

EXPLAINABLE SHARED CONTROL  
IN ASSISTIVE ROBOTICS

MARK ZOLOTAS

*Thesis submitted for the degree of Doctor of Philosophy*

Supervised by PROFESSOR YIANNIS DEMIRIS  
Personal Robotics Lab  
Department of Electrical and Electronic Engineering  
Imperial College London

January 2021



## COPYRIGHT DECLARATION

---

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a [Creative Commons Attribution-NonCommercial-Share Alike 4.0 International Licence](#) (CC BY NC-SA).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that; you credit the author, do not use it for commercial purposes and share any derivative works under the same licence.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

## ORIGINALITY DECLARATION

---

I hereby declare that this thesis and the work herein detailed, was composed and originated by myself, except where appropriately referenced and credited.

*London, January 2021*

---

Mark Zolotas



## ABSTRACT

---

Shared control plays a pivotal role in designing assistive robots to complement human capabilities during everyday tasks. However, traditional shared control relies on users forming an accurate mental model of expected robot behaviour. Without this accurate mental image, users may encounter confusion or frustration whenever their actions do not elicit the intended system response, forming a misalignment between the respective internal models of the robot and human. The Explainable Shared Control paradigm introduced in this thesis attempts to resolve such model misalignment by jointly considering assistance and *transparency*.

There are two perspectives of transparency to Explainable Shared Control: the *human's* and the *robot's*. Augmented reality is presented as an integral component that addresses the human viewpoint by visually unveiling the robot's internal mechanisms. Whilst the robot perspective requires an awareness of human "intent", and so a clustering framework composed of a deep generative model is developed for human intention inference.

Both transparency constructs are implemented atop a real assistive robotic wheelchair and tested with human users. An augmented reality headset is incorporated into the robotic wheelchair and different interface options are evaluated across two user studies to explore their influence on mental model accuracy. Experimental results indicate that this setup facilitates transparent assistance by improving recovery times from adverse events associated with model misalignment. As for human intention inference, the clustering framework is applied to a dataset collected from users operating the robotic wheelchair. Findings from this experiment demonstrate that the learnt clusters are interpretable and meaningful representations of human intent.

This thesis serves as a first step in the interdisciplinary area of Explainable Shared Control. The contributions to shared control, augmented reality and representation learning contained within this thesis are likely to help future research advance the proposed paradigm, and thus bolster the prevalence of assistive robots.



## ACKNOWLEDGEMENTS

---

Without question, the very first person I must thank is my supervisor: Yiannis. I cannot think of anyone who has played a more profound role in my education, from the time I had my undergraduate interview with him when I was just 17, all the way through to this PhD thesis submission. Over this entire period, he has encouraged me to persevere whenever I faced adversity, inspired confidence whenever I was in doubt, and most importantly, provided an honest and insightful ear whenever I simply needed someone to talk to. Thank you Yiannis, I am forever grateful and proud to have been one of your students.

To members of the Personal Robotics Lab, thank you all for being the best colleagues anyone could ever dream of having. I would especially like to thank Tobi, Antoine and Hyung Jin, who have been invaluable mentors during my PhD, as well as incredible friends. I also wish to acknowledge a certain group of squash regulars, Fan, Daryl, Ahmed, Urbano and Josh, who entertained my need to vent both on and off the court – thanks guys, I really needed that. Last but not least, to Vini and Rodrigo (a.k.a. the “Wheel-chair Dudes”), I will always cherish the countless unforgettable memories we shared, be it in the lab or all the way off in China!

I am also sincerely grateful to various other members of the Electrical and Electronic engineering department. Thank you Jeremy for persuading me to set off on this academic journey in the first place; your wit, humour and passion for science was not lost on me. To Max, Kristel and yet again, Yiannis, thank you for placing your faith in me as a teaching assistant, which has been one of the highlights of my PhD. I would also be remiss if I did not thank Joan, who made administrative procedure a breeze and blessed our research group with coffee breaks! As for everyone else in the Intelligent System and Networks group, thank you all so much for being willing to take part in my crazy experiments (seriously... thank you).

Likewise, I wish to thank the EPSRC for their financial support throughout my PhD. In addition to awarding me a Doctoral Training Partnership, they extended this support for the last few months given the global circumstances, and so for that I am utterly grateful.

The PhD has been no easy endeavour, but it would have been near impossible if not for the select few who are my support system. To my loving

family, Eileen, Les, Hector and Ken, thank you for cheering me on every step of the way, it means the world to me. To my wonderful flatmates, Alex (my Goonie) and Omar (my Fiona), thank you for making it possible to escape the everyday stress and call a place home for the first time since leaving Athens. To my legendary buzzkills, Huss, Moh and Petro, thanks for all the laughs, fond memories and most of all, for always having my back... even in making this thesis.

Finally, to my dear mum, Marilyn, for whom there are no words to truly express my love and gratitude. All I will say is that you have been an unparalleled role model for everything in my life – this PhD included.

## CONTENTS

---

1	INTRODUCTION	15
1.1	Research Questions . . . . .	17
1.2	Thesis Contributions . . . . .	17
1.3	Thesis Overview and Organisation . . . . .	19
2	BACKGROUND	21
2.1	Shared Control . . . . .	21
2.1.1	Conditionally Assistive Robots . . . . .	22
2.1.2	Goal-Oriented Frameworks . . . . .	22
2.1.3	Learning-Based Frameworks . . . . .	23
2.1.4	Methods of Evaluation . . . . .	24
2.2	Augmented Reality in Human-Robot Interaction . . . . .	26
2.3	Model Reconciliation . . . . .	27
2.4	Human Intention Estimation . . . . .	29
2.4.1	Understanding Human Intent . . . . .	29
2.4.2	What is Intent? . . . . .	30
2.4.3	Computational Approaches . . . . .	31
2.5	Deep Generative Models in Robotics . . . . .	35
2.6	Closing Remarks . . . . .	36
3	SHARED CONTROL FOR ASSISTIVE ROBOT NAVIGATION	39
3.1	Application Setting: “Smart” Wheelchairs . . . . .	39
3.2	Assistive Robot Architecture . . . . .	40
3.3	Control-Sharing Methodology . . . . .	42
3.3.1	Intention Estimation . . . . .	42
3.3.2	Assistive Control . . . . .	44
3.4	Current Limitations . . . . .	47
3.5	Conclusions . . . . .	48
4	TOWARDS EXPLAINABLE SHARED CONTROL USING AUGMENTED REALITY	51
4.1	Terminology . . . . .	52
4.2	Explainable Robotic Wheelchair Assistance . . . . .	53
4.2.1	Augmented Reality Cueing System . . . . .	53
4.2.2	Experiments . . . . .	58
4.2.3	Discussion . . . . .	63

4.3	Explainable Shared Control Paradigm . . . . .	65
4.3.1	Guidelines for Transparency . . . . .	66
4.3.2	Instantiation for Assistive Robot Navigation . . . . .	68
4.3.3	Experiments . . . . .	70
4.3.4	Discussion . . . . .	77
4.4	Conclusions . . . . .	78
5	DISENTANGLED SEQUENCE CLUSTERING FOR HUMAN INTEN- TION INFERENCE . . . . .	81
5.1	Motivation . . . . .	82
5.2	Preliminaries . . . . .	84
5.2.1	Variational Autoencoders . . . . .	84
5.2.2	Variational Inference for Sequences . . . . .	86
5.3	Disentangled Sequence Clustering Variational Autoencoder . .	88
5.3.1	Clustering with Variational Autoencoders . . . . .	88
5.3.2	Model Specification . . . . .	91
5.3.3	Network Architecture . . . . .	93
5.3.4	Intention Inference . . . . .	93
5.4	Validation Setting: Moving MNIST . . . . .	94
5.4.1	Dataset and Implementation . . . . .	94
5.4.2	Evaluation Protocol . . . . .	95
5.4.3	Results . . . . .	96
5.4.4	Ablation Study . . . . .	98
5.5	Intention Inference on Robotic Wheelchairs . . . . .	100
5.5.1	Dataset . . . . .	100
5.5.2	Post-Processing . . . . .	101
5.5.3	Implementation . . . . .	101
5.5.4	Choosing K . . . . .	102
5.5.5	Evaluation . . . . .	103
5.5.6	Results . . . . .	104
5.5.7	Illuminating the Clusters . . . . .	106
5.6	Related Work . . . . .	107
5.7	Conclusions . . . . .	108
6	CONCLUSIONS AND FUTURE DIRECTIONS . . . . .	109
6.1	Overview of Thesis Contributions . . . . .	109
6.2	Outstanding Issues . . . . .	110
6.2.1	Addressing the Target Population . . . . .	110
6.2.2	General Applicability of Explainable Shared Control . .	111
6.2.3	Tracing Model Misalignment . . . . .	112

6.2.4	Communicating the Human Intention Inference Model .	112
6.2.5	Closing the Explainable Shared Control Loop . . . . .	113
6.3	Future Research . . . . .	113
6.3.1	Mutual Model Adaptation . . . . .	113
6.3.2	Hierarchical Intention Prediction . . . . .	114
6.3.3	Multisensory Modalities . . . . .	115
6.3.4	User Personalisation . . . . .	116
6.4	Epilogue . . . . .	116
BIBLIOGRAPHY		119
A	SOFTWARE PACKAGES	141
A.1	Localisation and Navigation Packages for Mobile Robots . . . .	141
A.2	Open-Source Contributions . . . . .	141
B	EYE-GAZE WHEELCHAIR	143
B.1	Motivation . . . . .	143
B.2	System Design . . . . .	143
C	AUGMENTED REALITY FOR DUAL-ARM ROBOTS	147
C.1	Explaining Affordable Robot Behaviours . . . . .	147
C.2	Links to Explainable Shared Control . . . . .	148
D	AUTHOR'S PUBLICATIONS	149

