Faculty of Engineering and Information Technology

A Methodology for Operationalising the Robot Centric HRI Paradigm:

Enabling Robots to Leverage Sociocontextual Cues During Human-Robot Interaction

A thesis submitted for the degree of
Philosophiae Doctor (PhD)

Sonja Caraian

Supervisors

Principal Supervisor	*Alternate Supervisor*
Dr. Nathan Kirchner	Dr. Alen Alempijevic

Senior Lecturers, School of Electrical, Mechanical and Mechatronic Systems
Center for Autonomous Systems, University of Technology Sydney

Examiners

Prof. Dr. Vanessa Evers
Professor of Human Media Interaction
University of Twente, Enschede, Netherlands

Takayuki Kanda
Senior Research Scientist
ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan

October 2015

Certificate of Original Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signed,

Production Note:
Signature removed prior to publication.

Sonja Caraian

Date: 14/10/15

Acknowledgments

The work presented in this thesis was funded by the Australian Research Council (ARC) through my having received an Australian Postgraduate Award (APA) which paid for my fees and scholarship; the University of Technology, Sydney (UTS); the Centre for Autonomous Systems (CAS); and RobotAssist.

Completing this thesis has been one of the biggest challenges I have faced, and it would not have been possible without the help, guidance, support and love of those around me. First and foremost, I wish to thank my principal supervisor, Dr. Nathan Kirchner. He has been a tremendous mentor for me, encouraging and enabling me to become both a better researcher and to grow as a person. His guidance, support, patience, immense knowledge and enthusiasm, and maddening attention to detail have been invaluable throughout this thesis work.

My sincere gratitude also goes to Dr. Alen Alempijevic, who, as my alternate supervisor, has been an inspiration and role model, and his technical know-how has been invaluable. I thank him for his encouraging and constructive feedback.

I also wish to thank Dr. Teresa Vidal-Calleja for her guidance and counsel, Dr. Brad Skinner and Dr. Gavin Paul for taking the time to listen to and encourage me, and Prof. Gamini Dissayanake, who, as the director of CAS, has enabled me and other researchers to learn and grow.

Getting through this thesis required more than academic support, and I would also like to thank all of my friends who supported and guided me in finding my way. My gratitude and appreciation for their friendship is endless.

Finally, none of this would have been possible without my family. Words cannot express how grateful I am to my parents and brother. This thesis stands as a testament to a lifetime of unconditional love and support.

Contents

List of Figures

List of Tables

List of Acronyms

Acronym	
ANOVA	Analysis of Variance
C	Compliant
CAS	Centre for Autonomous Systems
D	Dimension
DOF	Degree-of-freedom
FEIT	Faculty of Engineering and Information Technology
FPYE	Face Plane Yaw Estimation
GP	Gaussian process
GUI	Graphical User Interface
H	Hypothesis
HHI	Human-Human Interaction
HL	Human-Likeness
HRI	Human-Robot Interaction
HSS	Head-to-Shoulder Signature
HYE	Head Yaw Estimation
JA	Joint Attention
L	Looking
!L	Not Looking
NC	Non-Compliant
PCA	Principal Component Analysis
RGB	Red Green Blue
RGB-D	Red Green Blue Depth
ROS	Robot Operating System
RMS	Root Mean Squared
RQ	Research Question
UTS	University of Technology, Sydney

Abstract

A Methodology for Operationalising the Robot Centric HRI Paradigm:

Enabling Robots to Leverage Sociocontextual Cues During Human-Robot Interaction

Sonja Caraian

October 2015

The presence of social robots in society is increasing rapidly as their reach expands into more roles which are useful in our everyday lives. Many of these new roles require them to embody capabilities which were typically not accounted for in traditional Human-Robot Interaction (HRI) paradigms, for example increased agency and the ability to lead interactions and resolve ambiguity in situations of naïvety. The ability of such robots to leverage sociocontextual cues (i.e. non-verbal cues dependent on the social-interaction space and contextual-task space in order to be interpreted) is an important aspect of achieving these goals effectively and in a socially sensitive manner.

This thesis presents a methodology which can be drawn on to successfully operationalise a contemporary paradigm of HRI – Kirchner & Alempijevic's Robot Centric HRI paradigm – which frames the interaction between humans and robots as a loop, incorporating additional feedback mechanisms to enable robots to leverage sociocontextual cues. Given the complexities of human behaviour and the dynamics of interaction, this is a non-trivial task. The Robot Centric HRI paradigm and methodology were therefore developed, explored and verified through a series of real-world HRI studies ($n_{total} = 435 = 16 + 24 + 26 + 96 + 189 + 84$).

Firstly, by drawing on the methodology, it is demonstrated that sociocontextual cues can be successfully leveraged to increase the effectiveness of HRI in both directions of communication between humans and robots via the paradigm. Specifically, cues issued by social robots are shown to be recognisable to people, who generally respond to them in line with human-issued cues. Further, enabling robots to read interaction partners' cues *in situ* is shown to be highly valuable to HRI, for example by enabling robots to intentionally and effectively issue cues. In light of the finding that people will display HHI-predicted sociocontextual cues such as gaze around robots, a novel head yaw estimation framework which showed promise for the HRI space was developed and evaluated. This enables robots to read human-issued gaze cues and mutual attention *in situ*.

Next, it is illustrated that a robot's effectiveness at achieving its goal(s) can be increased by adding to its ability to moderate the cues it issues based on information read from humans (i.e. increased interactivity).

Finally, the above findings are shown to generalise to other sociocontextual cues, social robots and application spaces, demonstrating that the developed methodology can be drawn on to successfully operationalise the Robot Centric HRI paradigm, enabling robots to leverage sociocontextual cues to more effectively achieve their goal(s) and meet the requirements of their expanding roles.

Chapter 1

Introduction

Can you tell when a friend might be sad, happy or frustrated just by looking at their face? How is it that, when you watch a television program in a foreign language, you are still able to capture important social information such as the status and rapport between people, and the overall atmosphere of interactions? When you see a directional indicator, such as someone pointing, why is it that your instinct is to look or move in that direction?

As a continuous source of information, sociocontextual cues (i.e. non-verbal cues dependent on the social-interaction space and contextual-task space in order to be interpreted) play a vital role in communication, for example as indicators of intention. As social robots move into positions in which they coexist in close proximity and work collaboratively with humans, they are increasingly assuming the role of interaction peers, with increased agency and the ability to lead interactions and resolve ambiguity in situations of naïvety. In order to effectively achieve their goal(s) and communicate in a socially sensitive manner, an important aspect of this is the capability of leveraging sociocontextual cues [81].

This thesis presents a devised methodology which can be drawn on to successfully operationalise a contemporary paradigm of Human-Robot Interaction (HRI) – Kirchner & Alempijevic's Robot Centric HRI paradigm – which positions robots as interaction peers by framing the communication during HRI as a loop, incorporating additional feedback mechanisms to enable robots to leverage sociocontextual cues to more successfully achieve their goals [81]. In light of the complexities of human behaviour and the dynamics of interaction, this is a non-trivial task.

This chapter introduces the research questions covered by this thesis, commencing with a discussion of the changing roles of social robots in society, which are increasingly viewed as interaction peers and expected to communicate in a socially sensitive manner. The new roles and requirements of robots provides the motivation for the development of a methodology to enable sociocontextual cues to be leveraged in HRI via the Robot Centric HRI paradigm. The following sections detail the objectives of this research, the approach taken to achieve said objectives, and the contributions arising from this work. Finally, an outline of the remaining chapters of this dissertation is presented.

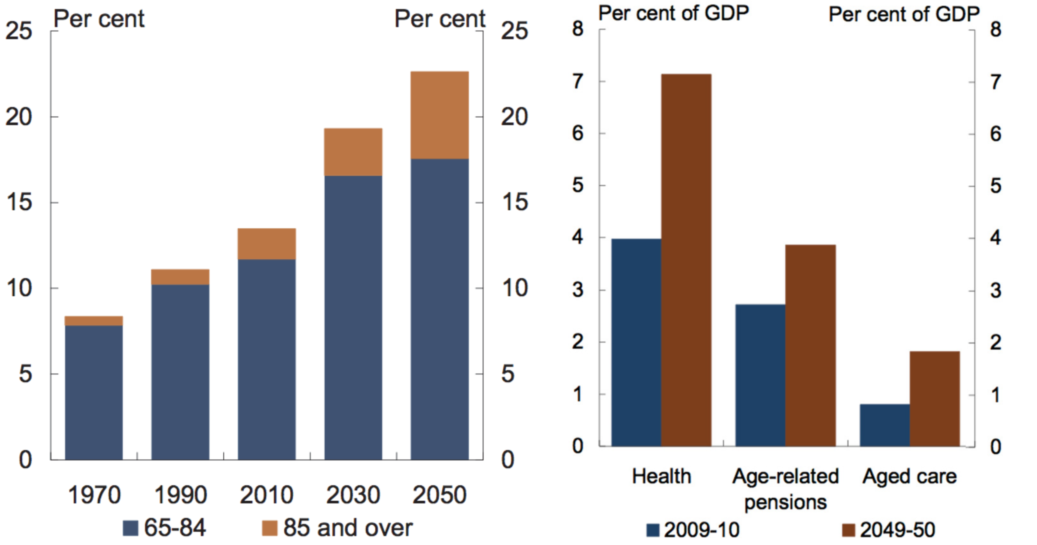
1.1 Leveraging Sociocontextual Cues to Increase the Effectiveness of HRI

1.1.1 The Increasing Presence of Social Robots in Society

The presence of robots in society is increasing rapidly as their reach expands from traditional industrial applications and research laboratories into more roles which are useful in our everyday lives. In many of these new roles it is necessary for robots to interact and work closely with humans, leading to the emergence of social robots: “physical entities embodied in complex, dynamic, and social environment[s] sufficiently empowered to behave in a manner conducive to their own goal(s) and those of their community” [30]. These autonomous or semi-autonomous robots interact and communicate with humans by following behavioural norms and rules attached to their roles by the people with whom they interact [12].

The aged-care industry is one example where a real need for such robots exists. Over the next few decades, globally ageing populations will significantly increase the strain on aged-care services. In Australia, the proportion of working-age people to older people (those aged 65 years and over) is projected to drop from the current ratio of 5:1 to 2.7:1 by 2050, as illustrated in Figure 1.1a. The resultant reduction in workforce availability to support the elderly is increasing associated aged care costs; from 2009–10 to 2049–50, for example, health spending on those aged over 65 years and over 85 years is expected to grow around seven-fold and twelve-fold, respectively [153], as shown in Figure 1.1b. This is increasing the necessity of residential aged-care and emphasising the concept of ageing-in-place, which it is widely recognised that socially interactive and humanoid assistive robots can play an important role in facilitating [81].

Another example of a potential application for social robots can be seen in the increasing levels of congestion and crowding on public transport systems, which is being driven by cities’ growing populations. For example, the Sydney Trains network has over 3 million annual journeys, with an average annual growth rate of 2.5%. Each weekday, there are over 1 million total journeys, resulting in trains being regularly filled to upwards of 130% of their average load factors (passengers as a percentage of seat capacity). Almost 160,000 of these passengers pass through Sydney’s Central Business District (CBD) stations during each of



(a) Proportion of the Australian population aged over 65, showing the projected increasing proportion of elderly people.

(b) Modelled projections of Australian government spending, showing increasing aged care costs.

Figure 1.1: Ageing populations, such as that in Australia, are raising aged-care costs and necessitating ageing-in-place.
Source: [153]

both the morning and afternoon peak periods, as shown in Figure 1.2, with a majority of this traffic (approximately 85%) concentrated at three of the eight stations [19]. Such high passenger volumes on networks running at their upper capacity results in significant congestion.

One potential method of addressing this issue is to guide people in physical space to increase the efficiency of passenger movement though such train stations, and public transport environments more widely. This would enable people to move more quickly through these environments, improving safety by easing congestion and reducing the dwell time of trains at stations. Given the large numbers of passengers and the often sprawling layout of many public transport environments, robotics has the potential to play a key role in achieving this goal. For example, a disembodied social robot built into the transport environment itself could direct people towards a less-crowded entrance to their platform, or oversee the disembarking and subsequent boarding of passengers from public transport.

The recognised potential of social robots is also driving their expansion into many other applications, such as education, entertainment, and domestic use, and roles such as assistants and companions [134]. Researchers in Japan, for example, have developed Paro, a robotic seal which is being used to provide physical and emotional support to the sick and elderly [141], shown interacting with the elderly in Figure 1.3. Such close interactions can also be seen in robots which work in collaboration with human nurses to lift medical patients and help disabled people move their limbs [111, 157]. Similarly, 'TIM' (Thought-controlled Intelligent Machine), a hands-free wheelchair aimed at quadriplegics, can be controlled via a combination of head movements and thoughts [118]. Such robots are also increasingly being deployed as receptionists, secretaries and teachers [60], taking on new societal roles which were previously unimaginable.

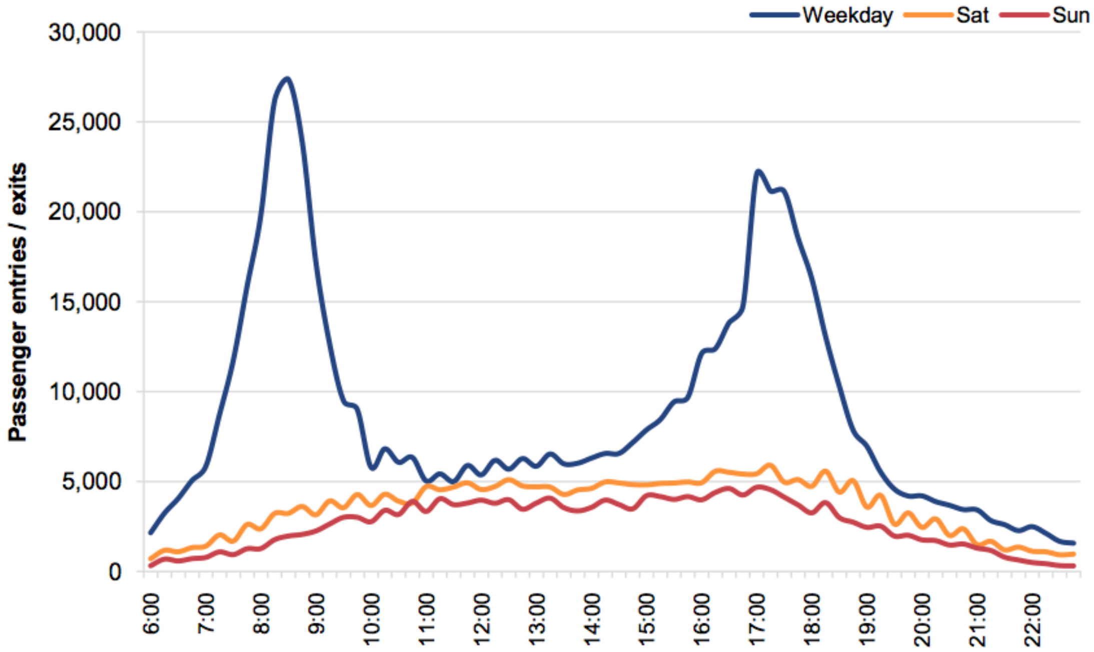


Figure 1.2: Sydney CBD station entries and exits by time of day and day type, showing high volume morning and afternoon peak periods.
Source: [19]

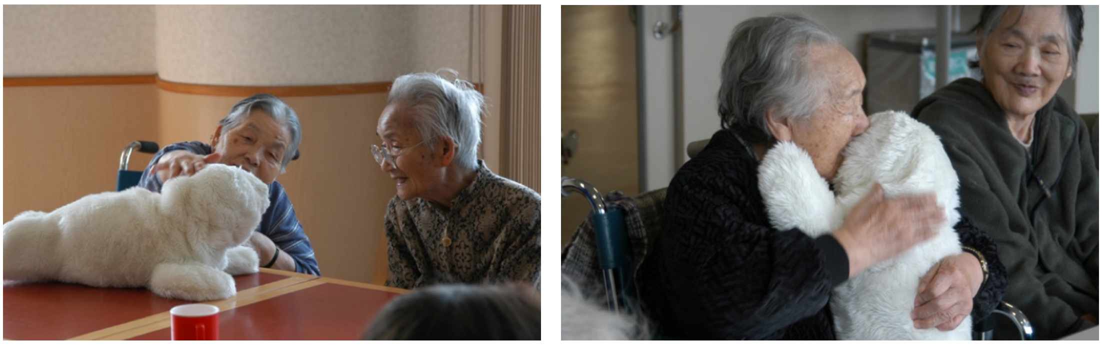


Figure 1.3: Socially interactive and assistive robots such as Paro, which interacts with the elderly in Japan, are emerging.
Source: [124]

1.1.2 Social Robots as Interaction Peers Leveraging Sociocontextual Cues

These changing roles of social robots and the growing importance of socially-appropriate behaviour to HRI is a sign of the significant evolution of the relationship between humans and robots, which were originally perceived as simply task completers, or tools. Traditional HRI paradigms often involved humans extending their influence on the environment through simply acting upon a machine, which would then complete a task [55]. This can be seen in examples such as [74], which presents a design philosophy for service robots which positions humans as 'operators' who command robots which have limited autonomy. The new roles which social robots are fulfilling, however, require them to embody new capabilities such as social sensitivity, autonomy, and intentional action, capabilities which were typically not accounted for in traditional HRI paradigms. Thus, in many of these new applications robots are assuming the role of interaction peers, having increased agency and the ability to lead interactions and resolve ambiguity in situations of naïvety.

In order to achieve an acceptable level of social integration as well as to more successfully achieve their goal(s) (such as instantiating interactions and/or resolving ambiguity) [81], robots in interaction peer roles must be capable of communicating in a socially sensitive manner [55, 130, 151, 166]. An important aspect of this is the capability of leveraging non-verbal cues [20, 84], which can enrich and increase the effectiveness and expediency of communication [81]. Research has shown that opening this non-verbal communication channel and integrating such cues into HRI can improve the understandability of HRI and the perception of the robot for the human subject, while also increasing the efficiency of human-robot task performance [17, 39, 65, 120].

As summarised in [81], however, it is known from psychology and behavioural science literature that while non-verbal cues have well defined and known meanings (social norms), further meanings are often ascribed to these cues based on the situation in which they are being interpreted (context); that is, the social-interaction space and contextual-task space – the sociocontext, a term coined by Kirchner & Alempijevic [81] – in which a non-verbal cue is being issued can affect whether it communicates a socially interpretable message. Thus, to convey a specific message, an appropriate non-verbal cue for the particular sociocontext (henceforth referred to as a sociocontextual cue, for brevity) should be leveraged.

1.1.3 Effects of Robot Human-Likeness on Sociocontextual Cues

To understand and effectively leverage such cues in HRI, however, the question first becomes: to what extent will the characteristics and effects of sociocontextual cues in HRI correspond to those of human-human cues? That is, to what extent can the psychology and behavioural science research on HHI cues be used to predict the outcomes of similar cues in HRI?

In exploring this question, it is first necessary to consider social robots' levels of anthropomorphism; that is, the extent to which such robots behave, respond and look like humans: their human-likeness (HL). Social robots' HL ranges from disembodied and overtly non-human robots (e.g. from the iRobot Roomba [69]), along a continuum to perfect androids completely indistinguishable from humans (e.g. Arnold Swarzenegger's Terminator). Today's social robots sit along different points of this continuum. As such, depending on the type of sociocontextual cue the particular social robot is attempting to issue, direct 'translation' of these cues from HHI to HRI is not guaranteed.

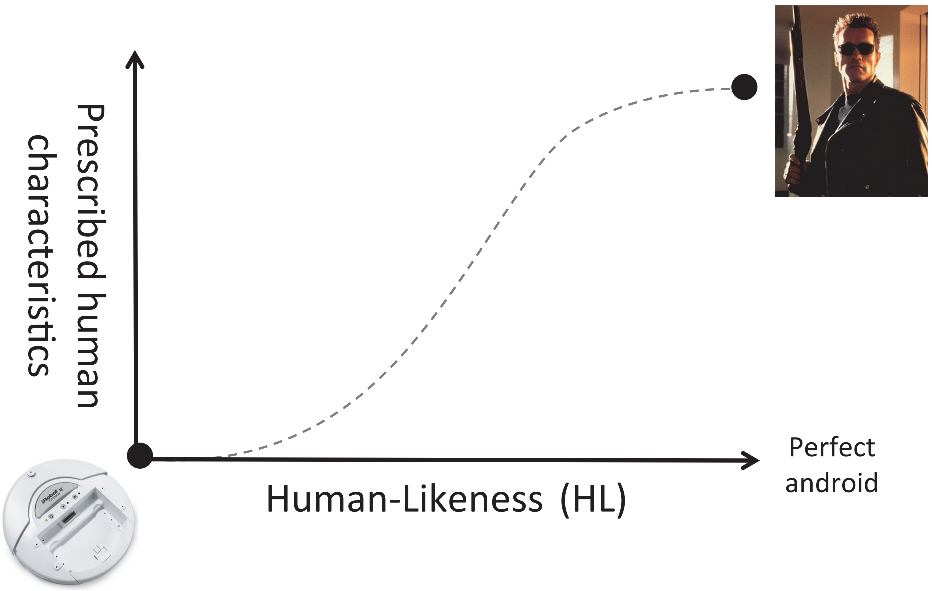


Figure 1.4: With increasing human-likeness, people prescribe robots a greater number of human characteristics.

However, if an appropriate cue is selected for the particular HL level of a robot, translation of sociocontextual cues from HHI to HRI is foreseeable [115]. As discussed in more detail in [31], as HL increases, people tend to prescribe robots a greater number of human characteristics [97], as depicted in Figure 1.4, leading to increased perception of the social robot's social capabilities and thus facilitating greater social understanding. When a balance is struck between people's expectations and the robot's capabilities, ambiguity and misinterpretations about the robot's ability and role can be avoided, facilitating humans' acceptance of social robots' mechanisms for communication and social interaction [31].

Consider, for example, HRI with higher-HL humanoid social robots. During HHI, bodily sociocontextual cues – "motion[s] of the body that contain information" [88] – are a continuous source of information about the feelings, mental state, personality and other traits of people [133]. This aspect of communication is used, both deliberately and unconsciously, to supplement and/or substitute for spoken language [28], in some cases carrying up to thirteen times the information of verbal cues [7, 103]. There are many bodily cues that humans display and are sensitive to, ranging from coarse, whole body cues to finer cues issued by single body parts, and can be static and/or dynamic, some examples of which are shown in Figure 1.5. Generally, they fall into the following categories: body orientation, posture, gestures, touching behaviour, facial expressions, and gaze behaviour [87]. For example, humans can issue sociocontextual cues via changes in physical distance during interaction, the location, orientation and configuration of their upper and/or lower body, shoulder slumps, head tilts, arm and hand movements, feet orientation, smiles or furrowing of the brow, or direction and length of gaze [102, 103].

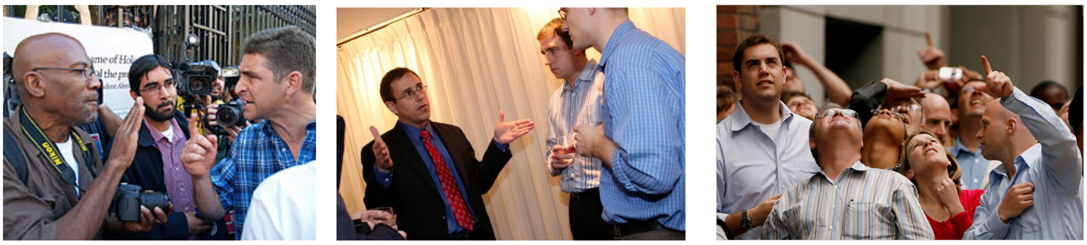


Figure 1.5: Non-verbal cues play an important role in communication.
Photo (a) source [168], (b) source [22], (c) source [32]

Higher-HL humanoid social robots are likely to be physically capable of emulating a growing number of these bodily sociocontextual cues during HRI. Given humans' innate reliance on bodily cues during interaction [7, 28, 133], and the growing perception of social robots as interaction peers, it is likely that humans will be increasingly expectant of and responsive to these cues during interaction with humanoid robots; thus the characteristics and effects of bodily cues in HHI seem likely to translate to HRI with such robots. Conversely, if a completely non-HL robot attempted to issue a bodily sociocontextual cue it seems less likely humans would recognise or respond to it in line with HHI cues. A simpler, sociocontextual directional indicator (such as an arrow), on the other hand, may be a more appropriate cue for a lower-HL robot to issue.

1.1.4 The Robot Centric HRI Paradigm and Robot Interactivity Through Sociocontextual Cues

Thus it seems reasonable that appropriate sociocontextual cues are likely to be able to be leveraged by social robots to aid in achieving their goal(s) and communicating in a socially sensitive manner. As previously mentioned, contemporary paradigms of HRI, such as that developed by Kirchner & Alempijevic [81], now frame the interaction between humans and robots as a loop, positioning robots as interaction peers: robots are no longer perceived as only task completers as in traditional master/slave-style HRI paradigms (e.g. [55, 74]). To achieve this, Kirchner & Alempijevic's Robot Centric HRI paradigm incorporates additional feedback mechanisms to account for growing social robot capabilities (particularly their ability to present sociocontextual cues designed to elicit particular behavioural responses to the human interaction partner) and the influence of the social feedback they generate by providing these cues to humans. These two key additional communication branches/mechanisms create a communication loop between humans and robots, giving robots increased agency and the ability to lead interactions and resolve ambiguity in situations of näivety.

The first branch of Kirchner & Alempijevic's paradigm, *Read*, shown in Figure 1.6, sees the robot able to sense behavioural sociocontextual cues displayed by the interacting human(s), including cues such as human presence and location, head pose, and gestures. As discussed by the authors, these cues can then be interpreted through a combination of contextual understanding and human behaviour-to-meaning mapping available from the fields of psychology and behavioural science. This sensing capability enables the robot to incorporate human-sensed information into a derived action plan, if necessary. Additionally, this *Read* capability increases the perception of the robot as a social entity [81]: it has been shown that robotic systems capable of adapting and responding to human social signals in polite, unintrusive, or persuasive manners are likely to be perceived as more natural, efficacious, and trustworthy [163]. This has been demonstrated in contexts such as education, where human teachers use pupils' social signals to inform the shape of their instructional messages: robotic agents able to learn context-dependent social behaviour and employ socially adept presentation language through the accurate sensing and interpretation of the social signals and context of the pupils are likely to be more successful [125, 163].

The second additional branch presented by Kirchner & Alempijevic, *Elicit*, which is also illustrated in Figure 1.6, indicates the ability of the robot to surreptitiously present sociocontextual cues in order to elicit particular behavioural responses from the interaction partner(s). This 'probing' of the user for information, which the robot can then add to its contextual understanding, can aid the robot in resolving ambiguity or eliciting specific human action [81]. This branch focuses on influence through sociocontextual cues because of their non-verbal nature, which makes their influence more surreptitious and implicit, and thus less susceptible to negative human response [113]: it has been demonstrated that humans often respond negatively to perceived robot-issued commands or task dictation [154]. Implicit interactions can also enable robots to be assistive when interacting humans are otherwise physically, socially, or cognitively engaged, or naïve to the robot's intentions [71, 84].

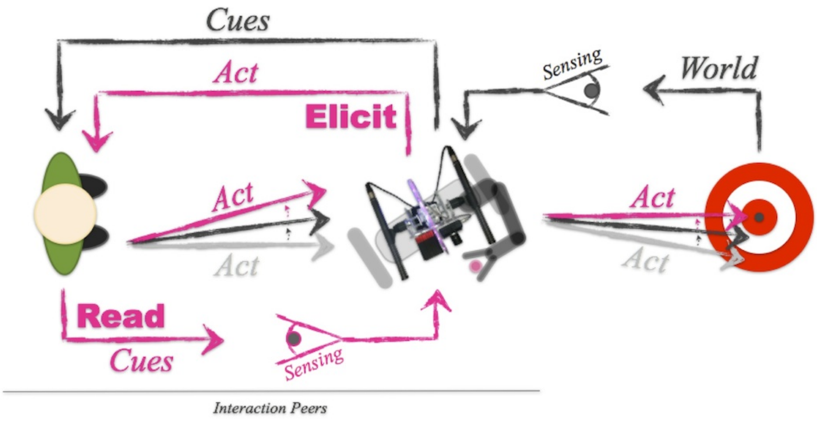


Figure 1.6: A contemporary Robot Centric HRI paradigm proposed by [81], which sees robots as interaction peers with increased agency.
Source: [81]

Further, depending on the design of the Robot Centric HRI paradigm implementation, different levels of robot interactivity can be achieved; that is, the potential of the robot to exhibit causal behaviour (respond in reaction to interaction with a human) [12]. In the context of the Robot Centric paradigm, a robot's interactivity is its ability to moderate the sociocontextual cues it issues based on the behavioural information it reads from humans: its ability to *Read*, then moderate its *Elicit* strategy based on this information and known behaviour-to-meaning mappings in such a way as to increase its effectiveness in achieving its desired outcome. Due to the piecemeal exploration of the Robot Centric HRI paradigm in [81], this concept has not yet been holistically tested. Kirchner & Alempijevic theorise, however, that the more interactivity a robot has, the more it will be able to operate as an interaction peer to achieve its goal(s).

Such a paradigm of HRI seems suitable to enable HHI sociocontextual cues to be effectively and predictably exploited during HRI, giving social robots the means to leverage such cues in order to more effectively achieve their goal(s). The question then becomes: how can this Robot Centric paradigm be successfully operationalised during real-world HRI? Given the complex nature of both human behaviour and the dynamics of interaction, this question is non-trivial. In order to achieve successful operationalisation, the target problem, the robot's goal(s) and the application space must be defined, and design and implementation of the *Read* and *Elicit* branches of the paradigm – as well as the necessary interactivity level of robot – must be considered and accomplished. It seems likely, however, that the steps of design and implementation of the paradigm could be formalised into a methodology which could be drawn on to operationalise the paradigm during real-world HRI, and hence also enable holistic verification of the paradigm itself.

1.2 Research Questions

This thesis explores the feasibility of developing such a methodology, which could be drawn on to successfully operationalise the Robot Centric HRI paradigm during real-world HRI outside of research centre environments, where lab-based assumptions do not always hold true. In doing so, the ability of social robots to leverage sociocontextual cues is also investigated, enabling them to more effectively achieve their goal(s) (such as instantiating interactions, shaping interaction participant roles and resolving ambiguity) [81] and meet the expectations of communicating in a socially sensitive manner (as required by their growing interaction peer role).

1.2.1 Methodology for Robot Centric HRI Paradigm Operationalisation

The primary research question thus becomes:

RQ: Methodology – *Can a methodology be developed which could be drawn on to successfully operationalise the Robot Centric paradigm during real-world HRI?*

1.2.2 Transferability of Sociocontextual Cues to HRI

In exploring this, it is first necessary to determine if the possibility does exist to successfully leverage sociocontextual cues in HRI via the Robot Centric paradigm; that is, if/how these cues will manifest in HRI. Thus the first subquestion to arise is:

RQ A: Sociocontextual cues in HRI – *To what extent will the characteristics and effects of sociocontextual cues in HRI correspond to those of human-human cues?*

This research question breaks down into two main areas of exploration, as the interaction peer role of social robots requires two-way, reciprocal interaction – the ability of the robot to *Elicit* via issuing cues, and *Read* human-issued cues. Thus, for sociocontextual cues to be reliably implemented and utilised during HRI it is necessary to explore both of these areas, which are further detailed below.

Elicit **- Can Robots Issue Cues?**

The first area requiring investigation is the *Elicit* branch of the Robot Centric HRI paradigm. In order for robots to effectively 'probe' interacting humans for information, this branch requires robots to present sociocontextual cues to human interaction partners to elicit specific responses. However, as social robots have varying levels of human-likeness, the following questions need to be explored in order to understand the characteristics and effects of these cues in this direction of communication between robots and humans:

RQ A.1: *Elicit* Feasibility – *Are today's social robots physically capable of issuing recognisable sociocontextual cues?*

RQ A.1: *Elicit* Response – *During real-world HRI, will people respond to social robot-issued cues in line with the way they respond to human-issued cues?*

Given their growing relevance as the roles of robots evolve, robot-issued sociocontextual cues are currently an area undergoing a great deal of HRI research. However, the focus of such research has generally been on its outcomes in HRI (e.g. [17]), rather than on whether humans perceive and respond to such cues equivalently in HRI as in HHI. For example, [59] specifically state that their experiment does not explore the underlying psychology of the gesture and gaze cues they employ. The difference of the work presented here is the focus on the extent to which psychology and behavioural science research on HHI cues can be used to predict the outcomes of similar cues in HRI, specifically during real-world HRI.

While some research touches on this HHI-HRI equivalency – for example, [114] explore whether a robot's gaze results in the outcome of people perceiving it more favourably and performing better on a task, as a human's gaze would – a key addition of the work presented here is consideration of the factor of human-likeness, and the relevance of selecting appropriate sociocontextual cues based on the particular HL level of a robot.

Read **- Can Robots Decipher Human-Issued Cues?**

The second consideration when exploring the characteristics of sociocontextual cues in HRI is the ability of the robot to *Read* human-issued cues. A combination of contextual understanding and human behaviour-to-meaning mapping from the fields of psychology and behavioural science could enable robots to interpret these cues and incorporate the information into derived action plans. This has the additional advantage of strengthening the perception of robots as interaction peers capable of communicating sensitively with humans. However, before this channel of communication can be leveraged in HRI, a number of questions need to be addressed:

RQ A.2: *Read* Value – *Will people display HHI-predicted sociocontextual cues around social robots? That is, is it valuable for such robots to have in situ cue detection?*

RQ A.2: *Read* Feasibility – *Will social robots be capable of detecting and interpreting these cues during real-world HRI?*

The reading of human-issued cues is also an ongoing research topic within the HRI community, and has led to a range of established techniques, for example of recognition of human facial expressions (an overview is given in [107]). Similarly to the *Elicit* branch, however, the key contribution of this work is exploring the underlying psychology of this, and whether people will display cues during HRI similarly to HHI.

The feasibility of robots detecting and interpreting these cues is also often limited to controlled and structured lab environments; the novelty of this work, conversely, is on whether this is possible during during real-world HRI outside of research centre environments, which generally lack such structure and constraints.

1.2.3 Robot Interactivity Moderating Effectiveness

With an understanding of the transferability of the characteristics and effects of sociocontextual cues to HRI, the next question to arise when exploring whether a methodology can be developed regards the design of the Robot Centric HRI paradigm itself:

RQ B: Interactivity – *Will a robot's effectiveness at achieving its goal(s) be increased by adding to its ability to moderate its* Elicit *strategy based on information gained through* Read*ing (i.e. increased interactivity)? That is, are both the* Read *and* Elicit *branches of the Robot Centric HRI paradigm valuable to a social robot in effectively achieving its goal(s)?*

While research within the HRI community is being carried out with regards to how to operationalise the *Elicit* and *Read* branches, the novel contribution of this work is exploring how these branches can be utilised to moderate a robot's level of interactivity in order to to increase its effectiveness at achieving its desired outcomes.

1.2.4 Summary

The above questions are explored in this thesis. As summarized in the sections above, these research areas of this work have been partially explored by others. However, the research questions posed in this work focus on novel areas not directly addressed by others' work. Initially, a methodology for operationalisation of the Robot Centric HRI paradigm during real-world HRI is devised. This methodology is subsequently drawn on to successfully operationalise the individual components of the paradigm, enabling exploration of the further research questions: first, the transferability of sociocontextual cues to HRI is investigated through a combination of literature and a number of real-world, social HRI studies in which an exemplar humanoid social robot issues an exemplar sociocontextual cue appropriate for its level of human-likeness. Next, the responses of people to such a cue are compared with the literature-predicted effects of similar human-issued cues, and a method of detecting an exemplar human cue is developed. The relationship between a robot's interactivity and its effectiveness is then explored with the exemplar humanoid robot and cue. Finally, the developed methodology and the findings from these explorations are shown to generalise to other cues and social robots in other application spaces.

1.3 Principal Contributions

The work presented in this thesis addresses the above research questions. The significant contributions are as follows:

1. **A methodology for operationalising Kirchner & Alempijevic's Robot Centric paradigm [81] during real-world HRI**

   Through the devised methodology, the paradigm can be operationalised to enable robots to leverage sociocontextual cues during HRI in order to more effectively achieve their goal(s) and meet the requirements of their Interaction Peer roles in a socially sensitive manner. In devising and validating the methodology, the following was also achieved:

   (a) **Demonstration that sociocontextual cues can be successfully leveraged during HRI via the Robot Centric HRI paradigm**

   Through literature and a series of experiments, it was demonstrated that sociocontextual cues are transferable from HHI to HRI, and can be leveraged during HRI via the Robot Centric paradigm. This was shown in a piecemeal fashion through exploration of the four topics below. These explorations drew together literature, methods and technology from others' work, and novel technology developed during the work presented in this thesis. The topics are summarised below, and the different permutations of topic exploration have been published in ***[J1]***, ***[J2]***, ***[C2]***, ***[C3]***, and ***[W1]*** listed in Appendix A.

   i. ***Elicit*** **Feasibility** – Sociocontextual cues issued by social robots are shown to be recognisable to people.

   ii. ***Elicit*** **Response** – It is demonstrated that people generally respond to robot-issued sociocontextual cues in line with human-issued cues, as outlined in literature from psychology and behavioural science. Thus social robots can successfully *Elicit* particular behavioural responses from interaction partners.

   iii. ***Read*** **Value** – It is established that people will display HHI-predicted sociocontextual cues around robots. For example, as per the predictions, no generalisable pattern of gaze behaviour towards robots is observable during real-world HRI; thus, enabling robots to *Read* interaction partners' sociocontextual cues *in situ* is highly valuable to HRI, for example by enabling robots to intentionally and effectively *Elicit*.

iv. ***Read* Feasibility** – To address the need established in (c) ***Read* Value**, a novel head yaw estimation framework which shows promise for this application was devised, enabling robots to *Read* exemplar human gaze cues and mutual attention *in situ*.

  - **Head yaw estimation framework for the HRI space**

    A HYE framework was developed which leverages the strengths of multiple HYE methods to achieve operation over the entirety of the HRI space while maintaining an HRI-suitable, landmark level of accuracy. Two key developments made this possible:

    - **Novel use of Gaussian processes to inherently fuse multiple HYE methods (including the novel method below), and hence leverage their strengths.**

    - **A novel HYE method which utilises the planar characteristic of people's faces to complement and extend the operation space of state of the art HYE methods within the HRI space.**

(b) **Deepened understanding of the Robot Centric HRI paradigm**

The work presented in this thesis also deepened the understanding of the Robot Centric HRI paradigm and demonstrated that it can be operationalised holistically during real-world HRI. In addition to the above explorations, this was achieved by investigating the relationship between social robots' interactivity – a concept Kirchner & Alempijevic speculatively proposed [81], but did not holistically test due to the piecemeal exploration in that work – and effectiveness at achieving their goal(s). It was demonstrated through a number of studies that increased levels of interactivity are beneficial to social robots: a robot's effectiveness at achieving its goal(s) is increased by adding to its ability to moderate its *Elicit* strategy based on information gained through *Read*ing (i.e. increased interactivity). This is further detailed in the publications ***[J1]*** and ***[C1]*** listed in Appendix A.

(c) **Demonstration of generalisability of the Robot Centric HRI paradigm and the devised methodology**

After holistic evaluation of the Robot Centric HRI paradigm with an exemplar sociocontextual cue and social robot, it was then empirically demonstrated that the findings of this thesis generalise to other cues, social robots and application spaces (as detailed in ***[C1]*** of Appendix A), and that the methodology can be drawn on to successfully operationalise the Robot Centric paradigm during real-world HRI.

1.4 Outline of Thesis

In order to conduct this research, background information surrounding the operationalisation of the Robot Centric HRI paradigm with an exemplar sociocontextual cue and humanoid social robot was first gathered. A methodology for operationalisation of the Robot Centric paradigm during real-world HRI was then developed, which was subsequently drawn on in a number of studies which were carried out with the exemplar cue and robot, including individual explorations of the *Elicit* and *Read* branches. This was followed by an investigation into the relationship between a robot's level of interactivity and its effectiveness at achieving its goal(s), specifically the value of the *Elicit* and *Read* branches. The developed methodology and the findings from the exemplar instance were then shown to generalise through an empirical evaluation with a sociocontextual cue and non-humanoid social robot distinct from the exemplar cue and robot.

The specific breakdown of this thesis is as follows:

Chapter 2 presents background information necessary to understand the work presented in this thesis. This begins with an outline of the suitability of gaze cues as an exemplar sociocontextual cue for further investigation in HRI with an exemplar humanoid social robot (as per **RQ A.1: *Elicit* Feasibility**). This is followed by foundational information about these cues in HHI, against which their effects in HRI can be compared. Next, background information necessary to explore transferring and operationalising gaze cues in the *Elicit* and *Read* directions of communication during real-world HRI is outlined. These exemplar cues are subsequently leveraged in Chapters 4–6 as the foundation for explorations of the Robot Centric HRI paradigm and the devised methodology (Chapter 3).

Chapter 3 sets out the methodology which was developed as part of this work to enable successful operationalisation of the Robot Centric paradigm during real-world HRI, including the *Elicit* and *Read* branches and the interactivity of the robot (as per **RQ: Methodology**). This methodology is leveraged in the subsequent chapters to successfully operationalise the Robot Centric HRI paradigm to empirically explore the remaining research questions.

Chapter 4 describes a study of an exemplar social robot's ability to *Elicit* via issuing gaze cues during real-world HRI. Firstly, the measures developed to enable this study are described, followed by details of the methodology used to explore whether humans will respond to robot-issued gaze cues in line with human-issued cues, as outlined in psychology and behavioural science literature (as per **RQ A.1: *Elicit* Response**). Empirical results are then presented, and the chapter ends with conclusions and a discussion of the limitations of the study.

Chapter 5 details an investigation into the *Read* branch of the Robot-Centric HRI paradigm. Firstly, a study focusing on understanding people's natural gaze behaviour towards social robots is discussed. From the resultant finding that *in situ* human gaze cue *Read*ing capabilities will have advantages to HRI (as per **RQ A.2: *Read* Value**), a head yaw estimation framework was developed (as per **RQ A.2: *Read* Feasibility**) and is next described. The methodology and results of an empirical evaluation of the framework are presented, then the chapter ends with conclusions and a discussion of the limitations of the framework.

Chapter 6 presents a study which explores whether a robot's effectiveness at achieving its goal(s) will be increased by adding to its ability to moderate its *Elicit* strategy based on information gained through *Read*ing (i.e. increased interactivity, **RQ B: Interactivity**), while simultaneously extending and addressing the shortcomings of the previous chapters' *Elicit* and *Read* studies. The methodology of exploring the value of the *Read* and *Elicit* branches of the Robot Centric HRI paradigm is presented; specifically, the wider effects of robot-issued cues are investigated, as well as the way people's gaze behaviour impacts on the effects and perceptions of such cues. Results of the study are discussed, then conclusions and a discussion of the study limitations are detailed.

Chapter 7 describes a real-world, externally valid social HRI study which demonstrates the generalisability of the developed methodology and the findings from the exemplar cue and robot presented in previous chapters. The methodology and results of an empirical evaluation with a sociocontextual cue, social robot and application space distinct from the exemplar instance are detailed, demonstrating that the methodology can be drawn on to successfully operationalise the Robot Centric paradigm during real-world HRI (as per **RQ: Methodology**).

Chapter 8 highlights the contributions of this thesis and draws conclusions from the findings. The limitations of the work presented in this dissertation are then discussed, and avenues for future work are proposed.

Chapter 2

Background and Aspects of the Transferability of an Exemplar Sociocontextual Cue to HRI

In order to address the research questions posed in this work, it is first necessary to narrow the wide exploration space of sociocontextual cues with social robots to an exemplar cue which can be thoroughly explored with an exemplar robot. This chapter begins in this way, outlining the suitability of sociocontextual gaze cues for further investigation in HRI. A foundational understanding of the importance, characteristics and effects of gaze cues during HHI is then given, providing a baseline to which gaze cues in HRI with an exemplar humanoid social robot can be compared. Next, background information necessary to explore transferring and operationalising gaze cues in the *Elicit* and *Read* directions of communication during real-world HRI with a humanoid social robot is outlined.

2.1 Bodily Sociocontextual Cues in HHI

During HHI humans display and perceive a wide spectrum of sociocontextual cues, from complex bodily cues to simpler, finer cues. In order to investigate the feasibility of developing a methodology to operationalise the Robot Centric HRI paradigm, beginning with exploring the characteristics and effects of sociocontextual cues in HRI, it was necessary to narrow the focus down to a single exemplar cue which could be thoroughly investigated in HRI. It is known that when and how cues are utilised during HHI varies with the situation in which the interaction takes place [110], necessitating the selection of a cue appropriate for the exemplar higher-HL, humanoid social robot which was available and thus utilised during a majority of the work presented in this thesis.

A bodily cue which is likely congruent with humanoid social robots' HL is gaze. Gaze is also sociocontextually important to many HHI social situations: humans are sensitive to the social significance of their gaze and the gaze of others, with studies of the relative weight of different sociocontextual cues during social interactions showing that facial and gaze behaviour play a major role [163]. This sensitivity begins in infancy: it has been demonstrated that infants as young as 3 months old can detect the direction of adults' gaze, and that this perception of deviated gaze induces corresponding shifts of infants' gaze direction [25, 63, 149]. This ingrained gaze behaviour, the typical observability and detectability of the head region [85], and the significance of gaze to social interaction, makes it an ideal exemplar selection for further investigation in HRI with the exemplar humanoid social robot.

2.1.1 The Role of Gaze Cues During Interactions

In order to explore the extent to which the characteristics and effects of gaze cues in HRI will correspond to those of human-human cues, it is first necessary to understand the functions of gaze during HHI, in order to establish a baseline against which gaze cues in HRI can be compared. It is known from the fields of psychology and behavioural science that gaze is an important cue for supporting successful human-human interaction during all stages of interactions. For example, during the initialisation of interactions, gaze can signal openness to and/or desire for interaction, as mutual gaze can indicate when each constituent is attending primarily to the other and that further interaction can proceed [48, 90].

Similarly, gaze has several main functions during situated interaction. As an indicator of attention direction, it can reveal attention on, level of interest in and intentions towards both people and objects in the environment, while mutual gaze indicates attentiveness to an established interaction partner [86, 90, 112].

Gaze also plays a vital role in shaping people's preferences for objects: when the 'mere exposure effect' (the more we look at something, the more we like it) and 'preferential looking' (we look more at things we like), interact in a positive feedback loop leading to a conscious decision (the 'gaze cascade effect') [41, 126, 137, 143, 173], preference influence can occur. Analysis of the 'gaze bias' that develops can be an accurate predictor of a person's eventual preference, as it has a number of characteristics which emerge during a decision-making process.

2.1.2 Dynamics of Gaze and the *Interaction Zone*

Given the perception of robots as interaction peers, it is reasonable to assume that during HRI with a higher-HL humanoid social robot gaze will have similar functions to the HHI functions discussed above. In order to examine the equivalency of gaze cues during both HRI initialisation and situated HRI, it is necessary to understand how these cues are employed in such HHI situations. It is known that an important moderating factor of cues in HHI is the three-dimensional physical configuration of interaction partners: the spacial arrangement of individuals impacts both the formation of cues and how they are interpreted by the addressee during both interaction initialisation and situated interaction [76, 94, 122, 123]. Therefore, a comprehensive analysis of sociocontextual cues in HRI cannot be achieved without understanding and accounting for this factor [13].

The framework which describes how spacial configuration affects communication is known as proxemics, a term coined by anthropologist E. Hall [56, 57]. Proxemic theory describes interpersonal spacial relationships between individuals, with a focus on how physical distance affects when and how people interact and communicate with those around them. In particular, the framework correlates physical distance to 'social distance', the level of comfort and familiarity between interaction partners. This social distance is categorised into four discrete distance zones, each of which is characterised by a progression of interactions ranging from highly intimate to public [57], moderating the functions of sociocontextual cues such as gaze. The zones, illustrated in Figure 2.1, consist of the intimate ($\sim$0–0.5m), personal ($\sim$0.5–1.2m), social ($\sim$1.2–3m), and public zone ($>\sim$3m).

Proxemic cues, that is, how people position themselves in these zones, are exploited during the initiation and course of interactions, as people greet and engage in situated social interaction only when another person is located within an appropriate proxemic zone [42, 98]. For example, subtle gaze cues are difficult to detect at public zone distances, and this zone is therefore generally outside the reach of interaction potential [57, 163]. On the other hand, a majority of situated interactions occur within the social zone, in which gaze cues and their perception play an important role in the initiation and course of interactions [54, 57].

In addition to this, individuals have a 30° wide "transactional segment" directly in front of them in which a majority of their activities occur [76], also depicted in Figure 2.1. During social interactions, individuals typically position themselves such that they are separated by social zone distances [57] and their transactional segments overlap [76]. Thus, when a person's transactional, forward-facing zone of attention [61] is projected over their proxemic zones, the overlap between the social proxemic zone and transactional segment can be considered a person's *interaction zone*, illustrated in Figure 2.1, within which social interactions are more likely to take place and gaze cues play a particularly important role in communicating during interactions. There are two key subdivisions within the *interaction zone*, in which distinct types of interaction tend to take place: interaction initiation likely occurs in the far-*interaction zone* ($\sim$2–3m in the x direction, $\pm\sim 1m$ in the y direction), while situated interactions are more likely to take place in the near-*interaction zone* ($\sim$1.2–2m in the x direction, $\pm\sim 0.5m$ in the y direction) [77, 119].

