cooperative Mixed Reality,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, pp. 1–21, Dec. 2019.

[48] M. A. Goodrich and A. C. Schultz, “Human-Robot Interaction: A Survey,” *Foundations and Trendső in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007. arXiv: 1011.1669v3 ISBN: 9781601980922.

[49] M. Minsky, “Telepresence,” *OMNI magazine*, 1980.

[50] J. V. Draper, D. B. Kaber, and J. M. Usher, “Telepresence,” *Human Factors*, vol. 40, pp. 354–375, Sept. 1998. Publisher: SAGE Publications Inc.

[51] B. G. Witmer and M. J. Singer, "Measuring Presence in Virtual Environments: A Presence Questionnaire," *Presence: Teleoperators and Virtual Environments*, vol. 7, pp. 225–240, June 1998.

[52] M. Lombard, T. B. Ditton, and L. Weinstein, "Measuring Presence: The Temple Presence Inventory," in *Proceedings of the 12th Annual International Workshop on Presence*, pp. 1–15, 2009.

[53] D. Hein, C. Mai, and H. HuSSmann, "The usage of presence measurements in research: a review," in *Proceedings of the International Society for Presence Research Annual Conference (Presence)*, pp. 21–22, The International Society for Presence Research Prague, 2018.

[54] S. Tachi, K. Tanie, and K. Komoriya, "Evaluation Apparatus of Mobility Aids for the Blind," 1980. Patent.

[55] L. Almeida, B. Patrao, P. Menezes, and J. Dias, "Be the robot: Human embodiment in tele-operation driving tasks," *IEEE RO-MAN 2014 - 23rd IEEE International Symposium on Robot and Human Interactive Communication: Human-Robot Co-Existence: Adaptive Interfaces and Systems for Daily Life, Therapy, Assistance and Socially Engaging Interactions*, pp. 477–482, 2014. ISBN: 9781479967636.

[56] H. Hoshino, K. Suzuki, T. Nishiyama, and K. Sawada, "Development of Tele-existence Visual Display System for Humanoid Robot Control," *International Conference on Artificial Reality and Telexistence (ICAT)*, 1999.

[57] S. Tachi, K. Minamizawa, M. Furukawa, and C. L. Fernando, "Telexistence - From 1980 to 2012," in *IEEE International Conference on Intelligent Robots and Systems*, pp. 5440–5441, 2012. ISSN: 21530858.

[58] M. Y. Saraiji, C. L. Fernando, Y. Mizushina, Y. Kamiyama, K. Minamizawa, and S. Tachi, "Enforced telexistence: teleoperating using photorealistic virtual body and haptic feedback," in *SIGGRAPH Asia 2014 Emerging Technologies*, (Shenzen, China), pp. 1–2, ACM, 2014.

[59] K. Tanaka, H. Nakanishi, and H. Ishiguro, "Physical Embodiment Can Produce Robot Operator's Pseudo Presence," *Frontiers in ICT*, vol. 2, May 2015.

[60] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *International Journal of Human Computer Studies*, vol. 77, pp. 23–37, 2015. Publisher: Elsevier.

[61] M. R. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, pp. 32–64, Mar. 1995. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[62] R. H. Mogford, "Mental Models and Situation Awareness in Air Traffic Control," *The International Journal of Aviation Psychology*, vol. 7, pp. 331–341, Oct. 1997.

[63] H. A. Yanco and J. Drury, ""Where am I?" Acquiring situation awareness using a remote robot platform," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, vol. 3, pp. 2835–2840, 2004. ISSN: 1062922X.

[64] F. T. Girotto, K. A. Durso, and S. Rawson, "Comprehension and Situation Awareness," in *Situational Awareness*, Routledge, 2011. Num Pages: 32.

[65] J. M. Riley, D. B. Kaber, and J. V. Draper, "Situation awareness and attention allocation measures for quantifying telepresence experiences in teleoperation," *Human Factors and Ergonomics In Manufacturing*, vol. 14, no. 1, pp. 51–67, 2004. Publisher: John Wiley & Sons, Ltd.

[66] M. R. Endsley, "Situation awareness global assessment technique (SAGAT)," in *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, (Dayton, OH), pp. 789–795, IEEE, 1988.

[67] A. K. Gardner, M. Kosemund, and J. Martinez, "Examining the Feasibility and Predictive Validity of the SAGAT Tool to Assess Situation Awareness Among Medical Trainees," *Simulation in Healthcare*, vol. 12, pp. 17–21, Feb. 2017.

[68] D. G. Jones and M. R. Endsley, "Can real-time probes provide a valid measure of situation awareness?," in *Proceedings of the Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium*, (Savannah, Georgia, USA.), pp. 245–250, SA Technologies, 2000.

[69] F. T. Durso, T. R. Truitt, C. A. Hackworth, J. M. Crutchfield, D. Nikolic, P. M. Moertl, D. Ohrt, and C. A. Manning, "Expertise and Chess: A Pilot Study Comparing Situation Awareness Methodologies," in *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness 1995* (M. A. Wise and N. Rahn-Virden, eds.), (Daytona Beach, FL), pp. 295–303, Embry-Riddle Aeronautical University Press, 1995.

[70] S. Loft, D. B. Morrell, and S. Huf, "Using the situation present assessment method to measure situation awareness in simulated submarine track management," *International Journal of Human Factors and Ergonomics*, vol. 2, no. 1, p. 33, 2013.

[71] M. Perry, "Distributed Cognition," in *HCI Models, Theories, and Frameworks*, pp. 193–223, Elsevier, 2003.

[72] M. R. Endsley, "A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 63, pp. 124–150, Feb. 2021.

[73] J. Z. Bakdash, L. R. Marusich, J. B. Kenworthy, E. Twedt, and E. G. Zaroukian, "Statistical Significance Filtering Overestimates Effects and Impedes Falsification: A Critique of Endsley (2019)," *Frontiers in Psychology*, vol. 11, 2020.

[74] R. M. Taylor, "Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design," in *Proceeding of the AGARD AMP Symposium on Situational Awareness in Aerospace Operations, CP478*, (Neuilly-sur-Seine), 1989. Publisher: NATO AGARD.

[75] M. Zimmerman, S. Bagchi, J. Marvel, and V. Nguyen, "An Analysis of Metrics and Methods in Research from Human-Robot Interaction Conferences, 2015 - 2021," in

2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 644–648, Mar. 2022.

[76] M. R. Endsley, S. J. Selcon, T. D. Hardiman, and D. G. Croft, "Comparative analysis of SAGAT and SART for evaluations of situation awareness," *Proceedings of the Human Factors and Ergonomics Society*, vol. 1, pp. 82–86, 1998.

[77] A. Finn and S. Scheding, "Developments and Challenges for Autonomous Unmanned Vehicles: A Compendium," *Intelligent Systems Reference Library*, vol. 3, 2010. ISBN: 9783642107030.

[78] M. A. Diftler, C. J. Culbert, R. O. Ambrose, R. Platt Jr., and W. J. Bluethmann, "Evolution of the NASA/DARPA Robonaut control system," *Robotics and Automation, 2003. Proceedings. ICRA '03. IEEE International Conference on*, vol. 2, pp. 2543–2548 vol.2, 2003. Publisher: IEEE ISBN: 1050-4729 VO - 2.

[79] A. Kristoffersson, S. Coradeschi, and A. Loutfi, "A Review of Mobile Robotic Telepresence," *Advances in Human-Computer Interaction*, vol. 2013, p. 17, Apr. 2013. Publisher: Hindawi Publishing Corporation.

[80] P. Ballou, "Improving pilot dexterity with a telepresent ROV," in *Proceedings of the Vehicle Teleoperation Interfaces Workshop, IEEE ICRA*, 2001.