## LIST OF FIGURES

---

Figure 1.1	Thesis overview . . . . .	19
Figure 2.1	Intent representation in robotic wheelchairs . . . . .	31
Figure 2.2	Block diagram of forward and inverse model . . . . .	34
Figure 3.1	System diagram of smart wheelchair . . . . .	41
Figure 4.1	Composite image of the wheelchair system’s Augmented Reality visualisations . . . . .	54
Figure 4.2	Schematic of our Augmented Reality system components . . . . .	55
Figure 4.3	Gridmap processing pipeline . . . . .	56
Figure 4.4	An overhead view of the experiment route . . . . .	59
Figure 4.5	Total time to completion for each trial . . . . .	61
Figure 4.6	Head rotation in the yaw direction during the reverse passageway section . . . . .	62
Figure 4.7	Overview of Explainable Shared Control . . . . .	65
Figure 4.8	Composite image of the Augmented Reality visualisations rendered during assistive robot navigation . . . . .	68
Figure 4.9	First-person perspectives of the Augmented Reality visualisations . . . . .	69
Figure 4.10	A floor plan of the navigation route . . . . .	71
Figure 4.11	Two participants navigating the office doorway entrance . . . . .	73
Figure 4.12	Average timing results across commonly occurring events that can result from model misalignment . . . . .	74
Figure 4.13	Time-to-completion per participant, with and without visualisations . . . . .	75
Figure 4.14	Joint angular gaze distribution across all participants along the horizontal and vertical axes . . . . .	76
Figure 5.1	Overview of the intention inference experiment on a robotic wheelchair . . . . .	83
Figure 5.2	Deep generative model graphs . . . . .	86
Figure 5.3	Computation graph for the Disentangled Sequence Clustering Variational Autoencoder . . . . .	88
Figure 5.4	Bouncing digits generated by drawing samples from different mixture components . . . . .	98

Figure 5.5	Validation set classification accuracy and conditional entropy on Moving MNIST . . . . .	99
Figure 5.6	Complete robot and network architecture for the robotic wheelchair experiment . . . . .	102
Figure 5.7	Plot of prior samples and predicted laser state reconstructions . . . . .	104
Figure 5.8	Assignment distribution of $y$ for $K=13$ with post-processed labels of intent . . . . .	106
Figure B.1	System diagram of eye-gaze controlled wheelchair . . .	144
Figure C.1	Augmented Reality view of a dual-arm collaborative robot . . . . .	147

## LIST OF TABLES

---

Table 4.1	Summary of the modified Wheelchair Skills Test assessment points . . . . .	60
Table 4.2	Summary of user ratings for initial system's visualisations . . . . .	63
Table 4.3	Summary of user ratings for final system's visualisations . . . . .	76
Table 4.4	Summary of general perceptions on visualisations . . .	77
Table 5.1	Performance on Moving MNIST test set . . . . .	97
Table 5.2	Test set metrics to determine number of clusters . . . .	103
Table 5.3	Performance on Wheelchair test set . . . . .	105

## LIST OF ALGORITHMS

---

Algorithm 3.1	Shared Control algorithm . . . . .	43
Algorithm 5.1	Sampling procedure to predict novel states from the inferred cluster . . . . .	92

## ACRONYMS

---

AI	Artificial Intelligence
AR	Augmented Reality
SC	Shared Control
XAI	Explainable Artificial Intelligence
XSC	Explainable Shared Control
HRI	Human-Robot Interaction
VAE	Variational Autoencoder
HMM	Hidden Markov Model
SSM	State Space Model
CNN	Convolutional Neural Network
MLP	Multilayer Perceptron
RNN	Recurrent Neural Network
GMM	Gaussian Mixture Model
ELBO	Evidence Lower Bound
LSTM	Long Short-Term Memory
VRNN	Variational Recurrent Neural Network
SLAM	Simultaneous Localization and Mapping
GMVAE	Gaussian Mixture Variational Autoencoder
DiSCVAE	Disentangled Sequence Clustering Variational Autoencoder
HMD	Head-Mounted Display
ROS	Robot Operating System
FoV	Field of View
IMU	Inertial Measurement Unit
LiDAR	Light Detection And Ranging

## INTRODUCTION

---

Assistive robots are one of the most promising avenues for enhancing the quality of life in people living with disability. Whether these robots occupy the form of a powered wheelchair, a feeding and drinking aid, or even a social companion, they all strive to somehow augment the independence of disabled people (Brose et al., 2010). Not only can such technologies accelerate improvements in quality of life, but they may also reduce the burden on available healthcare resources (Agree, 2014). Yet in spite of the potential rooted in assistive robots, their migration from controlled lab environments to widespread commercial use remains an open problem.

A predominant reason for the hindered pervasiveness of assistive robots lies in the complexity of providing assistance in general, be it through humans or robots. Assisting someone properly is a demanding task that must account for a multitude of factors, such as the frequently varying environment, the developmental changes experienced by the person being helped, as well as their personal initiative and authority (Demiris, 2009). Designing a system to accommodate these diverse conditions is a non-trivial task.

In the broader domain of Human-Robot Interaction (HRI), an auspicious trend for collaboration with robots is Shared Control (SC), which is the subject matter of this thesis. There are many definitions of SC in the HRI literature, but we adopt a well-accepted view of this construct as any task where a human and robot continuously exert control over a system to accomplish a common goal (Abbink et al., 2018). Relating back to the task of assistance, SC presents a propitious means of supporting disabled individuals in the operation of assistive robots.

However, the quality of assistance delivered through SC heavily depends on whether the person being assisted can form accurate mental models of the robot behaviour (Abbink et al., 2018; Goodrich and Olsen, 2003). Whenever robot actions do not align with human expectations, there is risk of causing obstruction and frustration rather than offering support (Nisbet, 2002). In the worst-case scenario, a person may even reject all robotic assistance due to the misalignment between their mental models of intended system behaviour and the robot's internal models. A method of reconciling mismatched agent models is thus required.

Within the field of Explainable Artificial Intelligence (XAI) and specifically planning, the problem of “model reconciliation” is regarded as a process of *explanation* (Chakraborti et al., 2017; Fox et al., 2017). Explanations here refer to model updates that help resolve any differences between a human’s expectations of robot plans and their actual representations. A popular way of producing these explanations is to visually expose latent robot representations to human observers via a mode of feedback, for which Augmented Reality (AR) is an increasingly prevalent choice of medium (Chakraborti, Sreedharan, Kulkarni and Kambhampati, 2018).

Motivated by the notion of using explanations to resolve model mismatch in SC, this thesis introduces the Explainable Shared Control (XSC) paradigm. The key objective of XSC is to establish *transparency* in the HRI, such that both the human and robot can interpret each other’s internal models in order to rectify any misalignment. In the context of this thesis, internal models responsible for the high-level planning and generation of goal-driven human behaviour are deemed “intentions” (Bratman, 1990; Tomasello et al., 2005). By assuming that these models are akin to the “intentions” of an agent, we regard a transparent interaction as one where there exists a communication channel of intent (Lyons, 2013; Lyons and Havig, 2014). For the *human* endpoint of this channel, the rationale behind robot actions must be visualised for explanation (Chakraborti, Fadnis, Talamadupula, Dholakia, Srivastava, Kephart and Bellamy, 2018). Whilst at the *robot* endpoint, intent must be implicitly inferred from sensory observations of overt human behaviour (Demiris, 2007; Goodrich and Olsen, 2003).

Our primary focus in this thesis is to satisfy these transparency requirements and thereby fulfil XSC. Addressing the human viewpoint, we present an AR interface that reveals the inner workings of an SC implementation for assistive robot navigation. From the robot angle, we derive an unsupervised clustering algorithm that can perform human intention inference without relying on any explicit indicators or labels of intent (e.g. predefined goal poses on a map). Both solutions for transparency are tested on a robotic – or “smart” – wheelchair, i.e. a powered wheelchair that has been extended to include a collection of sensors and an on-board computer (Simpson et al., 2004). This assistive mobile robot is the target application considered throughout the thesis.

## 1.1 RESEARCH QUESTIONS

Four research questions arise from this thesis on Explainable Shared Control:

1. What constitutes as an effective Shared Control methodology for assistive robot navigation?
2. How can an Augmented Reality Head-Mounted Display be integrated with a Shared Control system to expose its inner workings?
3. Can human users of Shared Control have their mental model accuracy improved by an Augmented Reality interface that visually explains the robot's internal mechanisms?
4. How can an interpretable model be developed for robots to infer human intentions without making any assumptions about specific task constraints?

## 1.2 THESIS CONTRIBUTIONS

The main contributions of this thesis to the field of robotics are:

- A Shared Control implementation tailored to help robotic wheelchair users safely navigate indoor environments. The Shared Control employs two processes: intention estimation and assistive control. User joystick commands are first analysed to yield an estimate of intent that is translated into a navigation goal for the mobile base. This goal then informs an obstacle avoidance routine on when and how to output assistive commands. We also supply the corresponding C++ implementation as an open-source Robot Operating System package.
- The first instance of an Augmented Reality Head-Mounted Display being integrated onto a smart wheelchair with built-in Shared Control. By rendering the internal state of the Shared Control onto the operator's view of the world via a head-mounted Augmented Reality interface, we aim to aid mental models in growing accustomed to the administered assistance.
- A list of objectives and guidelines that delimit the Explainable Shared Control paradigm, coupled with an Augmented Reality-based instantiation for robotic wheelchairs. Explainable Shared Control is conceptualised from the perspective of developing internal models (e.g. intention estimation) and designing Augmented Reality interfaces.