2.1.3 Mutual Gaze and Joint Attention

The social *interaction zone* described above is also the proximity at which HRI with social robots is likely to take place; humans are typically most comfortable carrying out situated human-robot interactions at far-personal and near-social zone distances, the larger distances being more comfortable with robots having greater humanoid appearance and higher levels of autonomy, such as social robots [67, 155]. As such, it becomes necessary to determine the gaze cues which typically support the common interaction types in this zone – interaction initiation and situated interaction – which are likely to play similarly important roles in HRI. Mutual gaze and joint attention (JA) are, respectively, two gaze cues which are particularly important to these interaction types. These are therefore convenient cues to explore in order to determine if such sociocontextual cues can be

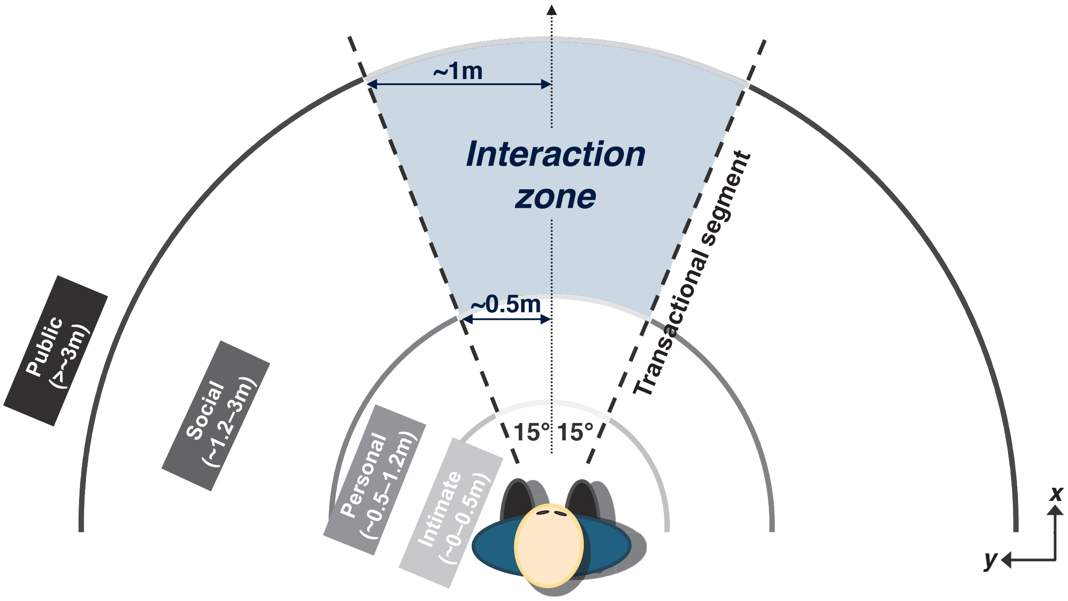


Figure 2.1: Hall's proxemic zones [57] depicted along with Kendon's transactional segment [76], the overlap of which is a person's likely *interaction zone*.

reliably implemented and utilised in HRI. To serve as a foundation for this HRI investigation, the characteristics and effects of mutual gaze and JA in HHI must be understood.

Mutual Gaze

Mutual gaze – gazing at another person's face [86] – is commonly used in HHI in the far-*interaction zone* as a means of establishing union between potential interaction partners during interaction initiation [144]. Mutual gaze indicates whether it is appropriate to initiate an interaction, signalling attentiveness and openness to interaction [26, 62]. You may be the recipient of another's gaze, for instance, because you are someone with whom they would like to interact [90].

However, this mutual gaze is unlikely to be steady or constant: it is known from studies in the fields of psychology and behavioural sciences that during social interaction gaze is directed at interaction partners repeatedly but for short periods [8]. This gaze behaviour (including frequency and duration of gaze) varies widely between individuals and depends on a number of mutually-moderating factors [34, 35, 150] which can be broken down into four broad categories [8, 86]:

1. *Personal* – age, sex, personality, culture, and clinical diagnoses such as autism (e.g. [37, 43])
2. *Experiential* – history and mood (e.g. [116])

3. *Relational* – interpersonal sentiment, intimacy and dependency, and perceptions of competence and power hierarchy, including aggression and dominance (e.g. [38, 58, 99])
4. *Situational* – interaction setting, physical proximity and orientation, task and motivation (e.g. information or response seeking, competition or cooperation), affective nature of the interaction, and intimacy of task/topic (e.g. [52, 92]).

The dynamics of mutual gaze in HHI are thus a complex function of individual and environmental variables which interplay to result in large gaze behaviour variations between individuals: the percentage of HHI encounter time spent gazing at an interaction partner can range from 28% to over 70% [75], with glance lengths in the range of 3–10s [8]. It is thus likely that similar mutual gaze characteristics will be displayed by humans during HRI, further necessitating *in situ* gaze *Read*ing capabilities for the robot, as further discussed in Section 2.3.

Joint Attention

While mutual gaze is particularly important during interaction initiation, it is also the establishing cue of joint attention (JA), a gaze cue commonly utilised during situated interactions in the near-*interaction zone*. An example of JA, the shared focus on an object or location, is depicted in Figure 2.2: the image shows three people who appear to be interacting, but also seem to be mutually engaged with and sharing focus on the salient object, the laptop. As can be appreciated by viewing the image, JA is an important tool in understanding others' minds, establishing reference for communication and/or marking desire for or intention to act on an object [14, 36].

The sensitivity of humans to the gaze of others, and the attention following that results, is the foundation of JA. It has been shown that infants as young as 3 months old can detect the direction of adults' gaze via mutual gaze, and that the perception of deviated gaze induces corresponding shifts of infants' attention direction [25, 63, 149], entering them into JA. Figure 2.3 shows young children engaged in JA, and this reflexive attention-orienting to a gazed-at location continues into adulthood: [45] and [46] demonstrated that adults respond more quickly to targets appearing at a location that is gazed at by a centrally presented face than to targets appearing at a location that is not gazed at, suggesting that the human brain may be specialised to shift attention in response to others' gaze.

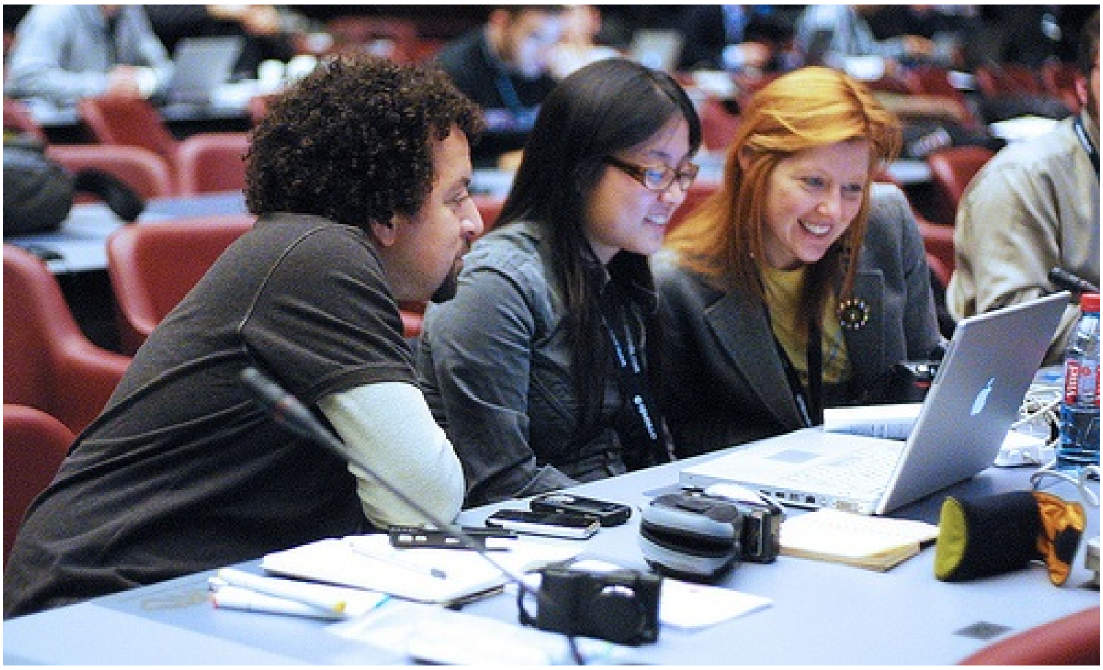


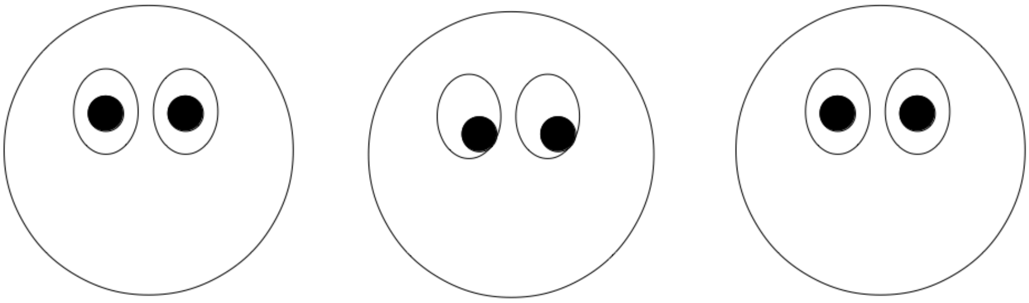
Figure 2.2: Three people engaged in joint attention, the shared focus on an object or location. Source: [145]

Joint attention can also influence preference by affecting the way we evaluate visual stimuli in the environment [14], as objects that fall under the gaze of others acquire properties that they would not display if not looked at [36]. For example, the direction of another person's gaze can provide information about relevant events and objects within the environment, signal which objects are of current value, and also transfer to the object the intentionality of the person looking at it [14, 15]. Thus, observational learning about specific objects in the world would be impossible without gaze following [36]. By triggering this enhanced information processing about objects in observers [131, 132], JA can affect the observers' evaluation and affective appraisal of objects in the environment. This results in objects that are looked at by others being more likeable than those that do not receive much attention [14, 15, 159]. This effect further translates into an increased preference for JA objects, as an enriching of the objects results from the intentionality of the perceived gaze (i.e. observing another person looking at an object) [15] and the tendency for people, on observing the actions of others, to spontaneously adopt the goals that may account for these actions [1].

However, it is also known that a specific sequence of gaze shifts must be displayed for this enriching effect and influence on object desirability to take place. The first step of this specific object-enriching cue is to signal an intention to communicate with an observer [159]. This can be achieved via a number of

Figure 2.3: Young children engaged in joint attention.
Source: [9]

means (e.g. verbally or through gestures), however via gaze this can be conveyed through mutual gaze from a communicator to an observer [156]. This mutual gaze at an observer is shown in Figure 2.4a. Next, to signal that an object is of value to the communicator, the gaze should be shifted from an observer towards the object, as attention naturally settles on desirable objects in the environment [15, 142]. This second, object-directed gaze step is shown in Figure 2.4b. Finally, gaze should shift back to the observer, in order to signal that the object is also of value to them [149] (shown in Figure 2.4c). The complete object-enriching cue can be appreciated by viewing Figure 2.4.



(a) Direct gaze (b) Object-directed gaze (c) Direct gaze

Figure 2.4: The three-step joint attention sequence to increase object desirability.

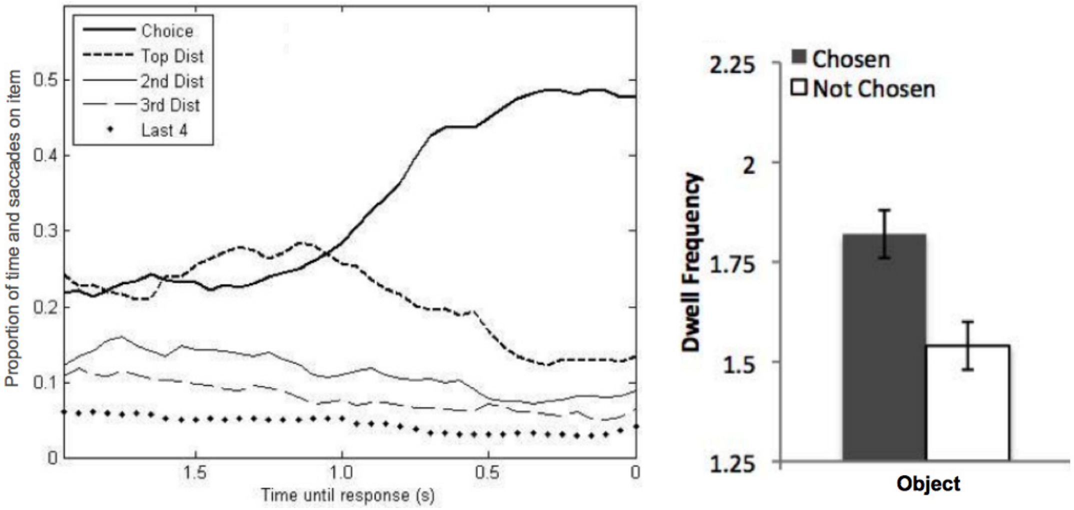
It has been found that only through observation of this specific sequence (henceforth called a JA cue, for brevity) – as opposed to direct gaze followed by gazing at the object (i.e. without looking back to the observer) – is object desirability increased.

2.1.4 Characteristic Effects of Joint Attention

A pre-requisite to investigating whether JA will operate as described above during HRI is an understanding of the effects of JA on observers in HHI, which can then be used as a baseline against which JA in HRI can be compared. Consequently, it becomes necessary to understand the typical characteristics of responses to JA in HHI. The first step is to consider a scenario in which the effects of JA are known and able to be observed. One such scenario is decision-making situations, as JA is known to have the power to influence preference. Thus, as a baseline for what characteristics and effects could be expected in HRI decision-making situations, an understanding of the characteristics of JA's effects in HHI decision-making situations, which manifest in gaze behaviour, becomes necessary.

As previously discussed, it is known from work in the fields of psychology and behavioural science that in human-human decision making scenarios, a person's gaze behaviour is a revealing characteristic: it contains information about attention direction and intentions towards objects in the environment [86], and plays a vital role in shaping people's preference for objects. This effect on preference can occur when the 'mere exposure effect' (the more we look at something, the more we like it) and 'preferential looking' (we look more at things we like), interact in a positive feedback loop leading to a conscious decision (the 'gaze cascade effect') [41, 126, 137, 143, 173].

The 'gaze bias' that develops can be an accurate predictor of a person's eventual preference, as it has a number of characteristics and patterns which emerge over the course of the decision-making process. As the decision moment approaches, for example, the breadth of the person's visual search decreases. This includes a decrease in the number of alternatives fixated on and the duration of each fixation. This search narrowing can be seen in Figure 2.5a, which illustrates how the proportion of time and saccades to unchosen objects decreases as the decision moment approaches. This is accompanied by a corresponding increase in the focus of the search, as the duration of each fixation and the proportion of total eye fixations on the chosen alternative rise [51, 135, 137, 143]. Finally, the number of saccades to the object which will be selected increases [100, 126]. These characteristics are shown in Figure 2.5b and Figure 2.5a: Figure 2.5b depicts how the frequency of looking (i.e. dwell frequency) at the preferred object is higher than the dwell frequency on the not preferred object, and Figure 2.5a shows how the proportion of time and saccades to the preferred object increases during search narrowing as the decision approaches.



(a) Plot of the proportion of gaze times and saccades which were directed at chosen and other (distractor) items during visual search of mock-company logos

(b) Dwell frequency of visual search of photographs during a two-alternative forced choice experiment

Figure 2.5: Visual search patterns during HHI decision-making scenarios. Image (a) source: [51], (b) source: [137]

Thus, such gaze characteristics are an indicator of preference in decision-making scenarios, and can be summarised in a number of measures:

1. Number of fixations on alternatives
2. Duration of fixations on alternatives
3. Proportion of total eye fixations on alternatives
4. Proportion of time fixated on alternatives.

These known characteristics and effects of JA in HHI are useful as a baseline for exploration of JA cues in HRI.

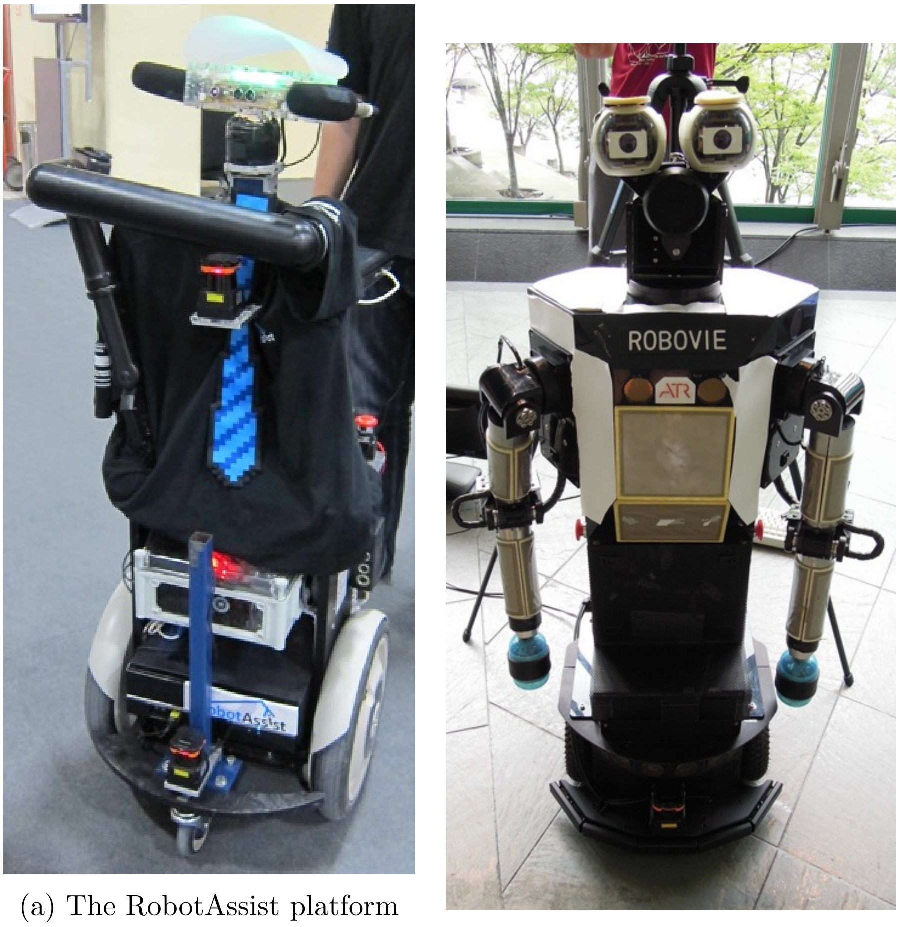
2.2 Feasibility of Robot-Issued Gaze Cues

As discussed above in Section 2.1, the importance and known characteristics of gaze cues in HHI makes them ideal for investigation in HRI in the *Elicit* and *Read* directions of communication. Beginning with robots *Elicit*ing particular responses from human interaction partners via issuing mutual gaze and JA cues, the first question which arises is: are today's social robots physically capable of issuing recognisable gaze cues (**RQ A.1: *Elicit* Feasibility**)? It thus first becomes necessary to understand the feasibility of social robots issuing these gaze cues during HRI.

2.2.1 Human-Likeness of Exemplar Humanoid Social Robots

As previously mentioned in Section 1.1, social robots sit along different points of the human-likeness continuum, ranging from disembodied and overtly non-human robots to perfect androids. Given the increased perception of social robots' social capabilities, which arises from increasing levels of HL, it is important to balance people's expectations and the robot's capabilities [31]. As such, it becomes necessary to investigate JA cues on a social robot with appropriate HL, for example a humanoid robot. Today's typical humanoid social robots have not progressed to the point where they look identical to and are indistinguishable from humans: Figure 2.6a and Figure 2.6b show two such typical robots from leading robotics research laboratories: Figure 2.6a depicts the RobotAssist platform from the Centre for Autonomous (CAS) Systems at the University of Technology, Sydney (UTS) [76], and Figure 2.6b shows Japan's ATR Intelligent Robotics and Communication Laboratories' Robovie platform [70, 72]. From examination of these images, non-human elements of the robots can be seen: for example, both sit on wheeled platforms rather than legs, use cameras as 'eyes', and have some degree of visible cabling and sensory hardware.

However, the HL of such robots is also apparent. They are recognisably humanoid in shape, with distinguishable upper and lower bodies, arms, heads and 'eyes' – many of the body parts necessary to issue bodily sociocontextual cues. The RobotAssist platform seen in Figure 2.6a, for example, has a 2 degree-of-freedom (DOF) head (pan-tilt) actuated by two servos, in which a Microsoft Kinect [105] is positioned. Similarly, the Robovie head contains two cameras as ‘eyes’ for binocular, stereo vision, which are actuated via 2×2 DOF (pan-tilt) for gaze control and/or the 3 DOF head (pan-tilt-roll). These configurations enable the sensors in the two robots’ heads to be moved quasi-independently from their respective platforms. While this capability is vital to the functionality of the robots, facilitating active sensing [3], it also gives them the physical ability to issue static and dynamic gaze cues. With such humanoid appearances and cue-issuing capabilities, combined with their perception as interaction peers, it is reasonable to assume that the HL of such robots will be such that bodily gaze cues they issue will have similar effects and characteristics to human-issued gaze cues during HHI (described in Section 2.1.3 and Section 2.1.4), as depicted in Figure 2.7.



(a) The RobotAssist platform [82]

(b) The Robovie platform [70, 72]

Figure 2.6: Two leading social robotics research platforms, illustrating their humanoid shape.

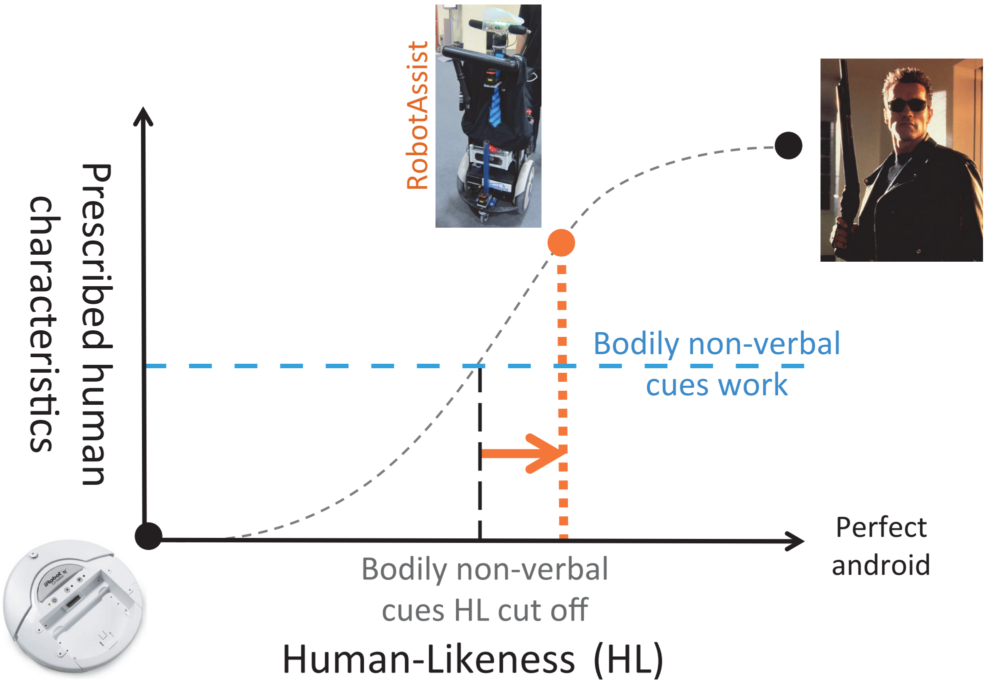
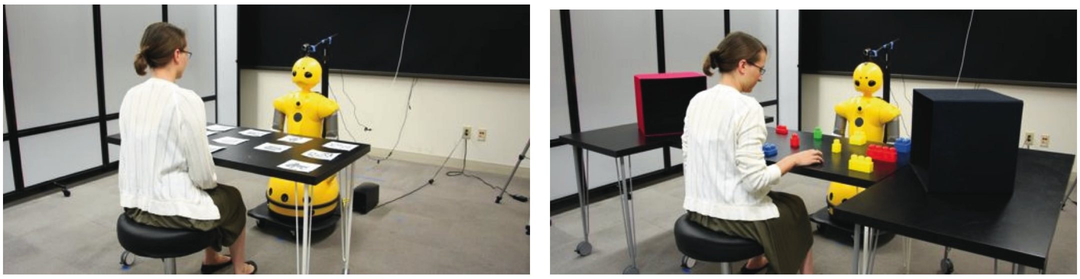

RobotAssist
Prescribed human characteristics
Bodily non-verbal cues work
Perfect android
Bodily non-verbal cues HL cut off
Human-Likeness (HL)

Figure 2.7: The humanoid shape and capabilities of the RobotAssist platform make it likely that gaze cues issued by the robot will have similar effects and characteristics to gaze cues issued during HHI.

2.2.2 Robot-Generated Joint Attention Cues

With humanoid social robots such as the RobotAssist platform having the physical potential of issuing gaze cues, in order to explore mutual gaze and JA in HRI it becomes necessary to understand whether these cues will be recognisable by humans when they are issued by robots. In fact, JA cues for robots is an established concept, and a great deal of research has been done into the physical operationalisation and outcomes of JA in human-robot interaction and collaboration (for example in [27, 29, 59, 66, 68, 73, 128]). As such, human-recognisable JA cues for robots have been designed by others' as a necessity for their work. For example, Figure 2.8 shows a robot executing two of the three JA steps required to increase object desirability during an experiment carried out by [65]. In Figure 2.8a the robot engages in mutual gaze with the observer, then executes an object-directed gaze in Figure 2.8b. Not shown, but detailed in the paper, the robot then gazes back at the observer, thus completing the JA cue to increase object desirability.

Similarly, JA cues have previously been constructed for the RobotAssist platform as part of other work, detailed in [81]. The authors constructed the cues based on the three-step JA cue depicted in Figure 2.4 and findings in work such as



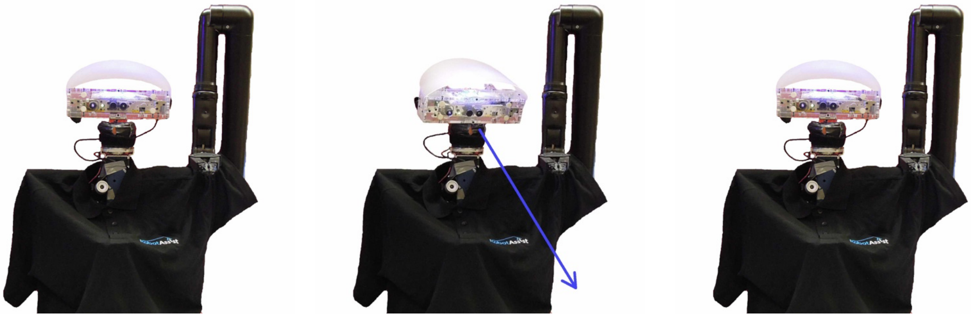
(a) Direct gaze from robot to human

(b) Robot object-directed gaze

Figure 2.8: A robot displaying two steps of the three step sequence to increase object desirability during an experiment.
Source: [65]

[15, 81, 142, 149, 159]. Figure 2.9a shows the RobotAssist platform executing the establishing action of the JA cue, which, as discussed above, consists of mutual gaze from the communicator to the observer. To achieve this, the robot orients its 'eyes' (in the sense of the entirety of its head, rotated about the vertical axis via the pan-DOF in the robot's neck) along the vector between itself and the participant. Following this, the attention guiding, object-directed cue is achieved by lowering the 'eyes' around 20°–30° (as shown in Figure 2.9b) and then again raising them (as in Figure 2.4), as can be seen by viewing Figures 2.9a, 2.9b and 2.9c sequentially. The authors' work has shown that for this action to be reliably perceived as JA the 'eyes' must move at approximately $150°/s$, as slower speeds can lead to the cue being interpreted as individualising [81].

Figure 2.10a and Figure 2.10b show, as per Figure 2.9b, what an observer would see if the RobotAssist platform issued the JA cue at a left or right object, respectively. As Figure 2.9 and Figure 2.10 illustrate, the RobotAssist platform is physically capable of issuing the recognisable JA cue necessary to increase object desirability (as per **RQ A.1: *Elicit* Feasibility**).

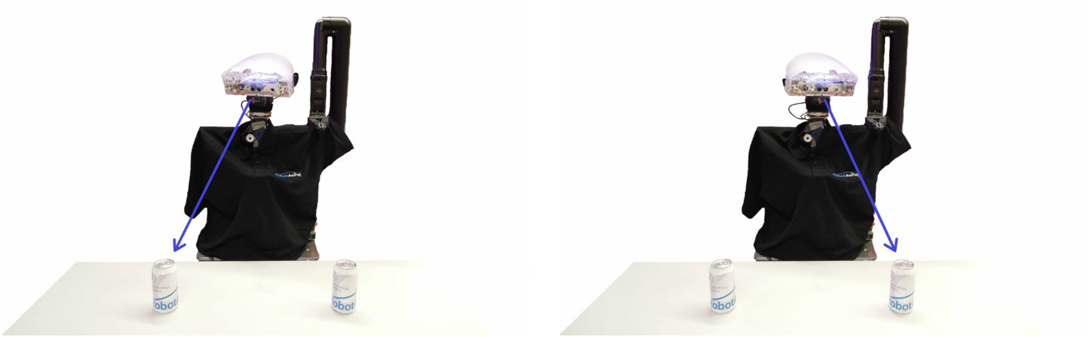


(a) Direct gaze

(b) Object-directed gaze

(c) Direct gaze

Figure 2.9: The RobotAssist platform executing the three-step JA sequence.



(a) Joint attention left (JAL) (b) Joint attention right (JAR)

Figure 2.10: The RobotAssist platform executing joint attention left (JAL) and joint attention right (JAR) cues.

The question then becomes, as per **RQ A.1: *Elicit* Response**: will people respond to RobotAssist platform-issued JA cues in line with they way they respond to human-issued JA cues? As discussed in Section 2.1.4, the characteristics and effects of JA in HHI decision-making situations are well known; in equivalent HRI decision-making scenarios, through observation of human interaction partners' gaze behaviour, the measures described in Section 2.1.4 can be used to determine if a robot's JA cues are operating equivalently. Thus, the characteristics of the influence of robot-issued JA cues on an observer are able to be investigated.

2.2.3 Summary

The shape and capabilities of humanoid social robots, such as the RobotAssist platform, mean such robots are physically able to issue human-recognisable bodily gaze cues such as the JA cue necessary to increase object desirability (as per **RQ A.1: *Elicit* Feasibility**). Literature from the fields of psychology and behavioural science outlined in Section 2.1.4 suggests that the influence of this JA cue on human interaction partners should manifest as an effect on their gaze behaviour, an indicator of preference. Thus, equivalent effects and gaze characteristics should be observed during JA in HRI in this exemplar situation. There is therefore a need to explore the extent to which JA in HRI will mirror HHI JA and thus the characteristics, if any, of its influence in HRI. As HRI is reciprocal, it then becomes necessary to understand whether it is possible for social robots to *Read* and interpret such gaze cues displayed by interacting humans.

2.3 Considerations Surrounding Sensing Human-Issued Gaze Cues

Literature presented thus far has given a baseline of expected human gaze behaviour around humanoid social robots, including large mutual gaze variations between individuals during interaction (Section 2.1.3) and gaze bias in decision-making scenarios when the robot is *Elicit*ing via JA cues (Section 2.1.4). In order to understand the equivalency of gaze cues in HRI, the reciprocal communication direction must also be understood; that is, the ability of the robot to *Read* these human-issued gaze cues.

The first question that arises is whether people will display these exemplary HHI literature-predicted gaze cues around robots during real-world HRI, necessitating *in situ* cue *Read*ing capabilities for robots (as per **RQ A.2: *Read* Value**)? Given the growing perception of robots as interaction peers and their increasing human-likeness and ability to communicate in socially sensitive manners, it is likely that the answer to this will be 'yes'. While this is still worth empirically determining, it is reasonable to assume that it will be necessary for robots to have the *in situ* ability to *Read* human-issued gaze cues, among other cues.

The question then becomes: will human-issued gaze cues be detectable and interpretable by social robots (as per **RQ A.2: *Read* Feasibility**)? It is foreseeable that in real-world HRI there will be situations in which a person's eyes (and hence their exact gaze) are not observable. However, gaze estimation is intrinsically linked with head pose (the orientation of a person's head in object space, which is more observable), as the perceived direction of gaze is highly influenced by the pose of the head [89, 169, 172]. Before sensitivity to eye orientation develops in infants, for example, they use head direction in isolation as a directional cue [23]; the relative contributions of head movement to gaze re-orientations is related to the do the eye movement which would have been required if the gaze shift had been performed without a head movement, as head movement amplitudes are subconsciously selected to return the eyes towards central position [146].

It is also known that the head contributes preferentially to the horizontal component of gaze shifts, with head movement amplitude decreasing and eye movement amplitude increasing as gaze shifts are directed away from the horizontal meridian. The slope of the generally linear relationship between head and gaze movement decreases both as gaze re-orientations are directed away from the horizontal plane, and as gaze direction becomes increasingly vertical. Instead, in such cases where gaze re-orientations are more vertical, eye movement is instead used to compensate for this lower head movement [44].

It is likely, though, that objects and areas of interest during HRI, at which gaze would be directed, will tend to be located on a relatively planar area, for example on counters, tables, or desks. Thus, in foreseeable social HRI applications, when the requirement is often to detect which landmarks gaze is directed at, gaze can be reasonably approximated through the use of head yaw alone [36, 42, 91, 138, 148, 171]. As a result, at far-*interaction zone* distances when gaze cues are likely to be limited to interaction initiation, it is necessary for robots to be capable of detecting mutual gaze via head yaw estimation (HYE). During situated interactions when interaction distance decreases into the near-*interaction zone*, more accurate HYE is necessary [147] to detect cues such as JA and gaze bias.

In order to enable HYE which can successfully operate over the entirety of the *interaction zone* to *Read* these human-issued gaze cues, it first becomes necessary to determine which tools can be leveraged to achieve this. This begins with identifying facial features suitable for HYE given the capabilities of Kinect sensors, especially in the far-*interaction zone*, to later fuse the complementary information of these features.

2.3.1 Head Features Exploitable for Head Yaw Estimation

The human head has a number of features which may be suitable for HYE in the *interaction zone* to enable robots to *Read* interaction partners' gaze cues. Many of these features have been previously used for HYE, as summarised in [112], including:

- *Eyes* – location of inner and outer corners and size (e.g. [2, 53])
- *Nose* – nostril and tip location (e.g. [101])
- *Lower face* – chin, jaw line, or mouth (e.g. [140])
- *General head shape* – width or roundness (e.g. [21, 93])
- *Ears* – for profile face detection (e.g. [161]).

HYE via a combination of features is also common, e.g. eyes and mouth [95], eyes, nose and mouth [47], eyes and nose [152, 161], and cylindrical head model and eye locations [158].

2.3.2 Characteristics of Sensors Available for Head Yaw Estimation

In order to exploit these features for HYE, they must first be detected in the data generated by sensors suitable for social robots. Many of these sensors have sensing characteristics which can effect how robustly this feature detection can take place, characteristics which should be understood before the determination of features which can be utilised to achieve HYE.

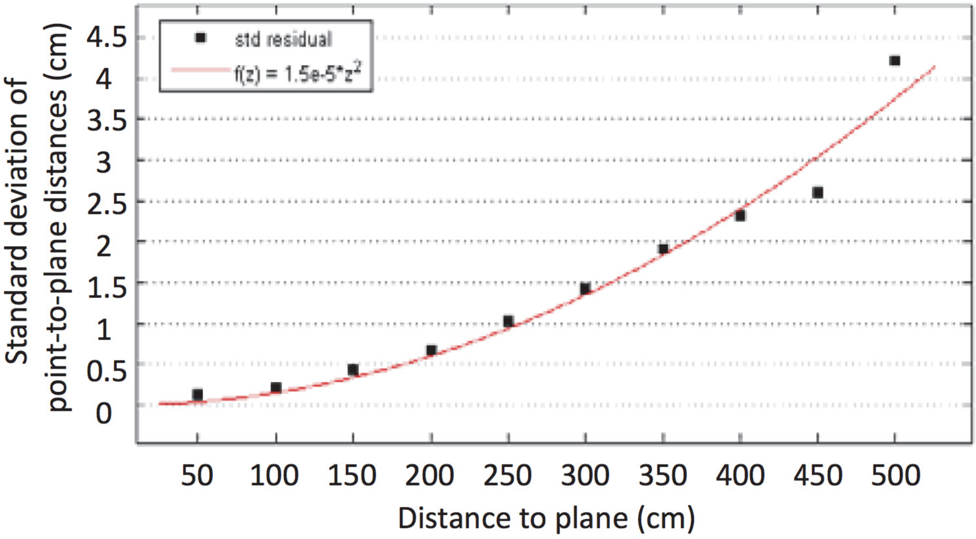


Figure 2.11: Standard deviation of plane-fitting residuals at different distances of the plane to the sensor. The best fit quadratic curve is plotted in red.
Source: [79]

One such sensor is the Microsoft Kinect, which is commonly used in today's social robots, including the RobotAssist platform. According to its specifications, the operating range of the Kinect is ∼0.5–5m, and [70] give an overview of the geometric qualities of its depth data in this range through analysing both the resolution (points per unit area density) and accuracy of the points. The resolution of the Kinect's depth data on the XY plane (perpendicular to the camera axis) is determined by the resolution of its infrared camera; that is, its pixel size. The constant 640×480 pixels of the depth image mean that the point density will decrease with increasing distance of the object from the sensor; specifically the point density p is known to be inversely proportional to the square of the distance from the sensor, Z, as in, $p \propto \frac{1}{Z^2}$. Thus, at a range of 2m, the depth resolution is ∼1cm, while at 5m the resolution is ∼7cm [78].

In addition to decreasing depth measurement resolution at greater distances, the random error e of the measurements increases. This random error is known to increase proportionally with square distance from the sensor according to the formula $e \propto Z^2$. As shown in Figure 2.11, this error e ranges from a few millimetres at $0.5m$ from the sensor up to $\sim 4cm$ at the maximum range of $5m$, where it is further influenced by the low resolution of the depth measurements [78].

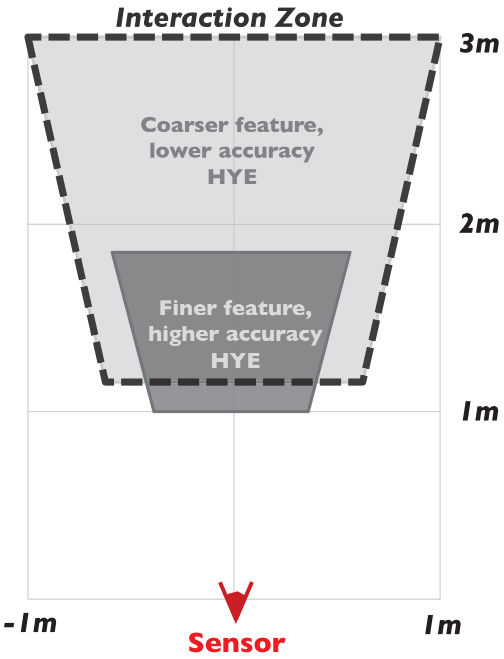
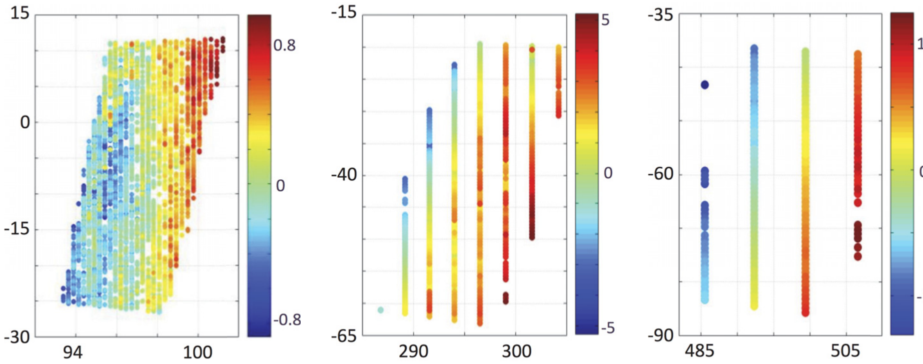


Figure 2.12: Typical operation spaces of coarser and finer facial feature HYE methods.

The lower point resolution and increasing measurement error of the Kinect data make it difficult to observe finer facial features at larger distances. For example, at a $2m$ range from the sensor, well within the *interaction zone*, the depth point resolution of $\sim 1cm$ in combination with the measurement error of $\sim 0.7cm$ result in an uncertainty of $\sim 1.7cm$; robustly detecting eye corners or nose tips in such data would be a considerable challenge. Thus, the usability of methods which rely on these features for HYE will be decreased at these distances as the data becomes increasingly featureless.

As a result, while they tend to have higher accuracy, the usability of methods which rely on finer facial features is decreased at far-*interaction zone* distances as the data becomes increasingly featureless. The operation space of such methods is therefore typically limited to the near-*interaction zone*, as depicted in Figure 2.12.

However, these geometric data qualities can also be exploited for far-*interaction zone* HYE; at larger distances, the combined effect of the random error and low depth resolution result in surfaces perpendicular to the sensor becoming stratified and appearing more featureless and planar [79], as shown in Figure 2.13. As a result, features such as the facial plane become artificially exaggerated in the data at larger distances, are therefore more likely to be observable. Thus, a HYE method based on these features, while similarly susceptible to noise at larger distances and possibly less accurate, is likely to be usable in a wider area of the *interaction zone* than those based on finer facial features, as illustrated in Figure 2.12.



(a) 1m (b) 3m (c) 5m

Figure 2.13: Point cloud of a planar surface at different distances from the sensor, projected on the YZ plane. Colours represent distance to the best-fit plane in centimetres.
Source: [79]

Thus, HYE methods based on finer and coarser facial features have distinct strengths: accuracy vs coverage of the *interaction zone*, respectively. As such, it would be ideal to fuse, and in doing so exploit any correlations between, such complementary data in order to leverage these strengths into a head yaw estimate which operates across the entirety of the *interaction zone*. Such an estimate would fulfil the HRI requirements of detection of coarser, mutual gaze in the far-*interaction zone*, and finer, more accurate gaze detection during situated interactions in the near-*interaction zone*.

2.3.3 Leveraging Multiple Imperfect Head Yaw Estimates

Before this can be achieved, it becomes necessary to understand ways in which this complementary data could be fused. A number of tools exist which could be used to achieve this fusion, including Bayesian fusion and Gaussian processes.

Bayesian Fusion

Bayesian fusion is a technique which probabilistically fuses input entities in a statistically sound manner. In the intended application, Bayesian fusion could be used in such a way as to actively select a more reliable head yaw estimate in a particular area on the basis of the uncertainty associated with each estimate. For example, in areas where a finer feature, higher accuracy head yaw estimate is available, its estimate could be 'trusted' over the coarser feature, lower accuracy head yaw estimate, maintaining the estimate's higher accuracy. When no finer feature estimate is available, the coarser feature HYE can be used instead.

Gaussian Processes

A second technique which could be used for fusion is Gaussian processes (GPs) [127], powerful non-parametric Bayesian learning techniques. GPs are commonly used for both regression and classification. However, as discussed in [162], GPs are also increasingly used as a method to achieve data fusion (for example [33, 50, 160]): learning a joint model of multiple input entities enables a single inference to be made, inherently 'fusing' the input data in such a way as to exploit a relationship between the input entities through the regression model. This is a key difference between utilising GPs to achieve fusion and employing Bayesian fusion, which does not incorporate such a relationship during fusion unless it is specifically modelled.

GPs are stochastic processes in which any finite set of training data and test data are jointly Gaussian distributed. Statistical inference is used to learn dependencies between points in the input (training) dataset, thus incorporating and handling uncertainty and incompleteness in a statistically sound manner. In addition to reducing the amount of training data required to adequately train the model, this yields continuous representations such that inferences can be made between sparse data [160] and noise is inherently filtered out, an advantageous characteristic given the levels of Kinect data error discussed in Section 2.3.2.

Generally, the primary drawback of using GPs is their typical computation time; this is especially relevant in an HYE application when close to real-time performance is desirable. The complexity of GPs is $O(N^3)$, where N is the number of training data points. While the focus of this work was on validating the proposed framework, in the proposed application this computational online use bottleneck could be addressed in a number of ways, for example by pre-computing and storing K^{-1}, which is independent of query points, or by selectively training the model. This involves first training the model with a limited training dataset, then selectively adding training points by first testing them through the model to determine if they add useful information to the training (as indicated by their level of covariance) [121].

Thus, GPs are a viable solution for developing a joint model – and in this way 'fusing' – finer and coarser feature head yaw estimates and their relationship into a single estimate.

2.3.4 Summary

The perception of social robots as interaction peers makes it likely that humans will display HHI literature-predicted gaze cues (as outlined in Section 2.1.3 and Section 2.1.4) around social robots, necessitating that robots have *in situ* cue *Read*ing capabilities (as per **RQ A.2: *Read* Value**). Thus, there is a need for HYE which operates over the entirety of the *interaction zone*, depicted in Figure 2.1, while maintaining levels of accuracy necessary to detect gaze cues during real-world HRI across the entire zone. During real-world HRI, where people's eyes and hence exact gaze direction may not always be detectable, literature shows that head yaw can be used as an indicative measure of gaze direction. Many features of the human head could be used for head yaw estimation, however the reliability of distinguishing different facial features in data such as that from the commonly used Kinect depends on the interaction distance. In the *interaction zone* it would be advantageous to leverage multiple HYE methods which rely on different facial features. This data could then potentially be fused into a single head pose estimate via the use of a tool such as Bayesian fusion or Gaussian processes, giving a robot the necessary ability to *Read* both mutual gaze in the far-*interaction zone* and JA and gaze bias cues in the near-*interaction zone.*

2.4 Conclusion

This chapter has discussed the background of gaze cues in HHI – one example of sociocontextual cues – and aspects of transferring these cues to HRI in both the *Elicit* and *Read* directions of communication. Gaze cues have been selected as an exemplar cue for further exploration into whether sociocontextual cues can be reliably implemented and utilised in HRI due to: a) the necessity to narrow the wide exploration space of sociocontextual cues down to a single cue; b) the sociocontextual significance of gaze cues to social interaction; and c) their congruency with the higher-HL social robot available for this work.

In Chapter 3, a methodology is detailed for operationalising the Robot Centric HRI paradigm – including the *Elicit* and *Read* branches and the interactivity of the robot – to enable robots to leverage sociocontextual cues such as the exemplar gaze cues. In the subsequent chapters the exemplar gaze cue information outlined in this Chapter 2 is leveraged as a baseline against which the methodology is validated, and to enable exploration of sub-research questions **RQ A: Sociocontextual cues in HRI** and **RQ B: Interactivity**.

Finally, in order to demonstrate that the findings and methodology generalise, a distinct sociocontextual cue and social robot are utilised during the study detailed in Chapter 7.

Chapter 3

Methodology for Robot Centric HRI Paradigm Operationalisation

This chapter presents, in response to **RQ 1: Methodology**, a methodology which can be leveraged to successfully operationalise the Robot Centric paradigm during real-world HRI, a non-trivial task. In addition to addressing the primary research question, this methodology will also facilitate exploration of the sub-research questions **RQ A: Sociocontextual cues in HRI** and **RQ B: Interactivity**, as detailed in subsequent chapters. This is achieved by drawing on the information of Chapter 2, where the importance, characteristics and effects of exemplar sociocontextual cues during HHI were outlined, serving as a baseline against which such cues during HRI with an exemplar humanoid social robot can be compared.

3.1 Introduction

The primary research question of this work, **RQ: Methodology**, poses the question of whether a methodology can be developed which could be drawn on to successfully operationalise the Robot Centric paradigm during real-world HRI. A devised methodology could address this question, and also facilitate exploration of the sub-research questions: firstly, through isolated operationalisation of the *Elicit* and *Read* branches via the methodology, **RQ A: Sociocontextual cues in HRI** can be investigated. Next, **RQ B: Interactivity**, concerning whether a robot's effectiveness at achieving its goal(s) can be increased by greater levels of interactivity, can be addressed through leveraging the methodology to design different levels of interactivity for a robot. Thus, a methodology for operationalisation of the paradigm was devised, as detailed below.

3.2 Methodology for Paradigm Operationalisation

As mentioned in Section 1.1, operationalising the Robot Centric paradigm during real-world HRI is a non-trivial task, given the complex nature of both human behaviour and the dynamics of interaction.

To achieve successful operationalisation, the devised methodology consists of four main stages, as depicted in Figure 3.1. Initially, the target problem and robot goal(s) must be defined in order to determine what the robot is attempting to achieve in the application space and interaction. Next, the application space in which the human-robot interaction will take place must be defined in order to understand and account for any external influences on the interaction and/or human behaviour which may exist. Following this, the Robot Centric HRI paradigm itself must be designed for the particular interaction, including the *Read* and *Elicit* branches and the robot's interactivity, thus ensuring that the branches are successfully activated and an appropriate interactivity level is achieved. Finally, the implementation must be designed to ensure that factors and practicalities which may affect the operationalisation are considered. These stages are described in more detail in the following sub-sections.

3.2.1 Target Problem and Robot Goal Definition

The initial stage of the methodology is to define the problem which interaction between human(s) and a social robot is attempting to address, and the robot's goal(s) in order to address that problem. Through this, an understanding of what the robot is trying to achieve in the application space and interaction – and why – can be determined. For example, a socially assistive robot may not be physically able to retrieve one of two objects on a table; its goal may then be to influence a person towards selecting the particular object it can reach. To address the issue of congestion in a public transport environment such as a train station, a social robot's goal may be to guide people through physical space to increase the efficiency of passenger movement.

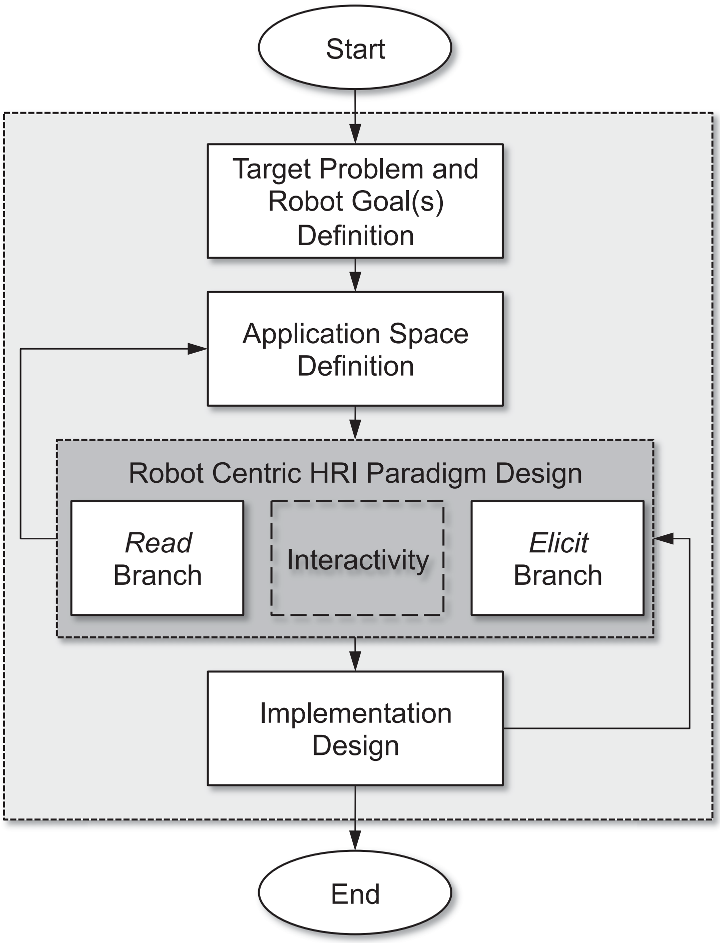

Start
Target Problem and Robot Goal(s) Definition
Application Space Definition
Robot Centric HRI Paradigm Design
Read Branch
Interactivity
Elicit Branch
Implementation Design
End

Figure 3.1: Process flow of the methodology for operationalisation of the Robot Centric HRI paradigm.