[81] R. B. Wynn, V. A. Huvenne, T. P. Le Bas, B. J. Murton, D. P. Connelly, B. J. Bett, H. A. Ruhl, K. J. Morris, J. Peakall, D. R. Parsons, E. J. Sumner, S. E. Darby, R. M. Dorrell, and J. E. Hunt, "Autonomous Underwater Vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience," *Marine Geology*, vol. 352, pp. 451–468, June 2014. Publisher: Elsevier.

[82] M. Hiltz, C. Rice, K. Boyle, and R. Allison, "Canadarm: 20 years of mission success through adaptation," *International Symposium on Artificial Intelligence, Robotics and Automation in Space*, vol. 1, pp. 1–8, 2001.

[83] N. A. Radford, P. Strawser, K. Hambuchen, J. S. Mehling, W. K. Verdeyen, A. S. Donnan, J. Holley, J. Sanchez, V. Nguyen, L. Bridgwater, R. Berka, R. Ambrose, M. Myles Markee, N. J. Fraser-Chanpong, C. McQuin, J. D. Yamokoski, S. Hart, R. Guo, A. Parsons, B. Wightman, P. Dinh, B. Ames, C. Blakely, C. Edmondson, B. Sommers, R. Rea, C. Tobler, H. Bibby, B. Howard, L. Niu, A. Lee, M. Conover, L. Truong, R. Reed, D. Chesney, R. Platt, G. Johnson, C. L. Fok, N. Paine, L. Sentis, E. Cousineau, R. Sinnet, J. Lack, M. Powell, B. Morris, A. Ames, and J. Akinyode, "Valkyrie: NASA's First Bipedal Humanoid Robot," *Journal of Field Robotics*, vol. 32, no. 3, pp. 397–419, 2015.

[84] M. Duke, S. Hoffman, and K. Snook, "Lunar Surface Reference Missions: A Description of Human and Robotic Surface Activities," *NASA Technical Report (TP- ⋯*, no. July, 2003.

[85] D. M. Hassler, C. Zeitlin, R. F. Wimmer-Schweingruber, B. Ehresmann, S. Rafkin, J. L. Eigenbrode, D. E. Brinza, G. Weigle, S. Böttcher, E. Böhm, S. Burmeister, J. Guo, J. Köhler, C. Martin, G. Reitz, F. A. Cucinotta, M. H. Kim, D. Grinspoon, M. A. Bullock, A. Posner, J. Gómez-Elvira, A. Vasavada, and J. P. Grotzinger, "Mars' surface radiation environment measured with the Mars science laboratory's curiosity rover," *Science*, vol. 343, no. 6169, 2014.

[86] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Advances in Psychology*, vol. 52, pp. 139–183, Jan. 1988. arXiv: 1011.1669v3 Publisher: North-Holland ISBN: 9780444703880.

[87] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," *Proceedings of the Human Factors and Ergonomics Society*, pp. 904–908, 2006. ISBN: 9780945289296.

[88] M. J. Boyle, "The legal and ethical implications of drone warfare," *International Journal of Human Rights*, vol. 19, no. 2, pp. 105–126, 2015. Publisher: Taylor & Francis.

[89] S. D. Keene, "Lethal and legal? The ethics of drone strikes," *Drone Warfare: Ethical Explorations*, pp. 1–36, 2016. ISBN: 9781634851046.

[90] J. Blitch, "Semi-autonomous tactical robots for urban operations," *IEEE International Symposium on Intelligent Control - Proceedings*, pp. 783–794, 1998. ISBN: 0780344235.

[91] J. C. Ralston and D. W. Hainsworth, "The Numbat: A Remotely Controlled Mine Emergency Response Vehicle," in *Field and Service Robotics*, pp. 53–59, London: Springer London, 1998.

[92] R. T. Laird, M. H. Bruch, M. B. West, D. A. Ciccimaro, and H. R. Everett, "Issues in Vehicle Teleoperation for Tunnel and Sewer Reconnaissance," *Quantum*, pp. 1–6, Apr. 2000. Publisher: US Dept of the Navy.

[93] B. M. Yamauchi, "PackBot: a versatile platform for military robotics," *Unmanned Ground Vehicle Technology VI*, vol. 5422, pp. 228–237, 2004.

[94] J. L. Burke, *Moonlight in Miami: Field Study of Human-Robot Interaction in the Context of an Urban Search and Rescue Disaster Response Training Exercise*. PhD thesis, University of South Florida, 2003. Publication Title: Human-Computer Interaction Issue: 1 ISBN: 6600103892 ISSN: 0737-0024.

[95] M. J. Micire, "Evolution and field performance of a rescue robot," *Journal of Field Robotics*, vol. 25, pp. 17–30, Jan. 2008.

[96] A. Davids, "Urban Search and Rescue Robots: From Tragedy to Technology," *IEEE Intelligent Systems*, vol. 17, no. 2, pp. 81–83, 2002.

[97] J. Casper and R. R. Murphy, "Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 33, no. 3, pp. 367–385, 2003.

[98] J. Yerushalmy, "Robot subs and an 8km winch: the equipment being deployed in search for Titan," *The Guardian*, June 2023.

[99] T. Lane, "A short history of robotic surgery," *Annals of the Royal College of Surgeons of England*, vol. 100, pp. 5–7, May 2018. Publisher: Royal College of Surgeons of England.

[100] A. Hemal and M. Menon, "Laparoscopy, robot, telesurgery and urology: future perspective," *Journal of Postgraduate Medicine*, vol. 48, no. 1, p. 39, 2002. Publisher: Medknow Publications.

[101] J. Heemskerk, N. D. Bouvy, and C. G. Baeten, "The end of robot-assisted laparoscopy? A critical appraisal of scientific evidence on the use of robot-assisted laparoscopic surgery," *Surgical Endoscopy*, vol. 28, pp. 1388–1398, Nov. 2014. Publisher: Springer New York LLC.

[102] A. M. Okamura, "Methods for haptic feedback in teleoperated robot-assisted surgery," *Industrial Robot: An International Journal*, vol. 31, no. 6, pp. 499–508, 2004. arXiv: NIHMS150003 ISBN: 6176321972.

[103] E. Paulos and J. Canny, "PRoP: Personal Roving Presence," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 296–303, 1998.

[104] E. Paulos and J. Canny, "Social tele-embodiment: Understanding presence," *Autonomous Robots*, vol. 11, no. 1, pp. 87–95, 2001. Publisher: Kluwer Academic Publishers ISBN: 0929-5593.

[105] M. Desai, K. M. Tsui, H. A. Yanco, and C. Uhlik, "Essential features of telepresence robots," *2011 IEEE Conference on Technologies for Practical Robot Applications, TePRA 2011*, pp. 15–20, 2011. ISBN: 9781612844824.

[106] A. Reis, R. Xavier, I. Barroso, M. J. Monteiro, H. Paredes, and J. Barroso, "The usage of telepresence robots to support the elderly," *TISHW 2018 - 2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing, Proceedings*, 2018. ISBN: 9781538684559.

[107] C. Neustaedter, G. Venolia, J. Procyk, and D. Hawkins, "To Beam or not to Beam: A study of remote telepresence attendance at an academic conference," in *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*, (San Francisco, USA), pp. 418–431, 2016.

[108] J.-J. Cabibihan, H. Javed, M. Ang, and S. M. Aljunied, "Why Robots? A Survey on the Roles and Benefits of Social Robots in the Therapy of Children with Autism," *International Journal of Social Robotics*, vol. 5, pp. 593–618, Nov. 2013.

[109] B. Scassellati, H. Admoni, and M. M. Matarić, "Robots for Use in Autism Research," 2012.

[110] C. L. van Straten, I. Smeekens, E. Barakova, J. Glennon, J. Buitelaar, and A. Chen, "Effects of robots' intonation and bodily appearance on robot-mediated communicative treatment outcomes for children with autism spectrum disorder," *Personal and Ubiquitous Computing*, vol. 22, pp. 379–390, Apr. 2018.