- Two user studies with a setup involving a smart wheelchair and a Head-Mounted Display, namely the Microsoft HoloLens<sup>1</sup>. The first evaluates the acceptance rate and learning curve of Shared Control for assistive navigation when complemented with an Augmented Reality immersive training regime. Results from this initial study highlight that graphical cues must be carefully designed to augment information acquisition and not induce distractions from the task-at-hand. The follow-up study then discerns the value of Explainable Shared Control at settling model misalignment during an indoor navigation trial. Users displayed faster traversal times for challenging events linked with poorly aligned mental models, and unlike before, at no expense of distracting or harming task performance.
- A deep generative model for unsupervised clustering on sequential data, termed the Disentangled Sequence Clustering Variational Autoencoder (DiSCVAE), which is utilised to make inferences about human intent. This generative model falls under the Variational Autoencoder (VAE) framework (Kingma and Welling, 2013; Rezende et al., 2014), allowing for efficient and scalable learning. Unlike previous VAEs in sequence modelling, the DiSCVAE simultaneously clusters *and* disentangles latent representations of sequential observations, granting explainable insight on its generative properties. Each cluster formed is indexed by a categorical variable and its mode can be used to infer discrete high-level features, e. g. “intentions”.
- Experimental analysis of the DiSCVAE performing two tasks: unsupervised classification and human intention inference. To validate the model’s capacity to discover classes from unlabelled sequences, we report results on Moving MNIST (Srivastava et al., 2015), a synthetic dataset for video representation learning. As for inferring intent from observed human behaviour, we provide results of the model applied to a Human-Robot Interaction dataset collected with a robotic wheelchair. Findings from this experiment demonstrate how clusters that signify intent can be learnt without relying on any explicit supervision.

Appendix D summarises the relevant publications derived from this thesis.

---

<sup>1</sup> <https://www.microsoft.com/en-us/hololens>

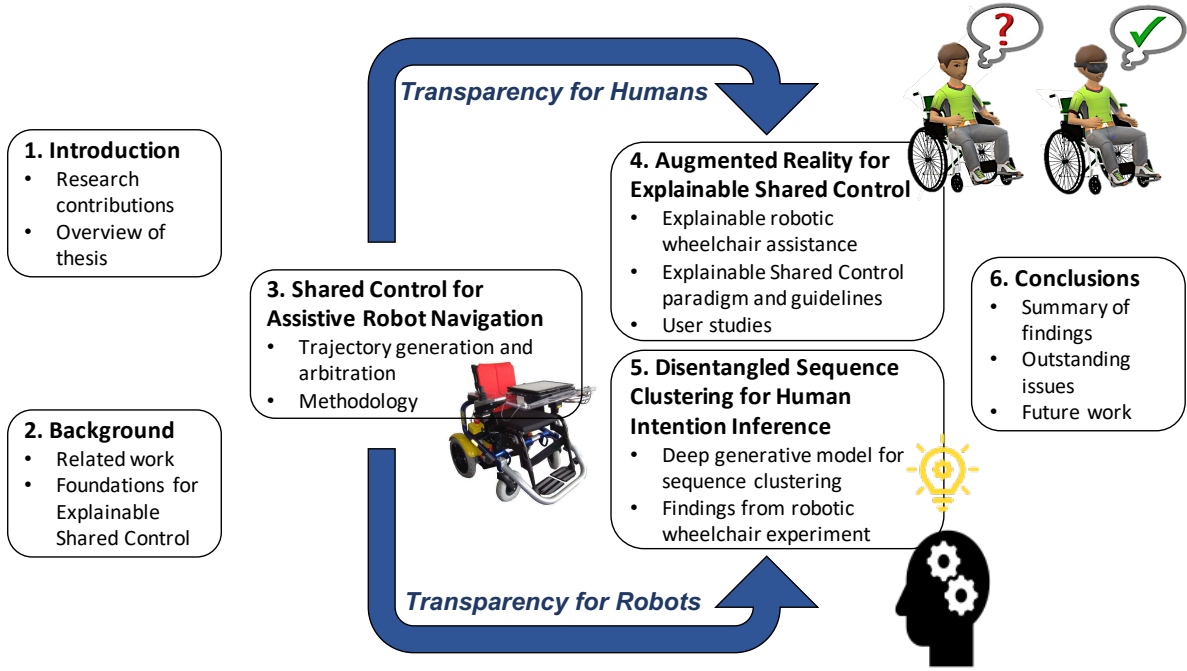


Figure 1.1: Thesis overview. Each block references a chapter, with the blue arrows distinguishing between the two perspectives of transparency in *XSC*.

### 1.3 THESIS OVERVIEW AND ORGANISATION

This thesis is comprised of six chapters in total (see Figure 1.1 for a visual overview), with supplementary material spanning across four appendices:

- **Chapter 2** reviews the foundations that underpin our approach to *XSC* in assistive robotics. In particular, prior works engaging in *SC* and *AR* are discussed in isolation before being framed in the context of model reconciliation. Various deep generative models and techniques for human intention estimation are then examined. Finally, connections are drawn between all the above.
- **Chapter 3** describes our *SC* methodology and architecture for assistive robot navigation in a smart wheelchair setting. There are two principal functions of the navigational assistance: trajectory generation and arbitration. Both functions link to the processes of intention estimation and assistive control, respectively, and are delineated as components of our proposed *SC* methodology.
- **Chapter 4** details a novel head-mounted *AR* system for robotic wheelchairs and introduces *XSC*, alongside a clarification of its terminology

and guidelines. Two [HRI](#) experiments in which subjects operated the [AR](#) Head-Mounted Display ([HMD](#)) system are evaluated to identify its effectiveness at improving mental model accuracy.

- **Chapter 5** proposes an unsupervised clustering approach for human intention inference using the [DiSCVAE](#). Assigning individual observations to the most probable component of this mixture model enables classes or intentions to be inferred from the formed clusters. To illustrate this capability, experimental results on a synthetic video dataset and an [HRI](#) dataset involving a smart wheelchair are additionally provided.
- **Chapter 6** concludes with a discussion summarising the findings of this thesis and its impact on assistive robotics, as well as any outstanding issues and potential research directions for future work.
- **Appendix A** outlines the software packages and open-source contributions that stem from this thesis.
- **Appendix B** investigates an eye-gaze controlled wheelchair as a non-invasive hands-free solution for people who do not possess the cognitive or motor capacity to steadily navigate an environment with a standard joystick device.
- **Appendix C** exemplifies an application of [AR](#) to project the intentions of a dual-arm collaborative robot, as opposed to explaining [SC](#).
- **Appendix D** is an account of all the peer-review publications originating from this thesis.

This chapter introduced the research objective of [XSC](#) to establish *transparency*, such that both the assistive robot and human understand each other's underlying "intent". The next chapter covers the related work that reinforces our endeavour for transparent [HRI](#).

## BACKGROUND

---

This chapter reviews the relevant foundations of Explainable Shared Control (XSC) for assistive robotics. In Section 2.1, Shared Control (SC) is introduced and situated within assistive robotics. Section 2.2 then explores Augmented Reality (AR) as a communication medium for Human-Robot Interaction (HRI) and Section 2.3 follows with an interrogation of the model reconciliation problem. Pertinent computational approaches to intention estimation are investigated in Section 2.4, alongside the interdisciplinary theories surrounding human intent. Section 2.5 presents state-of-the-art deep generative models and exemplifies their utility in robotics. Finally, Section 2.6 closes this chapter with remarks about how the reviewed topics form the basis of XSC in facilitating transparent and effective HRI.

### 2.1 SHARED CONTROL

The SC paradigm is widely regarded as any task in which a human operator and robot collaborate towards a common goal by continuously exerting control over a system (Abbink et al., 2018). The *continuous* element is vital as it compels participants of the SC to actively engage in the task-at-hand. We adopt this stance on SC and thus exclude from consideration a vast range of paradigms that rely on mode-switching mechanisms to slide between distinct levels of autonomy (Desai and Yanco, 2005; Dias et al., 2008).

We first present SC for its pivotal role in user-technology integration with assistive robots (e. g. robot-assisted mobility Cowan et al., 2012). Typical SC frameworks that comply with the aforementioned definition are then examined, which are categorised as either *goal-oriented* or *learning-based*. Frameworks that assume a known goal space exists for the specific task are referred to as goal-oriented, whilst learning-based SC mitigates the prerequisite for a goal representation by directly learning an assistive policy from human input data, e. g. their actions. Lastly, the remainder of this section is devoted to methods of evaluating effective SC.

### 2.1.1 Conditionally Assistive Robots

Assistive robotics is a subset of HRI that aims to bolster the autonomy of people living with disability. In spite of this promising outlook, designing robots to administer proper assistance is riddled with challenges. For instance, the unique requirements of different individuals and their disabilities must be taken into account, as well as adapted whenever there are variations in their capabilities (Brose et al., 2010; Cowan et al., 2012). Deriving policies for assistance is also complicated by the social and physical factors of continuously evolving environments (Agree, 2014; Cowan et al., 2012). To address these challenges and encourage improvements in a user’s developmental trajectory, the robot must facilitate *conditional assistance* (Demiris, 2009). In other words, the robot should balance out a user’s proficiency at independently completing the task with their need for support.