3.2.2 Application Space Definition

With an understanding of the problem and robot goal(s), it is next necessary to define the physical application space in which the interaction will take place, particularly its sociocontext (the contextual-task space and the social-interaction space [81]). There are two principal ways in which this can affect the interaction and design of the Robot Centric HRI paradigm:

1. *Other sources of thinking pattern and/or behaviour influence*

   Firstly, the application space sociocontext may result in particular influences on people's thinking pattern and/or behaviour, either promotion or suppression. For example, there may be physical obstacles which prompt people to walk along a certain path (a corridor closed for renovation, for instance), or social norms which discourage people from walking on the right side of a corridor (in left-hand-side driving countries).

An understanding of any such systematic influences in the environment is important to achieve effective and predictable operationalisation of the paradigm. If the sociocontext promotes a particular thinking pattern and/or behaviour which is in line with the robot goal(s), it may be that any influence the robot might have would be over-whelmed by the sociocontextual influences: for example, a robot directing passengers to take an alternative route when the direct route is already physically blocked. Thus, it may be unnecessary or of no value to put the robot in that application space at all.

On the other hand, the sociocontextual suppression of a thinking pattern and/or behaviour may be such that it prohibits the robot goal(s) from possibly being achieved. For example, a robot directing people to move right in a corridor may not be able to overcome the social norm of keeping left. In these cases, the robot goal(s) must be reconsidered (from Section 3.2.1), the application space itself must be altered, or the Robot Centric HRI paradigm design (discussed below in Section 3.2.3) must be configured to increase the effectiveness of the robot's ability to *Elicit* behaviour modification.

2. *Interacting people's familiarity with robots*

   A second consideration surrounding the application space is the familiarity of the people in the space with robots: it is known that the behaviour around robots of those familiar with robots (such as engineers) does not necessarily match the behaviour of those less familiar with robots (e.g. the wider-population) [31]. If the application space of the interaction is such that a large proportion of the interacting people will be familiar with robots – e.g. engineering/technology related – their behaviour and responses may not necessarily be in line with psychology and behavioural science predictions, which are more applicable to the general population. In such application spaces people's responses are thus likely to be less predictable, potentially compromising the effectiveness of the robot in achieving its goal(s). It may therefore be necessary to alter or reconsider the application space.

3.2.3 Robot Centric HRI Paradigm Design

The next stage of the methodology is to design the Robot Centric HRI paradigm, which is comprised of the *Read* and *Elicit* branches and the interactivity of the robot, as depicted in Figure 3.1. This process is further detailed in the following subsections.

Read **and** ***Elicit*** **Branch Design**

As previously discussed, the Robot Centric HRI paradigm positions robots as interaction peers more equal with humans in terms of agency and ability to lead interactions. Given this positioning of robots and humans on more equivalent levels, design of the *Read* branch (what the person is doing) and *Elicit* branch (what the robot is doing) of the paradigm have common core considerations, as shown in Figure 3.3, the perspective of which changes depending on whether the human (*Read*) or robot (*Elicit*) is the focus of the branch.

In designing the branches, these four core considerations are intention, behaviour, cues, and interpretation. These considerations are depicted in Figure 3.2, along with the flow between them in terms of designing the branches: the human or robot's intention is expressed in their behaviour, which manifests as cues, which must be interpretable to the other party (robot or human) by drawing on social norms (as detailed in psychology and behavioural science literature). These considerations and flow are further detailed below.

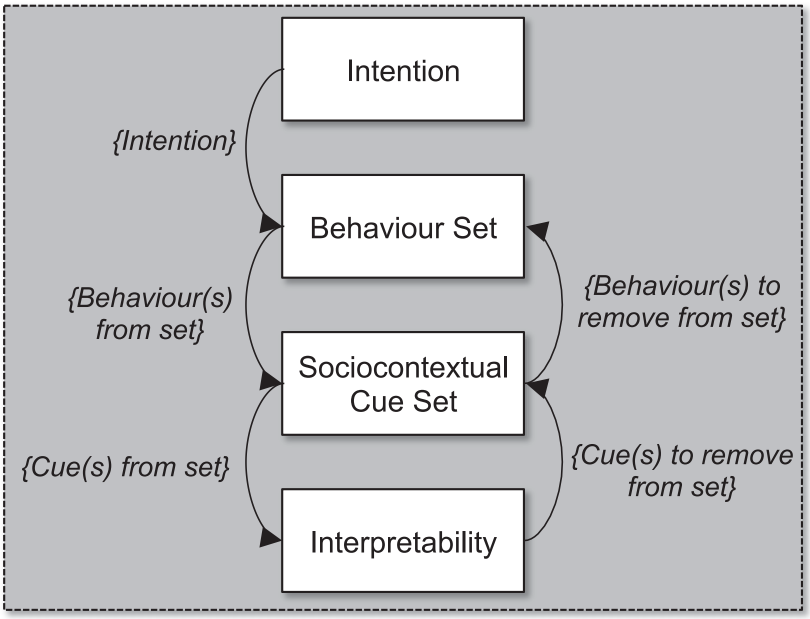


Figure 3.2: Process flow of design of the *Read* and *Elicit* branches of the paradigm.

- *Intention*

  Firstly, the intention of the human (*Read* branch case) or robot (*Elicit* branch case) must be considered. In a choice situation in which the robot intends the person to choose a particular object, for example, it may be useful for the robot to know which, if any, of the objects a person already prefers and intends to choose, as discussed in Chapter 2. Similarly, it may be useful for the robot to know the intended destination of a passenger in a public transport environment, if it intends to influence their movement.

- *Behaviour Set*

  Secondly, the set of behaviours which could be indicative of these intentions should be identified. Extensive literature from the fields of psychology and behavioural science shows that behaviours are expressions of underlying intentions and/or patterns of thinking and can also have the ability to influence others (as outlined in an exemplar instance in Section 2.1). This literature can be drawn on, in combination with the understanding of sociocontextual promotion/suppression of behaviours defined in Section 3.2.2, to identify intention-indicative behaviours which may be likely to be displayed by humans (*Read* branch case) or are appropriate to be displayed by the robot (*Elicit* branch case) in the intended application space.

  For instance, in a choice influence situation, a person's preference could be indicated by visual focus of attention (i.e. gaze) or tactile behaviour, while intended destination in a public transport environment could potentially be expressed via visual focus of attention, physical movement, or ticket details. Similarly, robot behaviour such as visual focus of attention or physical arm gestures could have the potential to influence preference in choice situations, while directional indicators or physical barriers could foreseeably result in path movement alterations.

  From these behaviour sets, a particular or several behaviours should be selected for the human or robot.

- *Sociocontextual Cue Set*

  Next, with intention-indicative and sociocontext-appropriate behaviour(s) selected to be detected from the human or displayed by the robot, the set of sociocontextual cues in which this behaviour could manifest should be considered by again drawing on literature from psychology and behavioural science; that is, the sociocontextual cues a human may display as a manifestation of their behaviour which could be *Read* by the robot, or, in the case of the *Elicit* branch, the sociocontextual cues which the robot could surreptitiously present back to the interaction partner in order to *Elicit* particular behavioural responses.

For example, a person's visual focus of attention, and hence intended choice, could manifest in cues such as eye gaze or head or torso orientation, while position or velocity may be manifestations of physical movement behaviour, which expresses intended destination. On the other hand, a robot's eye or head gaze cues may be sociocontextually appropriate to express its visual focus of attention, while directional indication behaviour could manifest as cues such as arrows or crosses.

From these cue sets, a single cue or multiple appropriate cues should be selected to *Read* from the human or for the robot to issue in order to *Elicit*. If no sociocontextually appropriate cue(s) can be identified, different behaviour(s) may need to be selected from the behaviour set identified above.

- *Interpretability*

  With cue(s) selected to be *Read* from the human or issued by the robot to *Elicit*, it then becomes necessary to determine if the cue will be interpretable by the other party in the interaction (robot or human).

  In the case of *Read*, the selected cue must be both sensible by the robot (what sensors/tools are available to *Read*? What are their capabilities? What software is available to interpret the data?) and meaningful in the application space (i.e. interpretable through the robot's contextual understanding and through human behaviour-to-meaning mapping available from the fields of psychology and behavioural science). For example, in a choice influence situation, a robot may need to *Read* both human presence in the *interaction zone* and gaze via depth camera data. Appropriate capabilities must be available to achieve this given the characteristics of typical depth camera data (as outlined in Section 2.3.2).

  Similarly, there are several factors which must be considered with regard to the *Elicit* cue. Firstly, the cue should be suitable for the particular social robot to issue in the application space, given its human-likeness. This has two key affects:

  1. The robot must have sufficient human-likeness – and thus humanoid characteristics – to be physically capable of issuing the selected cue. A joint attention cue, for example, requires it to be able to actuate its 'eyes' (or what can be perceived as its eyes) in the pan and tilt directions at the speed necessary for the meaning behind the cues to be correctly interpreted (as described in Section 2.2.2). In a public

transport environment, a lower-HL robot may only require icons to issue directional arrow cues.

2. It is ideal to have as information-rich a cue as possible, but the type of cue issued by the robot must remain congruent with its human-likeness, as a robot's HL is known to affect how its cues will be interpreted by interacting people [31, 106]. As HL increases, people prescribe robots a greater number of human characteristics [31, 97]; if a completely non-human robot attempted to issue a bodily sociocontextual cue such as joint attention, it seems less likely humans would recognise or respond to it in line with HHI cues, while the cues of perfect androids seem more likely to *Elicit* responses in line with human-issued bodily cues. Less human-like cues such as directional indicators, conversely, seem more likely to be recognised even when issued by less Human-Like robots. Therefore a higher-HL, humanoid social robot may plan to issue a joint attention cue to influence a person towards the intended object, as such cues are known to be capable of increasing preference for joint attention objects in decision making situations. On the other hand, a simpler directional indicator cue may be planned for a lower-HL disembodied social robot built into a transport environment, in order to influence people who are moving in an undesired direction.

In addition to human-likeness, a second factor affecting *Elicit* cue interpretation by the human is whether the cue is sufficiently sociocontextual and application space-appropriate; that is, whether it will communicate the specific intended message. From [81], which draws on literature from psychology and behavioural science, it is known that a simple cue which is heavily reliant on context will not only be enactable by a robot but will feasibly result in effective, expedient and surreptitious communication. Thus, unintended side-effects due to the context (different meanings can be ascribed to cues based on the situation in which they are being interpreted) and the complexities of behaviour can be minimised through selection of an appropriate cue.

If the selected *Read* or *Elicit* cue is unlikely to be interpretable, given the above, a different cue may need to be selected from the identified cue set.

Interactivity Design

Following the *Read* and *Elicit* branches, the next step of design of the Robot Centric HRI paradigm is to consider the interactivity of the robot (its ability to *Read*, and leverage the resultant information to moderate its *Elicit* strategy). Through different designs and activation sequences of the *Read* and *Elicit* branches of the paradigm, different levels of robot interactivity can be achieved, which may moderate the effectiveness of the robot's ability to achieve the desired behaviour via *Elicit* (and hence its goal(s)). For example, a traditional 'Task Completer' robot has low interactivity: without the ability to *Read*, such a robot is inherently unable to moderate its *Elicit*, and hence is only able to carry out a single type of *Elicit*; that is, its *Elicit* remains static.

On the other hand, greater levels of interactivity can be achieved by increasing the size of the cue set the robot is able to *Read* from the human, and/or the set of *Elicit* cues it can select from when moderating its *Elicit* strategy. This is depicted in Figure 3.3: the robot may be able to *Read* and interpret multiple cues (and hence behaviours) from the human, and leverage that information to moderate its *Elicit* strategy by drawing from a set of potential robot behaviours. These behaviours may be manifested as either a single robot cue (which is issued or not issued, in a binary fashion), or a set of human-interpretable cues from which the robot can select the most appropriate cue, given the situation. By displaying these cues, the robot attempts to lead the interaction by influencing the human's underlying intention, as also shown in Figure 3.3, which seems likely to manifest as a change in behaviour and cues.

An example of a higher level of robot interactivity can be seen in a study which investigated ensuring that a particular and unsuspecting member of a crowd is the recipient of a salient-item hand-over by a robot [81, 84]. In this case, sociocontextual cues were utilised to individualise the intended recipient and communicate the robot's intention (resolve ambiguity), influencing the participant to come forward to retrieve the object: through *Read*ing person location, the robot was able to physically direct its *Elicit* cues towards the intended recipient. Another study in which a robot was able to instantiate interaction with naïve passersby through issuing a combined physical presentation and gaze cue, achieved even higher interactivity: through *Read*ing a greater number of human cues (both person presence and position within the *interaction zone*), the robot was able to issue its cues at the appropriate time to influence the passerby to enter into an interaction, in one case responsively issuing cues as the participant approached the robot [81].

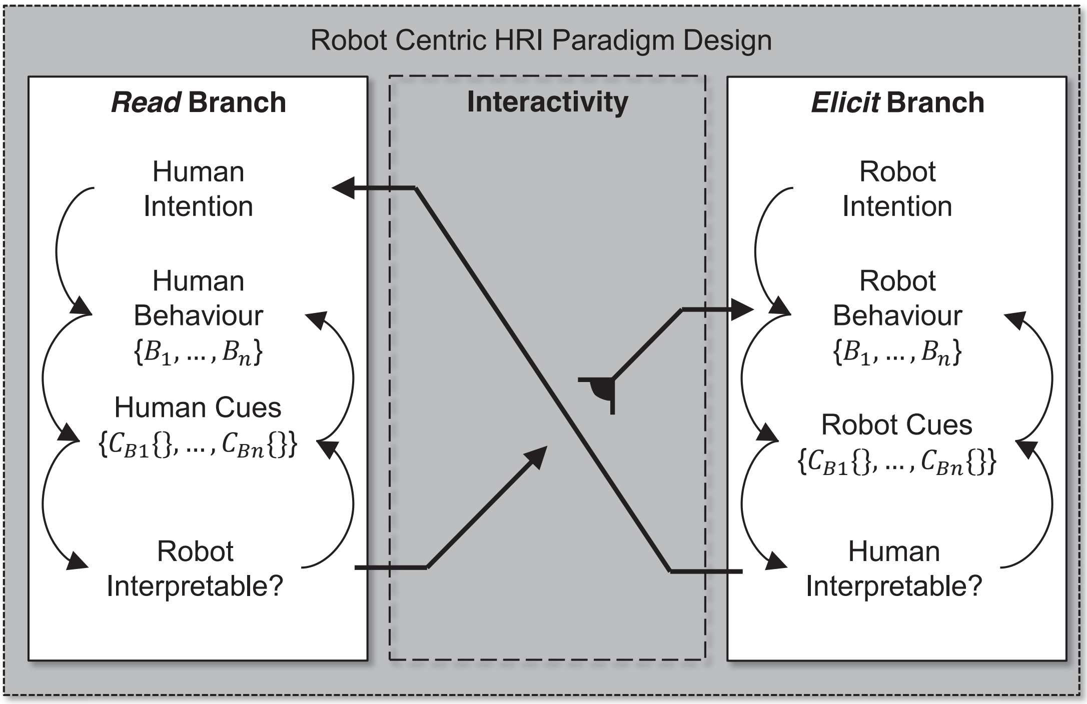


Figure 3.3: Detail of the process flow of the 'Design of the Robot Centric HRI Paradigm' stage of the methodology.

In a choice influence situation, similarly, a robot's interactivity could be increased by: a) *Read*ing both person presence and if the person is at gazing at a particular object (an object preference) and/or the robot, and; b) utilising the resultant information to moderate if/when (binary cue issuance) and/or to which object (a cue set) the robot will issue a joint attention *Elicit* cue. The importance of this ability will be shown in Chapter 6, where it is found that while JA cues in HRI can have effects in line with HHI JA cues, people have a greater tendency towards suspicion of, rather than compliance with, the robot, especially if they are not looking when the JA cue is issued. Thus, in the case where the robot is intentionally attempting to influence a person towards a particular object, if it is suspected that the person has already chosen the desired object (as indicated by *Read*ing their gaze direction through a method such as that presented in Chapter 5), the robot may be more likely to achieve its goal by not issuing a cue. However, if the person has not already chosen the desired object, a JA *Elicit* cue towards that object has the potential to positively influence people.

Likewise, in public spaces such as transport environments where a robot is attempting to influence passenger movement, Passenger Information (PI) systems incorporate the above characteristics, resulting in a range of fidelity and interactivity. Presently, Static and Dynamic PI systems are ubiquitous. At the information communications level, information appears to the viewer as being fixed and not readily changed in Static PI systems. Thus, Static PI systems map to the traditional paradigm for HRI where the robot/machine assumes the passive role of 'Task Completer' with low interactivity. On the other hand, Dynamic PI systems' information appears to the viewer as potentially changeable (i.e. an *Elicit* cue set exists). As demonstrated by the Robot Centric HRI paradigm, opportunity exists to change the fundamental paradigm for interaction via PI systems, and to leverage psychological and behavioural triggers to increase their interactivity, thus making PI systems more responsive; that is, to develop Responsive PI systems.

The above examples explain the hypothesis that the level of interactivity of the robot is correlated to its effectiveness at achieving its goal(s), where effectiveness is considered to be the ability of the robot to target its influence to achieve a specific desired outcome. However, this will be further empirically investigated in this work.

3.2.4 Implementation Design

With 'Robot Centric HRI Paradigm Design' complete, the final stage of the devised methodology is to design the implementation. It is here that factors and practicalities which may affect the operationalisation are considered. There are two key factors which must be taken into account:

1. *Configuration of the interaction*

   The configuration of the human and robot in the interaction is key to ensuring the *Read* and *Elicit* branches can be effectively operationalised. While minimal constraints are desirable, the configuration must be designed such that the person's behavioural cues can be successfully *Read*, and the person can witness and successfully interpret the robot's issued *Elicit* cues. For example, in a choice influence situation, the configuration must be such that the person can see the robot's joint attention cues and the robot, given the capabilities of available sensors, can sense the participant's gaze behaviour. In a public transport environment, it must be unambiguous where the robot is directing the person to go via its *Elicit* cue(s).

2. *Perception of robot autonomy*

   It is known that the perception of the level of autonomy of a robot can influence the way people interact with it [31, 80], with greater perceptions of autonomy likely to encourage greater 'natural' interaction by ensuring the people interact solely with the robot rather than with the robot's controllers. As such, during implementation of the paradigm it is desirable for the robot to appear autonomous in order to encourage interaction more likely to be aligned with behavioural expectations formed based on literature from the fields psychology and behavioural science.

3.3 Conclusion

This chapter has presented a methodology for operationalisation of the Robot Centric HRI paradigm, as per **RQ: Methodology**, a non-trivial task given the complexities of human behaviour and interaction dynamics. In the following chapters, this methodology is subsequently leveraged to successfully operationalise the Robot Centric HRI paradigm and empirically explore the remaining research questions of this work. In doing so, understanding of the transferability of sociocontextual cues such as exemplar gaze cues to HRI in the *Read* and *Elicit* directions of communication can be deepened (**RQ A: Sociocontextual cues in HRI**), and the relationship between robot interactivity and effectiveness can be investigated (**RQ B: Interactivity**).

Chapter 4

Elicit – Exploring the Effects of Robot-Issued Cues During Real-World HRI

With a methodology devised, developed and detailed in Chapter 3 which can be leveraged to operationalise the Robot Centric HRI paradigm, including its individual branches, this chapter presents a study of a typical humanoid social robot's ability to *Elicit* via issuing gaze cues during real-world HRI. This social exploration of the characteristics and effects of joint attention in HRI is an attempt to empirically evaluate qualitative human gaze behaviour in order to determine if humans will respond to exemplar humanoid robot-issued gaze cues in line with how they respond to human-issued cues, as outlined in Chapter 2. As per **RQ A.1: *Elicit* Response**, the findings support the hypothesis that exemplar cues such as joint attention gaze cues are transferrable to HRI, and that today's social robots can successfully *Elicit* particular behavioural responses from interaction partners, as necessitated by their interaction peer role.

4.1 Introduction

As previously discussed, as robots move into interaction peer roles it becomes necessary to understand the extent to which the characteristics and effects of their gaze cues will correspond to those of human-issued cues (as per **RQ A: Sociocontextual cues in HRI**), enabling them to *Elicit* particular behavioural responses from interaction partners. The first research question to arise – whether today's social robots are physically capable of issuing recognisable gaze cues such as JA (as per **RQ A.1: *Elicit* Feasibility**) – has been addressed in the literature presented in Section 2.2.2: the shape and capabilities of exemplar humanoid social robots, such as the RobotAssist platform, mean such robots are physically able to issue human-recognisable cues such as the JA cue to increase object desirability.

The question then becomes: during real-world HRI, will people respond to such robot-issued JA cues in line with how they respond to human-issued cues (as per **RQ A.1: *Elicit* Response**)? If the response is in line, equivalent characteristics and effects should be observed during JA in HRI as in HHI; literature from the fields of psychology and behavioural science outlined in Section 2.1.4 suggests that the influence of JA cues on human interaction partners is likely to manifest as an effect on the human's gaze behaviour, an indicator of preference.

Drawing on the methodology detailed in Chapter 3 to individually operationalise the *Elicit* branch of the Robot Centric paradigm, an empirical evaluation of the characteristics and effects of JA in HRI was carried out to explore these effects.

4.2 Measures of Expected Human Gaze Behaviour in HRI Joint Attention Scenarios

A prerequisite of an empirical evaluation of the equivalency of exemplar JA in HRI to HHI was a set of quantitative measures to gauge and analyse whether human interaction partners were displaying the expected behaviour in response to robot-issued JA cues. As outlined in Section 2.1.4, during HHI decision-making scenarios, when susceptibility to choice influence is likely to exist, people develop a 'gaze bias' towards their preferred object [143]. This bias, an indication of choice, results in two main visual search patterns that develop over the course of the decision-making process: 1) the breadth of the visual search decreases, and; 2) the focus increases [51].

These search patterns are marked by two primary gaze characteristics known to be indicators of interest and preference. Firstly, time spent looking at a location: this is a measure of the time people spend gazing at the objects, interaction partners or other locations during an interaction. Secondly, the number of fixations, or times that gaze saccades to those locations. Given the known effects of JA, as outlined in Section 2.1.4, these gaze characteristics can be broken down into a number of measures to gauge the effects of JA.

Gaze-bias towards objects

It is known from the literature in Section 2.1.4 that the JA of an interaction partner on an object assigns to that object properties that it would not display were it not gazed at. Thus, a set of measures were developed to gauge whether a robot's JA cues can similarly increase the saliency of objects to interaction partners. As gaze naturally settles on interesting objects in the environment [15, 142], this can be coarsely indicated by an interaction partner's tendency to look at objects more when JA cues are issued by the robot than when not. Figure 4.1 illustrates the locations relevant to a measure of this behaviour in a scene.

The measures developed are:

- Percentage of the total time spent looking at the objects (OB), as highlighted in blue in Figure 4.1. The remaining time will by definition be spent looking at locations other than the objects (!OB), as highlighted in orange in Figure 4.1. Henceforth this measure will be abbreviated as *%**t**(OBvs!OB)*.

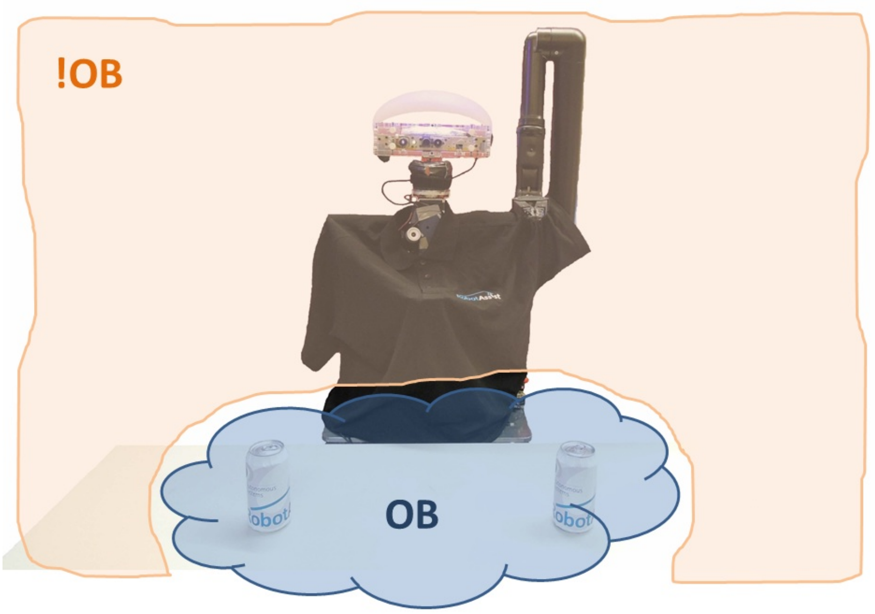


Figure 4.1: Object (OB) and Not Object (!OB) locations in a scene.

- Percentage of the total number of saccades that were directed at the objects (OB). Henceforth this measure will be abbreviated as *%**s**(OBvs!OB)*.

Gaze-bias towards JA object

As outlined in Section 2.1.4, JA towards an object can go further than simply increasing observers' general interest in objects. Through triggering enhanced information processing about the specific JA object in observers [132, 131], JA can affect the observers' evaluation and affective appraisal of that object, potentially influencing their choice. Thus, measures were also developed to gauge these finer effects of the influence of robot-issued JA cues on observers' interest; specifically, whether these cues increase the interest of interaction partners in a specific object to which a a robot issues a JA cue, rather than other, non-JA objects. Figure 4.2 illustrates the relevant locations in a scene. It is important to note that a third location, neither of the objects, also exists in this scene. This area will henceforth be referred to as 'Other'.

The measures are:

- Percentage of the total time spent looking at the JA object (JA). Figure 4.2 illustrates the JA object (highlighted in blue). Henceforth this measure will be abbreviated as *%**t**(JAvsTotal)*.

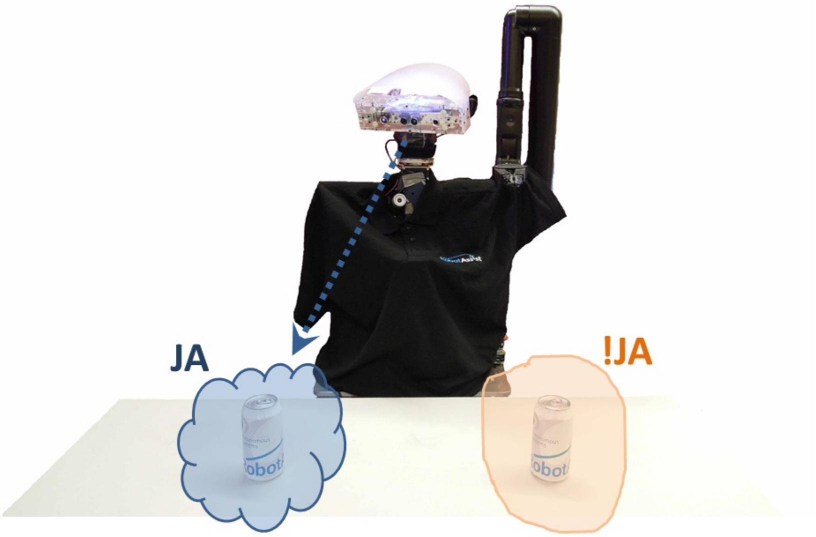


Figure 4.2: Joint attention (JA) and Not joint attention (!JA) objects in a scene.

- Percentage of the total saccades that were directed at the JA object. Henceforth this measure will be abbreviated as *%**s**(JAvsTotal)*.
- Percentage of the total time spent looking at the non-JA object (!JA). Figure 4.2 illustrates the !JA object (highlighted in orange). Henceforth this measure will be abbreviated as *%**t**(!JAvsTotal)*.
- Percentage of the total saccades that were directed at the !JA object. Henceforth this measure will be abbreviated as *%**s**(!JAvsTotal)*.

Gaze-bias towards chosen object

As gaze naturally settles on desirable objects in the environment [15, 142], measures were developed to quantify the 'gaze bias' that literature predicts will develop towards preferred objects [143] in decision-making situations. Thus, human interaction partners will likely develop a tendency to look more at the object they will eventually choose (vs towards other objects) over the course of their decision making process. Figure 4.3 illustrates locations relevant to these measures in a scene. It is important to note that a third location, neither of the objects, also exists in this scene. This area will henceforth be referred to as 'Other'.

The developed measures are:

- Percentage of the total time spent looking at the chosen object (CH). Figure 4.3 illustrates the CH object (highlighted in blue). Henceforth this measure will be abbreviated as *%**t**(CHvsTotal)*.

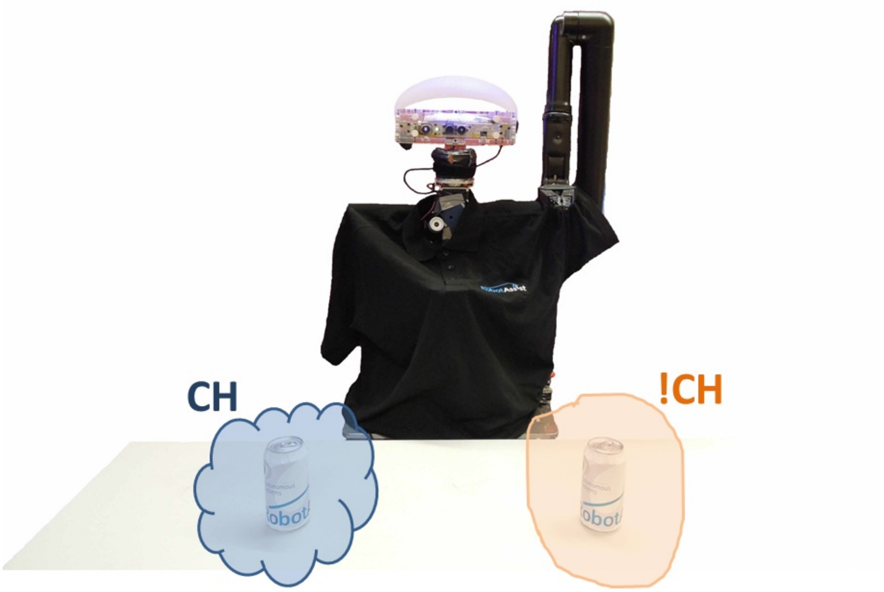


Figure 4.3: Chosen (CH) and Not Chosen (!CH) objects in a scene.

- Percentage of the total saccades that were directed at the CH object. Henceforth this measure will be abbreviated as *%**s**(CHvsTotal)*.
- Percentage of the total time spent looking at the not CH object (!CH). Figure 4.3 illustrates the !CH object (highlighted in orange). Henceforth this measure will be abbreviated as *%**t**(!CHvsTotal)*.
- Percentage of the total saccades that were directed at the !CH object. Henceforth this measure will be abbreviated as *%**s**(!CHvsTotal)*.

An additional measure of this phenomenon can be found by considering only the looks at the objects, rather than all looks as above.

These measures are:

- The time spent looking at the object which will be chosen (CH) as a percentage of time spent looking at the objects. The remaining time will by definition be spent looking at the not chosen object (!CH). Henceforth this measure will be abbreviated as *%**t**(CHvs!CH)*.
- The number of saccades that were directed at the object that will be chosen (CH) as a percentage of the total saccades directed at objects. The remaining saccades will by definition be directed at the not chosen object (!CH). Henceforth this measure will be abbreviated as *%**s**(CHvs!CH)*.

4.3 Empirical Evaluation of the Effects of Joint Attention During HRI

With quantitative measures developed, an empirical evaluation was carried out to explore the extent to which people respond to robot-issued JA cues in line with how they respond to human-issued cues. By drawing on the developed methodology discussed in Chapter 3, the experiment scenario was constructed in order to *Elicit* participant gaze behaviour which could be analysed in terms of the measures outlined above in Section 4.2. However, *Elicit*ing this gaze behaviour is non-trivial: a situation must be constructed which is sufficiently controlled to effectively observe the behaviour, while at the same time maintaining a reasonable level of 'natural' interaction with the robot and hence external validity, i.e. the extent to which the results can be more widely generalised. This was achieved through the design and execution of the experimental method detailed below, which was published in ***[C3]*** and ***[W1]***, listed in Appendix A.

4.3.1 Hypotheses

From the understanding of typical human behaviour during JA in decision-making scenarios, the following hypothesis and predictions were developed:

Hypothesis – The robot's presentation of JA cues will have similar characteristics and effects to JA cues issued during an HHI decision-making situation.

Predictions – Upon presenting such cues towards human interaction partners in an HRI decision-making scenario, the robot will be able to successfully *Elicit* HHI-literature predicted gaze and choice behaviour. Specifically, it is predicted that:

1. *OB gaze bias*: The robot's JA cues will lead participants to look more at the objects than when no cue is issued.

2. *JA object gaze bias*: The robot's JA cues at a specific object (the JA object) will lead participants to look more at that object than the other, !JA object.

3. *CH object gaze bias*: Over the course of their decision-making process, the participant will develop a tendency to look more at the object they eventually choose than towards other objects in the environment, i.e. they will develop a gaze bias towards their chosen object.

4.3.2 Participants

There were 16 participants in the experiment (14 female, 2 male). The participants were robot-naïve students from faculties of UTS besides the Faculty of Engineering & IT (predominantly the Faculty of Design, Architecture and Building), in order to increase the external validity of the experimental results as well as the likelihood that their responses would be in line with those predicted by literature in psychology and behavioural science (as discussed in Section 3.2.2).

4.3.3 Setting and Setup

As outlined in Section 2.1.4, the effects of JA are predictable and able to be observed in HHI decision-making situations, where JA is known to have the potential to influence preference. In order to enable the exploration of the equivalency of JA in HRI, an HRI decision-making scenario was constructed; specifically, a forced-choice paradigm with a dichotomous response format.

The RobotAssist platform (Figure 2.6a), which is approximately 1.4m tall with a 0.35m radius, was utilised during the experiment. As can be seen from the figure, the robot is equipped with a wheeled base and anthropomorphic upper body. The upper body consists of a six degree-of-freedom manipulator and a perspex head which encloses a Microsoft Kinect sensor and multi-coloured lights. The head is mounted on two servos which can rotate it in the pan ($-110° < \theta < 64°$) and/or tilt ($-45° < \theta < 80°$) directions, enabling the robot to execute the JA cue described in Section 2.2.2. Not shown in the figure are two USB speakers, which are mounted facing forwards underneath the robot's shirt.

The RobotAssist platform was positioned behind a table in the corner of an office within the Faculty of Engineering & IT, as pictured in Figure 4.4a. The office was relatively small with few distractions (other sources of mental state and/or behaviour influence, as discussed in Section 3.2.2) for the participants. Positioned on the table were two objects: soft drink cans. To ensure the two cans were both identical and neutral, and thus avoid pre-existing preference bias on the part of the participants – another potential source of mental state and/or behaviour influence (towards a brand they may prefer, for example) – the cans were both covered with large RobotAssist stickers. Participants' positioning directly in front of the table, and hence the robot, ensured it was possible for them to witness the robot's cue (as prescribed in the methodology, Section 3.2.4).

To give the impression of complete robot autonomy, as also specified in Section 3.2.4, the Wizard-of-Oz experimenter controlling the robot was positioned outside the office, completely concealed behind a large partition (shown in Figure 4.4b) and observing the participant via a live feed from a webcam mounted on the robot's chest. This impression of robot autonomy reduced the required involvement/intervention by experimenters in the experiment execution, encouraging natural interaction on behalf of the participants.

4.3.4 Experimental Conditions

The experiment was a 3×2 design: *Robot Attendance: None / Joint attention object / Not joint attention object* × *Participant's Selection: Selected attended object / Didn't*. It had two conditions:

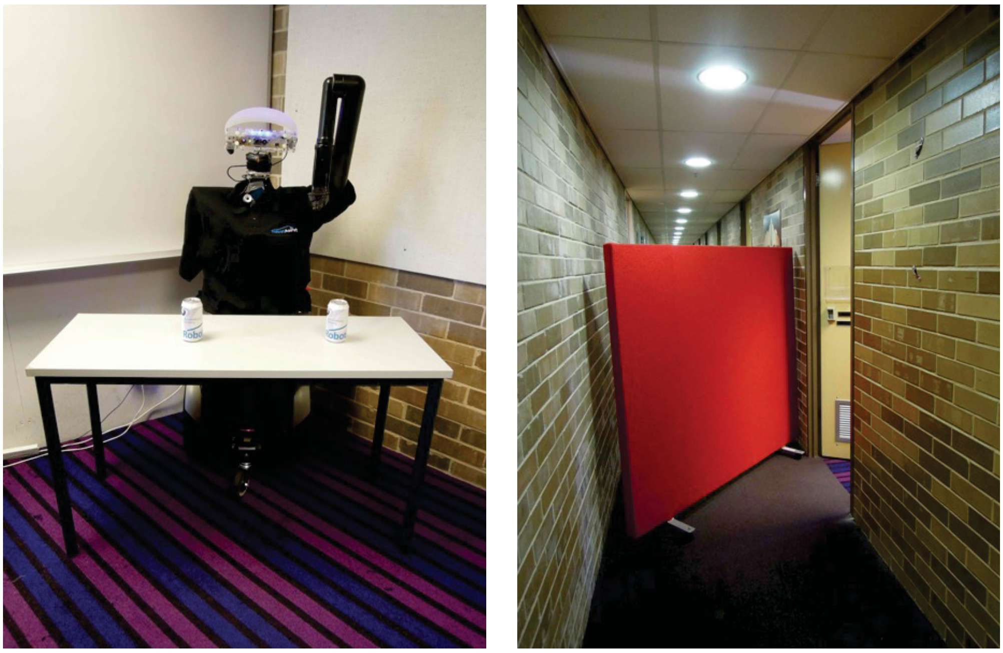
Control (C) - In this baseline condition the robot did not perform any gaze cues and its head remained stationary throughout the experiment.

Joint Attention (JA) - In the JA condition the robot issued one joint attention cue (as described in Section 2.2.2) to the participant. There were two cues in the JA cue set:

- *Joint Attention Left (JAL)* - The robot performed the joint attention cue at the Left object (as shown in Figure 2.10a).
- *Joint Attention Right (JAR)* - The robot performed the joint attention cue at the Right object (as shown in Figure 2.10b).

4.3.5 Procedure

The experimental procedure was a key element of ensuring that natural behaviour was elicited and that measurable results were gathered during the experiment. The experiment was conducted on a weekday during semester time at UTS. The solicitation of participants was carried out by an experimenter, who approached students in one of the university's buildings. In order to exclude engineers familiar with robots, potential participants were pre-screened to determine if they were from the Faculty of Engineering & IT. If they were not, they were shown a flyer stating that a robotics experiment was being carried out and they would be given a free chocolate and soft drink in exchange for participating.



(a) The RobotAssist platform setup within an academic's office

(b) Partition concealing the Wizard-of-Oz experimenter

Figure 4.4: Experiment setup.

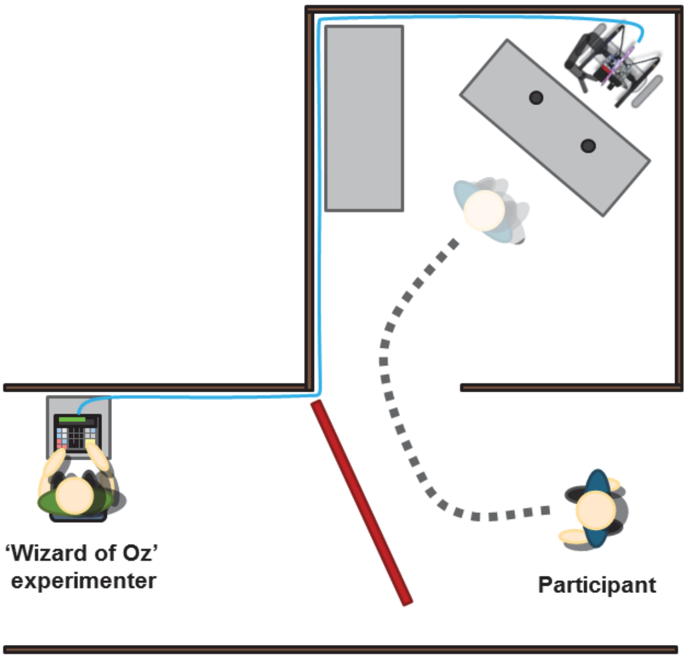


Figure 4.5: Experiment scenario.

If they agreed to partake in the experiment, the participant was led to the office where the robot was positioned, during which time they were not given any additional information. Instead, just outside of the office door, they were asked to read brief instructions. These instructions thanked them for participating in the study, and directed them to enter the office and approach the robot, at which time the robot would give them further instructions and begin a simple interaction. The instructions also informed them that if they did not understand the robot, they could ask it to repeat its instructions. This procedure ensured the participants had no explicit directions about how to interact with the robot, and thus interacted naturally.

When the participant entered the office, as illustrated in Figure 4.5, the experimental procedure shown in Figure 4.6 began. This commenced with the random selection of the experimental condition (Control, JAL or JAR), then the robot stating, "*Thank you for coming. In a moment I am going to need you to choose which drink you'd like. You can think about which drink you'd like now.*" This is subsequently referred to as the Initial Statement (IS) stage, and made the participant aware that they would be required to make a choice during the experiment, encouraging the display of the desired decision-making and JA gaze behaviour. The IS stage was followed by a $2s$ wait (Initial Wait (IW) stage) during which the participants could inspect the objects. It is in this stage that participants were expected to develop a decision-making 'gaze bias' towards their object of choice.

After this wait, if in the Joint Attention condition, the robot would issue either the JAL or JAR cue (Glancing (GL) stage), then wait another $2s$ (Post Glance Wait (PG) stage). It was expected that the effects of the robot-issued JA cue would manifest in the GL and this PG stage. Following this pause (or immediately after the first wait, in the C condition), the robot requested the participant to "*Please point at the drink you would like. When you have selected a drink, I will hand it to you*" (Asking for Choice (AC) stage), and their choice was manually recorded by the remotely observing Wizard-of-Oz experimenter.

This completed the relevant stages of the experiment and triggered a hard-coded pickup sequence, which was included purely for the entertainment of the participants. When the robot's arm had retrieved the drink, the robot stated "*Please get ready for me to hand you the drink*", before extending the drink to the fixed $\{x, y, z\}$ handover location and releasing it. To end the experiment, the robot thanked the participant and informed them that they could keep the drink if they wished.

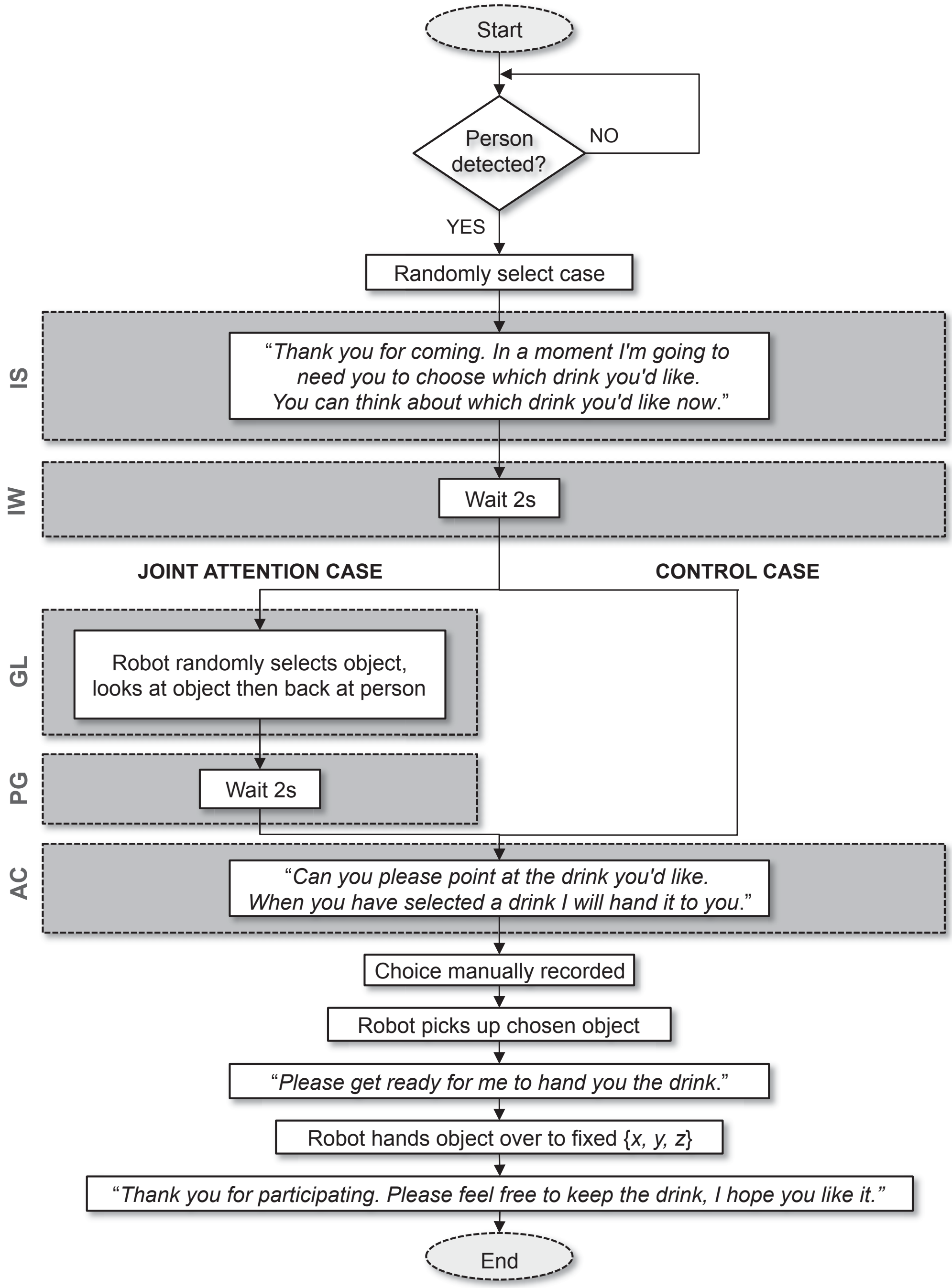

Start
Person detected?
NO
YES
Randomly select case
IS
"Thank you for coming. In a moment I'm going to need you to choose which drink you'd like. You can think about which drink you'd like now."
IW
Wait 2s
JOINT ATTENTION CASE
CONTROL CASE
GL
Robot randomly selects object, looks at object then back at person
PG
Wait 2s
AC
"Can you please point at the drink you'd like. When you have selected a drink I will hand it to you."
Choice manually recorded
Robot picks up chosen object
"Please get ready for me to hand you the drink."
Robot hands object over to fixed {x, y, z}
"Thank you for participating. Please feel free to keep the drink, I hope you like it."
End

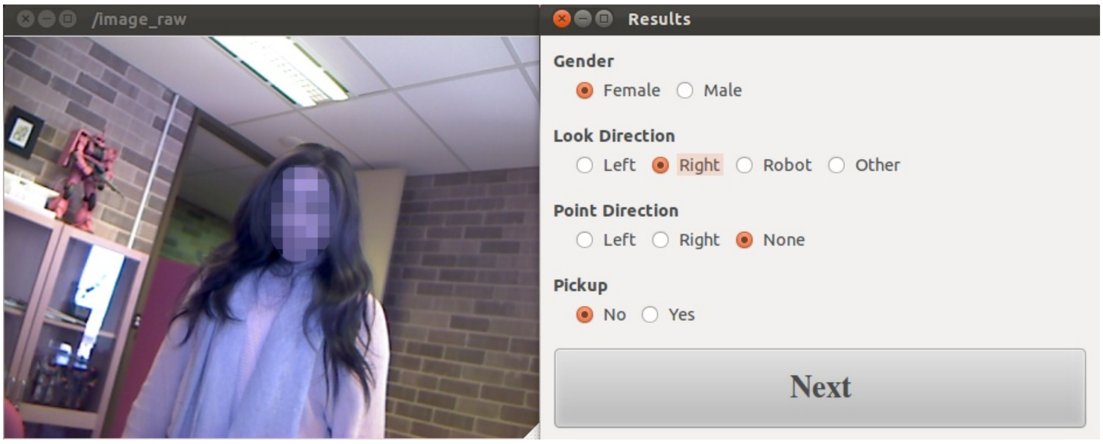
Figure 4.6: Experimental procedure.

4.3.6 Measurement

The type of cue performed by the robot was the only independent variable. The dependent variables involved the quantitative measures of participant gaze behaviour outlined above in Section 4.2. Data from the chest-mounted USB camera was coded post-hoc to quantify these measures.

4.3.7 Results

There were 16 trials (8 Control and 8 JA [5 JA Right and 3 JA Left]) available for analysis. The experiment was run between the hours of 10.15am–12pm, and approximately one hour of camera (∼13,900 frames) and RGB-D person detection data of the trials was autonomously collected by the robot during the experiment. A Robot Operating System (ROS) script was written that automatically replayed the camera images frame by frame along with a Graphical User Interface (GUI) which was designed and coded as part of this work. Figure 4.7 shows an example of a single frame from the replayed experiment data. On the left the chest-mounted USB camera image is displayed, while the developed GUI on the right enables information about the participant's gender (Male or Female), gaze direction in the frame (Left object, Right object, Robot or Other), point direction in the frame (Left object, Right object or None) and whether they had picked up an object (Yes or No) to be coded. The *Next* button recorded the entered data and refreshed the image with the next recorded frame. Also logged for each frame, but not shown in the GUI, were the current time-stamp, experiment condition (Control or JA), experiment state (IS, IW, GL, PG, AC), the robot look direction (Left object or Right object), and the person's choice.



Gender
● Female ○ Male

Look Direction
○ Left ● Right ○ Robot ○ Other

Point Direction
○ Left ○ Right ● None

Pickup
● No ○ Yes

Figure 4.7: The ROS coding image and Graphical User Interface.

The participant trial distribution is shown in Table 4.1. While there was a similar gender representation per trial in both conditions, there was a strong overall bias towards female participation. This bias seemed (subjectively) to be consistent with the demographics of the business students who are often around the hallway area in which they were approached during semester time, and who may have been recruited for the study.