[111] S. Elbeleidy, D. Rosen, D. Liu, A. Shick, and T. Williams, "Analyzing Teleoperation Interface Usage of Robots in Therapy for Children with Autism," in *Interaction Design and Children*, (Athens Greece), pp. 112–118, ACM, June 2021.

[112] S. Elbeleidy, A. Jadhav, D. Liu, and T. Williams, "Robot Teleoperation Interfaces for Customized Therapy for Autistic Children," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 750–753, 2022.

[113] M. Niemelä, P. Heikkilä, H. Lammi, and V. Oksman, "Shopping Mall Robots -Opportunities and Constraints from the Retailer and Manager Perspective," in *Social Robotics* (A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He, eds.), Lecture Notes in Computer Science, (Cham), pp. 485–494, Springer International Publishing, 2017.

[114] A. Tuomi, I. P. Tussyadiah, and P. Hanna, "Spicing up hospitality service encounters: the case of Pepper," *International Journal of Contemporary Hospitality Management*, vol. 33, pp. 3906–3925, Jan. 2021. Publisher: Emerald Publishing Limited.

[115] N. E. John, A. Rossi, and S. Rossi, "Personalized Human-Robot Interaction with a Robot Bartender," in *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22 Adjunct, (New York, NY, USA), pp. 155–159, Association for Computing Machinery, July 2022.

[116] J. Sung, H. I. Christensen, and R. E. Grinter, "Robots in the wild: understanding long-term use," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, HRI '09, (New York, NY, USA), pp. 45–52, Association for Computing Machinery, Mar. 2009.

[117] T. Stivers and J. Sidnell, "Introduction: Multimodal interaction," vol. 2005, pp. 1–20, Dec. 2005. Publisher: De Gruyter Mouton Section: Semiotica.

[118] B. Parkinson, "Emotions are social," *British Journal of Psychology*, vol. 87, no. 4, pp. 663–683, 1996. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8295.1996.tb02615.x.

[119] G. A. van Kleef, A. Cheshin, A. H. Fischer, and I. K. Schneider, "Editorial: The Social Nature of Emotions," *Frontiers in Psychology*, vol. 7, 2016.

[120] B. De Gelder, "Why bodies? Twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3475–3484, 2009. Publisher: The Royal Society.

[121] P. Warren, "Prosody and Parsing: An Introduction," *Language and Cognitive Processes*, vol. 11, pp. 1–16, Apr. 1996. Publisher: Informa UK Limited.

[122] T. W. Fong, F. Conti, S. Grange, and C. Baur, "Novel interfaces for remote driving: gesture, haptic, and PDA," in *Mobile Robots XV and Telemanipulator and Telepresence Technologies VII*, vol. 4195, pp. 300–311, Mar. 2001. ISSN: 0277786X.

[123] J. Drury, J. Scholtz, and H. Yanco, "Awareness in human-robot interactions," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483)*, vol. 1, pp. 912–918, 2003. ISSN: 1062-922X.

[124] A. Y. Elkady, V. Babariya, J. Joy, and T. Sobh, "Modular design and implementation for a sensory-driven mobile manipulation framework," *Journal of Intelligent and*

Robotic Systems: Theory and Applications, vol. 62, pp. 355–381, June 2011. Publisher: Springer Netherlands.

[125] H. Martins, I. Oakley, and R. Ventura, "Design and evaluation of a head-mounted display for immersive 3D teleoperation of field robots," *Robotica*, vol. 33, pp. 2166–2185, Dec. 2015. ISBN: 02635747.

[126] C. Pittman and J. J. LaViola, "Exploring head tracked head mounted displays for first person robot teleoperation," in *Proceedings of the 19th international conference on Intelligent User Interfaces - IUI '14*, (New York, New York, USA), pp. 323–328, ACM Press, 2014.

[127] G. Adamides, G. Christou, C. Katsanos, M. Xenos, and T. Hadzilacos, "Usability guidelines for the design of robot teleoperation: A taxonomy," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 2, pp. 256–262, 2015.

[128] C. McGinn, A. Sena, and K. Kelly, "Controlling robots in the home: Factors that affect the performance of novice robot operators," *Applied Ergonomics*, vol. 65, pp. 23–32, Nov. 2017.

[129] K. Yamazaki, A. Yamazaki, H. Kuzuoka, S. Oyama, H. Kato, H. Suzuki, and H. Miki, "GestureLaser and GestureLaser Car: Development of an Embodied Space to Support Remote Instruction," *Ecscw ' 99*, no. September, pp. 239–258, 1999. ISBN: 978-94-011-4441-4.

[130] A. Mora, D. F. Glas, T. Kanda, and N. Hagita, "A teleoperation approach for mobile social robots incorporating automatic gaze control and three-dimensional spatial visualization," *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, vol. 43, pp. 630–642, May 2013.

[131] N. Diolaiti and C. Melchiorri, "Teleoperation of a mobile robot through haptic feedback," *Proceedings: HAVE 2002 - IEEE International Workshop on Haptic Virtual Environments and their Applications*, pp. 67–72, 2002. Publisher: IEEE ISBN: 0-7803-7635-8.

[132] N. C. Mitsou, S. V. Velanas, and C. S. Tzafestas, "Visuo-haptic interface for teleoperation of mobile robot exploration tasks," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp. 157–163, IEEE, Sept. 2006. ISSN: 1-4244-0565-3.

[133] S. Hughes, J. Manojlovich, M. Lewis, and J. Gennari, "Camera control and decoupled motion for teleoperation," *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics\cite*, vol. 2, pp. 1339–1344, 2003. Publisher: IEEE ISBN: 0-7803-7952-7.

[134] A. Sanna, F. Lamberti, G. Paravati, and F. Manuri, "A Kinect-based natural interface for quadrotor control," *Entertainment Computing*, vol. 4, pp. 179–186, Aug. 2013. ISBN: 9783642302138.

[135] J. Sugiyama and J. Miura, "A wearable visuo-inertial interface for humanoid robot control," in *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 235–236, 2013. ISSN: 21672148.

[136] M. H. Draper, H. A. Ruff, J. V. Fontejon, and S. Napier, "The Effects of Head-Coupled Control and a Head-Mounted Display (HMD) on Large-Area Search Tasks," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 26, pp. 2139–2143, 2002.

[137] D. Kent, C. Saldanha, and S. Chernova, "A Comparison of Remote Robot Teleoperation Interfaces for General Object Manipulation," *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pp. 371–379, 2017. ISBN: 9781450343367.

[138] A. Singh, S. H. Seo, Y. Hashish, M. Nakane, J. E. Young, and A. Bunt, "An interface for remote robotic manipulator control that reduces task load and fatigue," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp. 738–743, 2013. ISSN: 1944-9445.

[139] G. Du and P. Zhang, "Markerless human-robot interface for dual robot manipulators using Kinect sensor," *Robotics and Computer-Integrated Manufacturing*, vol. 30, no. 2, pp. 150–159, 2014. Publisher: Elsevier ISBN: 07365845.

[140] D. Kruse, J. T. Wen, and R. J. Radke, "A sensor-based dual-arm tele-robotic system," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 1, pp. 4–18, 2015. ISBN: 9781479915156.

[141] L. Fritsche, F. Unverzag, J. Peters, and R. Calandra, "First-person tele-operation of a humanoid robot," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, vol. 2015-Decem, (Seoul, Korea), pp. 997–1002, IEEE, 2015. ISSN: 21640580.

[142] T. Rodehutskors, M. Schwarz, and S. Behnke, "Intuitive bimanual telemanipulation under communication restrictions by immersive 3D visualization and motion tracking," in *IEEE-RAS International Conference on Humanoid Robots*, vol. 2015-Decem, pp. 276–283, 2015. ISSN: 21640580.