One viable route for building “conditionally assistive” robots is to incorporate SC (Demiris, 2009). A robot employing SC will only correct – not override – noisy or unsafe inputs, hence respecting user desires for independence (Nisbet, 2002). Moreover, robots that supply assistance on an as-needed basis reinforce the user’s confidence and personal growth, which is crucial for preventing debilitating effects on development, e.g. “learned helplessness” (Abramson et al., 1978; Seligman, 1972). Assistive robots undertaking SC are also adept at handling complex environments where human intervention may be necessary (Abbink et al., 2018).

SC holds great potential for engineering a variety of assistive robots, but the focus of this thesis will be on its application to powered mobility. Examples of mobile robotic aids include “smart” wheelchairs (Simpson et al., 2004), wearable exoskeletons and artificial limbs (Cowan et al., 2012). In the following, we motivate our SC methodology for robotic wheelchairs (see Chapter 3) by elaborating on how SC is generally realised across different robot architectures.

### 2.1.2 Goal-Oriented Frameworks

Goal-oriented SC frameworks consist of two core processes: intention estimation and arbitration (Losey et al., 2018). A typical interaction cycle of these processes will first involve the robot recognising user intentions from a pre-defined set of task-specific goals in order to select appropriate control commands. An arbitration phase then considers both the user’s and robot’s

individual control inputs before finalising the output commands that best align with the overall SC objective, e. g. to conditionally assist.

A popular scheme of SC is to perform intention estimation and arbitration sequentially in isolation, which is termed “predict-then-blend” (Javdani et al., 2015). In this scheme, the robot predicts or estimates a user’s goal with respect to the task and chooses an assistive action to help achieve this goal. A blend of the user and robot control inputs is subsequently composed to regulate assistance in the SC, i. e. “policy-blending” (Dragan and Srinivasa, 2013). There is also usually a measure of confidence associated with predicted goals to inform the weighting in blending robot-user control (Carlson and Demiris, 2012; Dragan and Srinivasa, 2013; Huang and Mutlu, 2016). However, “predict-then-blend” methods rely on single point estimates of goals and suffer when the corresponding confidences are low, as the robot is left to either assist incorrectly or not assist at all.

To offer efficient assistance irrespective of the ambiguity surrounding human intent, many SC frameworks instead apply Bayesian reasoning to compute a belief over all task goals (Javdani et al., 2015). By maintaining this probabilistic belief, the robot can reason over an entire goal distribution to select actions that sustain assistance even when confidence levels are low (Javdani et al., 2015; Pellegrinelli et al., 2016). Furthermore, the uncertainty obtained from a predictive distribution can be exploited to determine how robot-user control is arbitrated (Jain and Argall, 2018, 2019). Although more robust than “predict-then-blend”, these probabilistic frameworks still assume that a discrete set of possible goals is known beforehand.

### 2.1.3 Learning-Based Frameworks

On the contrary, modern learning-based SC frameworks bypass the constraint of explicitly representing user goals and directly derive policies of assistance from sensory observations of human input. Rather than performing intention inference, these frameworks deploy learning algorithms to reveal end-to-end mappings between control inputs and assistive behaviour.

Learning by demonstration is a notable SC framework that learns an assistive policy from numerous rehearsals of “expert” guidance (Soh and Demiris, 2013, 2015). Soh and Demiris (2015) applied this framework in a smart wheelchair setting using a mixture model of sparse online Gaussian Processes (GPs) to learn “how” and “when” to help users drive. The model learnt this policy from observations of a supervisor administering assistance through a haptic device. Despite reported successes at reproducing expert-level aid,

there are a few drawbacks to this approach. First and foremost, the supervisor’s assistance may not always maintain consistency or necessarily agree with the primary user (Kucukyilmaz and Demiris, 2018). Another hindrance is that an external perspective on the task is a transformed frame of reference that could result in misguided control (Schettino and Demiris, 2019). The heavy reliance on an expert being available may also pose difficulties in data acquisition.

More recent machine learning frameworks for SC overcome these drawbacks by recovering assistive policies solely from observed user behaviour. For example, Reddy et al. (2018) demonstrated how human-in-the-loop reinforcement learning could enable versatile SC by decoding intended robot actions from user inputs without the assumption of an existing goal. Encapsulated in the learning algorithm was a decision-making component that determined how to provide assistance when conditioned on this implicit decoding of intent. Losey et al. (2019) instead extracted high-dimensional robot actions from a low-dimensional, human-controllable latent space via representation learning. These “latent actions” were beneficial in easing the SC of assistive robots. Within such frameworks, model training can also take place in simulation to alleviate the issue of data availability.

Executing SC in a learning-based manner is advantageous for real-world tasks without any prior knowledge about the human goals or policies for attaining them. Nevertheless, the low-level robot actions generated using the above frameworks do not corroborate a user’s high-level intent, e. g. their *plan* of action for achieving some goal (Pacherie, 2008; Tomasello et al., 2005). As a result, we probe the matter of human intention understanding further in Section 2.4.

#### 2.1.4 *Methods of Evaluation*

Another aspect of SC not yet discussed is how to determine its effectiveness at providing assistance. There is no clear consensus on how to evaluate SC, but a common standard is to identify its benefits over manual control (Abbink et al., 2018). Though this approach may prove enlightening when using traditional engineering metrics, like time-to-completion or fluency of control (Erdogan and Argall, 2017), it is unlikely to account for human factors, such as user preference (Ezeh et al., 2017) or cognitive workload (Carlson and Demiris, 2012; Ghorbel et al., 2018; Viswanathan et al., 2017). Moreover, Abbink et al. (2018) highlight that assessment of SC should extend beyond

the boundaries of the robot capabilities and task constraints, meaning the more realistic the experimental conditions, the better.

Given the thesis scope of powered mobility, we will now outline a corresponding means of evaluating SC on smart wheelchairs that abides by the above remarks. Indeed, metrics like task completion time, number of collisions and joystick command fluency are informative to smart wheelchair designers on the efficiency of SC versus manual control (Carlson and Demiris, 2012; Erdogan and Argall, 2017). However, human factors must also be evaluated using objective and subjective measures. For instance, cognitive workload is an important criterion that determines a user's attitude and likely acceptance of the resulting wheelchair assistance (Ezeh et al., 2017; Ghorbel et al., 2018; Viswanathan et al., 2017). Subjective questionnaires, such as the NASA-TLX (Hart and Staveland, 1988), are practical indicators of this metric, but there are also objective measures from human physiological data (e.g. eye gaze and head dynamics) that bear close ties with heightened workload (Doshi and Trivedi, 2012; Reimer and Mehler, 2011; Solovey et al., 2014). Therefore, the SC evaluation should encompass a range of subjective and objective metrics that gauge the positive impact on *both* the resulting control policy and human engagement.

Testing SC on smart wheelchairs that intend to push the boundaries of constrained lab conditions (Abbink et al., 2018) can also be achieved by referring to the Wheelchair Skills Test (WST) manual (Kirby et al., 2002). The WST is a scored set of skills that establish whether a person qualifies as sufficiently capable at operating a wheelchair during everyday activities (Kirby et al., 2002). Some example skills include applying brakes, traversing doorways, ascending or descending an incline, and performing 3-point turns. Designing user studies according to the WST thereby fulfils the “full spectrum of realistic situations and conditions” mandated by Abbink et al. (2018) on how to evaluate SC. The WST has also enjoyed prior success in the literature on SC for smart wheelchair studies (Ghorbel et al., 2018; Pineau et al., 2011).

Having presented a brief outlook on how to evaluate SC, specifically in assistive robot navigation, the only aspect left to address is communication (Losey et al., 2018). Information feedback in SC is especially relevant as complex “black-box” methods begin to emerge, such as those detailed in Reddy et al. (2018) and Losey et al. (2019). Accordingly, the next section investigates how AR can create a bridge of communication in HRI.

## 2.2 AUGMENTED REALITY IN HUMAN-ROBOT INTERACTION

Immersive technologies involving [AR](#) are a prominent way of expressing information to users about the inner workings of an intelligent agent. These technologies are designed to perform a range of capabilities suited for establishing an embodied interface and augmenting a user's natural perception of the Artificial Intelligence ([AI](#)). In this section, we concentrate on the physical case of robots and how [AR](#) has previously been utilised as a mode of information feedback in [HRI](#).

Dating back to the earliest applications of [AR](#) in robotics, there have been reports on its efficacy at guiding the control of human operators ([Azuma, 1997; Milgram et al., 1995, 1993](#)). In many collaborative situations that require the human to teleoperate the robot's end-effectors, [AR](#) serves as a communication channel between the two agents ([Azuma et al., 2001](#)). By overlaying the robot's perspective onto the operator's view and displaying the predicted effects of interacting with the surroundings, the user is capable of executing accurate remote control ([Azuma et al., 2001; Milgram et al., 1995](#)). With the pervasiveness of mobile [AR](#) systems prompted by improvements in wireless networking, these visualisation techniques have even migrated into navigation settings ([Carmigniani et al., 2011; Chatzopoulos et al., 2017](#)).