Table 4.1: JA experiment participant trial distribution.

Case	**Male**	**Female**	***Total***
Control	1	7	*8*
JA	1	7	*8*
Total	*2*	*14*	*16*

The stages of the experiment were grouped for planned contrast analysis (unless otherwise noted): *pre cue-issuing* (IS & IW) with *post cue-issuing* (AC), and *cue-issuing* (GL & PG).

Gaze towards objects

Figure 4.8 shows the pattern of visual attendance of the OB and !OB locations in the JA condition. A repeated measures Analysis of Variance (ANOVA) of the *%**t**(OBvs!OB)* measure with a planned contrast of the *pre* and *post cue-issuing* states against the *cue-issuing* states found that the participants developed a gaze bias towards the objects during the *cue-issuing* states, spending a significantly larger percentage of time attending the OB during these states on average, $\delta\bar{x} =$ 17.4%, F(1,7)=8.715, p=0.036.

Gaze towards JA object

Figure 4.9 shows the pattern of visual attendance of the JA and !JA objects. It was found through a repeated measures ANOVA of the *%**t**(JAvsTotal)* measure that JA participants did not display a significant JA object gaze bias during *cue-issuing* compared to *pre* and *post cue-issuing*, in contrast to what was expected.

To explore this further, an additional unplanned contrast of the GL against the PG state was conducted. This revealed a significant reduction in percentage of time attending the JA object in the PG state of $\delta\bar{x}$ = 20.3% below the GL state and $\delta\bar{x}$ = 8.8% below the *pre cue-issuing* stage, F(1,7)=8.168, p=0.024. Furthermore, a repeated measures ANOVA of the percentage time JA participants spent looking at the JA object and !JA object grouped showed that participants spent a significantly larger percentage of time, on average, attending the JA object than the !JA object during *cue-issuing* $\delta\bar{x} = 8.7\%$, F(1,7)=6.768, p=0.035.

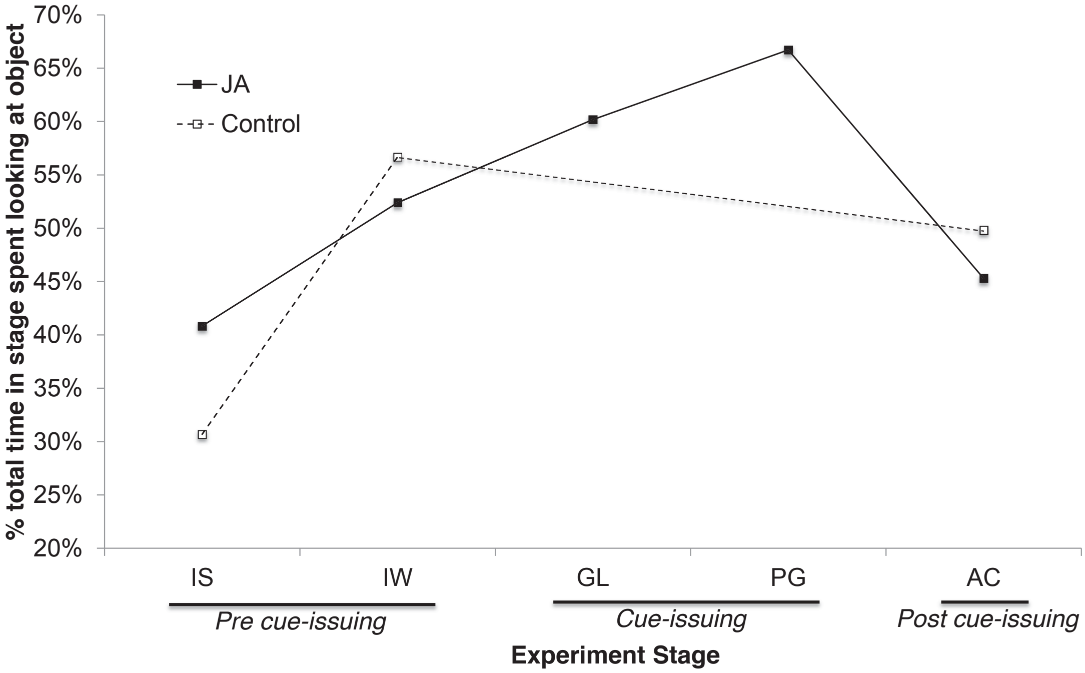


Figure 4.8: Participants' attendance to the OB and !OB locations during the experiment.

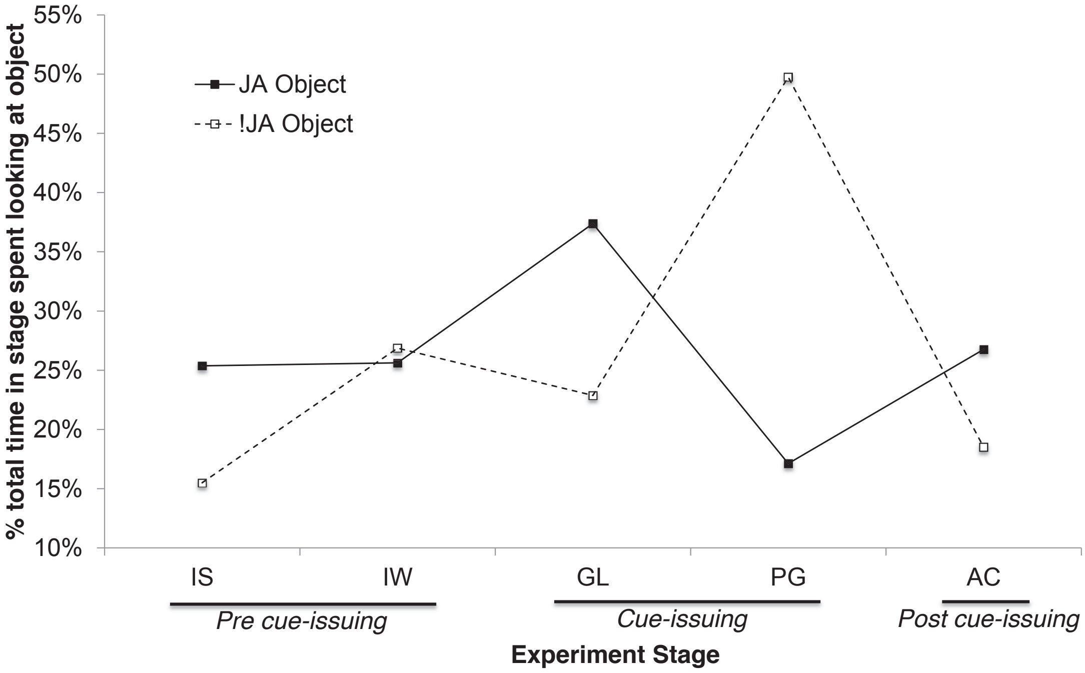


Figure 4.9: Participants' attendance to the JA and !JA objects during the experiment.

Gaze towards CH object

Figure 4.10 shows the pattern of visual attendance of the Chosen (CH) and Not Chosen (!CH) object. A repeated measures ANOVA of the *%**t**(CHvsTotal)* measure found that JA participants looked significantly more at the CH object in *cue-issuing* than *pre* and *post cue-issuing* ($\delta\bar{x} = 17.3\%$, $F(1,7)=6.078$, $p=0.043$). This gaze bias effect lasted until the end of the PG stage, when gaze at the CH object significantly declined ($\delta\bar{x} = 18.3\%$, $F(1,7)=9.674$, $p=0.017$) to approximately *pre-cue* levels. Further, a repeated measures ANOVA of the *%**t**(CHvsTotal)* and *%**t**(!CHvsTotal)* measures grouped found the CH gaze to be significantly different ($\delta\bar{x} = 8.6\%$, $F(1,7)=6.768$, $p=0.035$) from gaze at the !CH object, which remains relatively level.

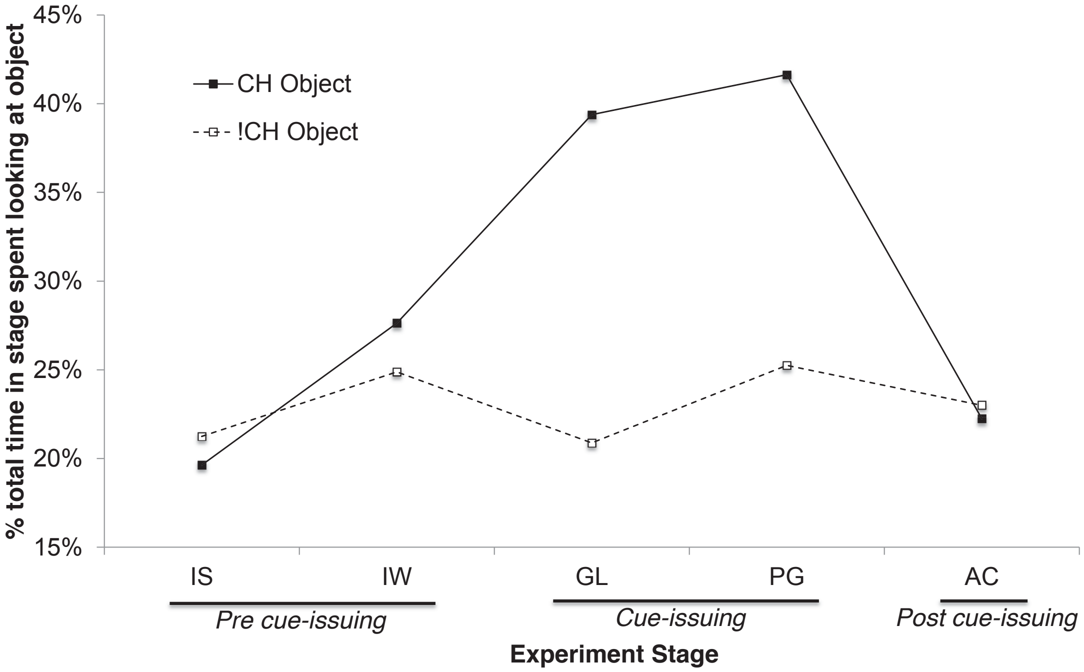


Figure 4.10: Participants' attendance of CH and !CH objects during the experiment.

A repeated measures ANOVA of the *%**s**(CHvsTotal)* measure was conducted with a planned contrast of the *pre* and *post cue-issuing* states against the *cue-issuing* states in the JA condition. It was found that the participants issued borderline-significantly smaller percentage of saccades during the *cue-issuing* states on average, $\delta\bar{x} = 17.4\%$ $F(1,7)=5.461$, $p=0.052$. A second planned contrast, of the *post cue-issuing* state against the *cue-issuing* states, revealed the participants issued a borderline-significantly larger percentage of saccades during the *post cue-issuing* states on average, $\delta\bar{x} = 22.3\%$, $F(1,7)=5.390$, $p=0.053$.

A repeated measures ANOVA of the *%**t**(JAvsTotal)* measure between the JA and Control condition found participants spent a significantly larger percentage of time attending the CH object during *post* than *pre cue-issuing* on average, $\delta\bar{x}$ = 9.1%, F(1,7)=8.808, p=0.010. There was also a significant interaction effect, with the difference being more pronounced in *post cue-issuing* (F(1,7)=8.808, p=0.010). This appears to have been driven by the Control condition, with Control participants spending $\delta\bar{x}$ = 16.3% more time attending the CH object than JA participants.

4.3.8 Discussion

The empirical results presented provide support for the hypothesis that the robot's presentation of JA cues will have similar characteristics and effects to JA cues issued during an HHI decision-making situation (as per **RQ A: Sociocontextual cues in HRI**). Support was found for the three predictions that participants in the JA condition would develop gaze biases towards 1) the objects in general; and, 2) the JA object and 3) their CH object in particular.

Gaze-bias towards objects

Participants in the JA condition spent a significantly larger percentage of time ($\delta\bar{x}$ = 17.4%, F(1,7)=8.715, p=0.036) attending the OB during the *cue-issuing* states than the *pre* and *post cue-issuing* states. This suggests that the robot-issued JA cue increased the saliency of the objects to the interaction partners, in line with HHI literature.

Gaze-bias towards JA object

Participants displayed a non-significant tendency towards looking at the JA object during the *cue-issuing* states. Somewhat interestingly, whilst the planned contrasts had anticipated that the participants' gaze would linger on the JA object for a period after the cue was issued in GL, the participants turned their gaze towards the !JA object soon after GL (shown in Figure 4.9). Specifically, immediately following the cue (PG), they looked significantly less ($\delta\bar{x}$ = 20.3%, F(1,7)=8.168, p=0.024) at the JA object compared to during the cue (GL). This unexpected finding potentially adds an opposing influence of the robot's JA cue (perhaps a suspicion of the robot's intentions), which will be further explored in Chapter 6. However, during *cue-issuing* states the participants spent a significantly larger percentage of time attending the JA object than the !JA object($\delta\bar{x} = 8.7\%$, F(1,7)=6.768, p=0.035), suggesting that the robot-issued JA cue does increase the participant's gaze towards the JA object compared to the !JA object.

Gaze-bias towards chosen object

Participants in the JA condition displayed a $\delta\bar{x} = 17.3\%$ gaze bias towards the object they would eventually choose (CH object) while the robot was issuing the JA cue (F(1,7)=6.078, p=0.043), gaze behaviour which was significantly different to that towards the !CH object (F(1,7)=6.768, p=0.035). In conjunction with the finding that during the *cue-issuing* states participants issued borderline-significantly smaller ($\delta\bar{x} = 17.4\%$, F(1,7)=5.461, p=0.052) percentage of saccades than during the *pre* and *post cue-issuing* states, this suggests an increase in the focus of the search. As it is known from HHI literature that this narrowing of the search develops over the course of the decision-making process, this suggests that JA participants' decisions may have been made during the *cue-issuing* states, as predicted by the HHI literature on gaze bias.

This is further supported by the finding that, in the *post-cue* stage when participants were asked for their choice, JA participants displayed significantly less gaze bias towards the CH object compared to Control participants ($\delta\bar{x} = 16.3\%$, F(1,7)=8.808, p=0.010). This suggests that, while JA participants may have made their decision during the *cue-issuing* stage, Control condition participants made theirs during the *post cue-issuing* stage.

4.4 Conclusions

The study presented in this chapter, which was published in ***[C3]*** and ***[W1]*** of Appendix A, focused on quantifying the characteristics and effects of JA in HRI. An empirical evaluation was conducted in which 16 participants took part in a forced-choice paradigm with a dichotomous response format, real-world HRI scenario in which the humanoid social robot issued JA cues to one of two identical objects.

It was found that the robot's presentation of a JA cue resulted in significant effects on gaze-based measures of influence, specifically the development of gaze biases towards the objects in general, and the JA object and chosen object in particular. These results suggest that participants responded to the robot-issued JA cue in line with how they respond to human-issued cues, as predicted by literature on the characteristics and effects of JA in HHI. This supports the hypothesis that the exemplar JA cue is transferable to HRI (as per **RQ A.1: *Elicit* Response**), and hence that today's social robots can successfully *Elicit* particular behavioural responses from interaction partners, as necessitated by their interaction peer role.

Additionally, verification of the prediction that the characteristics and effects of sociocontextual cues in HRI correspond to those of HHI cues also provides support for **RQ: Methodology**: the developed methodology was drawn on to successfully operationalise the *Elicit* branch of the Robot Centric paradigm during real-world HRI.

However, along with further exploration of these findings, this *Elicit* exploration would benefit from increased external validity and sample size in order to generalise and extend the understanding of the influence of robot-issued JA cues on surrounding humans.

Chapter 5

Read – Enabling Robots to Decipher Human-Issued Cues

Given exemplar social humanoid robots' ability to successfully *Elicit* via issuing gaze cues during real-world HRI (as per **RQ A.1: *Elicit* Response**), as detailed in Chapter 4, this chapter presents an exploration of the further ability of such robots to *Read* human-issued gaze cues during HRI in order to fulfil the requirements of their interaction peer role. Operationalisation of the *Read* branch is achieved by again drawing on the methodology for Robot Centric HRI paradigm operationalisation presented in Chapter 3. Firstly, in order to determine the value of *in situ* human gaze cue *Read*ing capabilities (as per **RQ A.2: *Read* Value**), a study focusing on understanding people's natural gaze behaviour towards robots is discussed. From the finding that no generalisable pattern of gaze behaviour was observable, a head yaw estimation framework was developed and is next detailed in this chapter. As head yaw is known to be generally indicative of gaze direction, through employing the framework typical social robots can successfully *Read* interaction partners' gaze behaviour *in situ* during HRI (as per **RQ A.2: *Read* Feasibility**), including both mutual gaze in the far-*interaction zone* and joint attention and gaze bias cues in the near-*interaction zone.*

5.1 Investigating Human Gaze Behaviour in the HRI Space

Thus far it has been shown in Chapter 4 that the gaze cues of exemplar social humanoid robots are perceived by observing humans as communication, and have similar characteristics and effects during HRI as HHI (as per **RQ A.1: *Elicit* Feasibility** and **RQ A.1: *Elicit* Response**). This suggests that exemplar JA cues can be leveraged by such robots to *Elicit* particular behavioural responses from interaction partners, and thus that sociocontextual cues can be reliably implemented and utilised during HRI in the robot to human direction of communication. For robots to communicate in a socially sensitive manner and fulfil the requirements of their interaction peer role, however, it is equally important for the reciprocal direction of communication to be addressed, i.e. the ability of such robots to *Read* gaze cues issued by interacting humans.

For example, consider the scenario depicted in Figure 5.1. In this situation, the robot would like to issue a JA cue to the water glasses on the table in order to bring them to the human's attention. However, in order for this *Elicit*ing to be successful, the cue must be witnessed by the human; that is, it must be issued when the human is engaging in mutual gaze with the robot. Literature on HHI outlined in Section 2.1.3 has given a baseline of expected mutual gaze behaviour around robots, suggesting that large mutual gaze variations will exist between individuals during interaction. Thus, without the ability to *Read* the gaze of the human and detect mutual gaze, the robot cannot guarantee that its cue will be witnessed.

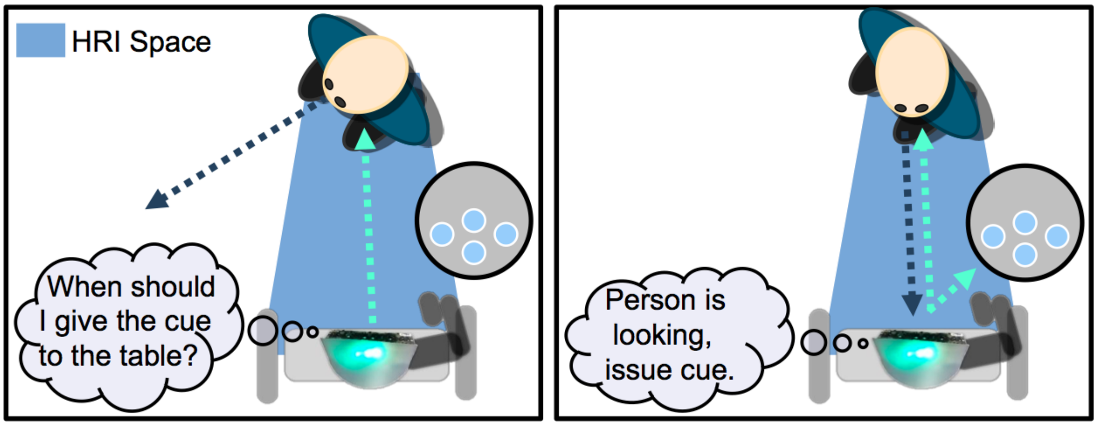

HRI Space
When should I give the cue to the table?
Person is looking, issue cue.

Figure 5.1: To ensure cues are witnessed, a robot needs an understanding of when people are gazing at it.

In order to address this issue, the first question which arises is whether people will actually display HHI-predicted gaze cues around robots, necessitating *in situ* gaze cue detection for robots (as per **RQ A.2: *Read* Value**). The empirical evaluation presented in Chapter 4 demonstrated that humans display HHI-predicted gaze bias in decision-making scenarios (as outlined in Section 2.1.4) when the robot is *Elicit*ing via JA cues, suggesting that such mutual gaze predictions will similarly exist in HRI. Thus, the question becomes: during real-world interactions in the HRI space, what natural mutual gaze behaviour, and hence attentiveness, will people display towards a robot? This question was examined in the experiment detailed below, which was published in ***[C2]*** of Appendix A.

5.1.1 Hypotheses

It is known from the literature presented in Section 2.1.3 that the dynamics of mutual gaze in HHI are a complex function of individual and environmental variables which interplay to result in large gaze behaviour variations between individuals. For example, the percentage of HHI encounter time spent gazing at an interaction partner can range from 28% to over 70% [75], with glance lengths in the range of 3–10s [8]. From this understanding of typical human mutual gaze behaviour during interactions, the following hypothesis was developed:

Hypothesis – There will be no generalisable pattern of participant mutual gaze behaviour towards a robot, necessitating online, *in situ* gaze estimation capabilities to enable robots to effectively *Elicit*.

5.1.2 Participants

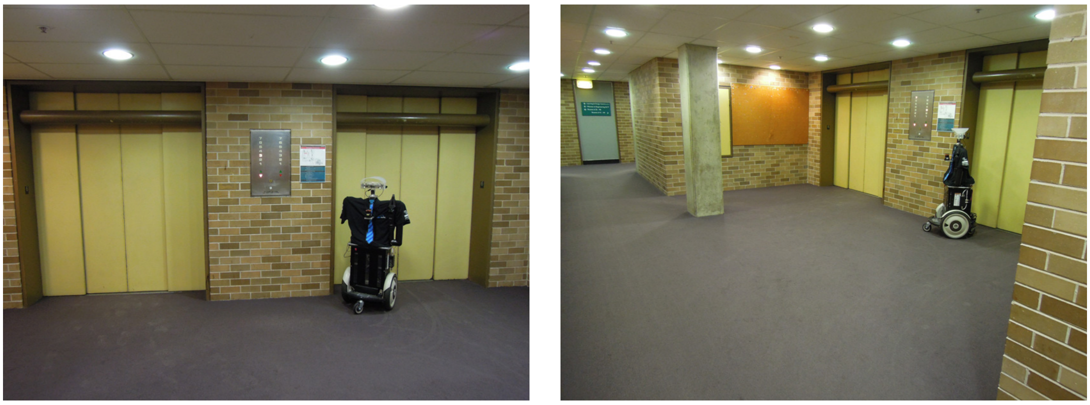
There were 24 unsolicited participants in the experiment (3 female, 21 male). They were passersby to the experiment location and no particular demographic was evident.

5.1.3 Setting and Setup

In order to explore participants' natural mutual gaze behaviour towards robots during real-world HRI, a semi-constrained interaction scenario was constructed. The RobotAssist platform was positioned adjacent to the buttons of a set of elevators within a UTS building, as shown in Figure 5.2. In order to give the impression of robot autonomy and encourage natural interaction on behalf of the participants, an experimenter was surreptitiously positioned at a laptop (giving the appearance of studying) around $30m$ away. The experimenter's sole role was to remotely trigger the data recording on the robot as participants approached the elevator foyer.

5.1.4 Procedure

Participants approached the elevator foyer from one of two directions in a nearby hallway, and the experiment began when they entered the foyer and the robot was within their line of sight, as depicted in Figure 5.3. The duration of the experiment was moderated by the length of time it took for the elevator to arrive, with $\bar{x} = 31.3s$ and $\sigma = 28.7s$. The experiment concluded when the participant exited the elevator foyer, either by getting on an elevator or by returning back down the hallway. The robot was stationary for the duration of the experiment, and thus issued no cues which would have affected people's natural mutual gaze behaviour.



(a) The RobotAssist platform setup in an elevator foyer.

(b) View of the experiment setup from one of the entry hallways.

Figure 5.2: The experiment setup.

5.1.5 Experimental Conditions

There was only one experimental condition, in which the robot remained stationary throughout the experiment.

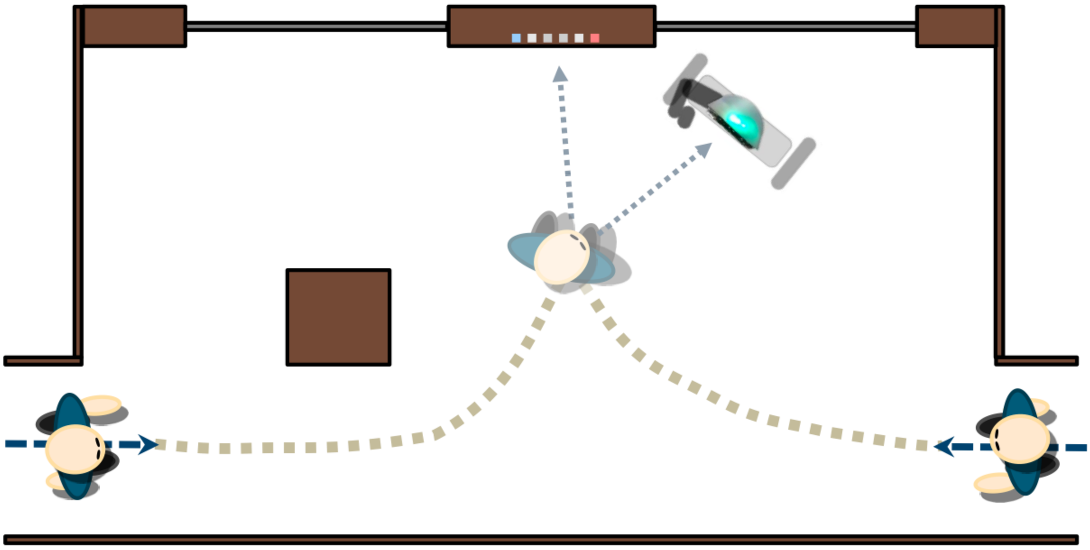


Figure 5.3: The experiment scenario.

5.1.6 Measurement

The participants' gaze behaviour was the only dependent variable. This included their gaze times and characteristics of Looking (L) or Not Looking (!L) at the robot. Data from a robot head-mounted Kinect camera was analysed post-hoc to quantify these measures.

5.1.7 Results

There were 24 trials available for analysis. The experiment was run over two weekdays approximately one week apart, between the hours of 10.15–11.30am and 11.00am–3.30pm, respectively. Approximately 14 minutes of camera (∼2,500 frames) and RGB-D person detection data of the trials was collected during the experiment from a Kinect mounted in the robot's head. The ground truth of participants' gaze behaviour was independently coded by three experimenters and averaged together post-hoc to quantify at which points participants were Looking and Not Looking at the robot: example images of participants Not Looking and Looking are shown in the top and bottom rows of Figure 5.4, respectively. Participants included in the results are those who were manually observed to have looked at the robot at some point during their interaction. Thus, the robot could have issued a cue directed at them which would have been witnessed.

Figure 5.5 shows the gaze patterns of the participants towards and away from the robot over the duration of their interactions. White represents points at which participants' gaze was directed towards the robot (L), and black represents the opposite (!L). Grey areas indicate the participant had exited the interaction and the trial had concluded. From the figure, it can be seen that there was a large degree of variation in the number ($\bar{x} = 3.5$, $\sigma = 2.4$) and length of looks at the robot ($\bar{x} = 2.1s$, $\sigma = 4.3s$), and the total percentage of time spent Looking at the robot over the course of the interactions ($\bar{x} = 24\%$, $\sigma = 22.7\%$).

The number of participants whose gaze was directed at the robot and the total number of participants still interacting over time are shown in Figure 5.6. It can be seen that, on average, approximately 50% of participants were looking at the robot at any point in time.

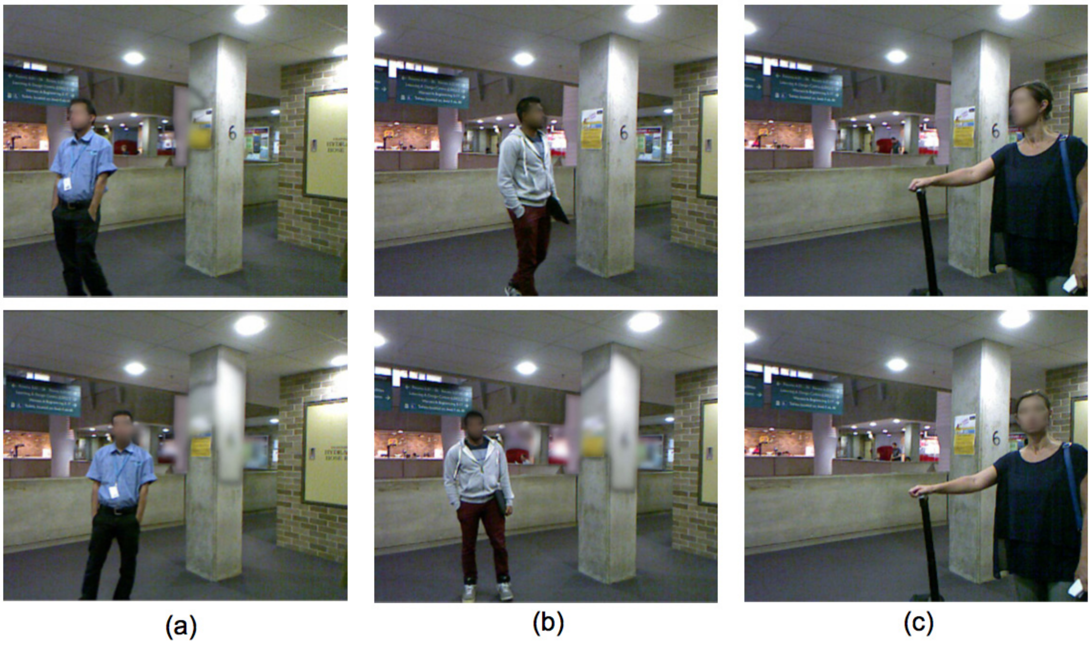


Figure 5.4: Participants Not Looking (top row) and Looking (bottom row) at the RobotAssist platform during the experiment.

5.1.8 Discussion

The results provide support for the hypothesis. No participant spent the entirety of their interaction Looking at the robot, and no readably generalisable pattern of Looking behaviour was observable between participants, as shown in Figure 5.5 and in the empirical results. For example, 11 participants were in their interaction for a length of time before their first look at the robot, and

22 participants carried out their final look at the robot a length of time before exiting the interaction; the interactions lacked a point at which a cue issued by the robot would have been reliably observed by all participants. Instead, as can be seen from Figure 5.6, at any one time only some of the people still in their interaction were Looking at the robot, and an issued cue was only likely to be observed by a maximum of approximately 50% of people. Similarly, it is difficult to determine a time when an action not intended to be witnessed could have been carried out (i.e. all participants were Not Looking).

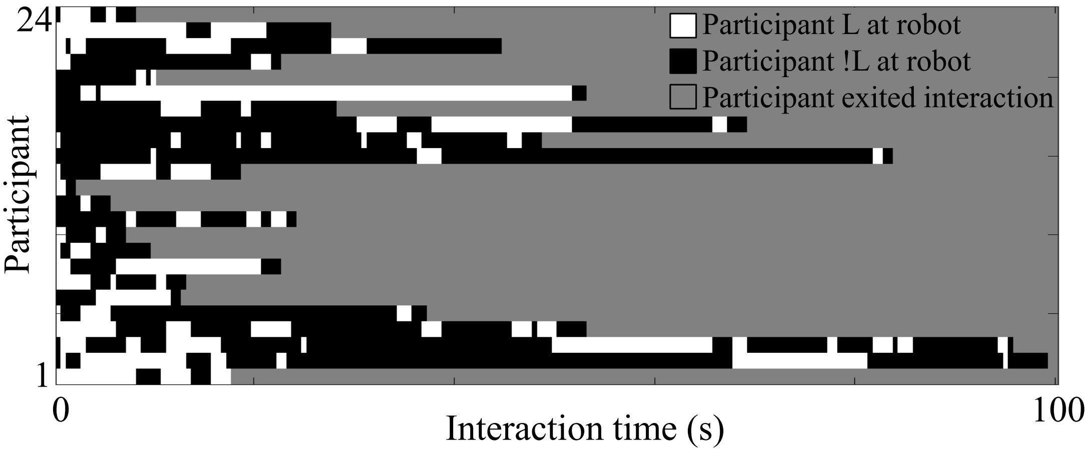


Figure 5.5: Looking patterns of participants at the robot over the course of their interactions.

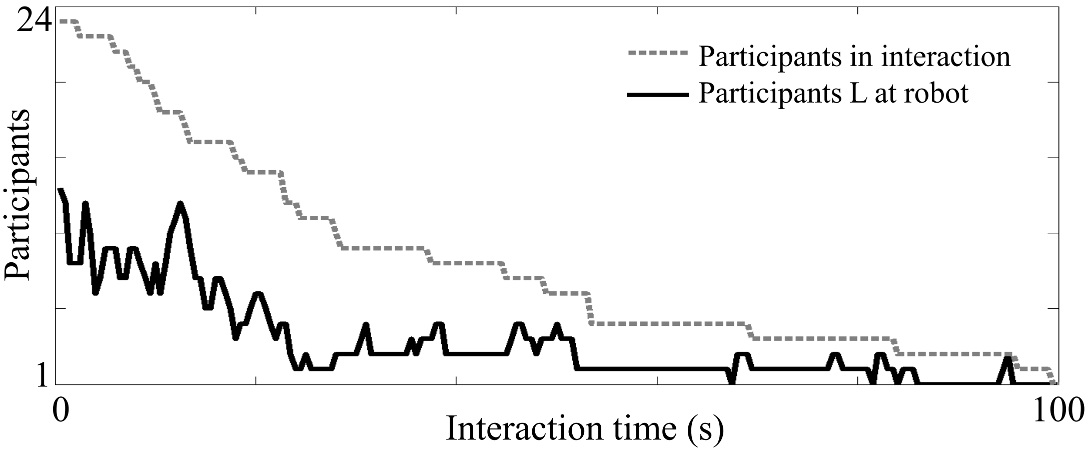


Figure 5.6: The number of participants looking at the robot and those still interacting over time.

5.1.9 Conclusions

This study, published in ***[C2]*** of Appendix A, focused on understanding people's natural gaze behaviour towards robots during unsolicited, real-world HRI. The results show that there is no observable generalisable pattern of gaze towards robots during real-world HRI. This suggests that the capability of detecting gaze behaviour *in situ* would be advantageous (as per **RQ A.2: *Read* Value**), for instance by enabling robots to effectively *Elicit*: issuing cues the robot wants reliably witnessed when the intended recipient is engaged in mutual gaze with the robot, for example.

5.2 Development of Head Yaw Estimation for the HRI Space

The results of the experiment presented above in Section 5.1 demonstrated that there is no observable generalisable pattern of mutual gaze towards robots during real-world HRI. Thus, it becomes necessary for robots to have an understanding of interaction partners' gaze, for example to ensure their issued cues are reliably witnessed, as well as to detect the gaze cues humans display in response to robot-issued JA cues (as investigated in Chapter 4). The question then arises: will human-issued gaze cues be detectable and interpretable by social robots (as per **RQ A.2: *Read* Feasibility**)?

Several considerations around enabling robots to have *in situ* gaze cue *Read*ing capabilities were discussed in the literature presented in Chapter 2. Firstly, it was shown in Section 2.1.2 that humans have an *interaction zone* in which a majority of their interactions take place, and therefore in which HRI is also likely to take place. Thus, this *interaction zone* can also be considered the HRI space.

As outlined in Section 2.1.3, the positioning of interaction partners within this *interaction zone* moderates how cues such as gaze are utilised: interaction initiation via mutual attention likely occurs in the far-*interaction zone* ($\sim$2–3m in the x direction, $\pm\sim 1m$ in the y direction), while situated interactions – where cues such as JA are often employed – are more likely to take place in the near-*interaction zone* ($\sim$1.2–2m in the x direction, $\pm\sim 0.5m$ in the y direction).

Additionally, during real-world HRI, where people's eyes and hence exact gaze direction may not always be detectable, literature shows that head yaw can be used as an indicative measure of gaze direction, as discussed in Section 2.3. Many features of the human head could be used for such head yaw estimation (Section 2.3.1). However, due to the characteristics of the commonly used Kinect data, the reliability of distinguishing different facial features in such data depends on the interaction distance (Section 2.3.2): finer features are more reliably detected in the near-*interaction zone*, however coarser facial features, while possibly less accurate, are likely to be usable for head yaw estimation in a wider area of the *interaction zone*.

It would therefore be ideal to incorporate such complementary data into a single model to leverage the individual strengths of multiple head yaw estimation (HYE) methods (accuracy vs coverage of the *interaction zone*) into a head yaw estimate which operates across the entirety of the *interaction zone*. This data could then potentially be fused into a single head yaw estimate, as discussed in Section 2.3.3. This would give a robot the ability to *Read* both mutual gaze in the far-*interaction zone* and JA and gaze bias cues in the near-*interaction zone* during real-world HRI, while maintaining necessary levels of accuracy across the entire zone (where the requirement is often to detect which landmarks gaze is directed at, as in Chapter 4).

5.2.1 Existing Head Yaw Estimation Approaches

A number of fusion systems currently exist which have been designed to overcome the limitations of independent HYE methods by fusing estimates from complementary approaches into a single result, leveraging the strengths of multiple sources [112]. Some examples of fusion systems include [10, 109, 139, 170]. In [139], for example, appearance template matching is fused with geometric cues. However, thus far many hybrid methods have focused on increasing estimation accuracy within a defined area (either near or far), rather than on operating over a greater proportion of the *interaction zone*. This makes them unsuitable for the intended application, where the requirement is for HYE which operates over the entirety of the *interaction zone*, depicted in Figure 2.1, while maintaining levels of accuracy necessary for HRI across the entire space.

As such, investigation into individual HYE approaches found that many currently exist which have the potential to be utilised in a developed HYE fusion framework. An overview of these methods can be found in [112], which can be broadly categorised into 2D (image-based) and 3D (depth data-based) techniques.

Methods based on 2D images can be further segregated into appearance model-based methods (e.g. [164]), which analyse the entire facial region, and feature-based methods, which localise specific facial features such as the eyes or nose (e.g. [21, 47]). Yet there are a number of disadvantages to 2D image-based methods which make them unsuitable for the target *interaction zone*, particularly the far-*interaction zone*, such as sensitivity to facial expression, identity and illumination variations, their limitation to discrete poses, and the low resolution of images which makes it necessary for the person to be fairly close to the sensor. Additionally, many feature-based 2D methods require the same facial features to be visible across different poses, or define pose-dependent features, limiting them to applications where near frontal images of people's faces are ensured [112, 152].

In response to some of these 2D image HYE challenges, the use of depth-sensing technologies and range data (such as that generated by the Kinect) has become more widespread. A state of the art example is the work of Fanelli et al. [40]. This work presents a real-time algorithm to estimate head pose from low quality depth data by learning a mapping between simple depth features and real parameters, such as 3D head position and rotation angles. The process and information flow of this method is depicted in Figure 5.7. As shown in the figure, discriminative random regression forests are used to classify depth image patches belonging to a person's head, making the method independent of torso pose, then to perform a regression in the continuous spaces of head positions and orientations to estimate head pose.

Discriminative random regression forests are used to classify depth image patches belonging to a person's head, making the method independent of torso pose, then to perform a regression in the continuous spaces of head positions and orientations to estimate head pose.

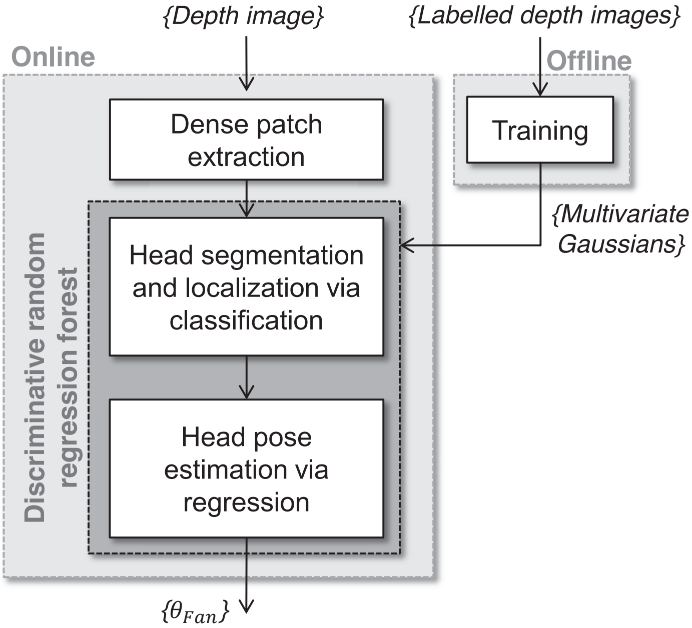

{Depth image}
{Labelled depth images}
Online
Offline
Dense patch extraction
Training
{Multivariate Gaussians}
Discriminative random regression forest
Head segmentation and localization via classification
Head pose estimation via regression
$\{\theta_{Fan}\}$

Figure 5.7: Process and information flow of the Fanelli et al. HYE method.

From a labeled head pose database of people positioned ∼$1m$ from the sensor and captured by the Kinect, the trees which constitute the forest are trained in order to jointly optimise their classification (of head vs not head points) and regression (of head pose) power. By maximising these two separate measures, a mean HYE accuracy of $\sim 5.7° \pm 15.2°$ is achieved when participants are located in similar regions to those in the training data, equaling or exceeding that of many other established methods [112]. This accuracy is suitable for a foreseeable set of HRI applications, in which the requirement is often to detect which landmarks gaze is directed at. Additionally, the Fanelli et al. method works on a frame-by-frame bases, does not require initialisation, and is capable of handling multiple people, large head pose variations, variations in appearance due to features such as facial hair or glasses, and partial occlusions. Thus, the Fanelli et al. method is suitable for inclusion in a developed fusion HYE framework, as its strength lies in accurate HYE in the near-*interaction zone*.

However, a key shortcoming of the Fanelli et al. method is that it only operates with HRI-suitable levels of accuracy over a limited section of the *interaction zone* (as depicted in Figure 2.12): to estimate head pose, the method relies on facial features which can be reliably detected within the near-*interaction zone*, but become more difficult to distinguish in the lower resolution data from far-*interaction zone* distances as the data becomes increasingly featureless.

To overcome this limited usefulness of the Fanelli et al. method, it is necessary to include a method of HYE which encompasses a wider area of the *interaction zone* in a developed HYE framework, for fusion with the Fanelli et al. estimate. However, while a variety of other head pose estimation methods exist, many have similar characteristics, as discussed in [112]: near-*interaction zone* accuracy is often prioritised at the expense of breadth of operation space (as seen in [95, 108], for example). This limits the additional useful information they would add if fused with the Fanelli et al. estimate.

5.2.2 Developed Head Yaw Estimation Framework

Thus, there was a need for HYE which operates over the entirety of the *interaction zone* while maintaining HRI-suitable, landmark levels of accuracy across the entire space. To address this need, a HYE framework was developed which fuses multiple HYE methods, including the Fanelli et al. method discussed above, and a novel method which is detailed below.

The framework, which is detailed in a publication currently under review (***[J2]*** of Appendix A), is depicted in Figure 5.8. It consists of two main stages: fusion input preparation (discussed below in Section 5.2.2), and data fusion. In order to evaluate both Bayesian and GP HYE fusion techniques, each of which have individual advantages, the data fusion was configured and tested in two ways. This is further detailed in Section 5.2.2.

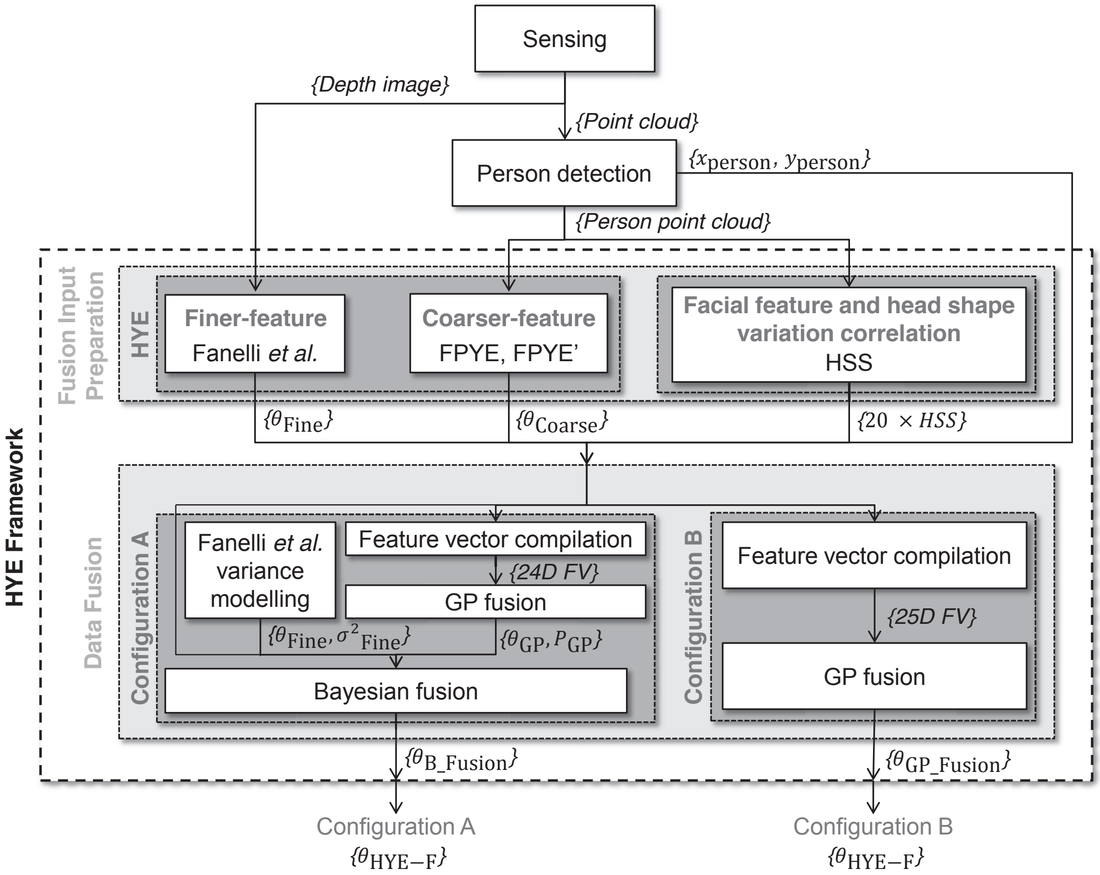

Sensing
{Depth image}
{Point cloud}
Person detection
{$x_{\text{person}}, y_{\text{person}}$}
{Person point cloud}
HYE Framework
Fusion Input Preparation
HYE
Finer-feature
Fanelli et al.
Coarser-feature
FPYE, FPYE'
Facial feature and head shape variation correlation
HSS
{θ_{Fine}}
{θ_{Coarse}}
{$20 \times HSS$}
Data Fusion
Configuration A
Fanelli et al. variance modelling
Feature vector compilation
{24D FV}
GP fusion
{$\theta_{\text{Fine}}, \sigma^2_{\text{Fine}}$}
{$\theta_{\text{GP}}, P_{\text{GP}}$}
Bayesian fusion
Configuration B
Feature vector compilation
{25D FV}
GP fusion
{$\theta_{\text{B\_Fusion}}$}
{$\theta_{\text{GP\_Fusion}}$}
Configuration A
{$\theta_{\text{HYE-F}}$}
Configuration B
{$\theta_{\text{HYE-F}}$}

Figure 5.8: Process and information flow of the HYE framework.

Fusion Input Preparation

Preparation of the data to be fused involves a number of steps, which are illustrated in Figure 5.8. Firstly, a sensing step outputs 3D depth data from the Kinect sensor in point cloud and depth image form. In order to achieve higher accuracy HYE in the near-*interaction zone*, the depth image is fed through a HYE step which leverages finer facial features. Here the Fanelli et al. method [40] is used to produce the first head yaw estimate for fusion, $\theta_{\text{Fine}} = \theta_{\text{Fanelli}}$.

A preliminary investigation verified that the operation space of the finer feature HYE step is limited, as discussed in Section 2.3.2 and depicted in Figure 2.12. In order to complement and extend this estimate's operation space, a coarser feature HYE step is also employed. To reduce the computational expense of the search for these features, people in the environment are first robustly detected and segmented, limiting the data which must be processed in the coarser feature stage: the data is passed through a previously developed person detection system [64] known to work in both the near- and far-*interaction zones* [85]. For development purposes, in this case the system was configured to detect just one person, however it is capable of detecting multiple people (having been shown to detect all 9 people in a crowded scene [83]). The output of this step is the person location (x_{person}, y_{person}) and a high-confidence person point cloud which can be used for HYE.

With the data reduced to the person of interest, the person point cloud is passed to the coarser feature HYE step. Further discussed below in *Coarser-Feature Head Yaw Estimation*, this stage employs a novel HYE method based on features which can be reliably detected in both the near- and far-*interaction zone*, complementing and extending the area of the *interaction zone* over which the finer feature HPE stage operates to produce a head pose estimate θ_{Coarse} for fusion. The operation space of this stage was also verified in a preliminary investigation: it was found that the method operates over a wider area of the *interaction zone*, as depicted in Figure 2.12.

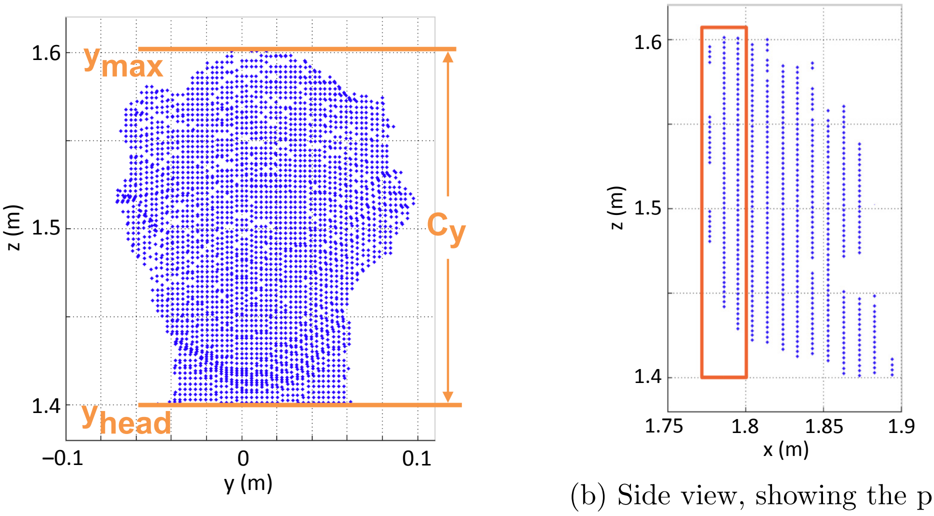
However, the features on which the finer and coarser methods rely for HYE will vary between individuals, resulting in inaccuracies in the θ_{Fine} and θ_{Coarse} head yaw estimates between people. It is reasonable to assume, though, that a contributor to this facial feature variance will be the shape of people's heads, and that the feature variance will in some way be proportional to this head shape. For example, facial plane characteristics will be similar between people with narrower heads, and likewise for those with wider heads. To increase the accuracy of the fused head yaw estimate, this association between facial features and head shape variation is also made available for data fusion via the method detailed below.