[143] S. Park, Y. Jung, and J. Bae, "A tele-operation interface with a motion capture system and a haptic glove," in *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence, URAI 2016*, pp. 544–549, 2016.

[144] N. J. Enfield, "On linear segmentation and combinatorics in co-speech gesture: A symmetry-dominance construction in Lao fish trap descriptions," *Semiotica*, vol. 2004, Jan. 2004.

[145] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: Guide to Algorithms and System Development*. Prentice Hall, 2001.

[146] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social Signal Processing: State-of-the-Art and Future Perspectives of an Emerging Domain," in *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, (New York, NY, USA), pp. 1061–1070, Association for Computing Machinery, 2008. event-place: Vancouver, British Columbia, Canada.

[147] D. Hirst and A. Di Cristo, *Intonation systems: a survey of twenty languages*. Cambridge University Press, 1998.

[148] K. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual Prosody and Speech Intelligibility," *Psychological Science*, vol. 15, pp. 133–137, Feb. 2004.

[149] D. Crystal, *Prosodic systems and intonation in English*, vol. 1. CUP Archive, 1969.

[150] G. Mohammadi and A. Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features," *IEEE Transactions on Affective Computing*, vol. 3, pp. 273–284, July 2012.

[151] J. Crumpton and C. L. Bethel, "Validation of vocal prosody modifications to communicate emotion in robot speech," in *2015 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 39–46, IEEE, June 2015.

[152] J. Lee, "Generating Robotic Speech Prosody for Human Robot Interaction: A Preliminary Study," *Applied Sciences*, vol. 11, p. 3468, Apr. 2021.

[153] P. J. Price, M. Ostendorf, S. ShattuckHufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *The Journal of the Acoustical Society of America*, vol. 90, pp. 2956–2970, Dec. 1991.

[154] M. C. Pennington and N. C. Ellis, "Cantonese Speakers' Memory for English Sentences with Prosodic Cues," *Modern Language Journal*, vol. 84, no. 3, pp. 372–389, 2000.

[155] M. Jilka, "Testing the contribution of prosody to the perception of foreign accent," in *New sounds 2000: proceedings of the Fourth international symposium on the acquisition of second-language speech*, (Amsterdam), pp. 199–207, 2000.

[156] P. Boula De Mareüil and B. Vieru-Dimulescu, "The contribution of prosody to the perception of foreign accent," *Phonetica*, vol. 63, pp. 247–267, Dec. 2006.

[157] S. Peppé, J. Maxim, and B. Wells, "Prosodic Variation in Southern British English," *Language and Speech*, vol. 43, no. 3, pp. 309–334, 2000.

[158] U. Gut and J.-T. Milde, "The Prosody of Nigerian English," in *Speech Prosody 2002, International Conference*, (Bielefeld, Germany), 2002.

[159] E. Couper-Kuhlen, "The prosody of other-repetition in British and North American English," *Language in Society*, vol. 49, pp. 521–552, Sept. 2020. Publisher: Cambridge University Press.

[160] P. Price, M. Ostendorf, S. ShattuckHufnagel, and N. Veilleux, "A methodology for analyzing prosody," *The Journal of the Acoustical Society of America*, vol. 84, pp. S99–S99, Nov. 1988.

[161] S. E. Brennan, A. Galati, and A. K. Kuhlen, "Two minds, one dialog: Coordinating speaking and understanding," in *Psychology of learning and motivation*, vol. 53, pp. 301–344, Elsevier, 2010.

[162] P. R. Doyle, J. Edwards, O. Dumbleton, L. Clark, and B. R. Cowan, "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*, (Taipei Taiwan), pp. 1–12, ACM, Oct. 2019.

[163] P. R. Doyle, L. Clark, and B. R. Cowan, "What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, (Yokohama Japan), pp. 1–14, ACM, May 2021.

[164] N. Dahlbäck, Q. Wang, C. Nass, and J. Alwin, "Similarity is more important than expertise: accent effects in speech interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (San Jose California USA), pp. 1553–1556, ACM, Apr. 2007.

[165] B. R. Cowan, P. Doyle, J. Edwards, D. Garaialde, A. Hayes-Brady, H. P. Branigan, J. Cabral, and L. Clark, "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*, (Dublin Ireland), pp. 1–8, ACM, Aug. 2019.

[166] V. Chidambaram, Y. H. Chiang, and B. Mutlu, "Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues," in *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 293–300, 2012.

[167] S. Yilmazyildiz, D. Henderickx, B. Vanderborght, W. Verhelst, E. Soetens, and D. Lefeber, "Multi-modal emotion expression for affective human-robot interaction," in *Proceedings of the Workshop on Affective Social Speech Signals (WASSS 2013), Grenoble, France*, 2013.

[168] K. S. Haring, Y. Matsumoto, and K. Watanabe, "How do people perceive and trust a lifelike robot," *Proceedings of the World Congress of Engineering and Computer Science 2013*, vol. I, pp. 23–25, 2013. ISBN: 9789881925237.

[169] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A Survey of Socially Interactive Robots: Concepts, Design, and Applications," tech. rep., Carnegie Mellon University, Pittsburgh, PA USA, Pittsburgh, PA, 2002.

[170] I. Straub, S. Nishio, and H. Ishiguro, "Incorporated identity in interaction with a teleoperated android robot: A case study," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp. 119–124, IEEE, Sept. 2010. ISSN: 1944-9445.

[171] M. F. Damholdt, M. Nørskov, R. Yamazaki, R. Hakli, C. V. Hansen, C. Vestergaard, and J. Seibt, "Attitudinal change in elderly citizens toward social robots: The role of personality traits and beliefs about robot functionality," *Frontiers in Psychology*, vol. 6, p. 1701, Nov. 2015. Publisher: Frontiers Research Foundation.

[172] A. Bruce, I. Nourbakhsh, and R. Simmons, "The role of expressiveness and attention in human-robot interaction," in *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 4, pp. 4138–4142, IEEE, 2002. ISSN: 10504729.

[173] H. Hüttenrauch, A. Green, M. Norman, L. Oestreicher, and K. S. Eklundh, "Involving users in the design of a mobile office robot," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 34, no. 2, pp. 113–124, 2004. ISBN: 1094-6977 VO - 34.

[174] D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "Teleoperation of multiple social robots," *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, vol. 42, no. 3, pp. 530–544, 2012. Publisher: ACM Press Place: New York, New York, USA ISBN: 9781605580173.

[175] Aldebaran, "NAO - Technical overview," 2017.

[176] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," *arXiv*, pp. 4779–4783, 2017. ISBN: 9781538646588.

[177] M. P. Aylett and Y. Vazquez-Alvarez, "Voice Puppetry: Towards Conversational HRI WoZ Experiments with Synthesised Voices," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, vol. 5, (New York, NY, USA), pp. 69–69, ACM, Mar. 2020.

[178] K. Kuwamura, S. Nishio, and S. Sato, "Can we talk through a robot as if face-to-face? Long-term fieldwork using teleoperated robot for seniors with alzheimer's disease," *Frontiers in Psychology*, vol. 7, p. 1066, July 2016. Publisher: Frontiers Research Foundation.

[179] S. Yilmazyildiz, G. Athanasopoulos, G. Patsis, W. Wang, M. Oveneke, L. Latacz, W. Verhelst, H. Sahli, D. Henderickx, B. Vanderborght, E. Soetens, and D. Lefeber,

"Voice Modification for Wizard-of-Oz Experiments in Robot-Child Interaction," in *Proceedings of the Workshop on Affective Social Speech Signals (WASSS 2013)*, (Grenoble), 2013.