Meanwhile in [SC](#), a key expectation of the robot is to regularly exchange information with its human partner ([Losey et al., 2018](#)). In many instances of [SC](#), the haptic channel is the selected modality of sensory feedback during this exchange ([Kucukyilmaz and Demiris, 2018; Losey et al., 2018](#)). However, force feedback only provides a limited user embodied experience and so [AR](#) possibly poses as a more transparent mode of relaying back information to interacting human partners. Head-Mounted Displays ([HMDs](#)) are particularly useful for this purpose due to their increased sense of presence and engagement over monitors or projectors ([Alshaer et al., 2017; Buttussi and Chittaro, 2018; Sibirtseva et al., 2018](#)). These headsets are not without flaws, but in the scope of [SC](#) where users must maintain attention and *actively* participate in the task-at-hand, they have been highlighted as superior to other feedback modalities ([Sibirtseva et al., 2018](#)).

Regardless of the positive prospects rooted in [AR](#) as a medium of exchange, explicating the hidden rationale of a robot is a non-trivial task. In [SC](#), the process of inferring robot intent is challenging for humans and often leads to a misalignment between a person's mental models of the expected behaviour and the robot's internal models ([Jain and Argall, 2019; Javdani](#)

et al., 2015). This phenomenon, known as model misalignment, echoes the need for *transparency* in the HRI (Lyons, 2013; Lyons and Havig, 2014).

### 2.3 MODEL RECONCILIATION

Model misalignment or “model reconciliation” (Chakraborti et al., 2017) is a well-known problem in Explainable Artificial Intelligence (XAI) and explainable planning (Fox et al., 2017). In this subject area, model reconciliation refers to the use of “explanations” or model updates (Chakraborti et al., 2017) to resolve differences in a human’s expectations of an artificially intelligent agent’s plan. Ergo, the aim behind generating explanations is to modify the human’s model of the world to agree with the agent’s model.

There is no absolute way of evaluating how explanations effectively foster this ‘agreement’, albeit many have investigated the matter. Perhaps the best point of reference is Gilpin et al. (2018), who claim that an explanation can be evaluated either in terms of *interpretability* or *completeness*. Given an explanation, the former reflects how well it helps humans understand the internals of an AI, whilst the latter relates to how accurately it describes the AI operation, such that humans can anticipate its behaviour (Clinciu and Hastie, 2019; Gilpin et al., 2018). Diagnosing these two measures can be performed subjectively, e.g. via post-instance questionnaires (Theodorou, 2019), or objectively through quantitative metrics, like monitored attention and cognitive load (Carlson and Demiris, 2009; Goodrich and Olsen, 2003), or predictive and descriptive accuracy (Murdoch et al., 2019). Bootstrapped with this protocol for evaluation, the next phase is to develop a means of reconciling mismatched models.

Intelligent systems within explainable planning tackle model reconciliation by adopting three qualities: trust, interaction and transparency (Fox et al., 2017). Trust refers to the user’s confidence in the capabilities and reliability of an AI, interaction relates to the user’s ability to query the AI, and transparency concerns the user’s clarity on the AI status (e.g. its goals and functionality Theodorou, 2019). These qualities are instrumental in the circulation of intelligent systems and fall under the wider umbrella of standards that strive towards ethical AI (Bryson and Winfield, 2017), e.g. on the transparency of autonomous systems<sup>1</sup>. Offering explanations that encompass these qualities can be cast as a problem of *visualisation*, where the “brain” of the AI should be externalised in order to align mismatched mod-

<sup>1</sup> <https://standards.ieee.org/project/7001.html>

els (Chakraborti, Fadnis, Talamadupula, Dholakia, Srivastava, Kephart and Bellamy, 2018; Wortham et al., 2017).

All three ingredients for model reconciliation and the notion of using visualisations to enable them are also heavily rooted in HRI. Although interaction is a natural precondition of this field, trust and transparency are also essential constructs for effective HRI that arise less naturally (Lyons, 2013). Inspired by the importance of trust and transparency in human-human relationships, the literature is rich in ways of manifesting these constructs for HRI (Hancock et al., 2011; Lyons, 2013; Lyons and Havig, 2014; Soh et al., 2019; Wortham et al., 2017) and SC (Alonso and de la Puente, 2018). Similar to XAI, a fitting means of injecting trust and transparency into the HRI is via an *interface*, specifically one that *shares intent* between robot and human partners (Lyons, 2013; Lyons and Havig, 2014). AR HMDs are interfaces that have had notable success at disambiguating intentions in HRI by using visualisations to externalise the robot’s mind (Chakraborti, Sreedharan, Kulkarni and Kambhampati, 2018; Sibirtseva et al., 2018; Walker et al., 2018).

Employing AR HMD interfaces that embody the preceding traits and visually explain the rationale behind any robotic assistance can thus hope to mitigate model misalignment in SC, provided that care is taken in the presentation of information. Otherwise, ineffective visualisations may exacerbate the dilemma, e. g. those that fail to exploit the surrounding space with graphical cues (Kim et al., 2018) or do not account for the influence of depth perception (Diaz et al., 2017). For appearance-constrained robots that lack anthropomorphic features, graphical objects used to signal intent warrant even greater scrutiny (Lyons and Havig, 2014; Walker et al., 2018). Additionally, the small Field of View (FoV) of most headsets bears the risk of misleading or distracting users (Sibirtseva et al., 2018). If the interface adheres to careful design considerations to avoid such issues, then AR HMDs propose one of the most auspicious trends for model reconciliation.

So far, we have looked at how AR interfaces share robot intent with humans for “robot-to-human” transparency (Lyons, 2013; Lyons and Havig, 2014), without considering the robot’s awareness of human states. Yet a completely explainable SC system must also establish “robot-of-human” transparency (Alonso and de la Puente, 2018; Lyons, 2013; Lyons and Havig, 2014), such that robots can interpret human states. The next section addresses this concept by examining how to equip robots with the ability to understand human intent.

## 2.4 HUMAN INTENTION ESTIMATION

Human intention estimation is a multidisciplinary subject that has been widely explored in HRI as a mechanism for enhancing collaboration between participating agents (Demiris, 2007; Jain and Argall, 2019; Losey et al., 2018). Whilst humans are naturally gifted with this ability early on in life (Tomasello et al., 2005), the computational process of deriving intent purely based on observable behaviour is a formidable task. We motivate this task by first outlining how humans understand intent in others. Prior computational approaches to intention estimation are then reviewed.

### 2.4.1 Understanding Human Intent

Humans formulate intent as a reaction to both implicit and explicit stimuli. Explicit signals from the physical surroundings of an individual could be responsible for eliciting external responses in their goal-directed behaviour, or an internal stimulus could instead be evoked by latent cognitive functions, such as the desires and beliefs of the person (Bratman, 1990; Cohen and Levesque, 1990). Detecting the impact of these dormant thought processes on the goals and intentions of other individuals is a far more challenging endeavour than if these processes were observable. Yet humans are remarkably proficient at the intention inference problem (Blakemore and Decety, 2001), with evidence of this ability emerging around the first year of life and gradually developing thereon until full-fledged competence is attained at just two years of age (Tomasello et al., 2005).

“Theory of mind” is a topic that lies at the boundaries of philosophy, neuroscience and cognitive psychology, in which the process of humans understanding one another’s minds is explored. Embedded in this topic is *simulation theory*, which offers a prominent explanation on how the neural and computational operations in the brain link overt movement with goal inference (Gallese and Goldman, 1998; Gallese et al., 2004; Hesslow, 2002). The operations responsible for modelling this sensorimotor loop in the brain are termed *internal models* (Wolpert et al., 2003, 1998).

Simulation theory of action is a broadly accepted justification for how humans ascribe intention (Blakemore and Decety, 2001; Grafton, 2009; Jeannerod, 2001). The simulationist perspective argues that people understand the mental states of others by performing the dual role of action observation followed by the simulation of action consequences (Blakemore and Decety, 2001; Hesslow, 2002; Jeannerod, 2001). In other words, by using our own

sensory-motor repertoire to internally generate goal-directed behaviour, we are able to infer action-intention mappings.

The discovery of “mirror” neurons in monkeys (Rizzolatti et al., 1996) and humans (Grèzes et al., 2003) also supplies neurophysiological evidence in favour of the biological motor system informing our perception of goal-directed behaviour in other beings. These neurons, which are excited by the recognition or execution of goal-directed movements, spurred on notions about how humans label action consequences according to their own mental simulations of such actions.

Another core element of a human being’s innate ability to identify high-level intentions from their perceptions of other people’s actions is *hierarchy*. Many studies from motor cognition suggest that different hierarchical intentional levels take place prior to any overt movement (Hamilton and Grafton, 2007; Pacherie, 2008; Wolpert et al., 2003). In simulation theory, efficient computational hierarchies have also been identified during the process of internally generating actions (Grafton, 2009).

Overall, simulation theory and the mirror neuron system have sufficient evidence supporting their close ties with estimating intent, making them a worthwhile source of inspiration for computational modelling.

#### 2.4.2 What is Intent?

Before proceeding any further, the terminology for “intent” used throughout this thesis should be resolved. With respect to intention and how it is distinguished from action goals, low-level *continuous* states of desire will generally be referred to as *goals* or *trajectories* (e.g. a vector of poses or control commands), whilst high-level *discrete* counterparts are *intentions* (Bratman, 1990; Cohen and Levesque, 1990; Tomasello et al., 2005). Goal or intention estimation will also adopt this distinction in semantic language. Moreover, *estimation* and *prediction* are considered as techniques of recognising either current or future intentions, respectively.