Finally, as the finer and coarser HYE steps operate over different areas of the *interaction zone*, the position of the person relative to the sensor ($x_{\text{person}}, y_{\text{person}}$), as determined by the person detection stage, is also sent to the data fusion stage. This location information adds useful information to the process, which is further discussed in Section 5.2.2.

Coarser-Feature Head Yaw Estimation

A novel method, Face Plane Yaw Estimation (FPYE), was developed which leverages the planar facial feature of people's heads for HYE; while no less susceptible to sensor noise at larger distances than finer facial features, the coarser facial plane is more likely to be reliably observable, making it a suitable feature for far-*interaction zone* HYE. As a person's facial plane is inherently perpendicular to their heads, determination of its orientation can be utilised to estimate head yaw.

In order to ensure reliable detection of the facial plane and independence from torso pose, the first step of the FPYE method is to segment the head region of the person. Operating on the high-confidence point cloud of the person from the person detection step, the person's point cloud data is searched to find the highest value, y_{max}. A second value $y_{head} = y_{max} - C_y$ is then calculated, which specifies the height threshold for the head. Figure 5.9a depicts the segmentation parameters: y_{max} is shown to be at the maximum point in the cloud, with y_{head} at the bottom of the segmented region. The shape height constant, C_y, defines the vertical size of the shape to be segmented.



(a) Frontal view, showing the segmentation parameters.

(b) Side view, showing the point stratification and planer appearance of the head point cloud at greater distances.

Figure 5.9: The segmented head point cloud.

The point cloud is then processed to remove pixels whose height is less than y_{head}. As the threshold height, y_{head}, is relative to the height of the person's point cloud, y_{max}, the segmentation is invariant to the height of the person. A constant value of $C_y = 0.2m$ was empirically determined in previous work to robustly capture the head region [64]. The output of this process is depicted in Figure 5.9a: the head region has been segmented from the person point cloud.

With this 3D representation of the person's head extracted, the orientation of the facial plane can be used to determine head yaw. The orientation of a person's facial plane is inherently perpendicular to their head yaw; thus, the normal of their facial plane should align with this head yaw. Figure 5.10a illustrates this normal on the point cloud of a person's head, viewed from the top down. This head yaw can be estimated by calculating the angle between the facial plane normal and the sensor's x-axis, θ_{FPYE}, as depicted in the figure. To calculate this angle, the coefficients of the facial plane in the x and y directions can be used, $\theta_{\text{FPYE}} = \arctan\left(\frac{y_{\text{Facial plane}}}{x_{\text{Facial plane}}}\right)$.

In calculating these coefficients, both the natural planar characteristics of a person's face – and the fact that at far-*interaction zone* distances the planar appearance of the face is also artificially emphasised by the characteristics of the Kinect sensor data, as discussed in Section 2.3.2 – is exploited. This phenomena can be seen in Figure 5.9b, which illustrates how the data stratification at larger distances results in the true depth features of a person's face being compressed into a flatter-than-reality representation in the x direction, as highlighted by the box drawn in the figure. This results in a greater density of points in the facial plane area, with minimal variance in the x direction and greater variance in the y and z directions.

This characteristic high point density and minimal variance in the x direction is exploited through Principal Component Analysis (PCA) to determine the coefficients of the facial plane in the 3D head point cloud. When applied to such a point cloud, PCA calculates the principle components of the data, i.e. the vectors with the greatest projection covariance. The first component is that with the largest possible variance (thus this component will 'explain' the largest part of the data variance), and subsequent components are each computed with the constraint that they are orthogonal to the previous component, and explain a maximum possible part of the remaining variance.

For example, consider when a person's face is oriented towards the sensor, and thus their facial plane is the primary visible plane, as in Figure 5.10a. In such a head point cloud, there will have a greater density of points in the facial plane area, and maximum variance in the y and z directions. The first two PCA components are thus likely to be aligned with the y- and z-axes, respectively, as there is less variance along the x-axis due to the point cloud's planar nature. Thus the plane formed by these first two principle components will align with the facial plane, and θ_{FPYE} can be calculated as the angle between the PCA-plane's normal and the sensor x-axis, as shown in Figure 5.10a.

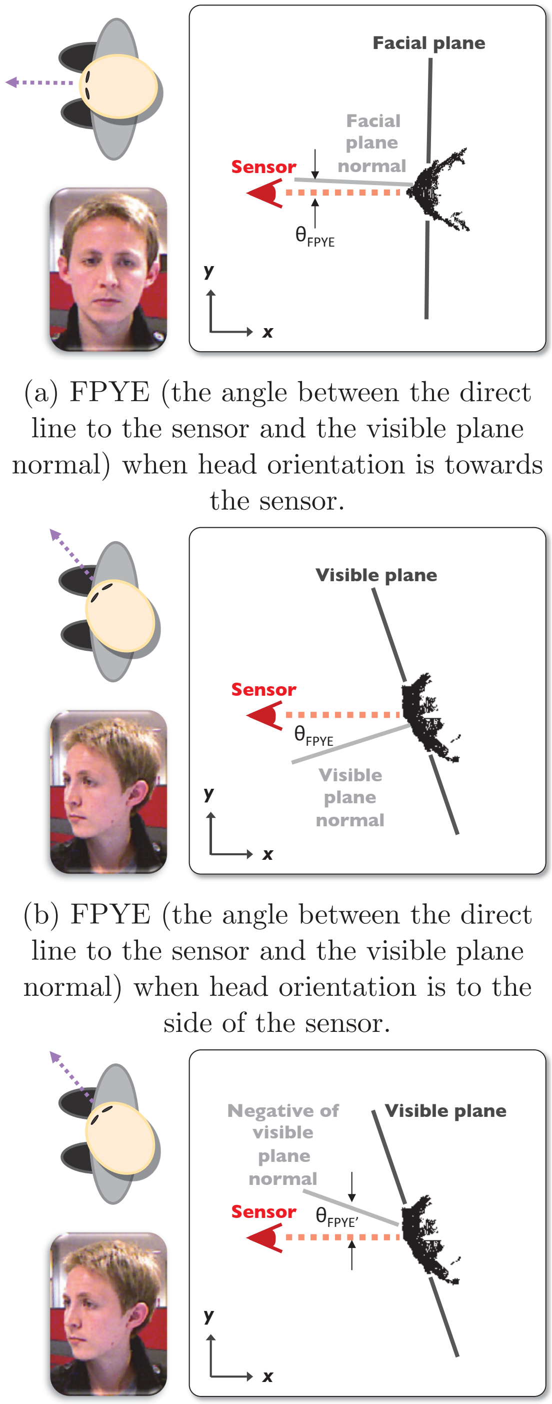
When the person's head is oriented towards the sensor, this estimate θ_{FPYE} is aligned with the true head yaw, as seen in Figure 5.10a. In this case, the principal plane visible to the sensor – and thus the plane with which the PCA coefficients will align – will be the facial plane area. However, with increasing rotation of the head from the sensor, different areas of the head become increasingly observable and begin to have greater point density than the facial plane. This can be seen in Figure 5.10b, where the primary plane visible to the sensor has become the side of the head. As a result, the principle visible plane, and hence PCA plane, no longer align with the facial plane. The angle of the PCA plane's normal to the sensor x-axis then begins to diverge from the true angle of the person's head yaw relative to the sensor. The FPYE method therefore has its highest accuracy when gaze is on the robot, i.e. mutual gaze. As previously discussed, this is an important attention cue, especially at far-*interaction zone* distances when people show signals of openness to interaction and where interaction initiation is likely to take place.

However, in the far-*interaction zone* it is also valuable to have a general indication of head yaw, for example to gauge people's level of interest in and intentions towards objects in the environment. Thus, the FPYE method was adapted into Face Plane Yaw Estimation' (FPYE'). As the head rotates and the PCA plane normal diverges from the true head yaw, at some degree of head rotation the angle of the normal to the sensor will cross over the 0° point into the negative. This can be seen in Figure 5.10b: though the true head rotation is in the positive direction, the angle of the PCA plane normal to the sensor is in the negative direction. After this point, with increasing head rotation in the positive direction, the normal angle will become increasingly negative, and vice versa. By taking the negative of this angle (i.e. the negative of the FPYE result, $\theta_{\text{FPYE'}} = -\theta_{\text{FPYE}}$), as shown in Figure 5.10c, the estimation error (compared to the FPYE method) is reduced.

This simple technique results in the FPYE' method having its highest accuracy when head orientations are away from the sensor, complementing the head yaw estimation range of the FPYE method. There is no rule to discern between these two cases: as discussed below in Section 5.2.2, the subsequent GP modelling of the data, including the coarse head yaw estimates θ_{Coarse}=$\{\theta_{\text{FPYE}}, \theta_{\text{FPYE'}}\}$, is able to inherently capture, via training, the relationship between these methods and the weights which should be assigned to each in different situations.

Facial Feature and Head Shape Variation Correlation

However, the facial features on which the FPYE and FPYE' methods rely for HYE will vary between individuals. This could result in discrepancies in readings between people, even when their true head yaw is similar. Yet it is likely that this



(a) FPYE (the angle between the direct line to the sensor and the visible plane normal) when head orientation is towards the sensor.

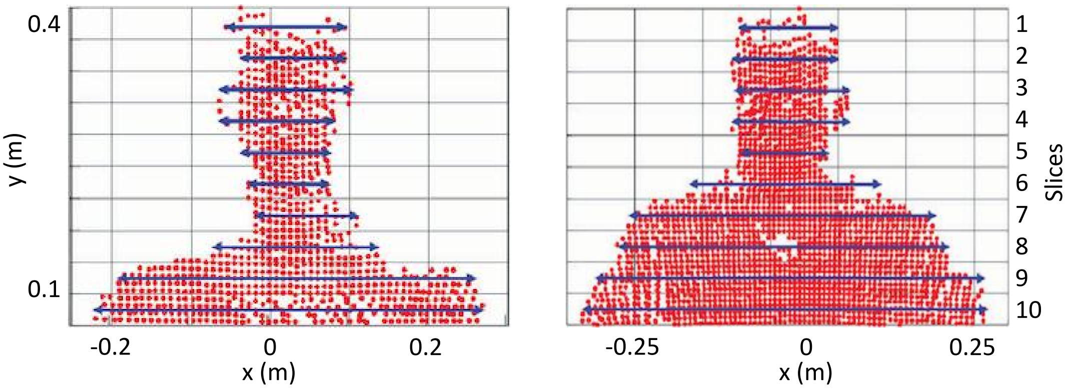
(b) FPYE (the angle between the direct line to the sensor and the visible plane normal) when head orientation is to the side of the sensor.

(c) FPYE' (the angle between the direct line to the sensor and the negative of the visible plane normal) when head orientation is to the side of the sensor.

Figure 5.10: Illustration of the visible plane and FPYE and FPYE' methods on a head point cloud viewed from the top down. The direct line between the sensor and head is indicated in red dashes, the visible plane is shown in solid black, and the visible plane normal ((a) and (b)) or its negative (c) in lighter gray.

facial feature variance will in some way be correlated with the particular individual from which the data arose. In order to 'calibrate' the readings between people, this correlation is exploited through our Head-to-Shoulder Signature (HSS) [85], shown in Figure 5.11, in the coarse-feature HYE method. Originally developed to achieve robust *in situ* person recognition, the HSS encapsulates individual-specific head, neck and shoulder region size and and shape information in a scale and viewing angle robust feature vector, and has been shown to be robust against a large spread of variations in appearance and clothing [83].

While [85] demonstrated the HSS as the basis for robust Support-Vector Machine-based person recognition, it is used here to enable the fusion to be responsive to differing facial characteristics between people, and their relationship with head yaw. While cases may potentially exist in which outliers are not handled by the HSS, it has been frequently used with large samples sizes (for example N = 420 [83]) and this issue has not appeared to be prevalent.



(a) Narrower head person.

(b) Wider head person.

Figure 5.11: Point clouds of two different people with 10 slice spans of the HSS illustrated. Source: [85]

Data Fusion

A number of entities are now available from the fusion input preparation stage of the framework: $\theta_{\text{Fine}} = \theta_{\text{Fanelli}}$, $\theta_{\text{Coarse}} = \{\theta_{\text{FPYE}}, \theta_{\text{FPYE'}}\}$, x_{person}, y_{person} and $20 \times$ HSS. This information is then passed to the fusion stage. As illustrated in Figure 5.8, two different fusion configurations were trialled to determine which configuration most successfully achieves head yaw estimation over the entirety of the *interaction zone* with HRI-suitable accuracy.

As both configurations incorporate GPs, they are first described in detail.

Gaussian Processes

The fusion of data via GPs can be carried out by learning a joint model of an input feature vector, $\mathbf{g}$, which varies slightly between the two fusion configurations. GPs are characterised by a mean $m(\mathbf{g})$ and covariance $k(\mathbf{g}, \mathbf{g}')$ function which together specify a distribution over functions; that is,

$$f(\mathbf{g}) \sim GP(m(\mathbf{g}), k(\mathbf{g}, \mathbf{g}')). \tag{5.1}$$

In the current context, the mean function $m(\mathbf{g})$ can be assumed to be zero by scaling the data such that it has a mean of zero. With corresponding head yaw outputs $f(\mathbf{g}) = h$, where $\mathbf{h}$ is the head yaw angle, and denoting groups of these points as $(\mathbf{G}, \mathbf{f}, \mathbf{h}) = (\{\mathbf{g}_i\}, \{f_i\}, \{h_i\})_{i=1}^{N}$ for the training set and $(\mathbf{G}_*, \mathbf{f}_*, \mathbf{h}_*) = (\{\mathbf{g}_{*,i}\}, \{f_{*,i}\}, \{h_{*,i}\})_{i=1}^{N}$ for the testing points, the joint Gaussian distribution with $m(\mathbf{n})$=0 is:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(G,G) & K(G,G_* \\ K(G_*,G) & K(G_*,G_*) \end{bmatrix}\right) \tag{5.2}$$

where $\mathcal{N}(\boldsymbol{\mu}, \mathrm{P})$ is a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance P, and K is used to denote the covariance matrix computed between all the points in the set. If head yaw outputs are assumed to have Gaussian noise ϵ and variance σ^2 such that $h = f(\mathbf{g}) + \epsilon$, the joint distribution becomes [104]:

$$\begin{bmatrix} \mathbf{h} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(G,G) + \sigma^2 I & K(G,G_*) \\ K(G_*,G) & K(G_*,G_*) \end{bmatrix}\right) \tag{5.3}$$

The covariance function models the relationship between the random variables corresponding to the given data [50, 160] and has a set of hyperparameters $\boldsymbol{\phi}$ associated with it. The covariance function selection plays a significant role in the success of the GP; the process of learning with a GP is the problem of learning these hyperparameters. After consideration of several covariance functions, the Squared Exponential (SE) covariance function k_{SE} was chosen. This covariance function is widely used [127] and has been successfully applied in a broad variety of applications (e.g. [5, 16, 4]). Additionally, it is a stationary covariance function; it is infinitely differentiable, resulting in a GP with this covariance function having mean square derivatives of all orders and thus being very smooth [127]. This is appropriate for the real-life variable being modelled by the GP, head yaw, the behaviour of which is inherently smooth.

After observing with the Automatic Relevance Determination (ARD) SE covariance function that all input features were being weighted with the same order of magnitude, the Isotropic Squared Exponential covariance function $k_{\text{SE-ISO}}$ was utilised. This is a less complex and therefore faster running version of the SE covariance function. It is parameterised as:

$$k_{\text{SE-ISO}}(\mathbf{g}, \mathbf{g}') = \sigma_f^2 \exp(-\frac{1}{2\ell^2}(\mathbf{g} - \mathbf{g}')^\top(\mathbf{g} - \mathbf{g}')) \tag{5.4}$$

where σ_f^2 is the signal variance and ℓ is the characteristic length-scale.

The learning of the hyperparameters was achieved by maximising the log of the marginal likelihood with respect to $\boldsymbol{\phi}$ [127]:

$$\log p(\mathbf{h}|G, \phi) = -\frac{1}{2}\mathbf{h}^T K_h^{-1}\mathbf{h} - \frac{1}{2}\log|K_h| - \frac{N}{2}\log 2\pi \tag{5.5}$$

where $K_h = K(G, G) + \sigma^2 I$ is the covariance matrix for the targets $\mathbf{h}$.

By training the model on a training dataset $D_{tr} = (g_i, h_i) \mid i = 1, \ldots, N$, the predictive distribution of the input 24-dimension feature vector g_i to the corresponding scalar training output h_i can be obtained. This training output is the ground truth (GT) head to sensor angle θ_{GT}, and this training data was gathered via the method outlined below. The predictive distribution is:

$$p(f_*|G_*, G, \mathbf{h}) = \mathcal{N}(\boldsymbol{\mu_*}, P_*) \tag{5.6}$$

where

$$\begin{aligned}\boldsymbol{\mu_*} &= K(G_*, G)[K(G, G) + \sigma^2 I]^{-1}\mathbf{h} \\ P_* &= K(G_*, G_*) - K(G_*, G)[K(G, G) + \sigma^2 I]^{-1} \ldots \\ &\qquad\qquad\qquad\qquad \ldots K(G, G_*) + \sigma^2 I\end{aligned}$$

Training Dataset for GP Model

In order to build the joint model of the input entities and fuse them into a single head yaw estimate, training data for the GP model was required in the form $D_{tr} = (g_i, h_i) \mid i = 1, \ldots, N$, as described above. To capture interpersonal variations, approximately N=2,400 training data points were gathered from 4 student subjects (3 male and 1 female).

As training data from the subjects was required over the entirety of the *interaction zone*, a methodology for estimating ground truth with a usable level of accuracy for social HRI was necessary. A wide range of head yaws relative to the sensor were therefore captured through the following procedure: the subject sat on a roller chair in front of a Kinect sensor at y direction offsets of $y_{\text{person}}{\approx}0m$, $y_{\text{person}}{\approx}0.5m$ and $y_{\text{person}}{\approx}1m$, as depicted in Figure 5.12.

At each offset, 15 tests were carried out. At the beginning of each test the person was directed to orient their head towards one of 15 signs posted $\approx 1.8m$ high on the wall directly in front of them and behind the Kinect. The signs were placed in order to give an accurate angle β measurement at a distance $6m$ from the wall ($1m$ from the sensor). The range of β at this distance was $\beta = -35° : 5° : 35°$, which was selected because of the primary goal of detecting mutual attention, when a person's head is oriented towards the robot. During tests, the person's head position relative to the sensor was extracted and the corresponding sign post's location was known, thus, the head orientation β could be computed.

Over a period of $\sim$4–6s, the subject then rolled the chair backwards from a distance of $x_{\text{person}}{\approx}1m$ from the sensor to $x_{\text{person}}{\approx}3m$, keeping their head oriented towards the appropriate wall sign during the movement. As the data from the $y_{\text{person}}{\approx}0.5m$ and $y_{person}{\approx}1m$ offsets is likely to be symmetrical and thus also correspond to $y_{\text{person}}{\approx}$–$0.5m$ and $y_{\text{person}}{\approx}$–$1m$ offsets, the entirety of the *interaction zone* is represented.

Though there may be uncertainties in the training data due to the above procedure (such as inexact β), such uncertainties are handled in a statistically sound manner by the GP model, which incorporates the uncertainty into the prior and learns the ground truth noise from the data during model training.

In order to verify the value of both the HSS and person location to the GP model, the training dataset was divided into training and testing data, and a number of 5-fold cross validations with different input feature vectors, $\boldsymbol{g}$, were run. The resulting root mean square (RMS) errors are shown in Table 5.1. It can be seen that the removal of the HSS, and both the HSS and person location, resulted in e_{RMS} increases of 1.5° and 2.4°, respectively, compared to the full 25-dimension feature vector. Similarly, the removal of θ_{Fanelli} from the model resulted in a higher e_{RMS} of 13.2°, demonstrating the value of including θ_{Fanelli} in the fusion, whether Bayesian or GP.

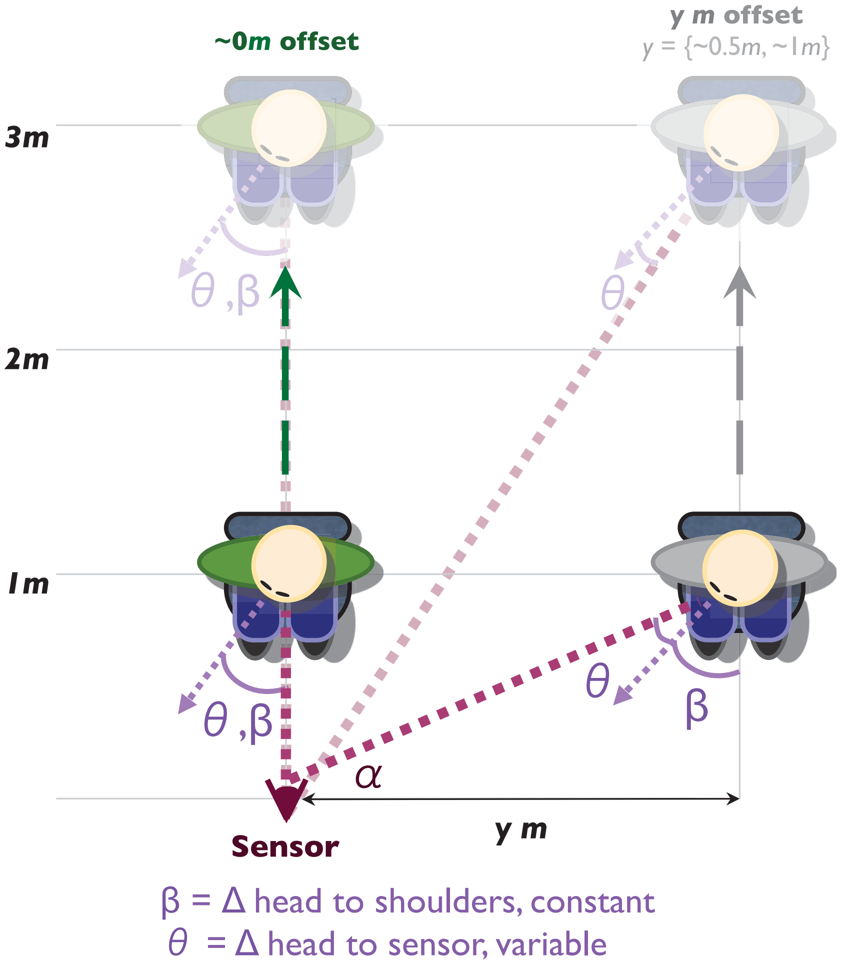

~0m offset
y m offset
y = {~0.5m, ~1m}
3m
2m
1m
θ ,β
θ
θ ,β
θ
β
α
Sensor
y m
β = Δ head to shoulders, constant
θ = Δ head to sensor, variable

Figure 5.12: Setup and procedure of GP model training data acquisition.

Bayesian Fusion

The first configuration utilises Bayesian fusion. The advantage of this approach is that the strengths of each of the methods can be maximised: in areas where θ_{Fine} is operating, its higher-accuracy estimate can be given larger weighting in the fusion by manually assigning the estimate a small variance. However, in the wider-*interaction zone*, when no θ_{Fine} estimate is available, a very large variance can manually assigned to θ_{Fine}, which results in other data being weighted more highly.

Table 5.1: 5-fold cross validation RMS model errors for different input feature vectors, $\boldsymbol{g}$.

Input feature vector, $\boldsymbol{g}$	$\boldsymbol{e_{RMS}}$
25D: $\{\theta_{\text{Fanelli}}, \theta_{\text{FPYE}}, \theta_{\text{FPYE'}}, x_{\text{person}}, y_{\text{person}}, 20 \times \text{HSS}\}$	11.5°
24D: $\{\theta_{\text{FPYE}}, \theta_{\text{FPYE'}}, x_{\text{person}}, y_{\text{person}}, 20 \times \text{HSS}\}$	13.2°
5D: $\{\theta_{\text{Fanelli}}, \theta_{\text{FPYE}}, \theta_{\text{FPYE'}}, x_{\text{person}}, y_{\text{person}}\}$	13.0°
3D: $\{\theta_{\text{Fanelli}}, \theta_{\text{FPYE}}, \theta_{\text{FPYE'}}\}$	13.9°

The first input to Bayesian fusion is a prior. In this fusion configuration, the prior was developed using the entities available from the fusion input preparation stage, excluding θ_{Fine}. These entities have a relationship in the wider-*interaction zone*: the HSS adds useful information about individuals' planar facial features, information relevant to θ_{FPYE} and $\theta_{\text{FPYE'}}$, and the person location relative to the sensor (x_{person}, y_{person}) is also useful as FPYE and FPYE' operate over a wide area of the *interaction zone.* To capture this relationship, the data is compiled into a 24-dimension input feature vector, $\mathbf{g}$=$\{\theta_{\text{FP}^2\text{E}}$, $\theta_{\text{FP}^2\text{E'}}$, x_{person}, y_{person}, $20 \times \text{HSS}\}$. A joint model of $\mathbf{g}$ is then learnt through a GP, inherently fusing the entities into a single head yaw estimate $\theta_{\text{GP}} = \mu_*$ with covariance matrix P_{GP}. This becomes the Bayesian fusion prior, $p(\theta|G) \sim \mathcal{N}(\mu^-, P^-)$, where $\mu^- = \theta_{GP}$, $P^- = P_{GP}$, and θ is the true head yaw relative to the sensor.

This GP fusion is similar to that carried out in the second framework configuration, detailed below in *Gaussian Process Fusion.* In the Bayesian fusion configuration, however, θ_{Fine} is left out of the GP fusion, resulting in a 24-dimension, rather than 25-dimension, input feature vector.

The second input to the Bayesian fusion is another measurement, θ_{Fine}. This is modelled as $p(\theta_{\text{Fine}}|\theta, G) = \mathcal{N}(\mu, \sigma)$, with $R = \sigma^2_{Fine}$. In this case, $\sigma^2_{Fine} = \sigma^2_{Fanelli}$, the noise variance of the $\theta_{\text{Fine}} = \theta_{\text{Fanelli}}$ estimate. The Fanelli et al. method does not inherently output a variance; instead, its true operating variance was experimentally characterised via training data to be $\sigma^2_{Fanelli} = 0.66^\circ$. This was calculated by averaging the variance of the method over 10 readings at 3 different locations throughout the *interaction zone.* As a result of this low uncertainty, this more accurate head yaw estimate is likely to be 'trusted' over the coarser feature based θ_{GP} estimate. Conversely, in areas where the Fanelli et al. method is not operating, it is assigned a large variance of $\sigma^2_{Fanelli} = 1000^\circ$ such that the GP estimate will be used.

The maximum a *posterior* estimator is then used to integrate the prior GP inference together with the θ_{Fine} measurement. It is formulated as

$$\underset{\theta}{\text{argmax}}\; p(\theta|\theta_{\text{Fine}}, G). \tag{5.7}$$

Then the posterior density is

$$p(\theta|\theta_{\text{Fine}}, G) \propto p(\theta_{\text{Fine}}|\theta, G) \times p(\theta|G) \tag{5.8}$$

computed as

$$\begin{aligned} P^{+} &= ((P^{-})^{-1} + (H^{\top}RH)^{-1})^{-1} \\ \mu^{+} &= \mu^{-} + P^{+}((P^{-})^{-1}\mu^{-} + (H^{\top}RH)^{-1}\mathbf{z}), \end{aligned} \tag{5.9}$$

where H is an $m \times n$ matrix that intrinsically selects part of the state that is observed through $\mathbf{z}$. The result is a probabilistic estimate $\{\mu^{+}, P^{+}\}$ with an updated mean and covariance function. The posterior is the output of the first fusion configuration of the framework, where $\theta_{\text{B\_Fusion}} = \mu^{+}$.

Gaussian Process Fusion

The second configuration incorporates all of the available input entities, and the relationship between them, into the data fusion. In this configuration, θ_{Fine} is directly included into the GP fusion, rather than being probabilistically fused afterwards. Therefore the data is compiled into a 25-dimension input feature vector, and $\mathbf{g} = \{\theta_{\text{Fanelli}},\ \theta_{\text{FPYE}},\ \theta_{\text{FPYE'}},\ x_{\text{person}},\ y_{\text{person}},\ 20 \times \text{HSS}\}$. Thus, given the trained joint model, and online data $\mathbf{g}$, the GP is able to model this data and produce a single head yaw estimate $\theta_{\text{GP\_Fusion}} = \mu_{*}$.

5.2.3 Online Head Yaw Estimation Framework Evaluation

This section presents qualitative and quantitative results from a number of experiments carried out to evaluate the developed HYE method and framework.

Operation Space Analysis

Firstly, in order to verify the predicted operation spaces of the fine and coarse feature HYE methods, $\theta_{\text{Fine}} = \theta_{\text{Fanelli}}$ and $\theta_{\text{Coarse}} = \{\theta_{\text{FPYE}},\ \theta_{\text{FPYE'}}\}$, their output readings from the readings of the 4 subjects in the training dataset were captured and visualised, as shown in Figure 5.13. In these graphs, the x-axes represent varying head yaws relative to the person's shoulders (β, shown in Figure 5.12). The y-axes represent the person's position x_{person} along the length of the *interaction zone*, as defined in Figure 2.1, and their position in the width of the *interaction zone* is represented by the three graphs. Raw reading value is represented by pixel colour, as per the colour scale below the plots. As readings were only available in 5° β increments, linear interpolation was carried out between the readings. The plot of readings from a perfectly accurate method would follow this colour scale, centred around $\theta = 0°$ (the head to sensor angle, shown in Figure 5.12): a gradual change from light blue to dark purple.

In order to capture the sensitivity of the methods to mutual gaze, their mean reading along the length of the *interaction zone* was calculated for when the head was theoretically oriented directly towards the sensor ($\theta = 0°$). This result is shown on the graphs, along with the line upon which the calculation is based: the lighter line represents the angle β at which the head would need to be rotated relative to the shoulders to remain oriented directly towards the sensor, and around which the colour scale should theoretically be centred. White space indicates where the methods did not operate.

It can be seen from these graphs that the fine feature Fanelli et al. HYE method only operates in the near-*interaction zone*, as predicted in Section 2.3.1 and Section 2.3.2. The coarse feature methods, on the other hand, operate over the wider *interaction zone*.

However, in areas where it is operating reliably ($x \leq \sim 1.5m$, $y \leq \sim 0.5m$), the reading pattern of the Fanelli et al. method is consistent with the theoretical reading pattern, and along the θ=0° line it has mean readings of 1.3° and 3.1° at $\sim 0m$ and $\sim 0.5m$ offsets, respectively. The FPYE method also has reasonable accuracy in these situations (0.8° and 5.8°, respectively). However, its readings deviate from the theoretical reading pattern as the head to sensor angle increases. This indicates, as predicted, that this method is most sensitive to mutual gaze.

To give a general indication of head pose where the fine feature method is not operating, the FPYE' method is useful: while its mutual gaze readings (5.8° and 5.8°, respectively) are not as accurate as the FPYE method, its reading pattern in the wider *interaction zone* is consistent with the theoretical pattern.

Furthermore, Figure 5.14 illustrates the methods' raw errors with the training data; that is, the difference between the mean of the methods' head yaw estimates and the calculated ground truth (as depicted in Figure 5.17). The complementary nature of the methods can be seen: in combination, the methods operate over the entirety of the *interaction zone* with reasonable levels of accuracy.

These results verify the predicted strengths of HYE methods based on finer and coarser facial features: accuracy vs coverage of the *interaction zone*, respectively. As such, the advantages of fusing such methods into a single head yaw estimate can be appreciated: this will exploit their individual strengths, achieving the desired operation space while maintaining HRI-suitable accuracy.

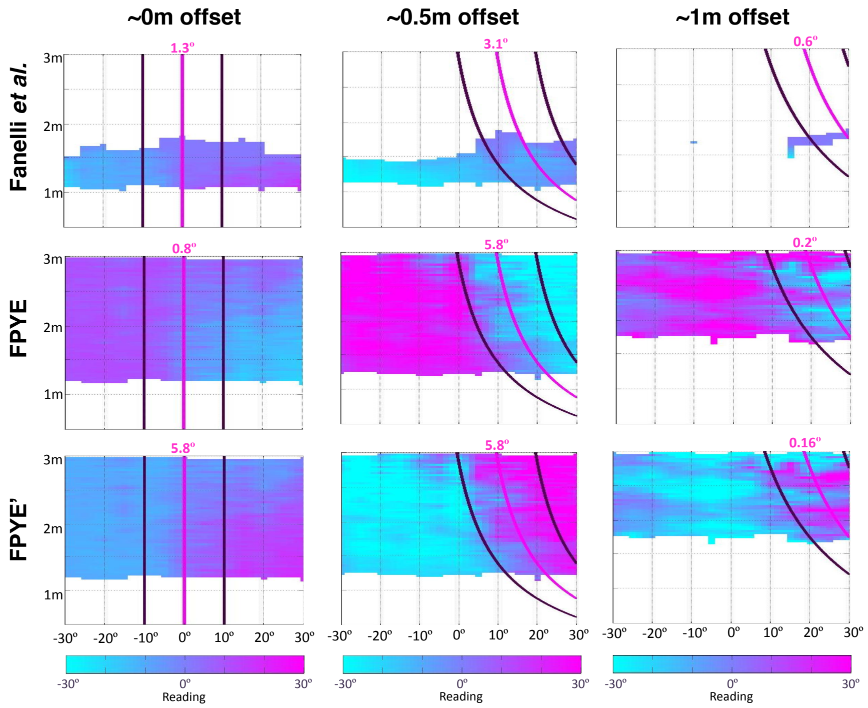


Figure 5.13: Raw readings of the HYE methods from the training data, with head to shoulder angle β on the graph x-axis and x_{person} on the y-axis. The light line represents the required head-to-shoulder angles β to achieve $\theta = 0$ at the particular offsets. White space indicates where the methods did not operate.

Data Fusion Configuration Comparison

Next, an evaluation of the two different data fusion configurations was undertaken to determine which is most suitable to achieve such HYE. A semi-controlled HRI experiment was carried out to achieve this, with constraints minimised as much as possible. Firstly, the core sensing system from our social robotic platform RobotAssist was disembodied in order achieve the necessary experimental setup, however we ensured the height and angle of the experimental sensing system was consistent with the setup had it been on the robot.

There were a total of 26 male participants aged between 22–45 in the experiment, which was carried out in the Centre for Autonomous Systems (CAS) at the University of Technology, Sydney, in two parts:

- Part A on Day 0, to evaluate operation and accuracy of the framework configurations over the *interaction zone*, with 25 participants (1 known from training and 24 novel participants).
- Part B on Day 15, to test the framework configurations' repeatability and robustness to interpersonal variations over time, with 19 repeat participants

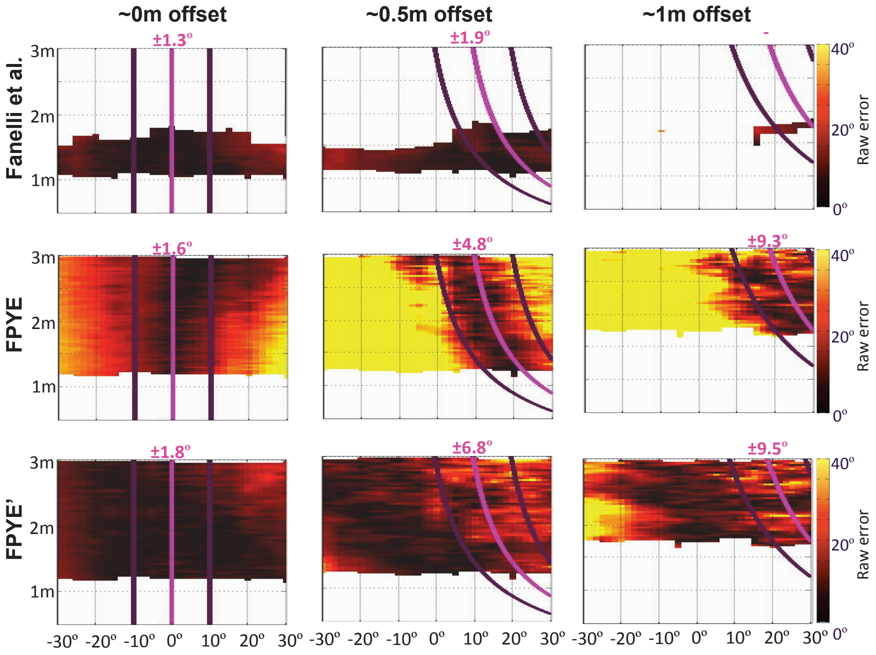


Figure 5.14: Raw error results of the HYE methods from the training data, with head to shoulder angle β on the graph x-axis and x_{person} on the y-axis, where raw error is the difference between the methods' readings and the calculated ground truth head angle relative to the sensor as depicted in Figure 5.17. The light line represents the required head-to-shoulder angles β to achieve $\theta = 0$ at the particular offsets. White space indicates where the methods did not operate.

from Part A (but none known to the GP) and 1 novel participant (also unknown to the GP).

No other selection criterion was employed other than asking the participants if they were willing to participate and were likely to be present at CAS on both experiment days.

The setting of the experiment, which was designed to be distinct from the training dataset acquisition setup in order to evaluate robustness to different data collection methods, is depicted in Figure 5.15a and its setup in Figure 5.15b. This was intentionally designed into the experiment as, in the real-world, it is not reasonable to expect training data from the people such frameworks will be operating with, in the exact scenarios in which the operation will take place.

As can be seen, three Microsoft Kinect sensors, each separated by a distance of $0.5m$, were setup within a $1.7m$ wide walkway between cubicles within CAS. As illustrated in Figure 5.15b, participants were instructed to walk between two pieces of tape on the floor directly in line with Sensor 1, the first ('Start') $\sim 3m$ from the sensors and the second ('Finish') $\sim 1m$ away.

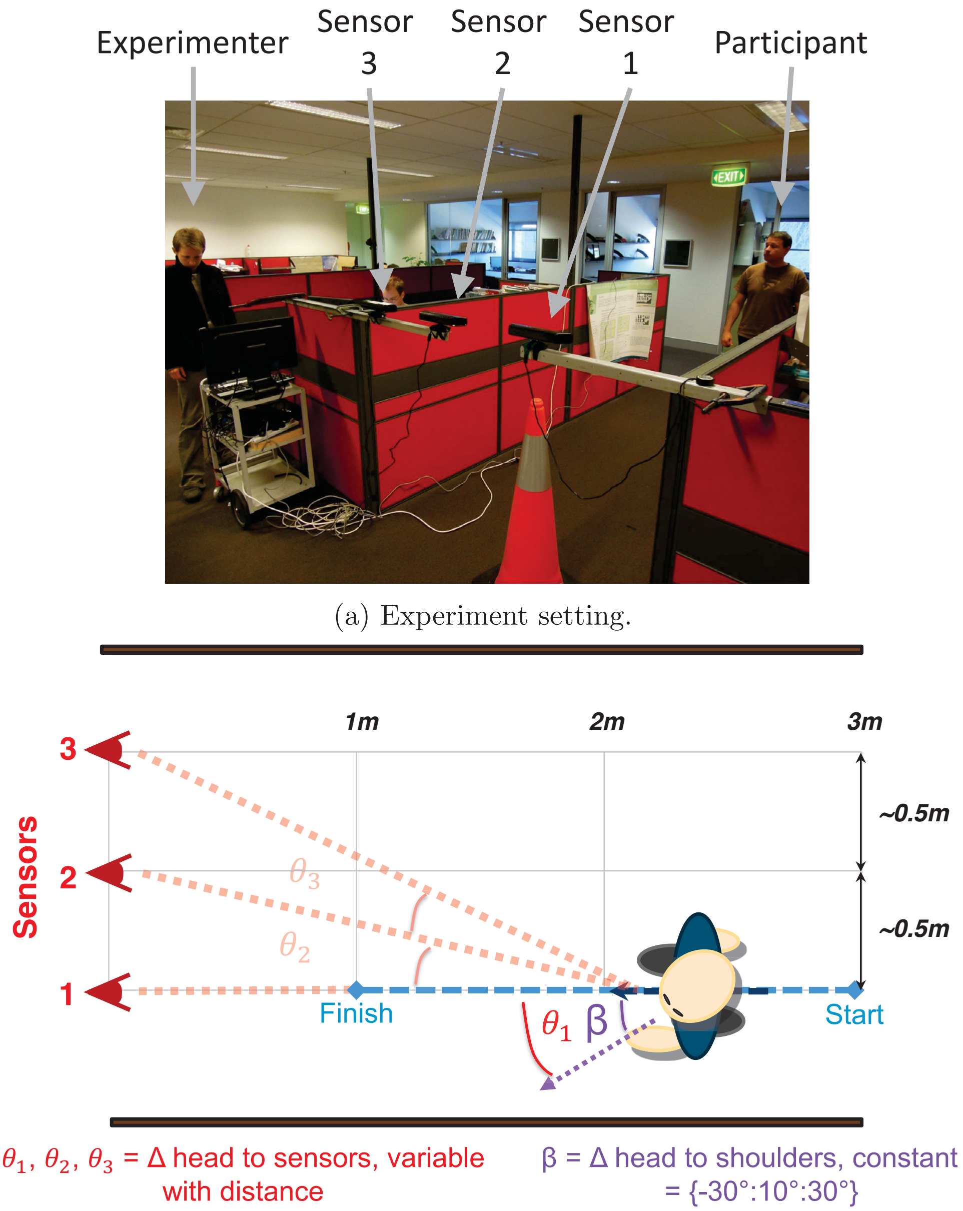
Three sensors were utilised in order to simplify the experimental procedure. Rather than putting the sensor in one position and requiring the participant to walk at $\sim 0m$, $\sim 0.5m$ and $\sim 1m$ offsets in three separate runs (as during training data acquisition), the participant walked in one line and the 'robot' (i.e. sensor) was put at the three offsets simultaneously. However, the data can be used as three independent measurements, rather than requiring the participant to repeat each test three times. Similarly to the training data acquisition method, this procedure ensures data is captured from participants over the entirety of the *interaction zone* as the data from the $y_{\text{person}} \approx 0.5m$ and $y_{\text{person}} \approx 1m$ offsets is likely to be symmetrical and thus also represent $y_{\text{person}} \approx -0.5m$ and $y_{\text{person}} \approx -1m$ offsets.

The participants were told that a gaze estimation system was being evaluated, and to take approximately $\sim$3–5s to traverse the distance. Similarly to the system training procedure detailed in *Training Dataset for GP Model*, at the beginning of each run the participant was directed to orient their head towards their choice of one of 7 signs posted on the wall, which was $\sim 6m$ behind the sensors. The signs corresponded to different head-to-shoulder orientations $\beta = -30° : 10° : 30°$ when the participant was positioned $1m$ directly in front the $\sim 0m$ offset sensor. This range of head orientations towards the sensor was selected due to the primary goal of detecting mutual attention, and β was computed from the head position and sign location as the distance was being traversed. Each person repeated the experiment between 3–5 times, choosing a different wall sign during each run. No further instructions were given. Data was collected from each participant beginning from when they were positioned at the 'Start' location and their head was oriented towards their chosen wall sign, and terminated when they had reached the 'Finish' position.

In the following subsections the data analysis methods and results of the two experiments are discussed.

Part A - Interaction Zone *Operation and Accuracy*

Part A of the experiment was carried out with 25 male participants (1 known from train and 24 novel) over a period of 3 hours, during which $\sim$7,500 data points were captured. One participant was known to the framework, having been used during training; however the experimental environment and procedure were significantly different from that of training data acquisition: the experiment was carried out in a corridor, a more dynamic and realistic environment, and


Experimenter
Sensor 3
Sensor 2
Sensor 1
Participant
1m
2m
3m
3
2
1
Sensors
θ_3
θ_2
θ_1
β
~0.5m
~0.5m
Finish
Start
θ_1, θ_2, θ_3 = Δ head to sensors, variable with distance
β = Δ head to shoulders, constant = {-30°:10°:30°}

(a) Experiment setting.

(b) Experiment setup.

Figure 5.15: Experiment setting and setup.

participants were walking freely rather than seated and rigidly restricted. To get an indication of the ability of the framework to accommodate both known and unknown people while also generally evaluating the framework, the remaining 24 participants were unknown.

The participants' paths were not heavily constrained; as can be seen from Figure 5.22a, which shows the paths traversed over the course the experiment, the participants were not uniformly positioned at $y_{\text{person}} = 0m$, $y_{\text{person}} = 0.5m$ and $y_{\text{person}} = 1m$ offsets from the sensors. Instead their offsets averaged $\bar{y}_{0m \text{ offset}}$

$= 0.04m$, $\bar{y}_{0.5m \text{ offset}} = 0.45m$, and $\bar{y}_{1m \text{ offset}} = 0.93m$ respectively. While these varying participant paths are a source of error, they also have the benefit of increasing the external validity of the experiment.

Results Analysis

Graphs of the methods' accuracy at varying head yaws relative to the person's shoulders and different positions of the person in the *interaction zone* were produced. These graphs were constructed similarly to those in Section 5.2.3, with the exception that pixel colour now represents the methods' errors (the difference between the methods' readings and the calculated ground truth head angle relative to the sensor), rather than raw reading, to clarify their accuracy. Participants' true paths were used to calculate accuracy during graph production.

In order to capture the sensitivity of the methods to mutual gaze, their mean accuracy along the length of the *interaction zone* was calculated for when the head was theoretically oriented directly towards the sensor as in mutual gaze ($\theta = 0$), and towards the vicinity of the sensor (the ranges $-10°<\theta<0°$ and $0°<\theta<10°$), indicating a person may be more open to interaction and/or more likely to witness robot-issued interaction initiation cues. These accuracy results, and the zones on which their calculations are based, are shown on the graphs. The lines represent the required head-to-shoulder angles β to achieve $\theta = -10°$, $\theta = 0$, and $\theta = 10°$ if the participant had walked at exactly the mean offset for each trial duration.

Firstly, to evaluate the characteristics of the fine and coarse feature HYE methods in the more realistic environment, their accuracy with the Part A experiment data was calculated. These results are shown in Figure 5.16, which depicts the methods' raw errors; that is, the difference between the mean of the methods' observations and the ground truth. In order to clarify this, this error is conceptually depicted in Figure 5.17. The figure shows an observation from one of the methods, with its mean and error range. The error is the difference between the mean of the method's observation and the ground truth.

Though Figure 5.16 illustrates that there is significant noise and generally higher inaccuracy than in the training data of Figure 5.14, it can be seen from the patterns that the FPYE and FPYE' methods operate as predicted: FPYE is most sensitive to head orientations towards the sensor (i.e. around the $-10°<\theta<10°$ region), while FPYE' gives a general indication of head yaw when head orientations are to the side of the sensor.

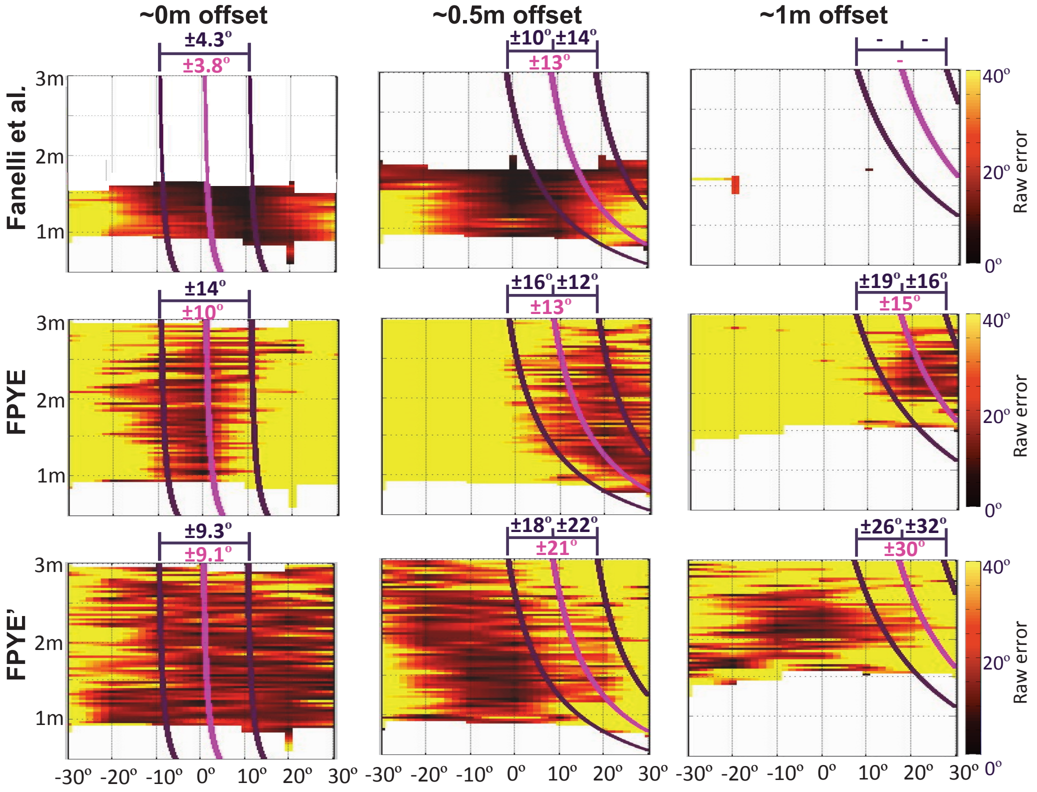


Figure 5.16: Experiment Part A individual HYE method raw error results, where raw error is the difference between the methods' readings and the calculated ground truth head angle relative to the sensor, as depicted in Figure 5.17. Head to shoulder angle β is on the graph x-axis and x_{person} on the y-axis. The lines represent the required head-to-shoulder angles β to achieve $\theta = -10°$, $\theta = 0$, and $\theta = 10°$ if the participant had walked at exactly the mean offset for the duration of each trial. White space indicates where the methods did not operate.