[180] A. Ezzerg, A. Gabrys, B. Putrycz, D. Korzekwa, D. Saez-Trigueros, D. McHardy, K. Pokora, J. Lachowicz, J. Lorenzo-Trueba, and V. Klimkov, "Enhancing audio quality for expressive Neural Text-to-Speech," arXiv, Aug. 2021. arXiv:2108.06270 [cs, eess].

[181] A. Gabryś, G. Huybrechts, M. S. Ribeiro, C.-M. Chien, J. Roth, G. Comini, R. Barra-Chicote, B. Perz, and J. Lorenzo-Trueba, "Voice Filter: Few-Shot Text-to-Speech Speaker Adaptation Using Voice Conversion as a Post-Processing Module," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7902–7906, May 2022. ISSN: 2379-190X.

[182] N. Guerouaou, G. Vaiva, and J.-J. Aucouturier, "The shallow of your smile: the ethics of expressive vocal deep-fakes," *Philosophical Transactions of the Royal Society B*, vol. 377, no. 1841, p. 20210083, 2022. Publisher: The Royal Society.

[183] M. Schröder, "Expressive Speech Synthesis: Past, Present, and Possible Futures," in *Affective Information processing* (J. Tao and T. Tan, eds.), pp. 111–126, Saarbrücken, Germany: Springer London, 2009.

[184] C. McGinn and I. Torre, "Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots," in *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2019-March, pp. 211–221, 2019. ISBN: 9781538685556 Publisher: IEEE.

[185] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. Te Boekhorst, "Human approach distances to a mechanical-looking robot with different robot voice styles," *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, pp. 707–712, 2008. Publisher: IEEE ISBN: 9781424422135.

[186] I. Torre, A. B. Latupeirissa, and C. McGinn, "How context shapes the appropriateness

of a robot' s voice," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 215–222, IEEE, 2020.

[187] Z. Yin, "An Overview of Speech Synthesis Technology," in *2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)*, pp. 522–526, 2018.

[188] P. Baggia, P. Bagshaw, M. Bodell, D. Z. Huang, L. Xiaoyan, S. McGlashan, J. Tao, Y. Jun, H. Fang, Y. Kang, and others, "Speech synthesis markup language (SSML) version 1.1," 2010. Publication Title: World Wide Web Consortium, Recommendation REC-speechsynthesis11-20100907.

[189] P. Baggia, "Speech Standards: Lessons Learnt," in *Human 4.0-From Biology to Cybernetic*, IntechOpen, 2020.

[190] C. Breazeal, "Emotive Qualities in Robot Speech," *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180).*, 2001. Publisher: IEEE.

[191] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," *arXiv*, pp. 6189–6193, 2019. Publisher: IEEE ISBN: 9781509066315.

[192] J. Crumpton and C. L. Bethel, "A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech," *International Journal of Social Robotics*, vol. 8, pp. 271–285, Apr. 2016.

[193] M. Schröeder, P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter, and E. Zovato, "Emotion Markup Language (EmotionML) 1.0," 2014.

[194] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive TTS Training with Frame and Style Reconstruction Loss," *arXiv*, vol. 14, no. 8, pp. 1–12, 2020. arXiv: 2008.01490.

[195] E. Velner, P. P. Boersma, and M. M. de Graaf, "Intonation in Robot Speech: Does it work the same as with people?," in *Proceedings of the 2020 ACM/IEEE International*

Conference on Human-Robot Interaction, (New York, NY, USA), pp. 569–578, ACM, Mar. 2020.

[196] H. Hill and E. Vaikiotis-Bateson, "Using Graphics to Study the Perception of Speech-in-Noise, and Vice Versa," in *Auditory-Visual Speech Processing (AVSP '05)*, pp. 63–64, 2005.

[197] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, pp. 572–585, May 2013.

[198] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can Objective Measures Predict the Intelligibility of Modified HMM-based Synthetic Speech in Noise?," in *Twelfth Annual Conference of the International Speech Communication Association*, pp. 1837–1840, 2011.

[199] K. E. Silverman, "On customizing prosody in speech synthesis: Names and addresses as a case in point," in *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993.

[200] N. Campbell, "Evaluation of speech synthesis," in *Evaluation of text and speech systems*, pp. 29–64, Springer, 2007.

[201] A. A. Sanderman and R. Collier, "Prosodic phrasing and comprehension," *Language and Speech*, vol. 40, no. 4, pp. 391–409, 1997. Publisher: SAGE Publications Sage UK: London, England.

[202] D. Von Hapsburg, C. A. Champlin, and S. R. Shetty, "Reception thresholds for sentences in bilingual (Spanish/English) and monolingual (English) listeners," *Journal of the American Academy of Audiology*, vol. 15, no. 1, pp. 88–98, 2004. Publisher: American Academy of Audiology.

[203] L. H. Mayo, M. Florentine, and S. Buus, "Age of second-language acquisition and

perception of speech in noise," *Journal of speech, language, and hearing research*, vol. 40, no. 3, pp. 686–693, 1997. Publisher: ASHA.

[204] K. J. van Engen, B. Chandrasekaran, and R. Smiljanic, "Effects of Speech Clarity on Recognition Memory for Spoken Sentences," *PLoS ONE*, vol. 7, no. 9, 2012. ISBN: 1932-6203.

[205] J. Bench, A. Kowal, and J. Bamford, "The BKB (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children," *British Journal of Audiology*, vol. 13, pp. 108–112, Jan. 1979.

[206] S. Sharma, R. Tripathy, and U. Saxena, "Critical appraisal of speech in noise tests: a systematic review and survey," *International Journal of Research in Medical Sciences*, vol. 5, p. 13, Dec. 2016.

[207] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, pp. 1085–1099, Feb. 1994.

[208] Etymotic Research, "The BKB-SIN Test Booklet," 2005.

[209] D.-J. Shin and P. Iverson, "Training Korean second language speakers on English vowels and prosody," *The Journal of the Acoustical Society of America*, vol. 19, pp. 060048–060048, 2013.

[210] I. Hove and V. Dellwo, "The effects of voice disguise on f0 and on the formants," *Proceedings of IAFPA 2014*, 2014.

[211] R. P. Meier, K. Cormier, and D. Quinto-Pozos, *Modality and structure in signed and spoken languages*. Cambridge University Press, 2002.

[212] N. S. Pollard, J. K. Hodgins, M. J. Riley, and C. G. Atkeson, "Adapting human motion for the control of a humanoid robot," *EEE International Conference on Robotics and Automation*, vol. 2, pp. 1390–1397, 2002. ISBN: 0-7803-7272-7.

[213] F. Zuher and R. Romero, "Recognition of human motions for imitation and control of a humanoid robot," in *Proceedings - 2012 Brazilian Robotics Symposium and Latin American Robotics Symposium, SBR-LARS 2012*, pp. 190–195, IEEE, Oct. 2012.

[214] L. P. Poubel, S. Sakka, D. ehaji, and D. Creusot, "Support changes during online human motion imitation by a humanoid robot using task specification," in *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 1782–1787, 2014. ISSN: 10504729.

[215] J. Koenemann, F. Burget, and M. Bennewitz, "Real-time imitation of human whole-body motions by humanoids," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 2806–2812, 2014. ISBN: 978-1-4799-3685-4.

[216] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell, "A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception," *Computer Graphics Forum*, vol. 34, no. 6, pp. 299–326, 2015.

[217] S. Kohlbecher, K. Bartl, S. Bardins, and E. Schneider, "Low-latency combined eye and head tracking system for teleoperating a robotic head in real-time," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10*, (New York, New York, USA), p. 117, ACM Press, 2010.

[218] R. Cambuzat, F. Elisei, G. Bailly, O. Simonin, and A. Spalanzani, "Immersive teleoperation of the eye gaze of social robots assessing gaze-contingent control of vergence, yaw and pitch of robotic eyes," in *50th International Symposium on Robotics, ISR 2018*, pp. 232–239, 2018.