For additional clarity, an intention is defined as containing both a *goal* and a *plan* of action for accomplishing it (Bratman, 1990; Tomasello et al., 2005). A plan is a practical means of fulfilling intended human behaviour in a future-directed context, e.g. a trajectory of control commands to manoeuvre a robot arm. Goals are the desired outcomes of pursuing this plan, which also occupy the *continuous* space of values in the  $\mathbb{R}$  domain, e.g. the target pose of the robot arm. Ultimately, an intention is viewed as an abstract composition of these two elements that can be identified by its index or la-

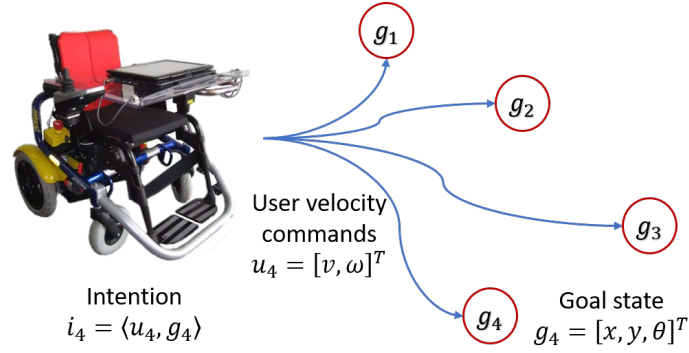


Figure 2.1: Diagram of how intentions are represented. An intention  $i_x$  is a tuple of user velocity commands  $u_x$  (the action plan) and their corresponding target state  $g_x$  (the goal). In this illustration, there are four available intentions based on the repertoire of available plans (indexed by  $x$ ).

bel across potential plans (held in the action repertoire [Wolpert et al., 2003](#)), hence its *discreteness* ([Pacherie, 2008](#)).

To provide a concrete instance of this terminology, we will now revert to our robotic wheelchair scenario. A multitude of works have explored intention estimation in robotic wheelchairs, with most adopting similar representations of user intent ([Carlson and Demiris, 2012](#); [Matsubara et al., 2015](#); [Narayanan et al., 2016](#); [Poon et al., 2017](#)). Intention is often defined in these works as a target wheelchair state  $g_x$  (the goal) or a set of states. In contrast, we explicitly frame intentions as a unification of these goals and the means for achieving them, i. e.  $i_x = \langle u_x, g_x \rangle$ , where  $u_x$  are the user’s intended velocity commands (the plan). An intention  $i_x$  is thus completely captured by its tuple constituents, as well as its index  $x$  amongst the collection of competing action plans (see Figure 2.1 for the visual depiction).

### 2.4.3 Computational Approaches

Intention estimation essentially revolves around an observing agent deriving a model to match an acting agent’s behaviour ([Demiris, 2007](#)). This model matching procedure can be enacted in either a *discriminative* or *generative* manner. The following describes previous computational approaches within these two classes, as well as biologically-inspired architectures linked to simulation theory.

#### 2.4.3.1 Discriminative Class

In the discriminative class, low-level features of the observed state are deciphered and characterised according to labels of intent that exist *a priori* for

the domain-specific task. Informative features are first obtained from a variety of sensory sources, such as eye gaze (Huang et al., 2015), visual captures of human activity (Koppula and Saxena, 2016), or even a fusion of multiple modalities (Doshi et al., 2011; Trick et al., 2019). These features are then classified using popular algorithms, e.g. support-vector machines (Doshi et al., 2011; Huang et al., 2015; Trick et al., 2019), conditional random fields (Koppula and Saxena, 2016), or neural networks (Nicolis et al., 2018), all of which have also been incorporated into anticipatory control systems for human-robot collaboration (Huang and Mutlu, 2016; Koppula and Saxena, 2016; Nicolis et al., 2018).

Alternatively, discriminative models can take a probabilistic stance by employing Bayesian reasoning to infer a conditional distribution over pre-existing representations of human intent (Jain and Argall, 2018, 2019; Javdani et al., 2015; Pellegrinelli et al., 2016). The resulting posterior distribution can be used to either forecast low-level trajectories of action patterns or classify predefined labels of desire (Losey et al., 2018). Despite the extra flexibility associated with this conditional distribution, discriminative algorithms are not capable of reproducing the perceived state as they do not model a joint distribution. In other words, they directly learn the parameters of the posterior and have no knowledge of the underlying observation space.

On the other hand, generative models represent a joint distribution that can be sampled from to reproduce state observations. As many theories on human intent are ascribed to our ability to internally reproduce or *simulate* the actions we observe (Blakemore and Decety, 2001; Grafton, 2009; Jeannerod, 2001), it is prudent to view a mathematical process with the same generative capacity as a fitting model. This generative capacity is particularly appealing for robots seeking to reproduce *embodiment* (Demiris, 2007), such as to drive a motor body during active event recognition (Ognibene and Demiris, 2013) and imitation (Lee et al., 2013; Wang et al., 2017). The discussion will now stray away from discriminative models and focus on generative alternatives.

#### 2.4.3.2 Generative Class

Generative approaches are a prominent class of probabilistic algorithms that recover a distribution over observable data by introducing latent random variables to capture any hidden underlying structure. Within the confines of human intention inference, the modelled latent space can then be presumed to represent all possible causal relations between intentions and observations of human behaviour (Demiris, 2007; Wang et al., 2013). The inference task

therefore boils down to iteratively learning the parameters of the generative process until the posterior distribution or unknown structure of the observed data is acquired.

Dynamic graphical models of stochastic processes are extensively utilised for intention inference in the generative schema. For instance, Hidden Markov Models (HMMs) and their generalisation, Dynamic Bayesian Networks (DBNs), are especially suitable when interpreting goal-directed actions from sequential data (Pentland and Liu, 1999). In this body of literature, many variants of DBNs have been used to reveal the connections between sensory data and motor behaviour, e.g. Hierarchical HMMs (Blaylock and Allen, 2006; Murphy and Paskin, 2002; Zhu et al., 2008), Growing HMMs (Vasquez et al., 2008, 2009) and Hidden semi-Markov Models (Tanwani and Calinon, 2017). Over the last decade, these parametric dynamics models are less commonly employed due to the hardship in designing them to perform inference over high-dimensional and non-linear human action spaces (Wang et al., 2013).

The advent of scalable learning algorithms has led to generative models that efficiently infer latent variables of “intent” from abundant sources of complex human behavioural data. Some early examples are dynamical GPs that encapsulate temporal dynamics in the latent space (Wang et al., 2008). These GPs can model intention-driven behaviour by encoding human actions via non-linear functional mappings (Matsubara et al., 2015; Wang et al., 2013). In recent years, the significant interest in deep neural networks that parameterise latent variable models has also transferred over to the intention inference domain (Hu et al., 2019, 2018). Nevertheless, the literature is still sparse in such deep generative models, which will be commented on in Section 2.5.

#### 2.4.3.3 Biologically-Inspired Architectures

It is also worthwhile to reflect on generative strategies that fall under the simulationist perspective of cognitive functions (Hesslow, 2002). From this point of view, internal models should be used to simulate motor control in harmony with anticipated intentions. Two prevailing types of internal models are *forward* and *inverse* models, which respectively depict feedback or feed-forward control (Wolpert et al., 1998). Figure 2.2 illustrates the properties of these coupled concepts, with inverse models akin to *controllers* or *behaviours* and forward models akin to *output predictors* (Karniel, 2002). These components are embedded in many control architectures that have been inspired by mirror neurons (Demiris and Khadhour, 2006; Wolpert et al., 2003).

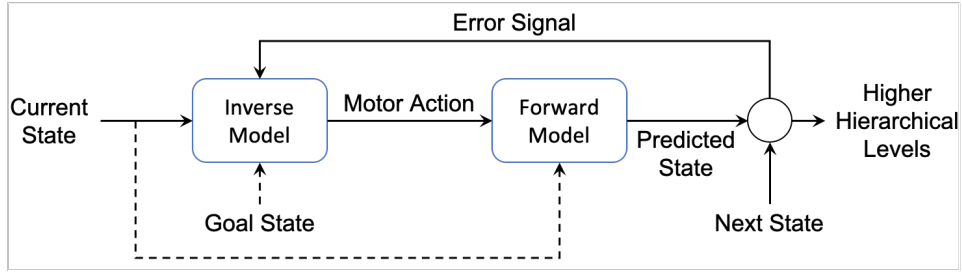


Figure 2.2: Block diagram of a paired forward and inverse model. The inverse model outputs an action given current and target state inputs, whilst the forward model predicts the next sensory state from the corresponding inverse model’s output. Dashed arrows denote optional inputs.

An appealing trait of controller-based architectures is their *hierarchical* layout. For example, Hierarchical Attentive Multiple Models for Execution and Recognition (HAMMER) (Demiris and Khadhour, 2006) is a biologically-inspired action recognition architecture that consists of paired inverse and forward models to initiate motor control. These internal models are the basic building units of HAMMER (shown in Figure 2.2) and are organised in a parallel manner, such that higher-level blocks act as top-down contributions to the lower-level blocks responsible for generating motor actions. This arrangement is markedly attractive for its plausibility in linking the recognition of future-directed intentions with compositions of lower layers (Hamilton and Grafton, 2007; Pacherie, 2008).

However, the “inverse problem” of directly extracting motor control from desired movements is intractable in these architectures for numerous reasons. First and foremost, there may exist non-linear many-to-one mappings of actions to sensations, which can yield non-convex one-to-many inverse images (Jordan and Rumelhart, 1992). In essence, this issue of *multimodality* translates into the fact that many diverse actions often fulfil the same intention. Furthermore, observing agents do not necessarily possess the same kinematic dynamics as those of the demonstrating agents. In this situation, a mapping between motor control policies of the different physical actors is required (Nehaniv and Dautenhahn, 2001).