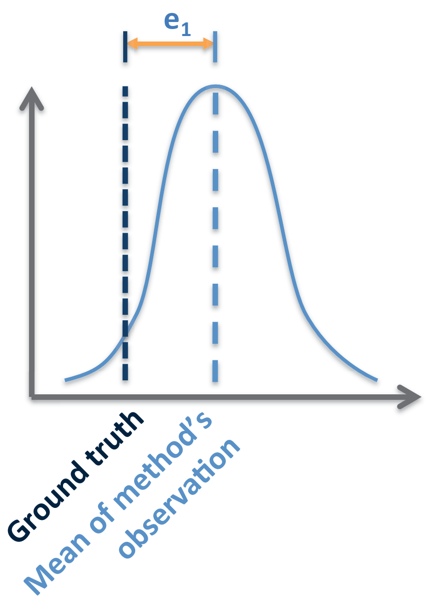


Figure 5.17: Conceptual depiction of the individual HYE methods' raw errors.

The near-*interaction zone* operation space of the Fanelli et al. method shown in Figure 5.16 corresponds to both the operation space prediction (illustrated in Figure 2.12), and the pattern witnessed with the training data (shown in Figure 5.13 and Figure 5.14). Surprisingly, however, the sensitivity of the Fanelli et al. method to mutual gaze in the near-*interaction zone* (especially between $x \approx$ 1–1.5m), which was observed in the training data, is not as evident in the Part A results. The decreased usefulness of the fine feature method in the more dynamic, real-world data has implications for the fusion, as discussed below. Despite this, however, the benefits of fusing these methods can be seen.

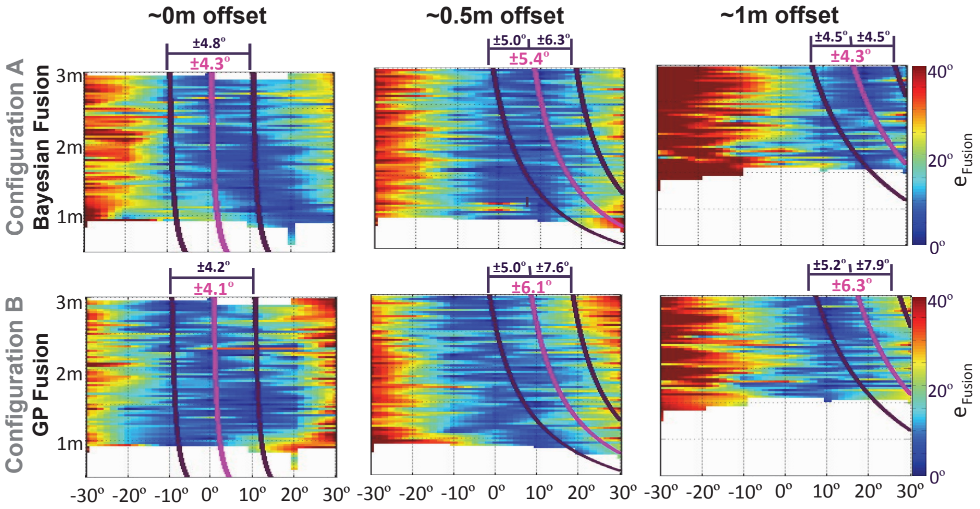


Figure 5.18: Experiment Part A fusion error results, with head to shoulder angle β on the graph x-axis and x_{person} on the y-axis, where raw error is the difference between the methods' readings and the calculated ground truth head angle relative to the sensor, as depicted in Figure 5.19. The lines represent the required head-to-shoulder angles β to achieve $\theta = -10°$, θ = 0, and $\theta = 10°$ if the participant had walked at exactly the mean offset for the duration of each trial. White space indicates where the fusion did not operate.

The results from the Bayesian and GP fusion framework configurations are shown in Figure 5.18. The error in this figure is then the difference between the fusion result and the ground truth. However, it is derived differently than the error of Figure 5.16: the data fusion statistically derives a result from the overlap between the standard deviations of the individual methods' observations. This concept is depicted in Figure 5.19, and it is this e_{Fusion} which is represented in Figure 5.18. From this, it can be seen that the fusion exploits the complementary nature of each of the individual methods' observations. As such, while the observations can have larger errors, the fusion result can have a smaller error than any of the individual methods. This complementary nature is not readably apparent in the raw errors shown in Figure 5.16. As a result, the fusion output will not necessarily resemble the pattern in the individual methods' results.

Data such as error standard deviations have been intentionally omitted from the presented results, as the focus of such data is intended to make comparisons, and would put focus on whether the developed framework and configurations improve on state of the art methods such as Fanelli et al. Instead, the focus of this work is on evaluating whether techniques can be developed to complement and overcome the limitations of such state of the art methods; hence the presented results are intended to evaluate whether the intended outcome of HRI-suitable *interaction zone* HYE has been achieved.

As such, Figure 5.18 demonstrates that fusion is a viable method of achieving HRI-suitable HYE throughout the *interaction zone*. As summarised in Table 5.2, the configurations have generally comparable results, with mean accuracies of ±4.1° and ±3.7°, respectively, when participants are looking directly at the sensor. Similarly, in the $-10°<\theta<10°$ area at $\sim 0m$ offset, their mean accuracies are ±4.5° and ±4.0°, respectively. Additionally, the uncertainty bounds (variance) produced by the framework show that the fused estimates $\theta_{\text{B\_Fusion}}$ and $\theta_{\text{GP\_Fusion}}$ are consistent. This demonstrates the ability of fusion methods to detect mutual gaze at larger, far-*interaction zone* distances.

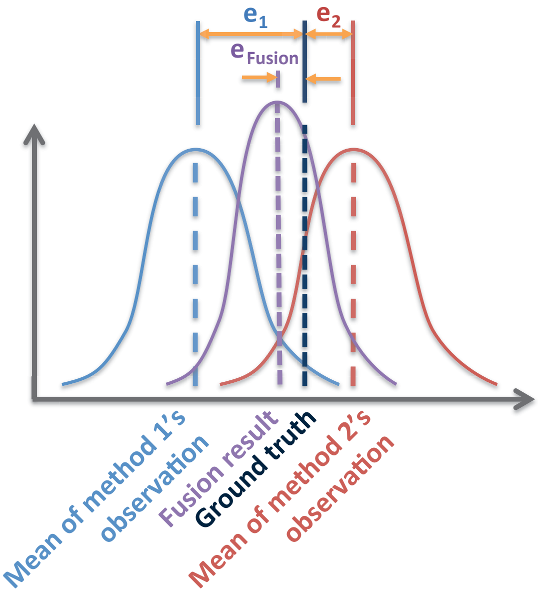

e_1
e_2
e_{Fusion}
Mean of method 1's observation
Fusion result
Ground truth
Mean of method 2's observation

Figure 5.19: Conceptual depiction of the fusion errors.

The configurations also give similar results at the other offsets along the $\theta = 0°$ line and in the $-10°<\theta<0°$ and $0°<\theta<10°$ areas, as shown in Table 5.2. Given these results are gathered from a number of people unknown to the system, who are also in motion, this demonstrates fusion methods' suitability for the unstructured and dynamic HRI space.

Table 5.2: Experiment Part A mean accuracies of the HYE methods along the $\theta = 0°$ line and within the $-10°<\theta<10°$ area at different offsets.

		HYE method				
		Fanelli *et al.*	FPYE	FPYE'	Bayesian fusion	GP fusion
0*m* offset	θ = 0°	±3.8°	±10°	±9.1°	±4.3°	±4.1°
	−10° < θ < 10°	±4.3°	±14°	±9.3°	±4.8°	±4.2°
0.5*m* offset	−10° < θ < 0 °	±10°	±16°	±18°	±5.0°	±5.0°
	θ = 0°	±13°	±13°	±21°	±5.4°	±6.1°
	0° < θ <10°	±14°	±12°	±22°	±6.3°	±7.6°
1*m* offset	−10° < θ < 0 °	–	±19°	±26°	±4.5°	±5.2°
	θ = 0°	–	±15°	±30°	±4.3°	±6.3°
	0° < θ <10°	–	±16°	±32°	±4.5°	±7.9°

However, there is a key difference between the Bayesian and GP fusion configuration results. The Bayesian fusion configuration preferentially 'trusts' the θ_{Fine} reading when it is operating, under the assumption that this reading will have higher accuracy than the θ_{GP} reading. As a result, the unexpected characteristic of the Fanelli et al. method readings in the near-*interaction zone* – specifically its decreased sensitivity to mutual gaze – directly becomes a characteristic of the Bayesian fusion output.

The GP fusion configuration, on the other hand, is able to mitigate this Fanelli et al. method weakness through incorporating input entity relationship into the fusion: though the sensitivity of the θ_{Fanelli} reading to mutual gaze is compromised, the FPYE method, while generally less sensitive, is still able to give a reasonable indication of mutual gaze. This is accounted for by the GP configuration, suggesting that such a configuration is more suitable to achieve the goal of HRI-suitable, *interaction zone* HYE. Thus, $\theta_{\text{HYE-F}} = \theta_{\text{GP\_Fusion}}$.

In both cases, however, it can be seen from the graphs that noise exists in the results data, for which there are a number of possible explanations. Firstly, the variance in the participants' paths could have led to the asymmetry of the readings around the $\theta = 0°$ line. Additionally, despite the inclusion of the HSS to reduce variance across individuals, it is reasonable to expect that the accuracy of the HYE framework will still vary slightly between people. Future work will be to reduce this framework inaccuracy; however, the results illustrate that in its current form the HYE framework achieves HRI-suitable accuracy throughout the *interaction zone* by building on the strengths of the data it fuses, especially in the GP fusion configuration.

Part B - Repeatability and Robustness to Interpersonal Variations

With the *interaction zone* operation and accuracy of the framework configurations validated, Part B of the experiment was carried out on Day 15 in order to evaluate their repeatability and robustness to interpersonal variations over time. In Part B, there were 19 repeat participants from Part A and 1 novel participant (though none of the participants were known to the GP, none having been used in the training data). Approximately 5,000 data points were gathered over a period of 4 hours. Again the participants' paths, shown in Figure 5.22b, were not heavily constrained and were similar to those traversed in Part A (Figure 5.22a), with slightly different mean offsets of $\bar{y}_{0m \text{ offset}} = 0.02m$, $\bar{y}_{0.5m \text{ offset}} = 0.43m$, and $\bar{y}_{1m \text{ offset}} = 0.86m$. Visual and qualitative representation of the results, shown in Figure 5.20 and Figure 5.21, were developed in line with Part A.

Results Analysis

The repeatability of the framework's configurations are comparable despite the time separation, as can be appreciated through visual comparison of the Part B (Day 15) accuracy graphs to those of Part A (Day 0) in Figure 5.16 and Figure 5.18. The numeric results, presented in Table 5.3, show that the differences of the framework configurations' accuracies on Day 0 and Day 15 of the experiment were minimal. The small magnitude of these accuracy variations reinforces the suitability of the framework – especially in a GP fusion configurations – for HRI applications, and suggests that they could simply be a result of the variations in the paths; the exact sources of these errors will be considered in future work.

These results demonstrate the framework's repeatability and robustness to interpersonal variations (for example changes in attire) over time.

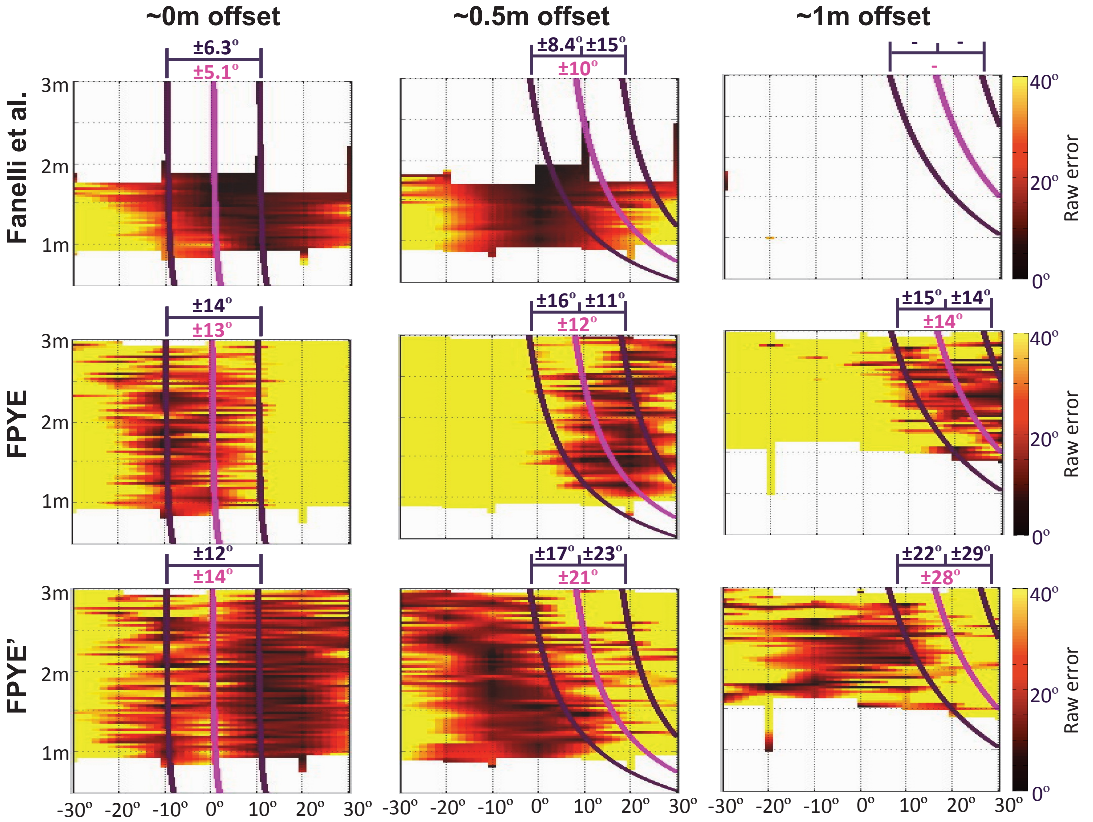


Figure 5.20: Experiment Part B individual HYE method raw error results, where raw error is the difference between the methods' readings and true head angle relative to the sensor, as depicted in Figure 5.17. Head to shoulder angle β is on the graph x-axis and x_{person} on the y-axis. The lines represent the required head-to-shoulder angles β to achieve $\theta = -10°$, $\theta = 0$, and $\theta = 10°$ if the participant had walked at exactly the mean offset for the duration of each trial. White space indicates where the methods did not operate.

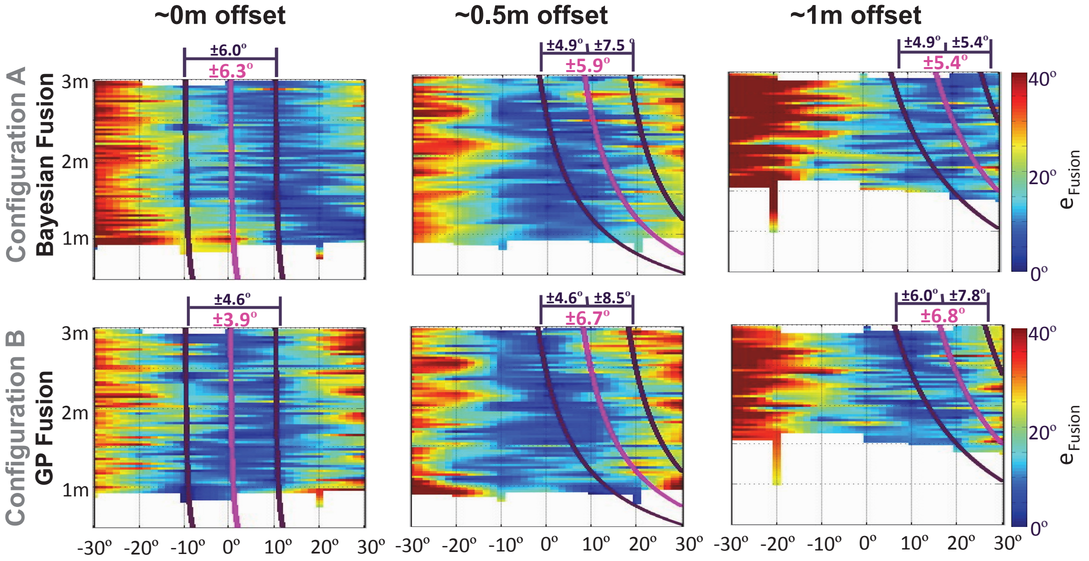
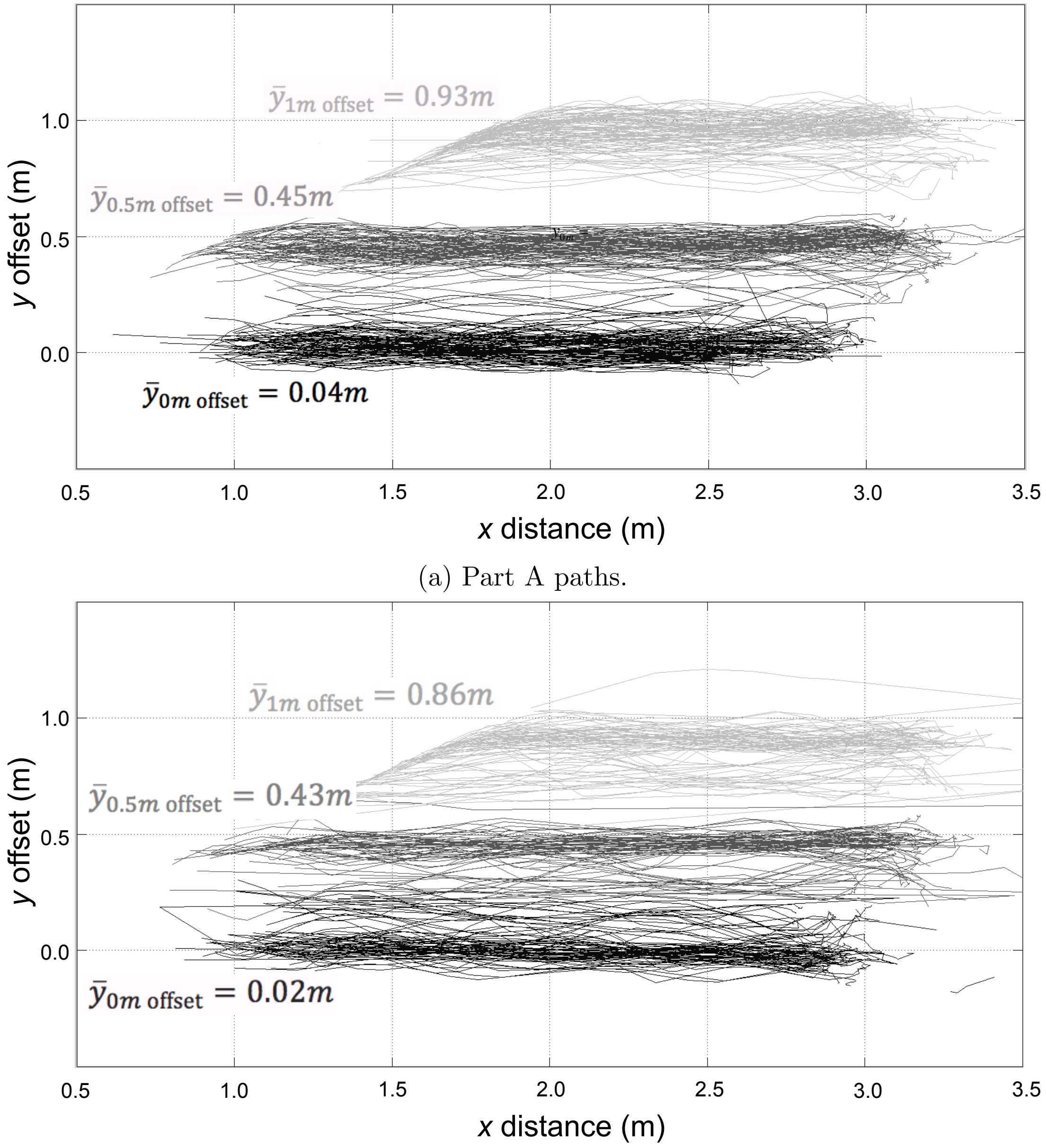


Figure 5.21: Experiment Part B fusion error results, with head to shoulder angle β on the graph x-axis and x_{person} on the y-axis, where raw error is the difference between the methods' readings and the calculated ground truth head angle relative to the sensor, as depicted in Figure 5.19. The lines represent the required head-to-shoulder angles β to achieve $\theta = -10°$, $\theta = 0$, and $\theta = 10°$ if the participant had walked at exactly the mean offset for the duration of each trial. White space indicates where the fusion did not operate.

Table 5.3: Experiment Part B mean accuracies of the HYE methods along the $\theta = 0°$ line and within the $-10°<\theta<10°$ area at different offsets

		HYE method				
		Fanelli *et al.*	FPYE	FPYE'	Bayesian fusion	GP fusion
0*m* offset	θ = 0°	±3.8°	±10°	±9.1°	±4.3°	±4.1°
	−10° < θ < 10°	±4.3°	±14°	±9.3°	±4.8°	±4.2°
0.5*m* offset	−10° < θ < 0 °	±10°	±16°	±18°	±5.0°	±5.0°
	θ = 0°	±13°	±13°	±21°	±5.4°	±6.1°
	0° < θ <10°	±14°	±12°	±22°	±6.3°	±7.6°
1*m* offset	−10° < θ < 0 °	–	±19°	±26°	±4.5°	±5.2°
	θ = 0°	–	±15°	±30°	±4.3°	±6.3°
	0° < θ <10°	–	±16°	±32°	±4.5°	±7.9°



(a) Part A paths.

(b) Part B paths.

Figure 5.22: Participant paths during the two experiment parts. The black, dark grey and light grey lines represent the paths of participants theoretically positioned at $0m$, $0.5m$ and $1m$ offsets, respectively. The true mean participant offsets are shown to the right of the graphs.

5.3 Conclusions

With no generalisable pattern of gaze behaviour towards robots observed during a real-world HRI study (**RQ A.2: *Read* Value**, published in ***[C2]*** of Appendix A), a head yaw estimation framework to detect gaze behaviour *in situ* was presented. The framework leverages the strengths of multiple HYE methods, including the novel Face Plane Yaw Estimation method, to achieve operation over the entirety of a person's *interaction zone* ($\sim$1.2–3m in the x direction, $\pm\sim1m$ in the y direction) while maintaining an HRI-suitable, landmark level of accuracy. Specifically, a joint Gaussian process model is learnt of the Fanelli et al. HYE method (which has impressive accuracy but does not operate over the entire target *interaction zone*) and the Face Plane Yaw Estimation method, inherently fusing them. The Face Plane Yaw Estimation method utilises the planar characteristic of people's faces, a feature which is distinguishable at greater distances than other finer facial features, to estimate head pose; as a result, the method both complements and extends of operation space of the Fanelli et al. method within the *interaction zone.* By incorporating the complementary methods into a single head pose estimate, the framework thus addresses the trade-off between accuracy and physical operation space inherent in many existing HYE methods.

The results of a two-part quantitative and qualitative, internally valid study showed the framework successfully operates over the entirety of the *interaction zone*, and gives a mean accuracy of ±3.7° when participants are positioned between ∼1–3m from the sensor and directly in front of it, and ±6.3° when they are offset by up to ∼1m left or right from the line of sight of the sensor. Additionally, the framework was shown to be repeatable and robust to interpersonal variations in appearance over time through testing with participants novel to the system, who were not included in the training dataset. These results provide support for **RQ A.2: *Read* Feasibility** that social robots are capable of detecting and interpreting human-issued cues such as gaze during real-world HRI, and are contained in ***[J2]*** of Appendix A, which is currently under review.

These results also provide further support for **RQ: Methodology**: the developed methodology was again drawn on to successfully operationalise the *Read* branch of the Robot Centric paradigm during real-world HRI.

However, preliminary data suggests a number of limitations of the framework. Firstly, hair is inconsistently detected in Kinect depth data due to its texture and density. The noise and inconsistent head shape which results reduces the uniformity of head-to-shoulder signature readings, especially of long-haired subjects and subjects looking downwards. This leads to variance in the GP model and thus large errors in the framework head yaw estimate. A similar phenomenon occurs when items such as hats are worn. To address this issue, a more flexible method of correlating facial feature variance with individuals will be adapted into the model.

Additionally, the behaviour of the HYE framework readings when a person is looking in the opposite direction from the robot, i.e. head yaws around 180°, is also unknown, particularly given the Kinect's inconsistency of hair detection. This should be further investigated and its implications considered.

Finally, the findings of this work would benefit from increased external validity through integration and testing of the HYE framework on a robotic platform during HRI, including its robustness to other rotational poses beyond head yaw and the validity of the assumption that landmark gaze detection is typically suitable for social HRI, particularly at far-*interaction zone* distances. Computationally efficient GP implementations could be used to enable this real-time, online performance (as done in [121] with a multi-output GP, for example).

Chapter 6

Interactivity – Demonstrating the Relationship Between Interactivity and Effectiveness

With the *Elicit* and *Read* branches of the Robot Centric HRI paradigm activated in isolation thus far, this chapter presents a study which extends this work by increasing the interactivity level of the exemplar humanoid social robot. Through doing so, the hypothesis from Chapter 3, that a robot's interactivity is correlated with its effectiveness at achieving its goal(s) (**RQ B: Interactivity**), is addressed. While reinforcing and overcoming the shortcomings of previous **RQ A: Sociocontextual cues in HRI** findings – such as lower external validity and sample size – the results of the study presented here within demonstrate that both the *Read* and *Elicit* branches of the Robot Centric HRI paradigm are valuable to social robots: through presenting participants with a dichotomous choice situation and twice asking them to choose between the objects, it is shown that *Elicit* strategy moderation is valuable to a robot in effectively achieving its goal(s), and information valuable to moderate *Elicit* can be gained through *in situ Read*ing.

6.1 Introduction

Up to this point, the *Elicit* and *Read* branches of the Robot-Centric HRI paradigm have been operationalised and investigated in isolation with an exemplar sociocontextual cue and robot, suggesting that the characteristics and effects of sociocontextual cues in HRI do correspond to those of human-human cues (**RQ A: Sociocontextual cues in HRI**). Firstly, Chapter 4 presented a small-scale, internally valid empirical study to explore if humanoid robot-issued JA cues would successfully activate the *Elicit* branch and whether people would respond to such robot-issued JA cues in line with they way they respond to human-issued cues. The results of this experiment suggested that JA activated the branch: the JA cue was both recognisable when issued by the RobotAssist platform (as per **RQ A.1: *Elicit* Feasibility**) and appeared to influence participants' decision-making behaviour (as per **RQ A.1: *Elicit* Response**). This implied that exemplar JA is transferable to HRI.

Next, the *Read* branch was explored, as discussed in Chapter 5. This began with investigation into whether people will actually display HHI-predicted gaze (and hence attentiveness) cues around robots, necessitating *in situ* gaze cue detection for robots (as per **RQ A.2: *Read* Value**). A real-world interaction study was conducted with the results suggesting that, similarly to HHI, there is no generalisable gaze pattern between people. Under the assumption that *in situ* gaze estimation would benefit HRI, existing systems for estimating head yaw, an indicator of gaze, were reviewed. It was found that many current head yaw estimation methods pursue accuracy at the expense of operation space, limiting their suitability for the wider real-world HRI space. Therefore, a HYE framework was developed to address this need for HYE which operates over the entirety of the HRI space, while maintaining levels of accuracy necessary for HRI across the entire space. Evaluation of the framework was then carried out, which demonstrated that it had internal validity and was potentially suitable for *in situ* gaze estimation applications (as per **RQ A.2: *Read* Feasibility**).

However, several shortcomings were identified in this previous *Elicit* and *Read* work. Firstly, in order to generalise understanding of the effects of JA in HRI, it was noted that the work of Chapter 4 would benefit from increased external validity and sample size, as it is not reasonable to generalise the results of such an experiment over the wider population due to the relatively controlled experimental setup and limited number of participants. Additionally, the experiment focused on a narrow range of JA effects, only considering the ability of the cue to influence participants towards the JA object (i.e. participants either ignore the robot's cue or are 'positively influenced' towards choosing the JA object). However, the possibility exists that JA could also have other effects, for example the opposite effect of 'reverse influencing' participants who had initially selected the JA object prior to the cue to change their mind to the other object instead. This possibility is especially relevant in light of the gaze pattern findings of Chapter 5: will robot-issued JA cues which are not directly witnessed have any influence on interaction partners, or is it reasonable to assume that such cues will not have any effect? That is, will one advantage of giving robots the ability to *Read* human gaze *in situ* be to enable them to intentionally issue cues, increasing the effectiveness and predictability of their *Elicit* outcomes?

These questions lead to **RQ B: Interactivity**: will a robot's effectiveness at achieving its goal(s) be increased by adding to its ability to moderate its *Elicit* strategy based on information gained through *Read*ing (i.e. increased interactivity)? That is, are both the *Read* and *Elicit* branches of the Robot Centric HRI paradigm valuable to social robots in effectively achieving their goal(s)? This breaks down into two key questions which correspond to the two paradigm branches: is moderating the *Elicit* strategy valuable to a robot in effectively achieving its goal(s)? Can information valuable to moderate *Elicit* be gained through *in situ Read*ing?

During experimental explorations presented thus far, the robot had no interactivity: the *Elicit* and *Read* branches of the Robot Centric HRI paradigm were activated in isolation. There remained a need, then, to construct and execute an experimental scenario which both: a) extended the prior *Elicit* investigation in order to explore a wider range of JA effects (such as reverse influence) to determine the value of *Elicit* strategy moderation, while also increasing external validity, and b) investigated how people's gaze behaviour impacts on the effects and perceptions of robot-issued JA cues, in order to explore the value of *in situ* gaze *Read*ing. In doing so, the following could be achieved: the findings of **RQ A: Sociocontextual cues in HRI** could be reinforced by overcoming the shortcomings identified in the previous experimental explorations; **RQ B: Interactivity** could be addressed, for the first time holistically operationalising the paradigm; and, through again drawing on the methodology developed in Chapter 3, its ability to enable successful operationalisation of the paradigm could be further verified (as per **RQ: Methodology**).

6.2 Exploration of the Value of the *Elicit* and *Read* Branches

To achieve the target outcomes outlined above, an experimental scenario was constructed and executed by drawing on the methodology presented in Chapter 3. Unlike in previous experiments, in this instance the robot was given a higher level of interactivity via the 'Robot Centric HRI Paradigm Design' stage of the methodology (Section 3.2.3). However, in order to ensure that the value of the individual *Elicit* and *Read* branches could still be isolated, this interactivity level was intentionally limited. Specifically, there was only one cue in each of the *Read* and *Elicit* cue sets: the robot was simply given the ability to *Read* person presence, and issue *Elicit* cues when the person was situated in the *interaction zone* (i.e. binary issuance).

During the study, participants were presented with a dichotomous choice situation and twice asked to choose between the objects. In the Control condition, the robot remained stationary throughout the experiment. However, in the JA condition, the robot issued a JA cue prior to the second request to choose.

The following sub-sections describe the hypotheses, participants, experimental design and procedure, and evaluation measures of the experiment, which is contained in ***[J1]*** of Appendix A.

6.2.1 Hypotheses

Hypotheses and predictions were developed based on the understanding of JA cues outlined in Chapter 2 and the previous work presented in Chapter 4 and Chapter 5.

Firstly, as discussed in Chapter 2, JA cues have the potential to positively influence observers' preferences by affecting their evaluation and affective appraisal of JA objects. However, as discussed in Section 1.1.4, the *Read* branch of the Robot Centric paradigm focuses on influence through sociocontextual cues because of their non-verbal nature, which makes their influence more surreptitious and implicit, and thus less susceptible to negative human response [113]: it has been demonstrated that humans often respond negatively to perceived robot-issued commands or task dictation [154]. This suggests that, at least in the case of HRI, there may be other potential outcomes of JA besides positive influence:

a negative human response to a robot-issued command – which a robot-issued JA cue could foreseeably be perceived as – could be suspicion and reverse influence. While the experiment detailed in Chapter 4 focused on a narrow range of JA effects, only considering positive influence, this possibility is especially relevant in light of the gaze pattern findings of Chapter 5: will robot-issued JA cues which are not directly witnessed have any influence on interaction partners, or is it reasonable to assume that such cues will not have any effect? That is, will one advantage of giving robots the ability to *Read* human gaze in situ be to enable them to intentionally issue cues, increasing the effectiveness and predictability of their *Elicit* outcomes?

From the above, the following hypotheses were developed:

H1 – The robot's presentation of a JA cue will have effects in line with JA cues issued during HHI.

H2 – *In situ* HYE will be advantageous for HRI.

From these, it was predicted:

P1 – When the robot issues a JA cue, participants will be influenced towards preferring the object towards which the cue is issued.

P2.1 – The influence of the robot's JA cue on participants will be impacted by whether the participant was directly looking (L) or not looking (!L) at the robot when the cue was issued, necessitating *in situ* HYE.

P2.2 – The HYE framework will show potential suitability for *in situ* head pose, and hence gaze, estimation in the target application.

6.2.2 Participants

There were 96 participants in the experiment (52 male and 44 female). Participants were sourced from throughout UTS via a number of methods, including flyers posted up around the university, notices in the university's staff newsletter, emails to university contacts, short presentations during lectures, and direct approach by experiment assistants throughout the university campus. In exchange for participation in the short robot interaction experiment, participants were offered a soft drink and small chocolate.

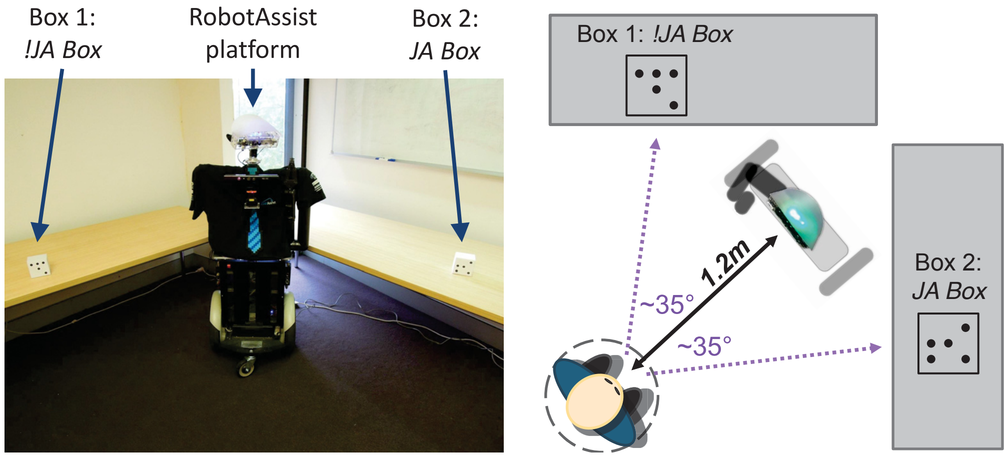
In order to increase the breadth of the participant demographic, and hence the external validity of the experiment, the only pre-requisite to participation was that the person not be from UTS' Faculty of Engineering & IT. This was to exclude those familiar with autonomous systems in general and robotics in particular, giving a more accurate representation of the wider population. Participants indicated their age range on a survey, and ages ranged from the lowest band (less than 18), to the highest (50 or older). No particular demographic was evident, and the participant group was deemed sufficiently representative of the general population by the experimenters for exploration of the research questions.

6.2.3 Setting

The RobotAssist platform (Figure 2.6a) was again utilised during the experiment. To minimise visual distractions for the participants, the experiment was staged in a bare meeting room. The robot was positioned adjacent to two long conference tables in one corner of the room, opposite to the corner from which participants entered. This setting is illustrated in Figure 6.1a, which presents a snapshot taken during the period of the experiment.

A decision-making scenario in which the effects of robot-issued JA cues could be gauged was presented to participants. A small white box ($\sim 80mm^3$) was situated on each of the tables (Box 1 on the table to the robot's right and Box 2 on the robot's left), creating a dichotomous choice situation. In order to prevent pre-existing preference bias on the part of the participants, the boxes were identically sized and 5 identical dots were printed on each of the boxes, similarly to dice. However, to ensure the participants would inspect the boxes (as outlined below in Section 6.2.5), the configuration of the dots was different between the two boxes. Figure 6.1b shows a diagrammatic representation of the setting in which the position of the participants relative to the robot and the boxes is shown.

In the JA condition, described below in Sect. 6.2.4, the robot issued a Joint Attention cue towards Box 2, henceforth referred to as the JA Box (and Box 1 as the !JA Box). The JA Box remained constant throughout all trials: as the focus in this study was on several potential outcomes, detailed below in Sect. 6.2.5, holding the JA Box constant removed the technical complexity of online detection of participants' gaze, which would have been necessary to enable the robot to counterbalance the conditions. As the head yaw estimation framework was still under development and the goal here was to evaluate it, this would have presented a significant challenge. Additionally, the chosen configuration enabled simplification of data normalisation, if required, during analysis.



(a) Experiment setting.

(b) Experiment setup.

Figure 6.1: The experiment setting and setup.

6.2.4 Experimental Conditions

There were two conditions in the experiment:

JA - Prior to the request for the participant to choose a box, the robot performed a JA cue towards the JA Box. The cue was performed as described in Section 2.2.2.

Control - The robot did not perform a JA cue, remaining stationary for the experiment duration.

All other acts were consistent throughout the trials.

6.2.5 Procedure

The experiment was a between-subjects design, and the procedure is illustrated in Figure 6.2. Prior to entering the experiment room, each participant was asked to read brief instructions. These instructions thanked them for participating in the study, and directed them to enter the room and move to the circle marked on the floor, which was approximately 1.2m in front of the robot. This circle was to ensure that: a) the participants would be visible to the person detection system, which was running online using the data from the head Kinect sensor; and b) the participants would be positioned such that it would be possible for them to witness the JA cue, in the JA condition.

The instructions further stated that once the participant was positioned in the circle, the robot would begin a simple interaction and give instructions. Participants were asked to remain in the circle for the duration of the experiment, and told that if something went wrong the robot's head lights would go red and that the participant should inform the experimenter straight away.

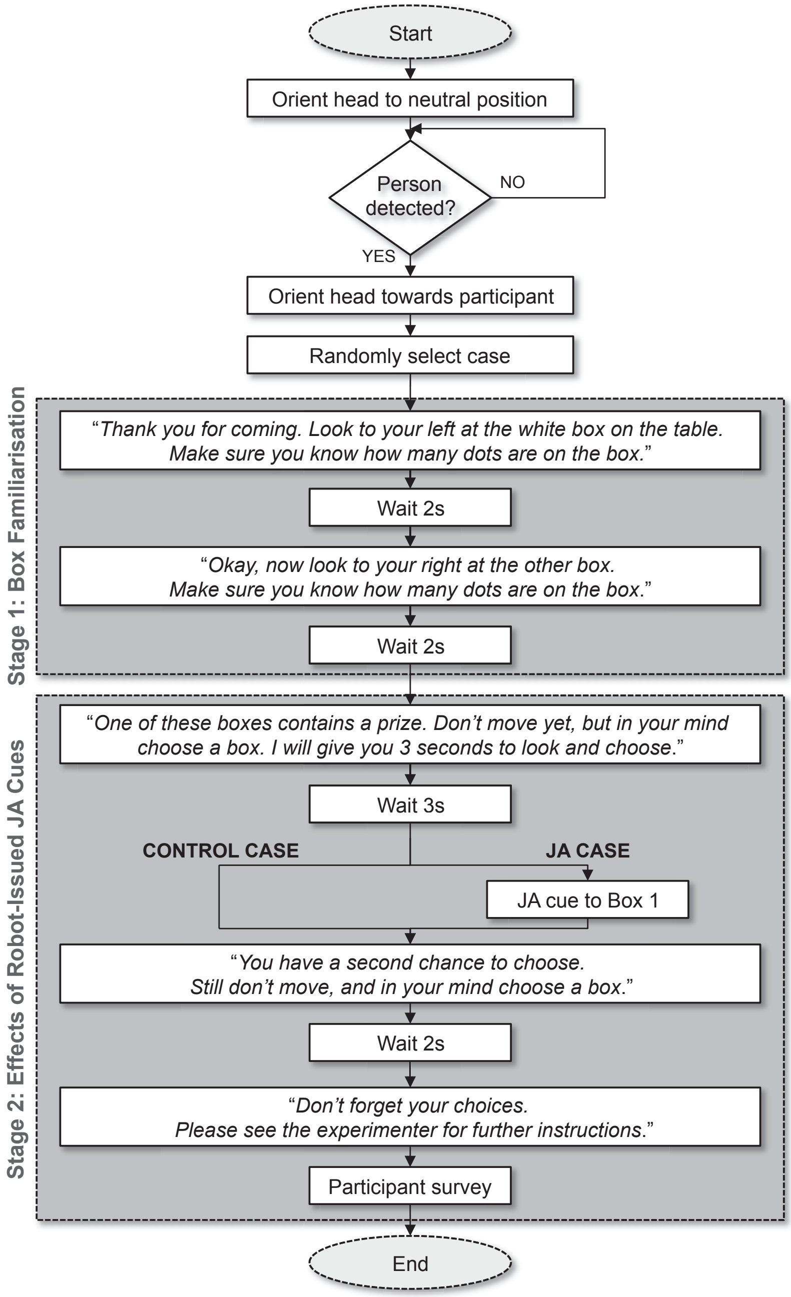

Start
Orient head to neutral position
Person detected?
NO
YES
Orient head towards participant
Randomly select case
Stage 1: Box Familiarisation
"Thank you for coming. Look to your left at the white box on the table. Make sure you know how many dots are on the box."
Wait 2s
"Okay, now look to your right at the other box. Make sure you know how many dots are on the box."
Wait 2s
Stage 2: Effects of Robot-Issued JA Cues
"One of these boxes contains a prize. Don't move yet, but in your mind choose a box. I will give you 3 seconds to look and choose."
Wait 3s
CONTROL CASE
JA CASE
JA cue to Box 1
"You have a second chance to choose. Still don't move, and in your mind choose a box."
Wait 2s
"Don't forget your choices. Please see the experimenter for further instructions."
Participant survey
End

Figure 6.2: Flowchart of the experimental procedure.

When they were ready, the participant then entered the experiment room, while the experimenter remained outside. The robot began with its head in a neutral gaze orientation downwards 10° and to the side 30°, looking away from the participant. Each trial commenced when the participant entered the experiment room and was successfully *Read* by the person detection system. This was designed to attempt to increase participants' impression of robot autonomy and intelligence, and encouraged natural interaction by reducing experimenter influence, ensuring the participants interacted solely with the robot. After a brief delay to allow time for the participant to move to the circle, the robot re-oriented its head towards the participant, looking directly at them. This movement was designed to demonstrate to the participant that the robot was capable of moving its head, increasing their awareness of, and sensitivity to, that channel of communication.

The experimental condition (Control or JA) was then randomly selected, after which the experiment proceeded into the first of two stages.

Stage 1 – Box Familiarisation

Stage 1 was designed to familiarise participants with the boxes, their location and their saliency to the experiment, as a small pre-experiment trial showed participants had difficulty locating the boxes without explicit direction. The stage was identical in both the Control and JA conditions. The robot began by stating, *"Thank you for coming. Look to your left at the white box on the table. Make sure you know how many dots are on the box."* This was followed by a $2s$ wait. At the conclusion of the wait period, the robot then said, *"Okay, now look to your right at the other box. Make sure you know how many dots are on the box"*, and another $2s$ wait began.

Stage 2 – Effects of Robot-Issued Joint Attention Cues

Stage 2 of the experiment then commenced, which was designed to facilitate box preference formation and thereby enable evaluation of the effect, if any, of JA on this preference. As outlined in Section 2.1.4, it is known that susceptibility to choice influence can exist during decision making situations. In such situations, JA cues can potentially influence choice, translating into an increased preference for the JA object. Thus, a choice scenario was constructed for the participants.

First, the robot stated, *"One of these boxes contains a prize. Don't move yet, but in your mind choose a box. I will give you three seconds to look and choose."* Participants were explicitly instructed to remain in the circle as the pre-experiment trial had shown that, contrary to the written instructions, participants had a tendency to move towards the boxes to inspect them more closely.

There were two objectives of making a request for Choice 1. The first was to increase the saliency of the boxes to the participant. Secondly, and more importantly, the robot's statement was designed to facilitate formation of an initial preference. During data analysis post-experiment, Choice 1 served as a baseline against which, for JA condition participants, the influence of the robot's JA cue could be explored.

After the $3s$ wait, in the JA condition the robot issued a JA cue to the JA Box, while in the Control condition the robot remained stationary. In both conditions the robot then stated, *"You have a second chance to choose. Still don't move, and in your mind choose a box"*, followed by another short wait. As it is known that the particular words used during repeated questioning, and repeated questioning itself, can affect preference (an overview can be found in [96]), this potential study variable was controlled by fixing the linguistics. Participants in the Control condition were also asked this question to ensure the effect on preference, if any, of this repeated questioning could be controlled for. However, JA participants' answers could be used to gauge the influence of the JA cue on their Choice 2, as compared to Choice 1. The robot then instructed the participant, *"Don't forget your choices. Please see the experimenter for further instructions."*

The above procedure created a number of choice permutations, enabling deeper exploration of the effects of JA than was possible during the prior study presented in Chapter 4. As illustrated in Figure 6.3, the potential outcomes, based on participants' Choice 1 and 2, are:

Positive influence - Participants who were influenced towards the JA Box, i.e. their Choice 1 was the !JA Box and Choice 2 was the JA Box.

Reverse influence - Participants who were influenced away from the JA Box, i.e. their Choice 1 was the JA Box and Choice 2 was the !JA Box.

Ignore - Participants whose Choice 1 was the !JA Box, and Choice 2 was also the !JA Box.

Reinforce - Participants whose Choice 1 was the JA Box, and Choice 2 was also the JA Box.

It is important to note that in the Control condition, the reference to the boxes as the 'JA Box' and '!JA Box' pertains only to their physical location, being right and left respectively, as no cue was issued in this condition.

In order to gather data about the choices, immediately following the interaction participants were led to an adjacent corridor outside the experiment room, where they completed a written survey which asked them to self-report their Choice 1 and 2 (referred to as 'Box 1' and 'Box 2' on the survey). In order to gauge whether the JA cue had an impact on participants' perceptions, 12 questions measured on a 7-point Likert scale – with 7 evenly spaced points from 'Strongly Disagree' to 'Strongly Agree' – were also asked, including ratings of features such as the intentions and intelligence of the robot, whether it had an agenda and personality, made a positive impression, and whether the participant perceived that it attempted to influence their choice.

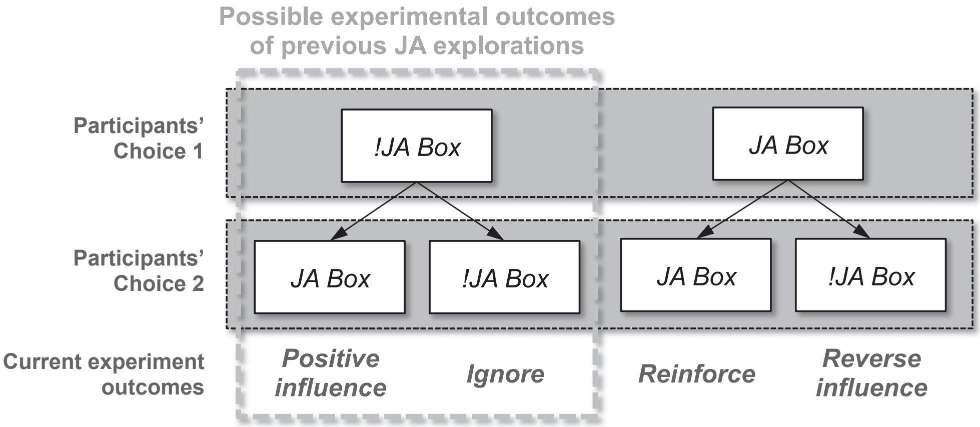


Figure 6.3: Diagrammatic representation of the possible experiment outcomes.

6.2.6 Measurement

Whether or not a JA cue was performed by the robot was the only independent variable. The dependent variables involved three measures, namely:

Participants' choice – This refers to participants' Choice 2 during Stage 2 of the experiment, as compared to their Choice 1. This measure was quantified through participants' self-reporting of their choices on the experiment survey.

Participants' perceptions – This refers to participants' perceptions about the interaction and robot, and was quantified through participants' self-reporting of their perceptions on the written survey post-experiment.

Usability of the HYE framework – This was a measure of the usability of the HYE framework in the application space; specifically, its ability to predict head pose. To explore this measure, the robot was configured to log all of the Kinect depth data, which was generated between 0.5–5Hz. This data was run through the HYE framework offline post-hoc to evaluate its usability.

6.2.7 Results

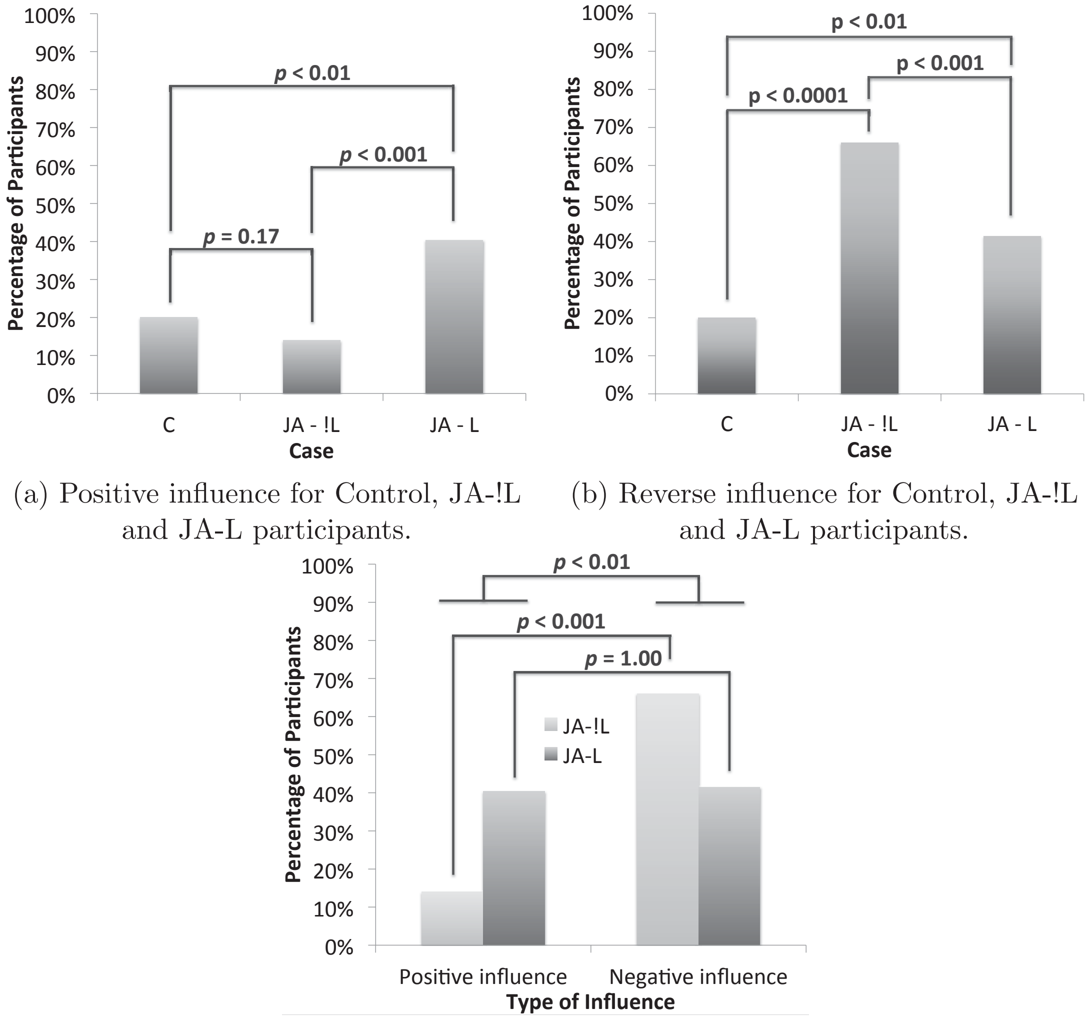
In total, 105 trials were conducted over a span of 2 days. Trials in which participants moved out of the drawn circle or did not completely fill-in their survey, or if there was a technical error with the robot during the trial, were not considered in the results, leaving 96 (50 Control and 46 JA) trial surveys for analysis. A total of ∼80 minutes of framework head pose estimates and RGB and depth camera images of the trials were autonomously collected by the robot during the experiment. On average, ∼1,160 RGB and ∼1,040 depth images were logged per trial. The results of participants' choices, their survey answers and the HYE framework are detailed below.