[219] E. de Wolf and J. Li, "Double Trouble: The Effect of Eye Gaze on the Social Impression of Mobile Robotic Telepresence Operators," in *Social Robotics* (A. R. Wagner, D. Feil-Seifer, K. S. Haring, S. Rossi, T. Williams, H. He, and S. Sam Ge, eds.), Lecture Notes in Computer Science, (Cham), pp. 357–368, Springer International Publishing, 2020.

[220] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 119–155, 2003. Publisher: Elsevier.

[221] L. Pessoa, "Do Intelligent Robots Need Emotion?," *Trends in Cognitive Sciences*, vol. 21, no. 11, pp. 817–819, 2017.

[222] D. A. Norman, "Emotional Design: Why We Love (or Hate) Everyday Things," *The Journal of American Culture*, vol. 27, no. 2, p. 234, 2004. ISBN: 0465051367.

[223] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and Autonomous Systems*, vol. 42, pp. 177–190, Mar. 2003. Publisher: North-Holland ISBN: 0921-8890.

[224] S. R. Fussell, S. Kiesler, L. D. Setlock, and V. Yew, "How people anthropomorphize robots," in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (Amsterdam, The Netherlands), pp. 145–152, IEEE, 2008.

[225] A. Mehrabian, *Silent messages*, vol. 8. Belmont, CA: Wadsworth, 1971.

[226] P. Yaffe, "The 7% rule: fact, fiction, or misunderstanding," *Ubiquity*, vol. 2011, no. October, p. 1, 2011. Publisher: ACM.

[227] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, pp. 45–60, Wiley, 1999.

[228] Y. T. Matsuda, T. Fujimura, K. Katahira, M. Okada, K. Ueno, K. Cheng, and K. Okanoya, "The implicit processing of categorical and dimensional strategies: An fMRI study of facial emotion perception," *Frontiers in Human Neuroscience*, no. SEP, 2013.

[229] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states," *Genetic, social, and general psychology monographs*, vol. 121, no. 3, pp. 339–361, 1995.

[230] A. Landowska, "Affective computing and affective learning ‐methods, tools and prospects," *EduAkcja. Magazyn Edukacji Elektronicznej*, no. June, pp. 16–31, 2013.

[231] J. A. Russell, A. Weiss, and G. A. Mendelsohn, "Affect grid: a single-item scale of pleasure and arousal.," *Journal of personality and social psychology*, vol. 57, no. 3, p. 493, 1989. Publisher: American Psychological Association.

[232] P. Ekman, "Expression and the Nature of Emotion," in *Approaches to Emotion*, pp. 319–344, Taylor & Francis, 1984.

[233] M. F. Jung, "Affective Grounding in Human-Robot Interaction," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. Part F1271, pp. 263–273, 2017. Publisher: ACM ISBN: 9781450343367.

[234] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement," 1978. Place: Palo Alto.

[235] R. Gockley, R. Simmons, and J. Forlizzi, "Modeling affect in socially interactive robots," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, (Hatfield, UK), pp. 558–563, IEEE, 2006.

[236] M. Nixon, P. Pasquier, and M. S. El-Nasr, "DelsArtMap: Applying Delsarte's aesthetic system to virtual agents," in *International Conference on Intelligent Virtual Agents*, vol. 6356 LNAI, (Berlin, Germany), pp. 139–145, Springer-Verlag, 2010. ISSN: 03029743.

[237] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010. Publisher: IGI Global.

[238] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[239] T. L. Chen, C.-H. King, A. L. Thomaz, and C. C. Kemp, "Touched by a robot: An investigation of subjective responses to robot-initiated touch," in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (Lausanne, Switzerland), pp. 457–464, IEEE, 2011.

[240] K. Shinozawa, B. Reeves, K. Wise, S. Lim, H. Maldonado, and F. Naya, "Robots as new media: A cross-cultural examination of social and cognitive responses to robotic and on-screen agents," in *Proceedings of Annual Conference of Internation Communication Association*, pp. 998–1002, 2003.

[241] C. Breazeal and B. Scassellati, "How to build robots that make friends and influence people," in *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289)*, vol. 2, (Kyongju, South Korea), pp. 858–863, IEEE, 1999.

[242] K. Kühnlenz, S. Sosnowski, and M. Buss, "The Impact of Animal-like Features on Emotion Expression of Robot Head EDDIE," *Advanced Robotics*, vol. 24, no. 8-9, pp. 1239–1255, 2010.

[243] L. Cañamero and J. Fredslund, "I show you how I like you-can you read it in my face?[robotics]," *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, vol. 31, no. 5, pp. 454–459, 2001. Publisher: Ieee.

[244] K. Berns and T. Braum, "Design concept of a human-like robot head," in *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, pp. 32–37, IEEE, 2005.

[245] D. Bazo, R. Vaidyanathan, A. Lentz, and C. Melhuish, "Design and testing of a hybrid expressive face for a humanoid robot," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, (Taipei, Taiwan), pp. 5317–5322, IEEE, 2010.

[246] C. Latta, N. Alvarado, S. S. Adams, and S. Burbeck, "An expressive system for endowing robots or animated characters with affective facial displays," in *The 2002 conference of the British Society for Artificial Intelligence and the Simulation of Behavior*, 2002.

[247] A. Beck, A. Hiolle, A. Mazel, and L. Cañamero, "Interpretation of emotional body language displayed by robots," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments - AFFINE '10*, (Firenze, Italy), p. 37, ACM Press, 2010.

[248] V. Sacharin, K. Schlegel, and K. R. Scherer, "Geneva Emotion Wheel Rating Study," *Poster presented at the International Society for Research on Emotion - ISRE*, 2012.

[249] K. R. Scherer, "What are emotions? And how can they be measured?," *Social science information*, vol. 44, pp. 695–729, Dec. 2005. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

[250] C. Becker-Asano and H. Ishiguro, "Laughter in social robotics-no laughing matter," in *Intl. Workshop on Social Intelligence Design*, pp. 287–300, 2009.

[251] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, (Boston, Massachusetts, USA), pp. 367–374, ACM, 2012.

[252] A. K. Coyne, A. Murtagh, and C. McGinn, "Using the Geneva Emotion Wheel to Measure Perceived Affect in Human-Robot Interaction," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, (New York, NY, USA), pp. 491–498, ACM, Mar. 2020.

[253] R. Savery, L. Zahray, and G. Weinberg, "Emotional musical prosody for the enhancement of trust: Audio design for robotic arm communication," *Paladyn, Journal of Behavioral Robotics*, vol. 12, pp. 454–467, Dec. 2021.

[254] J. Hagelbäck, O. Hilborn, P. Jeri, S. J. Johansson, C. A. Lindley, J. Svensson, and W. Wen, "Psychophysiological interaction and empathic cognition for human-robot cooperative work (PsyIntEC)," in *Gearing Up and Accelerating Cross-fertilization between Academic and Industrial Robotics Research in Europe*, pp. 283–299, New York, NY, USA: Springer, 2014.

[255] C. McGinn and K. Kelly, "Using the Geneva Emotion Wheel to Classify the Expression of Emotion on Robots," in *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 191–192, 2018. ISBN: 9781450356152.

[256] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao, "A facial expression emotion recognition based human-robot interaction system," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, pp. 668–676, Oct. 2017. Publisher: Institute of Electrical and Electronics Engineers Inc.

[257] D. H. Liu, *Performance of the Female: The Vtuber Phenomena.* Drexel University, 2022.

[258] C. Cao, M. Chai, O. Woodford, and L. Luo, "Stabilized real-time face tracking via a learned dynamic rigidity prior," *SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia 2018*, vol. 37, no. 6, 2018. ISBN: 9781450360081.