Recent advances in representation learning present promising computational solutions to bypass the “inverse problem” of earlier controller-based architectures, whilst sustaining the advantages of generative models and hierarchy (Bengio et al., 2013; Bengio and Delalleau, 2011). The following section presents state-of-the-art representation learning algorithms and applications to robotics.

## 2.5 DEEP GENERATIVE MODELS IN ROBOTICS

Deep generative models have achieved significant progress in addressing the fundamental challenges of representing complex latent structures over large and high-dimensional datasets. A seminal generative model is the Variational Autoencoder (VAE), which provides a scalable way of tractably learning latent variables over *multimodal* data spaces through approximate inference (Kingma and Welling, 2013; Rezende et al., 2014). Despite the representation power of VAEs, they are not naturally tuned for probabilistic inference or prediction. Consequently, recurrent frameworks are attractive adaptations of the original VAE, as they tailor to structured output prediction through a conditional generative model (Sohn et al., 2015). The Variational Recurrent Neural Network (VRNN) is a notable example for sequence modelling, which under evaluation demonstrated the expressive power of conditional distributions (Chung et al., 2015).

Many of these conditional VAEs have become increasingly popular in the robotics literature. In trajectory prediction, network architectures with this layout have been used to capture the dynamics of multiple interacting agents, such as basketball players (Ivanovic et al., 2018) or pedestrians perceived during autonomous driving (Lee et al., 2017). Key to both works is that their architectures could produce diverse predictions of multiple possible future trajectories, i.e. “multimodal” outcomes (Ivanovic et al., 2018; Lee et al., 2017). Aside from trajectory prediction, conditional VAEs have also triumphed in numerous other robotics scenarios, including imitation learning (Wang et al., 2017), human motion prediction and synthesis (Bütepage et al., 2017), as well as SC (Losey et al., 2019).

Although powerful, “black-box” algorithms are notoriously difficult to interpret and hence explain to end-users (Fox et al., 2017). As a result, a core research direction for deep generative models is to derive meaning behind the learnt latent space by disentangling its structure, i.e. recovering abstract concepts from independent factors of variation (Bengio et al., 2013). The most prevalent framework for learning disentangled representations is the VAE (Locatello et al., 2019; Tschannen et al., 2018). For instance, VAEs constructed using a hybrid of continuous-discrete variables have shown how particular latent dimensions manipulate meaningful generative properties of the data, e.g. handwriting style (Kingma et al., 2014) or speaker identity (van den Oord et al., 2017). The desirable qualities of disentanglement have spurred on new algorithms in the representation learning community for various purposes, like clustering (Dilokthanakul et al., 2016; Jiang et al.,

2017) and sequence modelling (Hsieh et al., 2018; Hsu et al., 2017, 2018; Yingzhen and Mandt, 2018).

An integral asset of disentangled representations lies in their *interpretability* (Locatello et al., 2019). The capability to interpret the latent space can facilitate abstract reasoning for practical outcomes, such as counting bouncing digits from video sequences (Kosiorsek et al., 2018) or determining a patient’s mortality risk from medical records (Fortuin et al., 2019). One prominent form of disentanglement that enhances the level of interpretability is to incorporate discrete information into the latent variable model (van den Oord et al., 2017). An exemplification of how discrete latent codes can augment a model’s interpretability is through *clustering* (topic of Chapter 5).

Nonetheless, very few deep generative models with disentangled latent variables have transferred over to the robotics domain. Hu et al. (2019) is a unique case, where a conditional VAE was adapted to intention inference by disentangling latent variables in a multi-agent driving scenario. In the same vein, our approach to human intention inference delineated in Chapter 5 incorporates discrete code into the VAE and employs a mixture prior to cluster the disentangled variables, enabling us to diagnose the learnt representation. This clustering framework is part of the venture to bridge the gap between representation learning and robotics.

## 2.6 CLOSING REMARKS

From this literary digest, three requirements have been identified as imperative for XSC. First, the SC must incorporate a suitable medium of information feedback to guide users into building precise mental models of expected robot behaviour. Of the available communication media, AR HMDs grant an auspicious means of demystifying the underlying SC. Next, the interface provided during the SC must adequately combat the issue of model misalignment. Reconciling mismatched mental models is feasible if the AR interface is designed to share robot intent and foster “robot-to-human” transparency (Lyons, 2013; Lyons and Havig, 2014). Finally, to complete the explainable paradigm, “robot-of-human” transparency must also exist (Alonso and de la Puente, 2018; Lyons, 2013; Lyons and Havig, 2014). By equipping robots with a powerful yet interpretable method of inferring human intent during SC, the robot can also better communicate back to users its own intent (Chang et al., 2018).

The remaining chapters of this thesis will divulge how each of these requirements are fulfilled to establish XSC. Chapter 3 develops an SC methodo-

logy for an assistive robot, namely a robotic wheelchair. After hinting at how model misalignment occurs in this standard setup, an [AR HMD](#) interface is then designed in [Chapter 4](#) and integrated onto the robotic wheelchair for “robot-to-human” transparency. User studies are also included to evaluate the impact on model reconciliation. Lastly, [Chapter 5](#) contributes a deep generative model for human intention inference that can be disentangled and thereby explained to users. Each of these components yields our proposed [XSC](#) paradigm.



## SHARED CONTROL FOR ASSISTIVE ROBOT NAVIGATION

---

This chapter addresses our first research question:

“What constitutes as an effective Shared Control methodology for assistive robot navigation?”

The purpose of this chapter is to propose a Shared Control (SC) methodology suited for assistive robot navigation. Section 3.1 first motivates the application setting of “smart” wheelchairs and the capacity for SC to bolster the independent mobility of disabled individuals. In Section 3.2, we introduce the assistive robot architecture used throughout this thesis. Section 3.3 follows with a description of the SC deployed on this target platform. The limitations of our methodology are then highlighted in Section 3.4, particularly from the viewpoint of model misalignment and how a medium of communication is necessary. Lastly, Section 3.5 summarises the chapter and its relevance to material covered later in this thesis.

Some of this chapter’s content has previously been published in [Zolotas and Demiris \(2019\)](#), specifically Sections 3.2 and 3.3.

### 3.1 APPLICATION SETTING: “SMART” WHEELCHAIRS

Independent mobility plays a significant role in our everyday activities and quality of life, irrespective of the age group ([Agree, 2014](#); [Iezzoni et al., 2001](#); [Metz, 2000](#)). Traditional mobility aids that incorporate “smart” characteristics from other technical domains, such as robotics, are an auspicious means of ensuring that disabled individuals can also exercise mobility. A typical example of such a platform for assistive robot navigation are “smart” wheelchairs ([Simpson, 2005](#); [Simpson et al., 2004](#)). These mobile robots are powered wheelchairs that have been extended to include a collection of sensors and an on-board computer, allowing for a more intelligent way of ensuring safety and control ([Leaman and La, 2017](#); [Simpson et al., 2004](#)).

Millions of disabled individuals who otherwise cannot operate a standard powered wheelchair are forecast to benefit from possessing smart wheelchairs ([Simpson, 2008](#)). [Simpson \(2005\)](#) presented a thorough review on the

main features of these wheelchairs, ranging from the hardware sensors for collision avoidance, to the input methods for navigation and the operating modes for path planning. Whilst such features have significantly benefitted many, a more recent review by [Leaman and La \(2017\)](#) noted that human factors are still not adequately taken into account during the development of smart wheelchairs. Some examples of human factors include building trust in the Human-Robot Interaction (HRI) and personalising the device to solely administer the level of assistance required by each user ([Leaman and La, 2017](#)).

Numerous engineering efforts have made progress in upgrading user-technology integration for smart wheelchairs by addressing these human factors. For example, prior research has explored the diversification of user control interfaces, as not all patients possess the cognitive or motor capacity to steadily and consistently navigate an environment using the traditional joystick ([Fehr et al., 2000](#)). Unconventional input methods, such as brain-machine interfaces ([Carlson and Del R. Millan, 2013](#)), head motion ([Li et al., 2016](#)) and even eye gaze ([Ktena et al., 2015](#))<sup>1</sup>, have begun to appear amongst modern smart wheelchairs. Yet regardless of how user-friendly an interface is, these wheelchairs still require sophisticated controllers to modulate the complex behaviour of patients with severe disabilities ([Fehr et al., 2000](#); [Viswanathan et al., 2017](#)). One pertinent solution is to adjust any noisy and unpredictable input signals by engaging in SC and offering *conditional assistance* ([Demiris, 2009](#)).

### 3.2 ASSISTIVE ROBOT ARCHITECTURE

A system diagram of the smart wheelchair used for all experimentation in this thesis is displayed in Figure 3.1. Originally designed by [Sarabia and Demiris \(2013\)](#); [Soh and Demiris \(2012\)](#), we have since regularly committed enhancements to this assistive robot architecture. On the hardware front, we have integrated two new technologies into the architecture: an eye tracker and an Augmented Reality (AR) Head-Mounted Display (HMD) (see Chapter 4 for more information). From the software perspective, two notable contributions are the intention inference algorithm (see Chapter 5) and the SC employed, which is the topic of this chapter. The blue box in Figure 3.1 contains software processes developed in this thesis as part of the SC.

Apart from these contributed changes, the wheelchair continues to be controlled using a joystick with a circuit board that enables an Arduino UNO to

<sup>1</sup> We also contribute an eye-gaze controlled smart wheelchair in Appendix B.