Influence of the Joint Attention Cue on Choice

Participants' choice data, as reported on their survey, was first normalised to control for systematic effects. This was carried out through examination of the Control condition participants, who were not issued a cue and hence served as a baseline. Theoretically, 50% of the Control participants should have chosen each of the two boxes both times they were instructed to choose. However, it was found that an external bias existed towards the JA (right-hand) Box: ∼60% of Control participants were in Group B, having selected the JA Box as their Choice 1. This is perhaps due to the general preference of right-handed people (who are a majority within the wider population [6]) for the right side [174].

In order to explore prediction ***P2.1***, participants in the JA condition were manually subdivided into those who were Looking or !Looking at the robot when the JA cue was issued. A ROS script was written which automatically output the RGB images captured of each participant in the 1.5s leading up to the time at which the JA cue was issued. An experimenter reviewed these images and identified whether the participants' head was oriented towards the robot at any point during the 1.5s; if so, they were classified as L (and vice versa).

This method of subdividing was chosen based on the, at times, slow $0.5Hz$ frame rate of the Kinect RGB camera, which was not fast enough to reliably capture eye saccades to and/or from the robot, which can be in the order of 20–100ms [11]. However, it is known that the amplitude of head orientation, an indicator of gaze, is in the order of $\sim$15° and $\sim$30° when the target is at $\sim$30° and $\sim$45°, respectively, and that for head re-orientations of $\sim$25°, the head rotates at a mean velocity of $\sim$50°/s [44]. In light of these gaze characteristics and the RGB camera frame rate, examining head pose in a $\sim$1.5s window (which gives an average of $\sim$3–4 image frames) gives a reasonable estimate of whether participants were likely looking or not looking when the cue was issued.



(a) Positive influence for Control, JA-!L and JA-L participants.

(b) Reverse influence for Control, JA-!L and JA-L participants.

(c) Positive and reverse influence for JA-!L and JA-L participants.

Figure 6.4: Results of the influence of the JA cue on choice.

Figure 6.4a shows the normalised results of positive influence for Control, JA-!L and JA-L participants. A series of 2×2 Fisher Exact Probability tests were carried out to analyse the differences of distribution between these groups. This revealed that a significantly higher percentage of JA-L participants were positively influenced than Control participants, $p < 0.01$. Similarly, a significantly higher percentage of JA-L participants were positively influenced than JA-!L participants, $p < 0.001$. Finally, there was no significant difference ($p = 0.17$) in the percentage of Control vs JA-!L participants who were positively influenced.

The normalised results of reverse influence are shown in Figure 6.4b. A number of 2×2 Fisher Exact Probability Tests revealed significant differences between all groupings: a significantly higher percentage of JA-L participants were reverse influenced than Control participants, $p < 0.01$; a significantly higher percentage of JA-!L participants were reverse influenced than Control participants, $p < 0.0001$, and; a significantly higher percentage of JA-!L participants were reverse influenced than JA-L participants, $p < 0.001$.

Figure 6.4c illustrates the normalised results of positive and reverse influence for JA-L and JA-!L participants. The percentage of JA-!L participants who were reverse influenced was found to be significantly higher than those positively influenced through a 2×2 Fisher Exact Probability Test, $p < 0.001$. For JA-L participants, this grouping was not significant ($p = 1.00$). Another 2×2 Fisher Exact Probability Test also revealed a significant difference between positive and reverse influence in JA participants in general (the mean of JA-L and JA-!L participants): a significantly higher percentage of JA participants were reverse influenced than positively influenced, $p < 0.01$.

Survey of Participant Perceptions

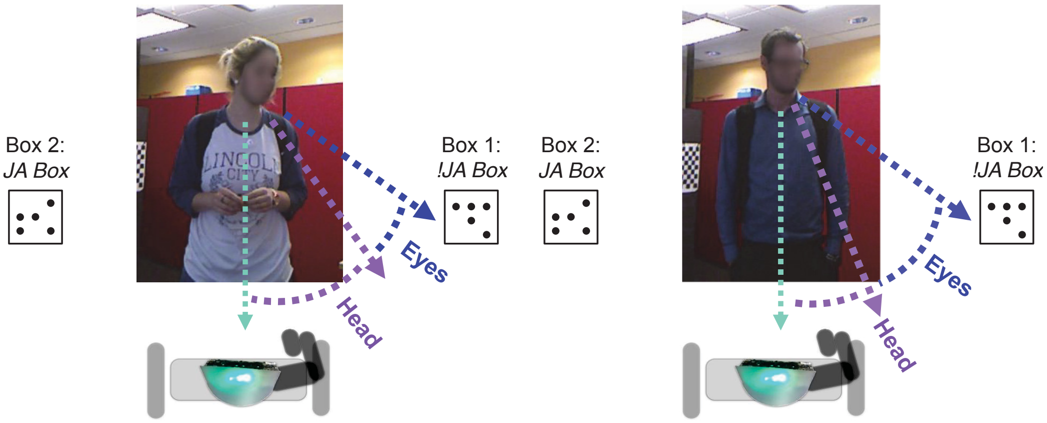
From the 7-point Likert scale survey, a question of particular relevance was whether participants perceived that *"The robot attempted to influence your choice"*. A series of planned independent t-tests were carried out to analyse the results of this question. Both JA participants in general ($\bar{x}$=*5.00*, SD = 2.09), and JA-L ($\bar{x}$=*5.41*, SD = 2.03) and JA-!L participants ($\bar{x}$=*4.42*, SD = 2.11) in particular, felt significantly more that the robot was attempting to influence them, compared to Control participants ($\bar{x}$=*3.24*, SD = 1.66) ($t[64] = 3.39$, $p < 0.01$, $t[52] = 3.57$, $p < 0.001$ and, $t[47] = 1.69$, $p = 0.049$, respectively). There was no significant difference between JA-L and JA-!L participants ($t[27] = 1.28$, $p = 0.11$).

Head Yaw Estimation Framework

The Kinect depth data captured during the experiment was run through the HYE framework post-hoc, outputting estimates of participants' head yaw in each frame. In order to have an indication of the true direction of participants' gaze, data from Stage 1 of the experiment (Box Familiarisation) was examined. In this stage, the robot directed participants to look at each of the boxes, sequentially, for $2s$ each. The boxes were positioned ∼35° left and right from the direct line between the participants and the robot. Under the assumption that participants followed these instructions, their head yaw during these periods can be considered a 'ground truth' against which the head yaw estimates can be compared.

However, during these directed look periods, it was manually observed that a range of patterns of the relative contribution of the head and eyes to the horizontal re-orientation of gaze existed amongst participants. Participants generally fell into two main groups: 1) those who predominantly used their head (∼25–30°) to re-orient their gaze, and used their eyes for the remainder of the distance (as illustrated in Figure 6.5a); and 2) those who predominantly directed their eyes towards the boxes, and used their head for a minority (∼10°–15°) of the re-orientation (shown in Figure 6.5b). These patterns are in line with literature on gaze, which shows that the relative contribution of the head to gaze re-orientation generally increases linearly from ∼60–80% for gaze shift amplitudes greater than ∼25° [44]. Additionally, as the boxes were positioned below both the participants' and the robot's head heights, similar patterns of head and eye contributions existed with the vertical component of participants' gaze, with many participants predominantly moving their head.

The ability of the HYE framework to accurately estimate these head contributions showed unexpected limitations in such a real-world application. It was found that the framework did not give usable results for participants who predominantly used their head to orient their gaze downwards at the boxes; previous internal evaluations of the framework had involved participants always maintaining a level head yaw relative to the sensor. Upon investigation of this characteristic of the framework, the source of the issue was found to be that the downward orientation of people's face planes compromised the reliability of one of the facial features currently used in the estimate. The exact sources of these errors, and a solution, will be considered in future work.



(a) Predominant head movement, with minority eye contribution.

(b) Predominant eye movement, with minority head contribution.

Figure 6.5: Participants' two predominant patterns of contribution of the head and eyes to the horizontal re-orientation of gaze towards the boxes.

However, when participants who maintained level head yaw were considered, the results of the framework were improved. Figure 6.6a shows an exemplar result from a participant who predominantly used their head to horizontally re-orient their gaze towards the boxes. From manual observation of the participant's RGB images, the ground truth of when the participant was looking at the robot and at the boxes was approximated and is shown as dashed lines. The head pose estimates from the framework are shown as solid lines, and the location of the boxes is also indicated. Figure 6.6b shows, similarly, results from an exemplar participant who predominantly used their eyes to re-orient their gaze.

These graphs illustrate the current characteristics of the framework. Firstly, though the framework estimate does not follow participants' head rotation to the full extent when the rotation angle exceeds ∼15°, the framework is able to give a coarse estimate of when participants are not looking at the robot. However, the framework is able to more reliably indicate when a person is looking at the robot.

6.2.8 Discussion

The empirical results presented in this paper provide support for both hypotheses, as well as both strengthening and deepening the findings of the previous *Elicit* and *Read* explorations.

Firstly, support was found for hypothesis ***H1***, that the robot's presentation of a JA cue will have effects in line with JA cues issued during HHI. Specifically, a significantly higher proportion of JA-L participants were positively influenced compared to Control participants ($p < 0.01$). This suggests, as per prediction ***P1***, that the robot's JA cue influenced the choice of JA participants who directly witnessed the cue towards the JA Box. This is in line with the HHI literature-predicted ability of JA cues to increase preference for JA objects during decision-making situations, reinforcing the previous *Elicit* branch findings (Chapter 4).

However, the behaviour of JA-!L participants was notably different to that of JA-L participants: a significantly lower percentage of JA-!L participants were positively influenced than JA-L participants ($p < 0.001$). This supports prediction ***P2.1***, that the influence of a robot's JA cue will be impacted by whether participants were L or !L when the cue was issued, suggesting that the ability of a robot to moderate its *Elicit* strategy will be valuable to achieve its goal(s). Furthermore, the result suggests that *Read*ing interaction partners' gaze *in situ* can provide the information necessary to achieve this moderation: issuing a cue when a person is detected to be looking results in more effective communication, improving compliance with the robot's intentions.

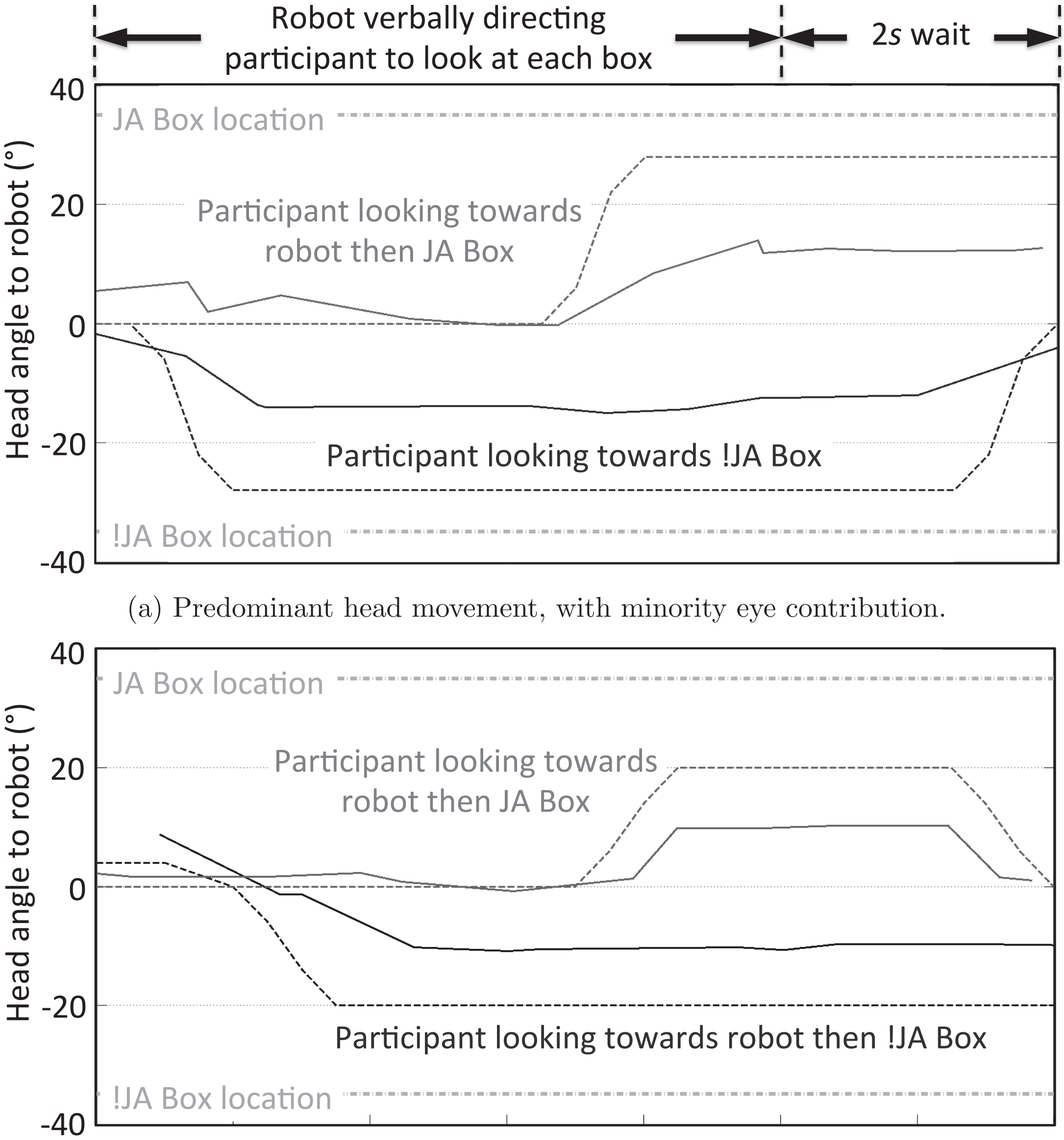
Building upon this, surprising results were uncovered during investigation of other effects of JA in HRI, specifically the outcome of reverse influence. For example, among JA participants in general, the percentage of participants reverse influenced was found to be significantly higher than those positively influenced ($p < 0.01$). This suggests that, while positive influence and compliance are potential outcomes of robot-issued JA cues, there is a greater trend towards perhaps being 'suspicious' of the robot and its intentions. This is reinforced by the finding that reverse influence on both JA-L and JA-!L participants was significantly higher than for Control participants ($p < 0.01$ and $p < 0.0001$, respectively); that is, JA condition participants who had initially selected the JA Box tended to change their mind away from it towards the !JA Box, regardless of whether they directly witnessed the cue or not.

The extent of reverse influence was also found to be impacted by participants' looking behaviour, further verifying ***P2.1***. Participants who did not directly witness the JA cue showed a stronger trend towards suspicion than those who were looking when the cue was issued: a significantly higher percentage of JA-!L participants were reverse influenced than JA-L participants ($p < 0.001$). Furthermore, there was no significant difference between the percentage of JA-L participants who were positively and reverse influenced; that is, the magnitude of positive or reverse influence was similar between participants who were looking at the robot when the cue was issued, suggesting that a certain number of people will simply change their mind when they witness a cue. However, for those who were not looking, the magnitude of reverse influence was significantly greater than the magnitude of positive influence ($p < 0.001$).

The results of the survey of participants' perceptions added interesting findings to the above. As expected, the JA cue introduced a general suspicion of the robot: JA, JA-L and JA-!L participants all felt significantly more than Control participants that the robot attempted to influence them ($t[64] = 3.39$, $p < 0.01$, $t[52] = 3.57$, $p < 0.001$ and, $t[47] = 1.69$, $p = 0.049$, respectively). Surprisingly, however, JA-L and JA-!L participants' self-report of their perception of an attempt to influence was not significantly different ($t[27] = 1.28$, $p = 0.11$). This is in contrast to the above behavioural finding that JA-!L participants showed a stronger trend towards suspicion than JA-L participants ($p < 0.001$); that is, despite their perceptions being similar (feeling similar levels of suspicion), JA-L and JA-!L participants behaved differently: JA-L participants were more likely to act in line with the robot's cue than JA-!L participants.

Though JA-L and JA-!L participants' perceptions were similar, the support found for prediction ***P2.1*** reinforces findings from the previous work (Chapter 5) and suggests that giving robots greater levels of interactivity – in this case, the ability to *Read* human gaze *in situ* and moderate their *Elicit* strategies – will be advantageous to HRI (as per ***H2***). Issuing a JA cue at the 'right' time (i.e. when people were looking) was found to be effective at positively influencing people towards the JA object. A cue issued errantly (when people were not looking), on the other hand, had the opposite effect of influencing people away from the JA object, an effect which was also larger in magnitude than the positive effect. Thus, *in situ* HYE would be advantageous to gauge when people are looking at the robot.

Results of the HYE framework demonstrated that it is potentially suitable for this *in situ* HYE, as per prediction ***P2.2***. While the framework currently has limitations in less-constrained HRI situations, selecting ideal data demonstrated that its strength lies in detecting when people are looking at the robot. Given people's tendency towards suspicion of the robot, *Read*ing this mutual looking behaviour is valuable: by timing the issuance of cues based on this knowledge, robots can work to avoid arousing suspicion in interaction partners.



(a) Predominant head movement, with minority eye contribution.

(b) Predominant eye movement, with minority head contribution.

Figure 6.6: Approximate ground truth and HYE framework estimates when participants looked towards the JA and !JA Boxes with different relative contributions of head and eyes to gaze re-orientation.

6.3 Conclusions

The more semi-constrained, higher external validity social HRI study presented in this Chapter (and contained in ***[J1]*** of Appendix A) focused on furthering the findings and overcoming the limitations of the previous explorations of joint attention in HRI with a exemplar humanoid robot (Chapter 4 and Chapter 5), as well as addressing **RQ B: Interactivity**. The preferences of 96 participants in a dichotomous choice situation, in which a robot issued a JA cue, were compared. The wider effects of the cue (such as the possibility of the cue influencing people to prefer the JA object less), as well as they way people's gaze behaviour impacts on the effects and perceptions of such cues were explored.

It was found that while exemplar JA cues in HRI can have effects in line with HHI JA cues (reinforcing the findings around **RQ A: Sociocontextual cues in HRI**), people have a greater tendency towards suspicion of than compliance with the robot, especially if they are not looking when the JA cue is issued. This is a surprising and somewhat counter-intuitive result: intuition may suggest that if the cue wasn't observed it would not have any influence, when in fact the findings show that if the cue wasn't observed it has a significant effect in the direction opposite to that desired/intended by the robot.

This has implications for the HRI community when designing JA cues in HRI, particularly in cases where the robot is intentionally attempting to influence a person towards a particular object: if it is suspected that the person has already chosen the desired object, the robot goal(s) may be more successfully achieved if it does not issue a cue. If the person has not already chosen the desired object, on the other hand, a JA cue towards that object has the potential to positively influence people, however only if the person directly witnesses the cue. It was found that people are not generally cognisant of this difference in behaviour. Thus, *in situ* gaze estimation would be advantageous to HRI, enabling a robot to intentionally and effectively *Elicit* via JA cues.

These results suggest, as per **RQ B: Interactivity**, that both the *Read* and *Elicit* branches of the Robot Centric HRI paradigm are valuable: moderating the *Elicit* strategy is valuable to a robot in achieving its goal(s), and information valuable to moderate *Elicit* can be gained through *in situ Read*ing. Thus, greater levels of robot interactivity are likely to lead to the robot having greater ability to effectively achieve its goal(s), where effectiveness is considered to be the ability of the robot to target its influence to achieve a specific desired outcome.

Furthermore, through again leveraging the developed methodology discussed in Chapter 3, the its ability to enable successful operationalisation of the paradigm during real-world HRI has been further shown (**RQ: Methodology**).

While the developed HYE framework showed potential suitability for addressing the need for *in situ* gaze estimation, discussed above, the framework currently has limitations in real-world HRI situations, particularly when participants' heads are oriented downwards with respect to the robot. Future work should focus on addressing this issue, as the unconstrained movement of people's heads in real-world interaction scenarios means this pitch must be considered.

Chapter 7

Generalising the Methodology and Reinforcing and Deepening Previous Findings

With the research questions of this work having been explored and addressed utilising an exemplar sociocontextual cue and social robot, this chapter presents a study which demonstrates the generalisability of the work thus far. By again leveraging the developed methodology (Chapter 3), the Robot Centric HRI paradigm was operationalised in an online fashion during a real-world HRI study with a sociocontextual cue and social robot distinct from the exemplar cue and robot. Specifically, the relationship between a lower-HL, disembodied social robot's interactivity and the effectiveness of its influence on people in public spaces was investigated. The two-part study was conducted in both a major Australian public train station and a university where passersby encountered the robot, designed with various levels of interactivity, which attempted to influence their passage. The results of the study demonstrate that the findings of this work generalise to other sociocontextual cues, social robots and application spaces, and that the methodology can be drawn on to successfully operationalise the Robot Centric paradigm during real-world HRI, as per **RQ: Methodology**.

7.1 Introduction

Up until this point, the research questions posed in this thesis have been verified with an exemplar sociocontextual cue (joint attention gaze cues) and social robot (higher-HL humanoid). Firstly, a methodology for operationalisation of the Robot Centric HRI paradigm was developed (Chapter 3). By leveraging the methodology to operationalise and investigate the components of the paradigm during real-world HRI, including the *Elicit* and *Read* branches and the interactivity of the robot, **RQ: Methodology** (regarding the methodology's ability to enable operationalisation of the paradigm) has been inherently addressed.

Further, through employing the methodology, it has been found that joint attention gaze cues issued by a higher human-likeness humanoid robot are transferrable and valuable to HRI in both directions of communication between humans and robots: they are recognisable to people when issued by today's robots (Chapter 2, **RQ A.1: *Elicit* Feasibility**), who generally respond to them in line with human-issued cues (Chapter 4, **RQ A.1: *Elicit* Response**). Enabling robots to *Read* interaction partners' sociocontextual gaze cues *in situ* has been shown to be advantageous to HRI (Chapter 5, **RQ A.2: *Read* Value**). A novel head yaw estimation framework which shows promise for this application was thus devised (Chapter 5, **RQ A.2: *Read* Feasibility**).

Next, it has been demonstrated that a robot's effectiveness at achieving its goal(s) can be increased through greater levels of interactivity: information gained through *Read*ing can be leveraged to moderate a robot's *Elicit* strategy, enabling intentional and effective *Elicit*ing. This is especially relevant given the at times unexpected responses to robot-issued cues (Chapter 6, **RQ B: Interactivity**).

However, the question then becomes: will these findings from the exemplar cue and robot presented in Chapters 4–6 generalise to other cues and social robots in other application spaces? That is, does the methodology generalise?

7.2 Moderating a Lower Human-Likeness Social Robot's Ability to Influence Through its Interactivity

In order to address this question, the methodology presented in Chapter 3 was again leveraged: the paradigm was operationalised in an online fashion during a real-world HRI study with a sociocontextual cue and non-humanoid social robot distinct from the joint attention cue and RobotAssist platform, respectively.

In this exploration, which was published in ***[C1]*** of Appendix A, the *Elicit* and *Read* branches of the HRI paradigm were not individually explored, as they were with the exemplar cue and robot. Instead, investigation began with the relationship between interactivity and robot effectiveness: given findings from literature and the experimental explorations presented thus far, a reasonably solid body of evidence suggests that the *Elicit* and *Read* branches exist and can be leveraged during HRI with social robots.

To demonstrate that the findings thus far generalise, the question becomes: will increasing a different social robot's interactivity through the Robot Centric HRI paradigm result in an increase in the effectiveness of its ability to achieve its goal(s) (as per **RQ B: Interactivity**), similarly to the exemplar instance?

7.2.1 Design of the Robot Centric HRI Paradigm

In order to explore the above question, the Robot Centric HRI paradigm was designed to achieve two different levels of interactivity, as per the methodology described in Chapter 3. These designs were realised in a disembodied robot dissimilar from the RobotAssist platform. This robot was two-part: a sensing and computational component, and an actuation component. The following subsections describe the design of these parts, the interactivity of the robot, and the *Read* and *Elicit* branches used in the study.

Interactivity Design

In both cases the Robot Centric HRI paradigm was designed with successive activation of the *Read*-*Elicit*-*Read* branches. By moderating the cue sets in each of these branches, the interactivity of each of the designs was regulated, in line with Figure 3.3.

In the paradigm design for Part 1 of the study (described below in Section 7.2.2), there were two cues in the cue set for the first *Read*: both person presence in the public proxemic zone (at which point initial robot setup can be carried out, as the interaction has not commenced), and whether said person had entered the *interaction zone* (cue issuance should be triggered). These zones are depicted in Figure 2.1. Three cues were available for random selection for issuance in the *Elicit* cue set: Static, Dynamic, and Responsive (detailed below in Section 7.2.1). The final *Read* had a one-cue set: participants' change in movement.

The paradigm design for Part 2 of the study (Section 7.2.2) built on the Part 1 design with a key change: the addition of a *Read* of the person's entry position into the *interaction zone* (bringing the cue set size to three). Specifically, the cues issued in *Elicit* could be moderated based on this behavioural information to attempt to increase the likelihood of achieving its desired outcome, thus enabling a greater level of interactivity. The final *Read* was again of the participants' change in movement.

In order to realise these interactivity levels, it was then necessary to design the *Read* and *Elicit* branches of the paradigm.

Read Branch Design

The base cue that was *Read* during this study was that of people's presence-location, which was then utilised to *Read* a number of different participant behaviours depending on the paradigm design.

In the study Part 1 design, *Read* was achieved via Wizard-of-Oz. In the Part 2 implementation, a previously developed person detection and tracking system [64, 83] was implemented on 'Boxxie', where it had previously been evaluated by the wider research group. Boxxie, the sensing and computational component of the robot, is shown in Figure 7.2c, and was demonstrated to be capable of robust people detection, tracking, and counting system in public spaces such as train stations [83]. The system is a black perspex box with approximate dimensions of $400 \times 400 \times 100mm$, which was mounted on an $\sim 2m$ high pole. Enclosed within the box are a PrimeSense Carmine 1.08 3D sensor, a fit-PC3 computer, and an OceanServer Power Module that utilises two OceanServer $14.4V$ Lithium-Ion batteries. Boxxie is able to run standalone for ∼7 hours.

Elicit Branch Design

One sociocontextual cue commonly used to influence positional movement in public spaces is that of directional indicators, which are more congruent with lower-HL social robots (as depicted in Figure 7.1). There are a number of characteristics known to moderate the interpretability, and thus effectiveness of such indicators in influencing behaviour, with greater effectiveness being achieved when the indication is strong, unambiguous, and successfully attracts people's attention [129]. Two key characteristics are change (e.g. flashing) and colour [18, 167].

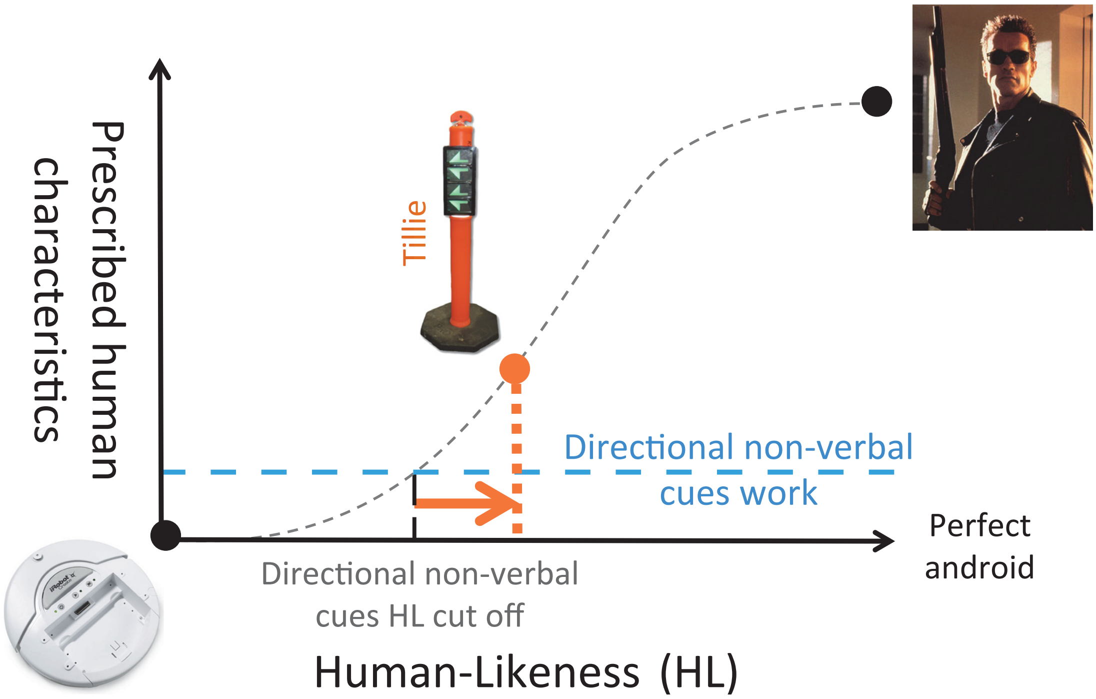

Prescribed human characteristics
Tillie
Directional non-verbal cues work
Perfect android
Directional non-verbal cues HL cut off
Human-Likeness (HL)

Figure 7.1: Directional indicators are likely to be more congruent with the lower-HL of Tillie, making them likely to be interpretable by interacting humans.

Firstly, flashing lights have been shown to be more conspicuous than constant lights [49, 165], as well as significantly increasing compliance with direction [117]. A frequency in the range of 2–5Hz results in greater noticeability [24, 136].

Colour and symbols can similarly affect the conspicuousness and meaning of directional indicators. By drawing on populations' colour stereotypes, colours' established symbolic meanings can be exploited. In Western cultures, for example, green and arrow symbols typically signal 'go', 'good' or safety, or direct movement in a certain direction [24, 167].

In order to issue cues with the above characteristics, a lower-HL influencing device was designed and built (the actuation component of the robot). Figure 7.2 shows the influencing device devised for use in this study. The device consists of an array of perspex screens, each with arrows etched into them. The levels of interactivity designed for this device were:

Static – The device is shown in Static mode in Figure 7.2b. As can be seen, the device shows no signs of activity. However, the arrows etched into perspex screens are visible – the information appears fixed and unchangeable.

Dynamic – Figure 7.2c shows the device in Dynamic mode. Here, internal illumination is used to give the effect that the green arrows are being projected onto the screen – the information appears potentially changeable.

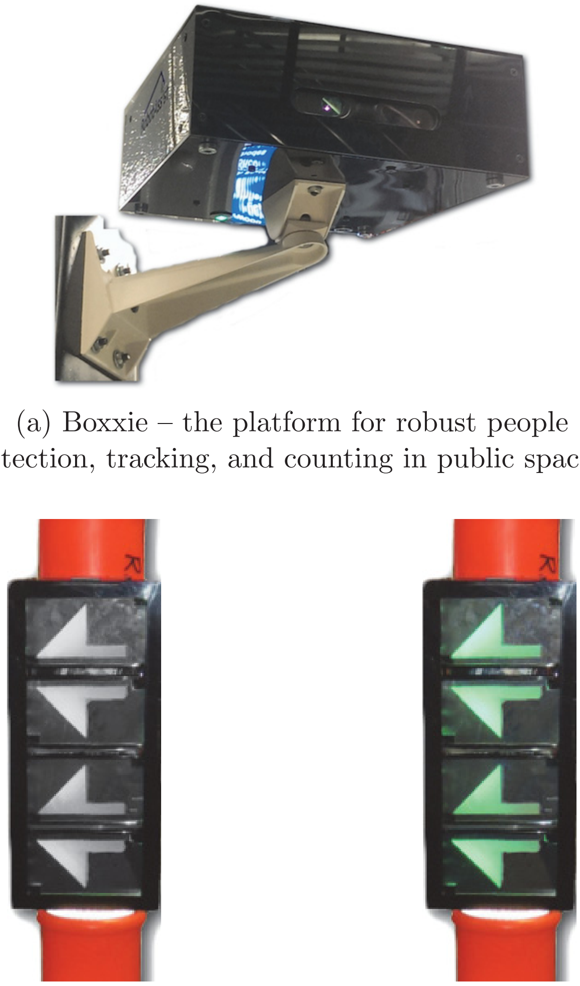
Responsive – A Responsive level of PI system was achieved through leveraging the psychological and behavioural trigger of an event congruent with physical entry into the *interaction zone*, which is known to evoke the perception of entering an interaction (as outlined in Section 2.1.2 and detailed in [57]). Specifically, while a person is in the public proxemic zone, the device remains in Dynamic mode. Then the device issues a cue – flashing several times between the states shown in Figure 7.2b and Figure 7.2c – upon the social trigger of *interaction zone* entry. A flashing frequency of $4Hz$ was selected.

7.2.2 Empirical Explorations

In order to explore the core question of whether the findings and methodology would generalise, and from the relationship between interactivity and influence effectiveness explored in Chapter 6, the following was hypothesised:

H1 – It is feasible to influence people's movement behaviour in public spaces using the Robot Centric HRI paradigm.

H2 – Increasing a robot's interactivity via the Robot Centric paradigm will result in an increase in the effectiveness of its ability to influence; that is, its ability to target its influence to achieve a specific desired outcome.



(a) Boxxie – the platform for robust people detection, tracking, and counting in public spaces.

(b) Influencing device – Static.

(c) Influencing device – Dynamic/Responsive.

Figure 7.2: The disembodied, two-part social robot utilised to generalise the findings of this work.

From these it was predicted:

P1 – Passenger information systems utilising the Robot Centric HRI paradigm (a level of interactivity) will have greater influence on participants than those utilising the traditional HRI paradigm (no interactivity).

P2 – *Read*ing an additional behavioural cue (i.e. increasing the cue set size) will yield insights valuable to moderate *Elicit* to increase the effectiveness of the robot's influence.

These were explored through a two-part study (total $n = 273$) carried out in both a public train station ($n = 84 + 105$) and a university ($n = 84$). Details of the study are presented below.

Part 1 – Influence in a Public Space

In order to evaluate the effect of the previously described influence, a field study with commuters at a major public train station was first conducted. As commuters moved within the train station, one of three levels of information systems – Static, Dynamic, and Responsive PI systems – attempted to influence their behaviour, and the subsequent effect was measured. The focus of this part of the study was on addressing ***H1***, however ***H2*** was also preliminarily explored. The following sub-sections describe the participants, experimental design and procedure, and evaluation measures.

Participants

There were 189 participants randomly selected and directly measured from a larger total number of passersby – 84 in Location 1 and a further 105 in Location 2. They were typical rail commuters. There was no remuneration for participation nor effort to recruit participants.

Setting

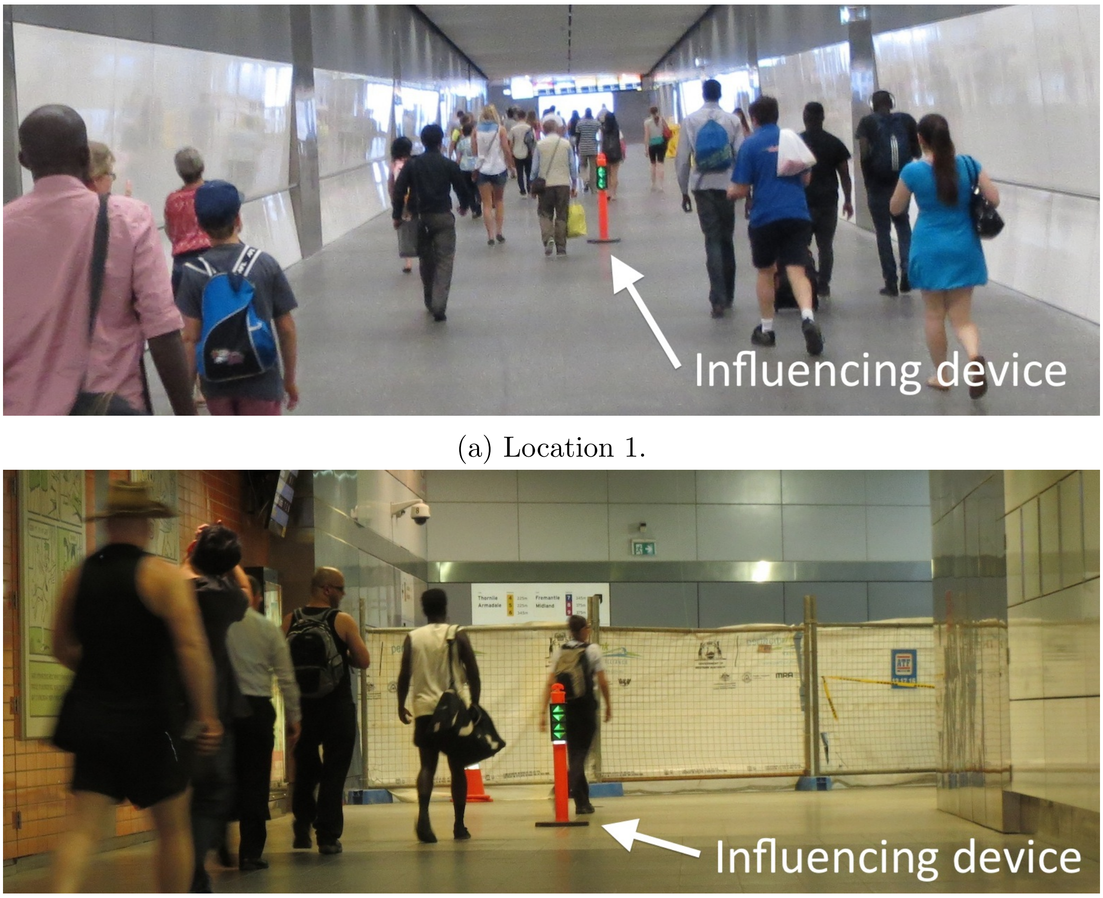
The influencing device shown in Figure 7.2a and Figure 7.2b was utilised, as depicted in Figure 7.3. The experiment was staged in Perth Central Station – a major public train station in Perth, Australia. Studies were conducted at two locations within the station: Location 1 was a long public thoroughfare corridor (shown in Figure 7.3a), and Location 2 was a blind corner subject to passenger flow cross over (shown in Figure 7.3b). As can be seen from the figure, in both situations the influencing device was placed at roughly the thoroughfare midpoint.

Experimental Conditions

The study was designed with three levels of information systems – Static, Dynamic, and Responsive – which were implemented as described in Section 7.2.1. All other acts/cues were consistent throughout the trials.

Procedure

A Wizard-of-Oz study was constructed. The influencing device was cycled through the three levels of information system, with four independent trials con-



(a) Location 1.

(b) Location 2.

Figure 7.3: Setting for the Part 1 study.

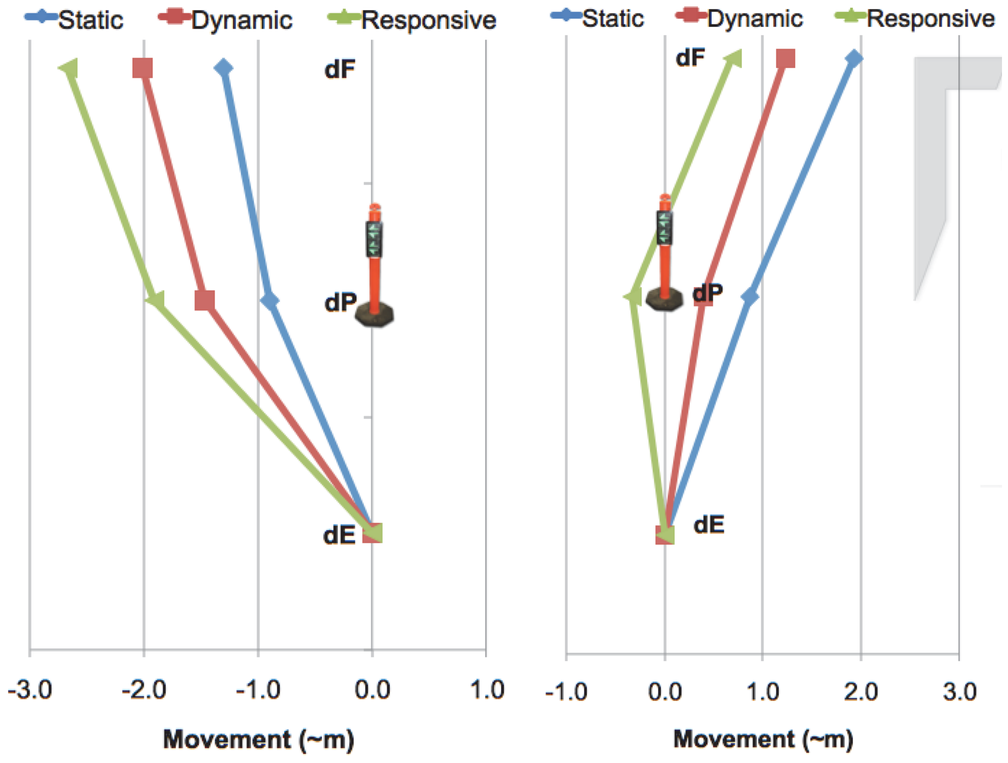
ducted at Location 1 (total of 84 trials) and five each at Location 2 (total of 105 trials) at each level. Each trial commenced with the influencing device being reset, and a commuter passerby being randomly selected by the experimenters. In the condition of Responsive the experimenters tracked the passerby and triggered the influencing device's cue as the passerby crossed into the *interaction zone*.

Measurement

Participants' change in distance from their originally measured position, and relative to a zero-axis which was parallel to the passage and ran through the influencing device was used as the measure, as shown in Figure 7.4. Three repeated measures were taken for each participant: the first, at the **E**ntry point of the *interaction zone* relative to the influencing device; the second, at the **P**ass point of the influencing device; and the third, at the **F**inal measure point which was the exit point of the influencing device's *interaction zone*.

Results

A total of 84 trials (28 trials for each of Static, Dynamic, and Responsive) were conducted at Location 1 and a total of 105 trials (35 trials for each of Static, Dynamic, and Responsive) were conducted at Location 2; three repeated measures were taken in each trial. A relatively steady stream of commuters flowed past during the trials, and approximately five commuters passed by per one selected to facilitate a trial. The experimenters did not attempt to control the number of participants or observers for the trials, and participants were randomly selected.



(a) Location 1.

(b) Location 2.

Figure 7.4: Influencing people towards the left.

Figure 7.4 shows the average of the three repeated measures for the Static, Dynamic, and Responsive conditions at Location 1 and 2. A mixed design ANOVA was performed for each location. The within subject main effect for the three measure points was significant in Location 1 and 2, $F = 67.64$, $p < 0.001$ and $F = 99.43$, $p < 0.001$ respectively. The between subject main effect for the three levels was also found to be significant in Location 1 and 2, $F = 259.44$, $p < 0.001$ and $F = 49.60$, $p < 0.001$ respectively. Pairwise comparisons were conducted between the three levels. Significant differences were found between Static and Dynamic (Location 1 – *mean difference* $= 0.63m$, $p = 0.018$, Location 2 – *mean* *difference* $= 0.54m$, $p = 0.05$), Static and Responsive (Location 1 – *mean difference* $= 1.18m$, $p < 0.001$, Location 2 – *mean difference* $= 1.175m$, $p < 0.001$), and Dynamic and Responsive (Location 1 – *mean difference* $= 0.55m$, $p = 0.039$, Location 2 – *mean difference* $= 0.64m$, $p = 0.017$). Pairwise comparisons also revealed significant differences between the **P** and **F** measure points (Location 1 – *mean difference* $= 0.56m$, $p < 0.001$, Location 2 – *mean difference* $= 0.94m$, $p < 0.001$), relative to measure point **E**.

Part 2 – Robot Interactivity and Influence Effectiveness

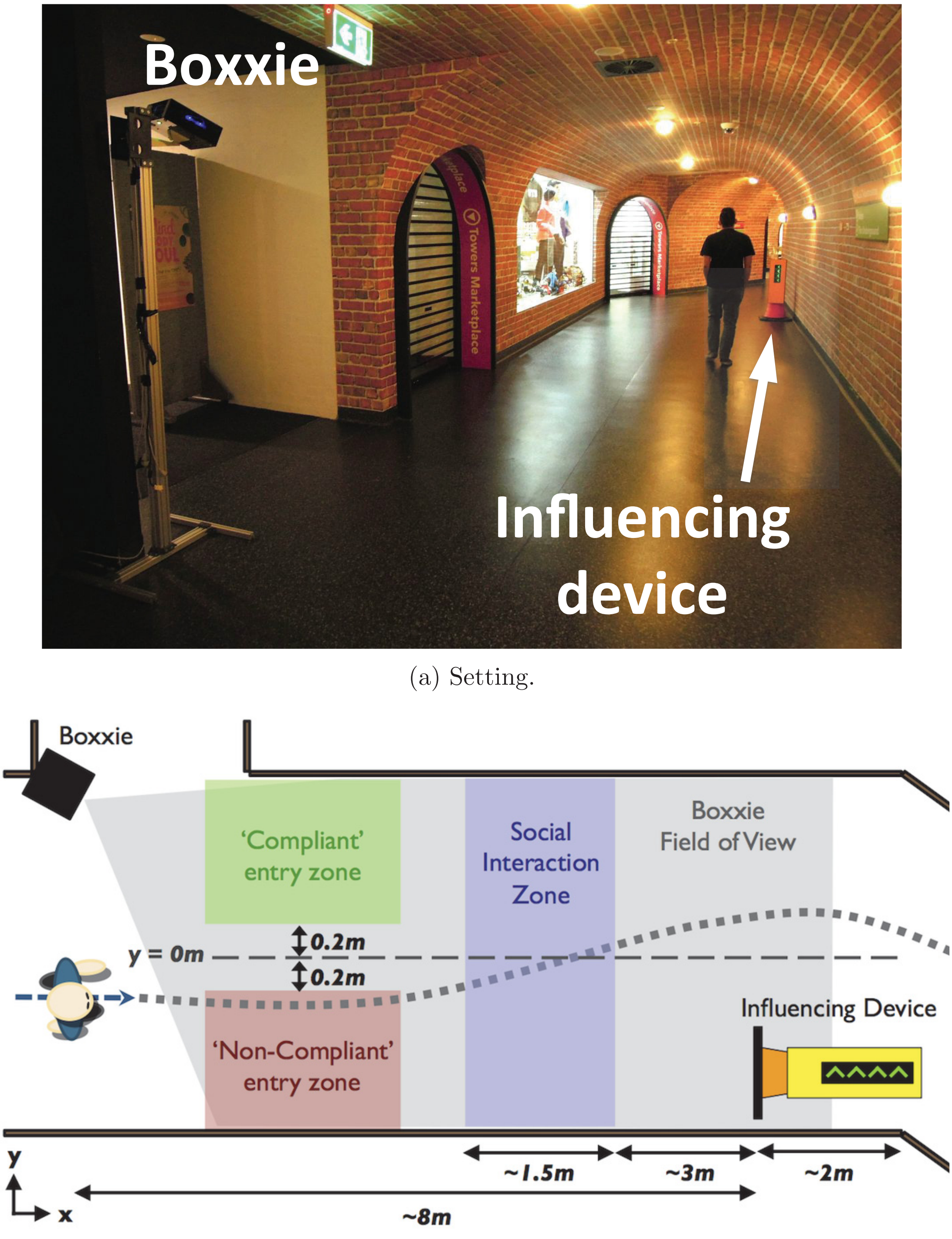
Part 2 of the study focused on more deeply exploring ***H2***. A field study was conducted with passersby in a university food court. As the passersby approached the influencing device, the information system presented as either Static or Responsive, depending on the passerby's initial behaviour, and attempted to influence. The subsequent effect was measured. Part 1 findings were also reproduced, in order to verify that the result was still valid in the different setting. The following sub-sections describe the participants, experimental design and procedure, and evaluation measures.

Participants

There were 84 unsolicited participants in the experiment. Participants were randomly selected passersby to the experiment location who were traveling towards the influencing device, and no particular demographic was evident. There was no remuneration for participation nor effort to recruit participants.

Setting

The experiment was staged in a long straight corridor with a blind corner in the university food court. This setting is illustrated in Figure 7.5a, which presents a snapshot taken during the period of the experiment. The influencing device was positioned $\sim 2m$ in front of the corner and against the right hand wall, from the point of view of the participants' approach direction. Boxxie was located $\sim 8m$ from the influencing device on the opposite wall of the corridor, with its field of view (FOV) directed out towards the influencing device. Figure 7.5b shows a diagrammatic representation of the setting in which the positions of Boxxie and the influencing device are shown, along with Boxxie's FOV.



(a) Setting.

(b) Setup.

Figure 7.5: The Part 2 study setting and setup.

Unbeknownst to participants, there were two entry zones into the experiment, which are also depicted in Figure 7.5b. Participants who entered on the left hand side of the corridor were termed to be initially 'Compliant' (C) with the desired influence behaviour. Participants on the right side of the corridor, on the other hand, were termed ‘Non-Compliant’ (NC). Participants who were moving down the centre of the corridor between these two zones ($-0.2m < y_{\text{person}} < 0.2m$) were considered neither C nor NC and were excluded from the experiment.