[259] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. NieSSner, and C. Theobalt, "State of the art on monocular 3D face reconstruction, tracking, and applications," *Computer Graphics Forum*, vol. 37, pp. 523–550, May 2018. Publisher: Blackwell Publishing Ltd.

[260] M. Grandgeorge, "Communication Between Humans: Towards an Interdisciplinary Model of Intercomprehension," in *Human-Robot Interaction: Evaluation Methods and Their Standardization*, pp. 3–19, Springer, 2020.

[261] A. Liu, G. Tharp, L. French, S. Lai, and L. Stark, "Some of What One Needs to Know About Using Head-Mounted Displays to Improve Teleoperator Performance," *IEEE Transactions on Robotics and Automation*, vol. 9, no. 5, pp. 638–648, 1993.

[262] S. Kratz, "Evaluating Stereoscopic Video with Head Tracking for Immersive Teleoperation of Mobile Telepresence Robots," *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pp. 4–5, 2015. Publisher: ACM Press Place: New York, New York, USA ISBN: 9781450333184.

[263] D. Sportillo, A. Paljic, and L. Ojeda, "On-Road Evaluation of Autonomous Driving Training," *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 182–190, 2019. Publisher: IEEE ISBN: 9781538685556.

[264] D. M. Johnson, "Introduction to Review of Simulator Sickness Research," *U.S. Army Research Institute for the Behavioral and Social Sciences*, no. April, pp. 1–70, 2005.

[265] F. Keyrouz and K. Diepold, "Binaural source localization and spatial audio reproduction for telepresence applications," *PRESENCE: Teleoperators and Virtual Environments*, vol. 16, no. 5, pp. 509–522, 2007. Publisher: MIT Press, One Rogers Street, Cambridge, MA 02142-1209, USA.

[266] J. Broderick, J. Duggan, and S. Redfern, "The Importance of Spatial Audio in Modern Games and Virtual Environments," in *2018 IEEE Games, Entertainment, Media Conference (GEM)*, (Galway, Ireland), pp. 1–9, IEEE, Aug. 2018.

[267] M. L. Hawley, R. Y. Litovsky, and J. F. Culling, "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 833–843, 2004. Publisher: Acoustical Society of America.

[268] S. Evans, C. McGettigan, Z. Agnew, S. Rosen, and S. Scott, "Getting the cocktail party started: masking effects in speech perception," *Journal of cognitive neuroscience*, vol. 28, pp. 483–500, Mar. 2016.

[269] K. Inkpen, R. Hegde, M. Czerwinski, and Z. Zhang, "Exploring spatialized audio & video for distributed conversations," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, (Savannah, Georgia, USA.), pp. 95–98, 2010.

[270] J. Ahrens, M. Geier, A. Raake, and C. Schlegel, "Listening and conversational quality of spatial audio conferencing," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, (Tokyo, Japan), Audio Engineering Society, 2010.

[271] M. Wong and R. Duraiswami, "Shared-Space: Spatial Audio and Video Layouts for Videoconferencing in a Virtual Room," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–6, IEEE, 2021.

[272] A. Torrejon, V. Callaghan, and H. Hagras, "Selectable Directional Audio for Multiple Telepresence in Immersive Intelligent Environments," in *2013 9th International Conference on Intelligent Environments*, pp. 181–187, IEEE, 2013.

[273] H. S. Ahn, S. McArdle, G. Sumner, and B. A. MacDonald, "Development of user interfaces for an internet-based forklift teleoperation system with telepresence," in *Proceedings of Australasian Conference on Robotics Automation*, (Melbourne, Australia), ARAA, 2014.

[274] A. Kristoffersson, S. Coradeschi, K. Severinson Eklundh, and A. Loutfi, "Sense of presence in a robotic telepresence domain," in *International Conference on Universal Access in Human-Computer Interaction*, (Orlando, Florida, USA), pp. 479–487, Springer, 2011.

[275] M. Álvarez, R. Galán, F. Matía, D. Rodríguez-Losada, and A. Jiménez, "An emotional model for a guide robot," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 5, pp. 982–992, 2010. Publisher: IEEE.

[276] T. Kishi, T. Otani, N. Endo, P. Kryczka, K. Hashimoto, K. Nakata, and A. Takanishi, "Development of expressive robotic head for bipedal humanoid robot," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Vilamoura, Algarve, Portugal), pp. 4584–4589, IEEE, 2012.

[277] M. Shayganfar, C. Rich, and C. L. Sidner, "A design methodology for expressing emotion on robot faces," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Vilamoura, Algarve, Portugal), pp. 4577–4583, IEEE, 2012.

[278] Y. Hu and G. Hoffman, "Using Skin Texture Change to Design Emotion Expression in Social Robots," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, (Daegu, Korea), pp. 2–10, IEEE, 2019.

[279] S. Song and S. Yamada, "Expressing emotions through color, sound, and vibration with an appearance-constrained social robot," in *Proceedings of the 2017 ACM/IEEE*

International Conference on Human-Robot Interaction, (Vienna, Austria), pp. 2–11, ACM, 2017.

[280] K. R. Scherer, V. Shuman, J. R. Fontaine, and C. Soriano, "The GRID meets the Wheel: Assessing emotional feeling via self-report," in *Components of emotional meaning: A sourcebook*, pp. 281–298, Oxford, UK: Oxford University Press, 2013. Publisher: Oxford University Press: Oxford, UK.

[281] P. Gendall, J. Hoek, and K. Gendall, "Evaluating the Emotional Impact of Warning Images on Young Adult Smokers and Susceptible Non-Smokers," *Journal of Health Communication*, vol. 23, pp. 291–298, Mar. 2018.

[282] O. Korovina, F. Casati, M. Baez, O. Berestneva, and R. Nielek, "Investigating crowdsourcing as a method to collect emotion labels for images," in *Conference on Human Factors in Computing Systems - Proceedings*, vol. 2018-April, (New York, New York, USA), pp. 1–6, ACM Press, 2018.

[283] N. Tschöpe, J. E. Reiser, and M. Oehl, "Exploring the uncanny valley effect in social robotics," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, (Vienna, Austria), pp. 307–308, ACM, 2017.

[284] M. B. Gurtman and A. L. Pincus, "The Circumplex Model: Methods and Research Applications," in *Handbook of Psychology*, pp. 407–428, Hoboken, NJ, USA: John Wiley & Sons, Inc., Apr. 2003.

[285] K. V. Mardia, *Statistics of directional data*. Cambridge, Massachusetts, USA: Academic Press, 1972.

[286] M. B. Gurtman and J. D. Balakrishnan, "Circular measurement redux: The analysis and interpretation of interpersonal circle profiles," *Clinical psychology: Science and practice*, vol. 5, no. 3, pp. 344–360, 1998. Publisher: Wiley Online Library.

[287] J. Sprenger, "The objectivity of subjective Bayesianism," *European Journal for Philosophy of Science*, vol. 8, no. 3, pp. 539–558, 2018. Publisher: Springer.

[288] K. M. Banner, K. M. Irvine, and T. J. Rodhouse, "The use of Bayesian priors in Ecology: The good, the bad and the not great," *Methods in Ecology and Evolution*, vol. 11, no. 8, pp. 882–889, 2020. Publisher: Wiley Online Library.

[289] J. K. Kruschke, *Doing Bayesian data analysis: A tutorial with R and BUGS*. Amsterdam, The Netherlands: Elsevier, 2011.

[290] P. Belin, "Voice processing in human and non-human primates," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, pp. 2091–2107, Dec. 2006.

[291] L. Mary, "Significance of Prosody for Speaker, Language and Speech Recognition," in *Extraction and Representation of Prosody for Speaker, Speech and Language Recognition*, pp. 1–18, Springer, 2012.

[292] P. Demonte, "HARVARD Speech Corpus—Audio Recording 2019," Oct. 2019.

[293] M. J. Hossain, S. M. A. Amin, M. S. Islam, and Marium-E-Jannat, "Development of robotic voice conversion for RIBO using text-to-speech synthesis," in *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)*, pp. 422–425, IEEE, Sept. 2018.

[294] J. James, B. T. Balamurali, C. I. Watson, and B. MacDonald, "Empathetic Speech Synthesis and Testing for Healthcare Robots," *International Journal of Social Robotics*, Sept. 2020.

[295] V. Seib, R. Memmesheimer, and D. Paulus, "A ROS-based system for an autonomous service robot," in *Robot Operating System (ROS)*, pp. 215–252, Springer, 2016.

[296] L. Taylor, A. Downing, G. A. Noury, G. Masala, M. Palomino, C. McGinn, and R. Jones, "Exploring the applicability of the socially assistive robot Stevie in a day center for people with dementia," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 957–962, IEEE, 2021.

[297] J. W. Peirce, "PsychoPy-Psychophysics software in Python," *Journal of Neuroscience Methods*, vol. 162, pp. 8–13, May 2007.

[298] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021.

[299] F. Eyssel and D. Kuchenbrandt, "Social categorization of social robots: Anthropomorphism as a function of robot group membership," *British Journal of Social Psychology*, vol. 51, pp. 724–731, Dec. 2012.

[300] A. Esposito, T. Amorese, M. Cuciniello, M. T. Riviello, and G. Cordasco, "How Human Likeness, Gender and Ethnicity affect Elders' Acceptance of Assistive Robots," in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pp. 1–6, IEEE, Sept. 2020.

[301] C. Klein and D. Allan, "Robot racism? Yes, says a study showing humans' biases extend to robots," 2019. Publisher: CNN.

[302] R. Sparrow, "Robotics Has a Race Problem," *Science, Technology, & Human Values*, vol. 45, pp. 538–560, May 2020.

[303] R. Tamagawa, C. I. Watson, I. H. Kuo, B. A. MacDonald, and E. Broadbent, "The Effects of Synthesized Voice Accents on User Perceptions of Robots," *International Journal of Social Robotics*, vol. 3, pp. 253–262, Aug. 2011.

[304] C. R. Wolf, "Virtual Platforms Are Helpful Tools but Can Add to Our Stress," May 2020.

[305] B. K. Wiederhold, "Connecting through Technology during the Coronavirus Disease 2019 Pandemic: Avoiding "zoom Fatigue"," *Cyberpsychology, Behavior, and Social Networking*, vol. 23, no. 7, pp. 437–438, 2020.

[306] J. N. Bailenson, "Nonverbal Overload: A Theoretical Argument for the Causes of Zoom Fatigue," *Technology, Mind, and Behavior*, vol. 2, Feb. 2021.

[307] A. Reis, M. Martins, P. Martins, J. Sousa, and J. Barroso, "Telepresence Robots in the Classroom: The State-of-the-Art and a Proposal for a Telepresence Service for

Higher Education," in *Technology and Innovation in Learning, Teaching and Education* (M. Tsitouridou, J. A. Diniz, and T. A. Mikropoulos, eds.), (Thessaloniki, Greece), pp. 539–550, Springer International Publishing, 2019.

[308] F. Hegel, C. Muhl, B. Wrede, M. Hielscher-Fastabend, and G. Sagerer, "Understanding Social Robots," in *2009 Second International Conferences on Advances in Computer-Human Interactions*, (Cancun, Mexico), pp. 169–174, IEEE, Feb. 2009.

[309] B. G. Glaser and A. L. Strauss, "Awareness Contexts and Social Interaction," *American Sociological Review*, vol. 29, no. 5, p. 669, 1964. Publisher: American Sociological Review.

[310] D. M. Wegner and T. Giuliano, "The Forms of Social Awareness," in *Personality, Roles, and Social Behavior*, pp. 165–198, New York, NY: Springer New York, 1982.

[311] M. C. C. Baranauskas, "Social awareness in HCI," *Interactions*, vol. 21, pp. 66–69, July 2014.

[312] S. Baron-Cohen and S. Wheelwright, "The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences," *Journal of Autism and Developmental Disorders*, vol. 34, no. 2, pp. 163–175, 2004. arXiv: 0162-3257/04/0400-0163/0 ISBN: 0162-3257.

[313] M. R. Endsley, "Design and Evaluation for Situation Awareness Enhancement," *Proceedings of the Human Factors Society Annual Meeting*, vol. 32, pp. 97–101, Oct. 1988. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[314] D. Wilson and T. Wharton, "Relevance and prosody," *Journal of Pragmatics*, vol. 38, pp. 1559–1579, Oct. 2006.

[315] D. Sperber and D. Wilson, *Relevance: Communication and cognition*, vol. 142. Citeseer, 1986.

[316] M. T. Orne, "Demand characteristics and the concept of quasi-controls," in *Introducing psychological research*, pp. 395–401, Springer, 1996.

[317] D. Krupke, S. Starke, L. Einig, J. Zhang, and F. Steinicke, "Prototyping of immersive HRI scenarios," in *Human-Centric Robotics- Proceedings of the 20th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines, CLAWAR 2017*, (Porto, Portugal), pp. 537–544, 2018.

[318] A. Albastaki, M. Hoggenmüller, F. A. Robinson, and L. Hespanhol, "Augmenting Remote Interviews through Virtual Experience Prototypes," in *32nd Australian Conference on Human-Computer Interaction*, (Sydney NSW Australia), pp. 78–86, ACM, Dec. 2020.

[319] M. Gorzel, A. Allen, I. Kelly, J. Kammerl, A. Gungormusler, H. Yeh, and F. Boland, "Efficient Encoding and Decoding of Binaural Sound with Resonance Audio," in *2019 AES International Conference on Immersive and Interactive Audio*, (York, UK), Audio Engineering Society, Mar. 2019.

[320] M. E. Bartlett, C. E. R. Edmunds, T. Belpaeme, and S. Thill, "Have I Got the Power? Analysing and Reporting Statistical Power in HRI," *ACM Transactions on Human-Robot Interaction*, vol. 11, pp. 1–16, June 2022.

[321] A. Kendon, "Spatial organization in social encounters: The F-formation system," in *Conducting interaction: Patterns of behavior in focused encounters*, 1990. Publisher: Cambridge University Press.

[322] H. Hüttenrauch, K. S. Eklundh, A. Green, and E. A. Topp, "Investigating spatial relationships in human-robot interaction," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, (Beijing, China), pp. 5052–5059, IEEE, 2006.

[323] A. Kiselev, M. Scherlund, A. Kristoffersson, N. Efremova, and A. Loutfi, "Auditory Immersion with Stereo Sound in a Mobile Robotic Telepresence System," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, (Portland Oregon USA), pp. 55–56, ACM, Mar. 2015.

[324] A. K. Coyne, K. Sapkota, and C. McGinn, ""Who said that?" Applying the Situation

Awareness Global Assessment Technique to Social Telepresence," *ACM Transactions on Human-Robot Interaction*, vol. 12, pp. 1–22, Apr. 2023.

[325] A. K. Coyne and C. McGinn, "The Effect of Human Prosody on Comprehension of TTS Robot Speech," in *Proceedings of the 32nd IEEE International Conference on Robot and Human Interactive Communication, RO-MAN*, (Busan, South Korea), IEEE, 2023.

[326] K. Mizumaru, ロボットからの社会的なはたらきかけのデザインに関する研究 *(Research on the design of social interaction from robots)*. PhD Thesis, Hokkaido University, 2023.

[327] H. R. M. Pelikan, S. Reeves, and M. N. Cantarutti, "Encountering Autonomous Robots on Public Streets," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, (Boulder CO USA), pp. 561–571, ACM, Mar. 2024.