Experimental Conditions

There were two conditions for the robot-issued cue – Responsive and Static. These conditions were randomly counterbalanced with the C and NC participants: in some trials the Static information system was presented to C participants and the Responsive cue was presented to NC participants, whilst in other trials this was reversed. All other acts were consistent throughout the trials.

Procedure

Each trial commenced with the random selection of a condition, and began when a participant walking down the corridor was detected by Boxxie as having entered the public proxemic zone and was *Read* as either C or NC, depending on which entry zone they were located in, as shown in Figure 7.5b. Depending on the condition, the influencing device was set to either Static or Responsive. The participant’s position was subsequently tracked via Boxxie, and, in the Responsive condition, the influencing device’s cue was triggered as they crossed into the *interaction zone.*

Measurement

As in Part 1, the participants’ change in distance from their originally measured position, and relative to a zero-axis which was parallel to and in the centre of the corridor, was again used as the measure. As the participants would have had to move in the negative direction to cut the corner, and the positive direction was in line with the attempted influence direction, a less negative change in distance equated to greater influence.

Two measures were taken for each participant: the first at the **E**ntry point of the *interaction zone* relative to the influencing device, and the second at the **F**inal detection point – at which they passed out of the range of the person detection system – which was approximately $1m$ past the influencing device.

Results

In total, 100 trials were conducted. Trials in which the participant was lost by the person detection system before reaching the influencing device were not considered in the results, leaving 84 trials for analysis. There were 56 C and 28 NC participants. A total of ∼2,700 person location readings were autonomously logged during the experiment, with an average of ∼32 person location readings logged per trial.

Figure 7.6 shows the average of the measure for C and NC participants in the Static and Responsive conditions. A two way ANOVA revealed a significant main effect between C and NC participants, $F = 1,614.91$, $p < 0.05$, *mean difference* $= 0.21m$, and a borderline significant main effect between Static and Responsive conditions, $F = 121.02$, $p = 0.058$, *mean difference* $= 0.058m$. The interaction effect was not significant.

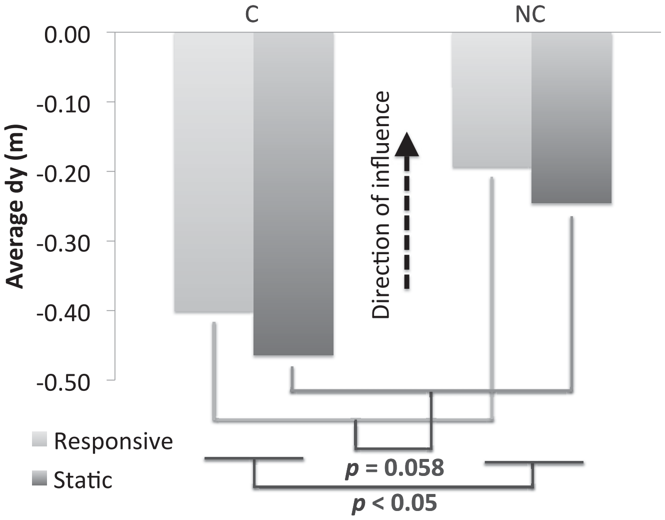


Figure 7.6: Influence reducing the extent of people cutting the corner in Part 2 of the study.

7.2.3 Discussion

The empirical results provide support for both hypotheses. Firstly, support was found for hypothesis ***H1***, that people's movement behaviour in public spaces can be influenced using the Robot Centric HRI paradigm. Specifically, participants in Part 1 of the study moved significantly in the direction of intended influence as they travelled towards and past the influencing device in both Location 1 ($F = 67.64$, $p < 0.001$) and Location 2 ($F = 99.43$, $p < 0.001$). There was also significant movement between the **P** and **F** measure points (Location 1

– *mean difference* $= 0.56m$, $p < 0.001$, Location 2 – *mean difference* $= 0.94m$, $p < 0.001$) relative to measure point **E**, suggesting that there was an ongoing influence effect.

Furthermore, the influence effectiveness was significantly different between the three levels in Part 1 of the study (Static, Dynamic, and Responsive) in both Location 1 ($F = 259.44$, $p < 0.001$) and Location 2 ($F = 49.60$, $p < 0.001$), with Dynamic significantly more effective than Static (Location 1 – *mean difference* $= 0.63m$, $p = 0.018$, Location 2 – *mean difference* $= 0.54m$, $p = 0.05$), and Responsive significantly more effective than Dynamic (Location 1 – *mean difference* $= 0.55m$, $p = 0.039$, Location 2 – *mean difference* $= 0.64m$, $p = 0.017$). This demonstrates, as per prediction ***P1***, that passenger information systems utilising the Robot Centric HRI paradigm (Responsive) will have greater influence on participants than those utilising the traditional HRI paradigm (Static and Dynamic). Furthermore, these results suggest that the influencing device's use of the Robot Centric HRI paradigm to enable a Responsive (i.e. more interactive) PI system saw it most able to influence participants into complying with its suggestions. This provides partial support for hypothesis ***H2***, that increasing levels of robot interactivity (from Static to Dynamic to Responsive) will result in an increase in the effectiveness of its ability to influence.

Prediction ***P1*** was further supported by the results from Part 2 of the experiment, which reproduced the results from Part 1 in order to verify that the findings were in line. Specifically, Responsive was found to result in borderline significantly greater influence compared to Static ($F = 121.02$, $p = 0.058$, *mean difference* $= 0.058m$). The borderline result is potentially due to the exclusion of participants who were neither C nor NC (i.e. in the centre of the corridor).

The results from Part 2 provide support for prediction ***P2*** that *Read*ing an additional behavioural cue will yield insights useable to moderate *Elicit* to increase the effectiveness of the robot's influence (in Part 2 of the study, an additional *Read* of the participant's entry position into the *interaction zone*). Specifically, a significant difference was found between the influence on C and NC participants ($F = 1,614.91$, $p < 0.05$), with NC participants influenced an average of $0.21m$ more than C participants. This result further supports hypothesis ***H2***, and has implications for the design of *Elicit* influence strategies. For instance, consider the case where 'too much' influence may have a negative repercussion. The robot, in that case, may refrain from presenting *Elicit* cues to 'more influenceable' people observed to be already near this threshold.

7.3 Conclusion

This chapter focused on exploring whether the findings from the exemplar cue and robot presented thus far would generalise to other cues and social robots in other application spaces; that is, whether the methodology for Robot Centric HRI paradigm operationalisation would generalise. In doing so, the previous findings regarding the relationship between a robot's interactivity and effectiveness at achieving its goal(s) were reinforced and deepened. To achieve this, a study was carried out utilising a sociocontextual cue and non-humanoid social robot distinct from the exemplar joint attention cue and RobotAssist platform, respectively.

The study, which was published in ***[C1]*** of Appendix A and which leveraged the methodology presented in Chapter 3, focused on quantitively investigating whether increasing the distinct robot's interactivity would result in a similar increase in the effectiveness of its influence (i.e. its ability to target its influence to achieve a specific desired outcome). A two-part study (total $n = 273$) was conducted in both a major Australian public train station ($n = 84 + 105$) and a university ($n = 84$). Passersby were exposed to a robot designed to influence their passage, which had various levels of interactivity.

It was found that – similarly to the exemplar instance – an increase of the distinct robot's interactivity lead to an increase in the robot's ability to influence: in this case, the passage deviation of passersby. As hypothesised, holistic implementation of the Robot Centric paradigm enabled more nuanced, predictable and systematic influence to be achieved and Responsive (i.e. more interactive) information systems had greater effectiveness, yielding larger influence. Thus, the methodology was drawn on to successfully operationalise the Robot Centric paradigm during real-world HRI, as per **RQ: Methodology**. These results suggest that the findings thus far, and the methodology, do generalise.

Chapter 8

Conclusions

Through literature and a series of both piecemeal and holistic experiments ($n_{total} = 435 = 16 + 24 + 26 + 96 + 189 + 84$), this thesis explored the feasibility of developing a methodology for successful operationalisation of the Robot Centric paradigm during real-world HRI. In investigating whether such a methodology was achievable, several questions arose. Firstly, the question of whether cues common in human-human interaction can be reliably implemented and utilised during real-world HRI was addressed. It was demonstrated that an exemplar higher human-likeness social robot's presentation of an exemplar JA cue to *Elicit* behavioural responses resulted in effects in line with human-human cues. The advantages of *in situ* sociocontextual cue *Read*ing capabilities were also demonstrated, and a head yaw estimation framework to detect gaze behaviour in the HRI space was developed. The relationship between robot interactivity and effectiveness at achieving its goal(s) was then shown in the exemplar instance, and through subsequently reproducing this relationship with a sociocontextual cue, social robot and application space distinct from the exemplar instance, the developed methodology and findings were shown to be generalisable.

Through the methodology, the Robot Centric HRI paradigm can thus be operationalised to give social robots the ability to leverage sociocontextual cues, enabling them to more effectively achieve their goal(s) (such as instantiating interactions, shaping interaction participant roles and/or resolving ambiguity) [81] and meet the expectations of communicating in a socially sensitive manner (as required by their growing interaction peer role).

8.1 Specific Conclusions on Contributions

8.1.1 Methodology for operationalisation of the Robot Centric paradigm during real-world HRI

A methodology for operationalisation of the Robot Centric paradigm during real-world HRI was developed. The four main stages of the methodology – target problem and robot goal(s) definition, application space definition, Robot Centric HRI paradigm design, and implementation design – enable the paradigm to be successfully operationalised, in spite of the complex nature of both human behaviour and the dynamics of interaction.

As a result of this methodology, the Robot Centric HRI paradigm can be leveraged to position social robots as interaction peers able to utilise sociocontextual cues to derive and/or achieve their goal(s) with greater effectiveness (as discussed below in Section 8.1.2). As a result, they are able to communicate in a more socially sensitive manner and, more importantly, meet the requirements of the new societal roles they are fulfilling through increased levels of agency and the ability to lead interactions and resolve ambiguity in situations of naïvety [81].

8.1.2 Demonstration that sociocontextual cues can be successfully leveraged during HRI via the Robot Centric HRI paradigm

Literature and a series of experiments were drawn on to demonstrate that sociocontextual cues are transferrable to HRI, and can be successfully leveraged in both directions of communication between humans and robots via the Robot Centric HRI paradigm, as detailed below and in ***[J1]***, ***[J2]***, ***[C2]***, ***[C3]***, and ***[W1]*** listed in Appendix A.

Elicit – The Effects of Robot-Issued Cues During Real-World HRI

A social exploration of the characteristics and effects of exemplar sociocontextual JA cues during HRI with an exemplar humanoid social robot was carried out. It was found that the robot's presentation of a JA cue resulted in significant effects on gaze-based measures of influence, specifically the development of gaze biases towards the objects in general, and the JA object and chosen object in particular. These results suggest that participants responded to the robot-issued JA cue in line with responses to human-issued cues, as detailed in psychology and behavioural science literature on the characteristics and effects of JA in HHI. In a subsequent study, this finding was shown to generalise to other sociocontextual cues, social robots and application spaces distinct from the exemplar instance.

This study demonstrated that sociocontextual cues are transferrable to HRI, and that social robots can successfully *Elicit* particular behavioural responses from interaction partners, as necessitated by their interaction peer role.

Read – Robots Deciphering Human-Issued Cues

Through a study focusing on understanding people's natural gaze behaviour towards robots during real-world HRI, it was shown that no observable generalisable pattern of gaze behaviour exists. To enable robots to intentionally and effectively *Elicit* during interactions – issuing cues when a person is paying attention, for example – it is therefore valuable for them to to have *in situ* cue *Read*ing capabilities.

To address this need, a head yaw estimation framework to detect gaze behaviour *in situ* was developed. The framework leverages the strengths of multiple HYE methods, including a novel method, to achieve operation in the HRI space with HRI-suitable, landmark levels of accuracy: an internal validation gave a mean accuracy of ±3.7° when participants are positioned between ∼1–3m from the sensor and directly in front of it, and ±6.3° when they are offset by up to ∼1m left or right from the line of sight of the sensor. Additionally, the framework was shown to be repeatable and robust to interpersonal variations in appearance over time. By incorporating the complementary methods into a single head pose estimate, the framework thus addresses the trade-off between accuracy and physical operation space inherent in many existing HYE methods.

8.1.3 Deepened understanding of the Robot Centric HRI paradigm

Kirchner & Alempijevic [81] speculatively proposed the concept of interactivity, i.e. a robot's ability to moderate its *Elicit* strategy based on information gained through *Read*ing. However, due to the piecemeal exploration of the Robot

Centric HRI paradigm in their work, the paradigm was never implemented in such a way as to explore interactivity and its relationship to a robot's effectiveness at achieving its goal(s). The work presented in this thesis investigates this relationship, deepening the understanding of, and demonstrating that, the Robot Centric HRI paradigm can be operationalised holistically during real-world HRI.

Specifically, a study was carried out which explored whether a robot's effectiveness at achieving its goal(s) would be increased by greater interactivity; that is, whether both the *Read* and *Elicit* branches of the Robot Centric HRI paradigm are valuable to social robots in effectively achieving their goal(s). Specifically, the wider effects of robot-issued JA cues (such as the possibility of the cue influencing people to prefer the JA object less), as well as how people's gaze behaviour impacts on the effects and perceptions of such cues, were examined. The preferences of 96 participants in a dichotomous choice situation in which a robot issued a JA cue were compared, which showed that while JA cues in HRI can have effects in line with HHI JA cues, people have a greater tendency towards suspicion of, rather than compliance with, the robot, especially if they are not looking when the JA cue is issued.

These results have implications for the design of sociocontextual cues in HRI in general, and JA cues in particular. Whether a robot is intentionally attempting to *Elicit* via JA cues, or issuing such cues inadvertently during functional actions, a JA cue towards that object has the potential to positively influence people towards a particular object. However, this is only the case if the person directly witnesses the cue; sociocontextual cues in HRI can also have unexpected wider effects on surrounding humans.

This finding demonstrates the value of the *Read* and *Elicit* branches, and hence the relationship between a robot's level of interactivity and its effectiveness: *Elicit* strategy moderation is valuable to a robot in effectively achieving its goal(s), and information valuable to moderate *Elicit* can be gained through *in situ Read*ing capabilities. These are further detailed in ***[J1]*** and ***[C1]*** listed in Appendix A.

While the ability of the developed head yaw estimation framework to address the identified need to *Read* human gaze during real-world HRI was found to be limited, it showed potential in the intended application, and its shortcomings should be addressed in future work.

8.1.4 Demonstration of generalisability of the Robot Centric HRI paradigm and the devised methodology

Following holistic verification of the Robot Centric HRI paradigm and devised methodology with an exemplar sociocontextual cue and social robot, this work was reinforced and demonstrated to generalise through the subsequent exploration of the relationship between a lower-HL, disembodied social robot's interactivity and the effectiveness of its influence on people in public spaces. A two-part study was conducted in both a major Australian public train station and a university, where passersby encountered the robot, designed with various levels of interactivity, which attempted to influence their passage. It was shown that higher interactivity led to greater effectiveness (as detailed in ***[C1]*** of Appendix A), and thus that the devised methodology can be drawn on to successfully operationalise the Robot Centric paradigm during real-world HRI, enabling robots to leverage sociocontextual cues to achieve their goal(s).

8.2 Future Research

Three main directions for future research were identified during this thesis. Firstly, while the developed head yaw estimation framework showed potential suitability for addressing the identified need for *in situ* gaze *Read*ing in the HRI space, the framework currently has limitations in real-world HRI situations. This is particularly the case when participants' heads are oriented downwards with respect to the robot: the unconstrained movement of people's heads in real-world interaction scenarios means this pitch must be considered. The prospect of addressing this issue, and therefore enabling robots to leverage this information in order to effectively moderate their *Elicit* strategy, is particularly appealing.

A second aspect for future work is further investigating the interactivity of robots, particularly other sociocontextual cues the robot could *Read* to more intelligently moderate its *Elicit* strategy. For example, it was shown in this work that an understanding of when people's gaze is directed at the robot can enable the robot to communicate intentionally to people in the environment through an understanding of when people's attention is directed at the robot and when it is not. The ability to *Read* other cues seems likely to add similarly valuable information to enable more effective *Elicit*.

Finally, other interesting directions for future research would be to further explore and understand the effects of other factors on a robot's ability to influence, including the congruency of a robot's cue with its human-likeness (which in this work were thus far matched fairly well) and people's habituation to the robot.

Appendix A

Publications and Other Outcomes

The work reported in this thesis has resulted in a number of peer reviewed publications and awards, which are described in this section. The Directly Related Publications section lists the publications which are directly related to the work presented in this thesis, while the Related section lists publications related, in some significant manner, to the work described here within. The Awards and Experience section describes the awards received for and experience gained through the work discussed in this thesis.

A.1 Directly Related Publications

Journal Articles

[J1] – **S. Caraian** and N. Kirchner. Effects of Robot-Issued Joint Attention Cues in HRI. *Journal of Human-Robot Interaction*, 2014. *Under review.*

[J2] – **S. Caraian**, N. Kirchner, A. Alempijevic and T. Vidal-Calleja. Gaze Estimation for Social Human-Robot Interaction. *Special Issue of the ACM Transactions on Interactive Intelligent Systems on 'New Directions in Eye Gaze for Interactive Intelligent Systems'*, 2015. *Under review.*

Conference Papers

[C1] – **S. Caraian**, N. Kirchner and Peter Colborne-Veel. Moderating a Robot's Ability to Influence People Through its Level of Sociocontextual Interactiveness. In *HRI '15: Proceedings of the 10th ACM/IEEE Conference on Human-Robot Interaction*, pp. 1–7, 2015.

[C2] – **S. Caraian** and N. Kirchner. Head Pose Behavior in the Human-Robot Interaction Space. In *HRI '14: Proceedings of the 9th ACM/IEEE Conference on Human-Robot Interaction*, pp. 132–133, 2014.

[C3] – **S. Caraian** and N. Kirchner. Influence of Robot-Issued Joint Attention Cues on Gaze and Preference. In *HRI '13: Proceedings of the 8th ACM/IEEE Conference on Human-Robot Interaction*, pp. 95–96, 2013.

Workshop Papers

[W1] – **S. Caraian** and N. Kirchner. Incidental Robot Gaze Behavior Inadvertently Influencing Choice. In *HRI '13: Proceedings of the 8th ACM/IEEE Conference on Human-Robot Interaction*, 2013.

A.2 Related Publications

Conference Papers

[RC1] – A. Alempijevic, **S. Caraian**, D. Egan-Wyer, G. Dissanayake, R. Fitch, B. Hengst, D. Hordern, N. Kirchner, M. Koob, M. Pagnucco, C. Sammut and A. Virgona. RobotAssist – RoboCup@Home 2011 Team Description Paper. In *RoboCup@Home Competition*, 2011.

[RC2] – **S. Caraian** and N. Kirchner. Robust Manipulability-Centric Object Detection in Time-of-Flight Camera Point Clouds. In *Proceedings of the 2010 Australasian Conference on Robotics and Automation*, 2010.

[RC3] – N. Kirchner, A. Alempijevic, **S. Caraian**, R. Fitch, D. Hordern, G. Hu, G. Paul, D. Richards, S. P.N. Singh and S. Webb. RobotAssist - a Platform for Human Robot Interaction Research. In *Proceedings of the 2010 Australasian Conference on Robotics and Automation*, 2010.

A.3 Awards and Recognition

[AR1] – **Asia-Pacific Chair** – HRI Pioneers Workshop, Bielefield, Germany, 2014. Participated in planning of workshop and oversaw advertisement of workshop in Asia-Pacific region.

[AR2] – **Best Presentation Award (Mechatronics, Robotics, and Health Technology)** – Faculty of Engineering & IT Engineering Research Showcase 2013. Awarded on basis of presentation content, style and confidence.

[AR3] – **Best Presentation Award (Mechatronics, Robotics, and Health Technology)** – Faculty of Engineering & IT Engineering Research Showcase 2012.

[AR4] – **Collaborative HRI Research** – ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan, 2012. Experimental collaboration in Human-Robot Interaction.

[AR5] – **Visiting Student Researcher (Invited)** – Stanford University, CA, 2011. Three months in Mechanical Engineering Faculty.

[AR6] – **Finalist in RoboCup@Home Competition** – Istanbul, Turkey, 2011. Member of RobotAssist team. Placed overall 4th in the world in second year of competition.

[AR7] – **Participant in RoboCup@Home Competition** – Singapore, 2010. Member of RobotAssist team. Earned a place in finals in first year of competition.

References

[1] H. Aarts, P. M. Gollwitzer, and R. R. Hassin. Goal Contagion: Perceiving Is for Pursuing. *Journal of Personality and Social Psychology*, 87(1):23–37, 2004.

[2] K. Alberto Funes Mora and J.-M. Odobez. Gaze Estimation from Multimodal Kinect Data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.

[3] A. Alempijevic, S. Caraian, D. Egan-Wyer, G. Dissanayake, R. Fitch, B. Hengst, D. Hordern, N. Kirchner, M. Koob, M. Pagnucco, C. Sammut, and A. Virgona. RobotAssist – RoboCup@Home 2011 Team Description Paper. In *RoboCup 2011 Humanoid League team descriptions*, 2011.

[4] A. Alempijevic, R. Fitch, and N. Kirchner. Bootstrapping Navigation and Path Planning Using Human Positional Traces. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1234–1239, 2013.

[5] M. Andriluka, S. Roth, and B. Schiele. People-Tracking-by-Detection and People-Detection-by-Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[6] M. Annett. The binomial distribution of right, mixed and left handedness. *Quarterly Journal of Experimental Psychology*, 19(4):327–333, 1967.

[7] D. Archer and R.M. Akert. Words and Everything Else: Verbal and Nonverbal Cues in Social Interpretation. *Journal of Personality and Social Psychology*, 35(6):443–449, 1977.

[8] M. Argyle and J. Dean. Eye-Contact, Distance and Affiliation. *Socieometry*, 28(3):289–304, 1965.

[9] Autistm Training Solutions. Play Skills.....Are We Missing Something? http://autismtrainingsolutions.wordpress.com/2009/09/25/play-skills-are-we-missing-something/.

[10] S. Ba and J.-M. Odobez. Multiperson Visual Focus of Attention from Head Pose and Meeting Contextual Cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, 2011.

[11] A.T. Bahill, M.R. Clark, and L. Stark. The main sequence a tool for studying human eye movements. *Mathematical Biosciences*, 24(3–4):191–204, 1975.

[12] C. Bartneck and J. Forlizzi. A Design-Centred Framework for Social Human-Robot Interaction. In *Ro-Man*, pages 591–594, 2004.

[13] S. Battersby, M. Lavelle, P. Healey, and R. McCabe. Analysing Interaction: A comparison of 2D and 3D techniques. In *Proceeding the LREC Workshop on Multi-Modal Corpora*, 2008.

[14] A. P. Bayliss, M. A. Paul, P. R. Cannon, and S. P. Tipper. Gaze cuing and affective judgments of objects: I like what you look at. *Psychonomic Bulletin and Review*, 13(6):1061, 2006.

[15] C. Becchio, C. Bertone, and U. Castiello. How the gaze of others influences object processing. *Trends in Cognitive Sciences*, 12(7):254–258, 2008.

[16] E.V. Bonilla, K.M.A. Chai, and C.K.I. Williams. Multi-task Gaussian Process Prediction. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, pages 153–160, 2008.

[17] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. Effects of Nonverbal Communication on Efficiency and Robustness in Human-Robot Teamwork. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2005.

[18] J.D. Bullough and N.P. Skinner. Influence of Visual Signal Flash Intensity and Duration on Perception. In *Suppression, Detection, and Signaling Research and Applications Symposium (SUPDET)*, pages 1–5, 2013.

[19] Bureau of Transport Statistics (BTS). Compendium of Sydney Rail Travel Statistics. Technical report, Transport for New South Wales (TfNSW), 2012.

[20] G. Castellano and C. Peters. Socially perceptive robots: challenges and concerns. *Interaction Studies*, 11(2):201–207, 2010.

[21] A. Caunce, D. Cristinacce, C. Taylor, and T. Cootes. Locating Facial Features and Pose Estimation Using a 3D Shape Model. *ISVC*, 1:750–761, 2009.

[22] K. Chiodo. Gesturing While Talking Helps Change Your Thoughts. `http://www.sott.net/article/220978-Gesturing-While-Talking-Helps-Change-Your-Thoughts`.

[23] V. Corkum and C. Moore. *Joint attention: Its origins and role in development*, chapter Development of joint visual attention in infants. Lawrence Erlbaum Associates, Hillsdale, NJ, England, 1995.

[24] G. D'Egidio, R. Patel, B. Rashidi, M. Mansour, E. Sabri, and P. Milgram. A study of the efficacy of flashing lights to increase the salience of alcohol-gel dispensers for improving hand hygiene compliance. *American Journal of Infection Control*, 42:852–855, 2014.

[25] B. D'Entremont, S. M. J. Hains, and D. W. Muir. A demonstration of gaze following in 3- to 6-month-olds. *Infant Behavior and Development*, 20(4):569–572, 1997.

[26] A. Deshmukh, G. Castellano, M.Y. Lim, R. Aylett, and P.W. McOwan. Ubiquitous Social Perception Abilities for Interaction Initiation in Human-Robot Interaction. In *ACM Multimedia 2010 Workshop - Affective Interaction in Natural Environments*, 2010.

[27] M. Doniec, G. Sun, and B. Scassellati. Active Learning of Joint Attention. In *6th IEEE-RAS International Conference on Humanoid Robots*, pages 34–39, 2006.

[28] D. Droeschel, J. Stückler, and S. Behnke. Learning to Interpret Pointing Gestures with a Time-of-Flight Camera. In *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Lausanne, Switzerland, 2011.

[29] D. Droeschel, J. Stuckler, D. Holz, and S. Behnke. Towards Joint Attention for a Domestic Service Robot - Person Awareness and Gesture Recognition using Time-of-Flight Cameras. In *IEEE International Conference on Robotics and Automation*, 2011.

[30] B.R. Duffy. *The social robot.* PhD thesis, Department of Computer Science, University College Dublin, 2000.

[31] B.R Duffy. Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42:177–190, 2003.

[32] M. Earls. Pointing and Gawking. `http://herd.typepad.com/herd_the_hidden_truth_abo/2008/12/pointing-and-gawking.html`.

[33] M.A. El-Beltagy and W.A. Wright. Gaussian Processes for Model Fusion. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, 2001.

[34] P.C. Ellsworth and A.M. Carlsmith. Effects of Eye Contact and Verbal Content on Affective Response to Dyadic Interaction. *Journal of Personality and Social Psychology*, 10(1):15–20, 1968.

[35] P.C. Ellsworth and L.M. Ludwig. Visual Behavior in Social Interaction. *The Journal of Communication*, 22:375–403, 1972.

[36] N.J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24:581–604, 2000.

[37] R.V. Exline. Explorations in the process of person perception: Visual interaction in relation to competition, sex, and need for affiliation. *Journal of Personality*, 31:1–20, 1963.

[38] R.V. Exline, S.L. Ellyson, and B. Long. *Nonverbal Communication of Aggression*, chapter Visual Behavior as an Aspect of Power Role Relationships. Plenum Press, New York, 1975.

[39] F. Eyssel, D. Kuchenbrandt, and S. Bobinger. Effects of Anticipated Human-Robot Interaction and Predictability of Robot Behavior on Perceptions of Anthropomorphism. In *HRI '11: Proceedings of the 6th ACM/IEEE International Conference on Human Robot Interaction*, 2011.

[40] G. Fanelli, T. Weise, J. Gall, and L. Van Gool. Real Time Head Pose Estimation from Consumer Depth Cameras. In *33rd Annual Symposium of the German Association for Pattern Recognition*, pages 101–110, 2011.

[41] R. Fantz. Visual Experience in Infants: Decreased Attention to Familiar Patterns Relative to Novel Ones. *Science*, 146(3644):668–670, 1964.

[42] M. Farenzena, A. Tavano, L. Bazzani, D. Tosato, G. Paggetti, G. Menegaz, V. Murino, and M. Cristani. Social interactions by visual focus of attention in a three-dimensional environment. In *Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis (PRAI*HBA)*, volume 30, pages 115–127, 2009.

[43] S.J. Frances. Sex Differences in Nonverbal Behavior . *Sex Roles*, 5(4):519–535, 1979.

[44] E.G. Freedman and D.L. Sparks. Eye-Head Coordination During Head-Unrestrained Gaze Shifts in Rhesus Monkeys. *Journal of Neurophysiology*, 77:2328–2348, 1997.

[45] C. K. Friesen and A. Kingstone. The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin and Review*, 5(3):490–495, 1998.

[46] C. K. Friesen and A. Kingstone. Abrupt onsets and gaze direction cues trigger independent reflexive attentional effects. *Cognition*, 87(1):1–10, 2003.

[47] A. Gaschler, K. Huth, M. Giuliani, I. Kessler, J. de Ruiter, and A. Knoll. Modelling State of Interaction from Head Poses for Social Human-Robot Interaction. In *Proceedings of the "Gaze in Human-Robot Interaction" Workshop held at the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012)*, 2012.

[48] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll. Social Behavior Recognition Using Body Posture and Head Pose for Human-Robot Interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[49] S. Gerathewohl. Conspicuity of Steady and Flashing Light Signals: Variation of Contrast. *J. Opt. Soc. Am. (JOSA)*, 43(7):567–568, 1953.

[50] M. Girolami. Bayesian Data Fusion with Gaussian Process Priors : An Application to Protein Fold Recognition. In *Workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology (PMSB)*, 2006.

[51] M. G. Glaholt, M.-C. Wu, and E. M. Reingold. Predicting preference from fixations. *PsychNology Journal*, 7(2):141–158, 2009.

[52] G.N. Goldberg, C.A. Kiesler, and B.E. Collins. Visual Behavior and Face-to-Face Distance During Interaction. *Socieometry*, 32(1):43–53, 1969.

[53] N. Gourier, D. Hall, and J. Crowley. Estimating Face orientation from Robust Detection of Salient Facial Structures. In *In Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*, 2004.

[54] E. Griffin. *A First Look at Communication Theory*, chapter Proxemic Theory of Edward Hall, pages 60–67. McGraw-Hill, London, 2008.

[55] V. Groom. What's the best role for a robot? Cybernetic models of existing and proposed human-robot interaction structures. In *Proceedings of the International Conference on Informatics in Control, Automation, and Robotics (ICINCO)*, Funchal, Portugal, 2008.

[56] E.T. Hall. A System for the Notation of Proxemic Behavior. *American Anthropologist*, 65(5):1003–1026, 1963.

[57] E.T. Hall. *The Hidden Dimension*. Doubleday, Garden City, NY, 1966.

[58] J.A. Hall, E.J. Coats, and L.S. LeBeau. Nonverbal Behavior and the Vertical Dimension of Social Relations: A Meta-Analysis. *Psychological Bulletin*, 131(6):898—924, 2005.

[59] J. Ham, R. Bokhorst, and J.-J. Cabibihan. The influence of gazing and gestures of a storytelling robot on its persuasive power. *International semester research project*, 2010.

[60] T. Hashimoto, M. Senda, and H. Kobayshi T. Shiiba. Development of the interactive receptionist system by the face robot. In *SICE Annual Conference*, 2004.

[61] C. Heath and P. Healey. Making Space for Interaction: Architects Design Dialogues. In *9th International Gesture Workshop*, pages 250–261, 2012.

[62] P. Holthaus, K. Pitsch, and S. Wachsmuth. How Can I Help? Spatial Attention Strategies for a Receptionist Robot. *International Journal of Social Robotics*, 3(4):383–393, 2011.

[63] B. Hood, J. Willen, and J. Driver. Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, 9(2):131–134, 1998.

[64] D. Hordern and N. Kirchner. Robust and Efficient People Detection with 3-D Range Data using Shape Matching. In *Proceedings of the 2010 Australasian Conference on Robotics and Automation*, 2010.

[65] C.-M. Huang and B. Mutlu. Robot Behavior Toolkit: Generating Effective Social Behaviors for Robots. In *HRI '12: Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, 2012.

[66] C.-M. Huang and A. Thomaz. Effects of Responding to, Initiating and Ensuring Joint Attention in Human-Robot Interaction. In *20th IEEE International Symposium on Robot and Human Interactive Communication*, 2011.

[67] H. Huettenrauch, K.S. Eklundh, A. Green, and E.A.Topp. Investigating Spatial Relationships in Human-Robot Interaction. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[68] M. Imai, T. Ono, and H. Ishiguro. Physical Relation and Expression: Joint Attention for Human-Robot Interaction. In *IEEE Transactions on Industrial Electronics*, volume 50, 2003.

[69] iRobot Corporation. iRobot Roomba. `http://www.iRobot.com/`.

[70] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu. Robovie: an Interactive Humanoid Robot. *Industrial Robot: An International Journal*, 28(6):498–504, 2001.

[71] W. Ju and L. Leifer. The Design of Implicit Interactions: Making Interactive Systems Less Obnoxious. *Design Issues*, 24(3):72–84, 2008.

[72] T. Kanda, H. Ishiguro, T. Ono, M. Imai, and R. Nakatsu. Development and evaluation of an interactive humanoid robot "Robovie". In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, volume 2, pages 1848–1855, 2002.

[73] F. Kaplan and V. Hafner. The Challenges of Joint Attention. *Fourth International Workshop on Epigenetic Robotics*, 7(2):135–169, 2006.

[74] K. Kawamura, R.T. Pack, and M. Iskarous. Design Philosophy for Service Robots. In *IEEE International Conference on Systems, Man and Cybernet-*

ics: Intelligent Systems for the 21st Century, volume 4, pages 3736–3741, 1995.

[75] A. Kendon. Some Functions of Gaze-direction in Social Interaction. *Acta Psyhologica*, 26:22–63, 1967.

[76] A. Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters.* Cambridge University Press, 1990.

[77] A. Kendon and A. Ferber. *Comparative ecology and behavior of primates*, chapter A Description of some human greetings, pages 591–668. Academic Press, London, 1973.

[78] K. Khoshelham. Accuracy analysis of kinect depth data. In *ISPRS workshop laser scanning*, pages 133–138, 2011.

[79] K. Khoshelham and S.J. Oude Elberink. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors*, 12(2):1437–1454, 2012.

[80] T. Kim and P. Hinds. Who Should I Blame?: Effects of Autonomy and Transparency on Human-Robot Interaction. In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 80–85, 2006.

[81] N. Kirchner and A. Alempijevic. A Robot Centric Perspective on the HRI Paradigm. *Journal of Human-Robot Interaction - Special Issue on HRI Perspectives and Projects from around the Globe*, 2(1):1–23, 2012.

[82] N. Kirchner, A. Alempijevic, S. Caraian, R. Fitch, D. Hordern, G. Hu, G. Paul, D. Richards, S.P.N. Singh, and S. Webb. RobotAssist – a Platform for Human Robot Interaction Research. In *Proceedings of the Australasian Conference on Robotics and Automation*, pages 1–8, 2010.

[83] N. Kirchner, A. Alempijevic, X. Dai, P.G. Plöger, and R. K. Venkat. A robust people detection, tracking, and counting system. In *Australasian Conference on Robotics and Automation*, pages 1–8, 2014.

[84] N. Kirchner, A. Alempijevic, and G. Dissanayake. Nonverbal Robot-Group Interaction Using an Imitated Gaze Cue. In *HRI '11: Proceedings of the 6th ACM/IEEE International Conference on Human Robot Interaction*, pages 497–504, 2011.

[85] N. Kirchner, A. Alempijevic, and A. Virgona. Head-to-shoulder signature for person recognition. In *Proceedings of IEEE International Conference on Robotics and Automation*, 2012.

[86] C. Kleinke. Gaze and Eye Contact: A Research Review. *Psychological Bulletin*, 100(1):78–100, 1986.

[87] M. Knapp, J. Hall, and T. Horgan. *Non-verbal Communication in Human Interaction.* Cengage Learning, Boston, MA, 2013.

[88] G. Kurtenbach and E. Hulteen. *The Art of Human Computer Interface Design*, chapter Gestures in Human-Computer Communications, pages 309–317. Addison-Wesley, Reading, MA, 1990.

[89] S. Langton, H. Honeyman, and E. Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & Psychophysics*, 66(5):752–771, 2004.

[90] S.R Langton, R.J. Watt, and I. Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in Cognitive Sciences*, 4(2):50–59, 2000.

[91] O. Lanz, R. Brunelli, P. Chippendale, M. Voit, and R. Stiefelhagen. *Computers in the Human Interaction Loop*, chapter Extracting Interaction Cues: Focus of Attention, Body Pose, and Gestures, pages 87–93. Springer, London, UK, 2009.

[92] M.H. Levine and B. Sutton-Smith. Effects of Age, Sex, and Task on Visual Behavior during Dyadic Interaction. *Developmental Psychology*, 9(3):400–405, 1973.

[93] Y. Li, H. Ai, C. Huang, and S. Lao. Robust Head Tracking Based on a Multi-State Particle Filter. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006.

[94] D. Linder. *Personal Space.* General Learning Press, Morristown, NJ, 1974.

[95] P. Liu, Reale M., and L. Yin. 3D Head Pose Estimation Based on Scene Flow and Generic Head Model. In *IEEE International Conference on Multimedia and Expo*, pages 794–799, 2012.

[96] T.D. Lyon. Questioning Children: The Effects of Suggestive and Repeated Questioning. *USC Law School, Olin Working Paper No. 99-24*, 1999.

[97] K. F. MacDorman and H. Ishiguro. The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies*, 7(3):297–337, 2006.

[98] N. Marquardt and S. Greenberg. Informing the Design of Proxemic Interactions. *Pervasive Computing*, 11(2):14–23, 2012.

[99] D.P. McAdams, R.J. Jackson, and C. Kirshnit. Looking, laughing, and smiling in dyads as a function of intimacy motivation and reciprocity. *Journal of Personality*, 5(3):261–273, 1984.

[100] A. McCoy and M. Platt. Expectations and outcomes: decision-making in the primate brain. *Journal of Comparative Physiology A*, 191(3):201–211, 2004.

[101] S. Meers and K. Ward. Head-Pose Tracking with a Time-of-Flight Camera. In *Proceedings of the Australasian Conference on Robotics and Automation*, 2008.

[102] A. Mehrabian. Some referents and measures of nonverbal behavior. *Behavior Research Methods and Instrumentation*, 1(6):203–207, 1969.

[103] A. Mehrabian. *Nonverbal Communication.* Aldine-Atherton, Chicago IL, 1972.

[104] A. Melkumyan and F. Ramos. A Sparse Covariance Function for Exact Gaussian Process Inference in Large Datasets. In *Proceedings of the 21st International Joint Conference on Artifical intelligence*, pages 1936–1942, 2009.

[105] Microsoft. Kinect. `http://www.xbox.com/en-us/kinect/`.

[106] T. Minato, M. Shimada, S. Itakura, K. Lee, and H. Ishiguro. Evaluating the human likeness of an android by comparing gaze behaviors elicited by the android and a person. *Advanced Robotics*, 20(10):1147–1163, 2006.

[107] S. Mitra and T. Acharya. Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, 37(3):311–324, 2007.

[108] L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell. Fast Stereo-Based Head Tracking for Interactive Environments. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 375–380, 2002.

[109] L.-P. Morency, A. Rahimi, and T. Darrell. Adaptive View-Based Appearance Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 803–810, 2003.

[110] D. Morris, P. Collett, P. Marsh, and M. O'Shaughnessy. *Gestures.* Stein and Day, NY, 1979.

[111] T. Mukai, S. Hirano, H. Nakashima, Y. Kato, Y. Sakaida, S. Guo, and S. Hosoe. Development of a nursing-care assistant robot RIBA that can lift a human in its arms. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010.

[112] E. Murphy-Chutorian and M. Manubhai Trivedi. Head Pose Estimation in Computer Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[113] B. Mutlu, C. Bartneck, J. Ham, V. Evers, and T. Kanda, editors. *Social Robotics: Proceedings of the 3rd International Conference on Social Robotics.* Springer, Amsterdam, Netherlands, 2011.

[114] B. Mutlu, J. Forlizzi, and J. Hodgins. A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In *6th IEEE-RAS International Conference on Humanoid Robots*, 2006.

[115] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita. Nonverbal Leakage in Robots: Communication of Intentions through Seemingly Unintentional Behavior. In *HRI '09: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 2009.

[116] M. Natale. Induction of mood states and their effect on gaze behaviors. *Journal of Consulting and Clinical Psychology*, 450:960, 1977.

[117] I. Nevo, M. Fitzpatrick, R.E. Thomas, P.A. Gluck, J.D. Lenchus, K.L Arheart, and D.J. Birnbach. The Efficacy of Visual Cues to Improve Hand Hygiene Compliance. *Journal of the Society for Simulation in Healthcare*, 5(6):325–331, 2010.

[118] J. S. Nguyen, T. H. Nguyen, and H. T. Nguyen. Semi-autonomous wheelchair system using stereoscopic cameras. In *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009.

[119] N. Nova. A Review of How Space Affords Socio-Cognitive Processes during Collaboration. *PsychNology*, 3(2):118–148, 2005.

[120] K. O'Brien, J. Sutherland, C. Rich, and C. Sidner. Collaboration with an Autonomous Humanoid Robot: A Little Gesture Goes a Long Way. In *HRI '11: Proceedings of the 6th ACM/IEEE International Conference on Human Robot Interaction*, 2011.

[121] M.A. Osborne, A. Rogers, S. Ramchurn, S.J. Roberts, and N.R. Jennings. Towards Real-Time Information Processing of Sensor Network Data using Computationally Efficient Multi-output Gaussian Processes. In *International Conference on Information Processing in Sensor Networks*, 2008.

[122] A. Ozyurek. *Language and Gesture*, chapter The influence of addressee location on spatial language and representational gestures of direction. Cambridge University Press, Cambridge, UK, 2000.

[123] A. Ozyurek. Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language*, 46:688–704, 2002.

[124] PARO Robots U.S., Inc. PARO. `http://www.parorobots.com/`.

[125] C. Pelachaud, V. Carofiglio, B. De Carolis, F. de Rosis, and I. Poggi. Embodied contextual agent in information delivering application. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent*, pages 758–765, 2002.

[126] R. Pieters and L. Warlop. Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of Research in Marketing*, 16:1–16, 1999.

[127] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, Boston, MA, 2006.

[128] P. Ravindra, S. De Silva, K. Tadano, A. Saito, S. Lambacher, and M. Higashi. The Development of an Assistive Robot for Improving the Joint Attention of Autistic Children. In *IEEE 11th International Conference on Rehabilitation Robotics*, 2009.

[129] J. Reason. Combating omission errors through task analysis and good reminders. *Qual Saf Health Care*, 11:40–44, 2002.

[130] B. Reeves and C. Nass. *The media equation: how people treat computers, television, and new media like real people and places.* Cambridge University Press, New York, NY, 1996.

[131] V. M. Reid and T. Striano. Adult gaze influences infant attention and object processing: implications for cognitive neuroscience. *European Journal of Neuroscience*, 21:1763–1766, 2005.

[132] V. M. Reid, T. Striano, J. Kaufman, and M. H. Johnson. Eye gaze cueing facilitates neural processing of objects in 4-month-old infants. *Neuroreport*, 15(16):2553–2555, 2004.

[133] V.P. Richmond and J.C. McCroskey. *Nonverbal behavior in interpersonal relations.* Allyn and Bacon, Needham Heights, MA, 3 edition, 1995.

[134] B. Robins, K. Dautenhahn, and P. Dickerson. Robots as Embodied Beings - Interactionally Sensitive Body Movements in Interactions Among Autistic Children and a Robot . In *Proceedings of the IEEE International Workshop on Robots and Human Interactive Communication*, pages 54–59, 2005.

[135] E. Russo and F. Leclerc. An eye-fixation analysis of choice processes for consumer nondurables. *Journal of Consumer Research*, 21(2):274–290, 1994.

[136] J.W. Scadding and N. Losseff. *Clinical Neurology.* Hodder Arnold, London, UK, 2011.

[137] E. R. Schotter, R. W. Berry, C. R. McKenzie, and K. Rayner. Gaze bias: Selective encoding and liking effects. *Visual Cognition*, 18(8):1113–1132, 2010.

[138] S. Sheikhi and J.-M. Odobez. Recognizing the Visual Focus of Attention for Human Robot Interaction. *Human Behavior Understanding: Lecture Notes in Computer Science*, 7559:99–112, 2012.

[139] J. Sherrah and S. Gong. Fusion of perceptual cues for robust tracking of head pose and position. *Pattern Recognition*, 34(8):1565–1572, 2001.

[140] J. Sherrah, S. Gong, and E.-J. Ong. Face Distributions in Similarity Space under Varying Head Pose. *Image and Vision Computing*, 19(12):807–819, 2001.

[141] T. Shibata, M. Yoshida, and J. Yamato. Artificial emotional creature for human-machine interaction. In *IEEE International Conference on Computational Cybernetics and Simulation*, pages 2269–2274, 1997.

[142] S. Shimojo, C. Simion, and C. Shimojo. Gaze bias both reflects and influences preference. *Nature Neuroscience*, 6(12):1317–1322, 2003.

[143] C. Simion. *Orienting and Preference: An Equiry into the Mechanisms Underlying Emotional Decision Making.* PhD thesis, California Institute of Technology, 2005.

[144] G. Simmel. *Introduction to the Science of Sociology*, chapter Sociology of the Senses: Visual Interaction, pages 356–361. University of Chicago Press, Chicago, 1969.

[145] SportLinguist. Baby Behaviors Around 9-12 Months Enable "Conversation". `http://sportlinguist.com/2011/02/22/baby-behaviors-around-9-12-months-enable-conversation/`.

[146] J. Stahl. Amplitude of human head movements associated with horizontal saccades. *Experimental Brain Research*, 126(1):41–54, 1999.

[147] G. Stephenson and D. Rutter. Eye-contact, distance and affiliation: A re-evaluation. *British Journal of Psychology*, 61(3):385–393, 1970.

[148] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel. From gaze to focus of attention. In *Lecture Notes in Computer Science, Proceedings of Third International Conference on Visual Information Systems (VISUAL99)*, pages 761–768, 1999.

[149] T. Striano and D. Stahl. Sensitivity to triadic attention in early infancy. *Developmental Science*, 8(4):333–343, 2005.

[150] K.T. Strongman and B.G. Champness. Dominance Hierarchies and Conflict in Eye Contact. *Acta Psyhologica*, 28:376–386, 1968.

[151] J. Stuckler and S. Behnke. Integrating indoor mobility, object manipulation, and intuitive interaction for domestic service tasks. In *Proceedings of 9th IEEE-RAS International Conference on Humanoid Robots*, 2009.

[152] Y. Sun and L. Yin. Automatic Pose Recognition of 3D Facial Models. In *19th International Conference on Pattern Recognition*, pages 1–4, 2008.

[153] W. Swan. Australia to 2050: Future Challenges. Technical report, The Treasury, Australia, 2010.

[154] D.S. Syrdal, K. Dautenhahn, K.L. Koay, and M. Walters. The Negative Attitudes towards Robots Scale and Reactions to Robot Behaviour in a Live Human-Robot Interaction Study. In *Proceedings of New Frontiers in Human-Robot Interaction, a symposium at the AISB2009 Convention*, 2009.

[155] D.S. Syrdal, K.L Koay, M.L Walters, and K. Dautenhahn. A personalized robot companion?- The role of individual differences on spatial preferences in HRI scenarios. In *Proceedings of the 16th IEEE International Conference on Robot & Human Interactive Communication*, 2007.

[156] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll. Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences*, 5:675–691, 2005.

[157] A. Tsukahara, Y. Hasegawa, and Y. Sankai. Standing-Up Motion Support for Paraplegic Patient with Robot Suit HAL. In *IEEE 11th International Conference on Rehabilitation Robotics*, 2009.

[158] R. Valenti, N. Sebe, and T. Gevers. Combining Head Pose and Eye Location Information for Gaze Estimation. *IEEE Transactions on Image Processing*, 21(2):802–815, 2012.

[159] A. van der Weiden, H. Veling, and H. Aarts. When Observing Gaze Shifts of Others Enhances Object Desirability. *Emotion*, 10(6):939–943, 2010.

[160] S. Vasudevan, F. Ramos, E. Nettleton, and H. Durrant-Whyte. Nonstationary dependent Gaussian processes for data fusion in large-scale terrain modeling. In *IEEE International Conference on Robotics and Automation*, 2011.

[161] T. Vatahska, M. Bennewitz, and S. Behnke. Feature-based Head Pose Estimation from Images. In *7th IEEE-RAS International Conference on Humanoid Robots*, 2007.

[162] T. Vidal-Calleja, D. Su, F. De Bruijn, and J. Valls Miro. Learning Spatial Correlations for Bayesian Fusion in Pipe Thickness Mapping. In *IEEE International Conference on Robotics and Automation*, 2014.

[163] A. Vinciarelli, M. Panticc, and H. Bourlarda. Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[164] M. Voit, K. Nickel, and R. Stiefelhagen. Neural Network-Based Head Pose Estimation and Multi-view Fusion. In *First International Evaluation Workshop on Classification of Events, Activities and Relationships*, pages 291–298, 2006.

[165] J.J. Vos and A. Van Meeteren. Visual Processes Involved in Seeing Flashes. In *International Symposium of Imperial College of London: The Perception and Application of Flashing Lights*, pages 3–16, 1971.

[166] A. Weiss, T. Scherndl, M. Tscheligi, and A. Billard. Evaluating the ICRA 2008 HRI challenge. In *HRI '09: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, pages 261–262, 2009.

[167] C.D. Wickens and J.G. Hollands. *Engineering Psychology and Human Performance.* Prentice-Hall, Upper Saddle River, NJ, 2000.

[168] Wikipedia. List of gestures. `http://en.wikipedia.org/wiki/List_of_gestures`.

[169] W.H. Wollaston. On the Apparent Direction of Eyes in a Portrait. *Philosophical Transactions of the Royal Society of London*, 114:247–256, 1824.

[170] J. Wu and M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recognition*, 41(3):1138–1158, 2008.

[171] Z. Yücel and A.A. Salah. Resolution of focus of attention using gaze direction estimation and saliency computation. In *Proceedings of the International Conference on Affective Computing and Intelligent Interfaces*, 2009.

[172] Z. Yücel, A.A. Salah, C. Meriçli, T. Meriçli, R. Valenti, and T. Gevers. Joint Attention by Gaze Interpolation and Saliency. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2012.

[173] R. Zajonc. Attitudinal Effects of Mere Exposure. *Journal of Personality and Social Psychology Monograph Supplement*, 9(2):1–27, 1968.

[174] E.B. Zurif and M.P. Bryden. Familial Handedness and Left-Right Differences in Auditory and Visual Perception. *Neuropsychologia*, 7:179–187, 1969.