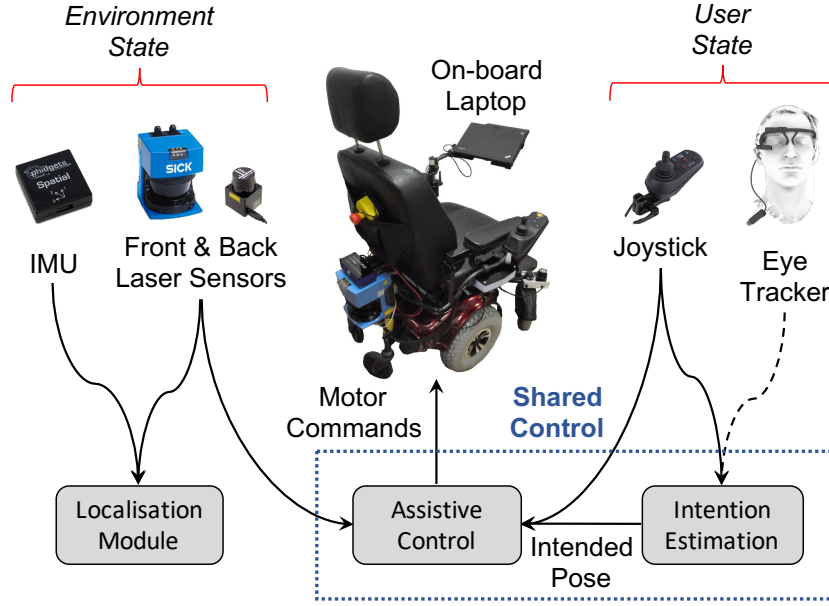


Figure 3.1: System diagram of the target smart wheelchair platform. The laser and IMU sensor data form the environmental state, whilst the joystick and eye tracker readings comprise the user state. Eye gaze data is only used in our work to assess the user’s cognitive state, but the dashed arrow indicates how it could also feed into the intention estimation module. Within the SC, the user’s joystick commands are evaluated to estimate an intended pose and then arbitrated through an assistive control procedure that outputs safe motor commands based on the latest laser sensor readings.

translate user-issued commands into motor signals. The mobile platform has a rectangular shape (1.0m×0.65m) and is equipped with two Hokuyo URG-04LX-UG1 Light Detection And Ranging (LiDAR) scanners at the front, and one SICK LMS200 LiDAR sensor at the back for full 360° Field of View (FoV). A Phidgets spatial 3/3/3 IMU is also equipped to improve the odometry estimate of the mobile setup. All the listed system components are developed atop the Robot Operating System (ROS) middleware running on the on-board laptop (Quigley et al., 2009).

In our proposed architecture, sensory signals perceived by the robot either relate to the user state or the environmental state. A user’s state is composed of the commands they input through the wheelchair’s built-in joystick, as well as their gaze direction determined by a wearable eye tracker (we use an open-source eye tracking platform made by Pupil Labs Kassner et al., 2014). Environmental data is acquired from the three attached laser rangefinders and the IMU, which help construct a map using the localisation module’s

Simultaneous Localization and Mapping (SLAM) component<sup>2</sup>. The localisation module then produces an estimate of the current robot state, i.e. the robot's pose on the constructed map. In the next section, we dive deeper into the SC.

### 3.3 CONTROL-SHARING METHODOLOGY

The SC methodology detailed in this section falls under the *goal-oriented* class of frameworks discussed in Section 2.1.2. Providing conditional assistance via this type of SC involves two core functions: intention estimation and arbitration (Demiris, 2007; Losey et al., 2018). Both these functions follow a user-centred design in that they were informed by the feedback of human subjects in a pilot study (Section 4.2). In other words, this single iteration of feedback helped answer the chapter's research question on what constitutes as effective SC. This section will now illuminate our answer, with Algorithm 3.1 containing the pseudocode for the complete SC implementation.

#### 3.3.1 Intention Estimation

In the scope of powered mobility, user intent is often expressed as a desired sequence of destination poses on a map. These are equivalent to the *goals*  $g$  characterised in Section 2.4.2, which do not necessarily capture the user's actions  $u$ , i.e. their *plans*. Chapter 5 adopts a novel perspective on this representation that completely encompasses the intention tuple  $i$ , but for now we settle on a trajectory of poses as a depiction of user intent. We take a *forward model* approach to generate this trajectory of intended poses by utilising both the robot kinematics and user-selected joystick actions to predict future robot states. The time horizon associated with this prediction is dependent on the simulation period  $T$ , as we examine in the following sections.

##### 3.3.1.1 Forward Models

A forward model can be developed in multiple ways. For instance, a forward model can be considered a “distal teacher” of the inverse model, whereby the error in output predictions allows adjustments to be made to the controllers (Jordan and Rumelhart, 1992; Karniel, 2002). Another portrayal of

<sup>2</sup> Additional details regarding the localisation module and other software nodes are provided in Appendix A.

**Algorithm 3.1:** Shared Control algorithm

---

**Inputs:** Robot state  $r_t$ ; User input command  $u_t$ ; Simulation period  $T$ ;  
**Output:** Assisted output command  $a_t$

*// Intention estimation by trajectory generation (Section 3.3.1)*

```

1  $\tau_1 \leftarrow r_t$ 
2 for  $t \in \{1, \dots, T-1\}$  do
3    $\tau_{t+1} \leftarrow \phi(u_t, [r_t])$  using Equation (3.4)
4 end
5 if  $\text{IsUnsafe}(\tau_{1:T})$  then
6   // Arbitrate (Section 3.3.2.2) if intended trajectory is unsafe, i.e.
7   // if IsUnsafe returns a non-empty vector of collision points  $c_o$ 
8    $s_t \leftarrow \text{ObstacleAvoidance}(\tau_{1:T}, u_t)$  (Mujahed et al., 2018)
9    $a_t \leftarrow \text{PolicyBlend}(s_t, u_t)$  using Equation (3.6)
10 else
11    $a_t \leftarrow u_t$ 
12 end
13 Function  $\text{IsUnsafe}(g_\tau)$ 
14   // Check for collisions along each robot edge (Section 3.3.2.1)
15    $c_o \leftarrow \emptyset$ 
16   foreach  $\text{edge } \mathcal{P}_e \in E$  do
17     foreach  $\text{obstacle } p_o \in O$  do
18        $p_i \leftarrow \text{Intersect}(\mathcal{P}_e, p_o)$ 
19       if  $p_i \neq 0$  then
20         // Update collision vector
21          $p_i^* \leftarrow \text{Transform}(p_i, g_\tau)$  using Equation (3.5)
22          $c_o.append(p_i^*)$ 
23     end
24   end
25   return  $c_o$ 

```

---

forward models is as “state estimators”, in which the objective is to predict state changes based on a set of initial conditions and inputs (Karniel, 2002). For different robot platforms, this state estimation procedure makes use of the robot dynamics to inform state changes, e. g. using differential-drive kinematics to estimate the change in state of a differential mobile base.

Forward models are functions  $\phi(\cdot, \cdot)$  that output the system’s predicted next state  $r'_{t+1}$ , given some input motor command  $u_t$  and optionally the current state  $r_t$  (Wolpert et al., 1998). A forward model equates to the following transformation:

$$r'_{t+1} = \phi(u_t, [r_t]). \quad (3.1)$$

A confidence function  $C(\cdot, \cdot)$  can also be defined as any function that assigns a reward or penalty  $\epsilon$  depending on how well the predictions of the forward models fare against the actual next state  $r_{t+1}$ :

$$\epsilon = C(r'_{t+1}, r_{t+1}). \quad (3.2)$$

There is no restrictive form on how to measure the similarity between two states, as it is yet again likely to be specific to the robot application, environment and task domain.

### 3.3.1.2 Trajectory Generation

In line with the “state estimation” outlook on forward models, the trajectory generation process for estimating user navigational intent takes place as follows. Given the wheelchair’s mobile base adheres to differential-drive kinematics, its motion is constrained by the equation (Minguez et al., 2016):

$$-\dot{x} \sin \theta + \dot{y} \cos \theta = 0, \quad (3.3)$$

where  $r_t = (x, y, \theta)$  represents the robot state, i. e. its location  $(x, y)$  and orientation  $\theta$  at time  $t$ . The predicted next state  $r'_{t+1} = (\dot{x}, \dot{y}, \dot{\theta})$  is thus computed as:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} \cos \theta & 0 \\ \sin \theta & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v \\ \omega \end{bmatrix}, \quad (3.4)$$

with motor command  $u_t = (v, \omega)$  at time  $t$  denoting input linear  $v$  and angular  $\omega$  velocities.

This operation is repeatedly performed over a simulation period of  $T$  to output a trajectory  $\tau_{1:T} = (\dot{x}_{1:T}, \dot{y}_{1:T}, \dot{\theta}_{1:T})$ . For local path primitives in wheelchair navigation,  $T$  is likely to be around 20 discretised timesteps (Poon et al., 2017), e. g. a two-second duration if operating at a rate of 10Hz. Simulating trajectories in this manner offers a trace of recognised intent, albeit the constant velocity assumption here means this trace is only a snapshot representation. Once trajectory  $\tau_{1:T}$  is obtained, it is then passed through an assistive control process for safe navigation.

### 3.3.2 Assistive Control

The ensuing phase determines how to best assist a user in accordance with their estimated intentions. In our assistive navigation context, we break this

process down into two independent stages: collision avoidance and arbitration. We will now outline both stages and compare them to the SC presented in Soh and Demiris (2012).

### 3.3.2.1 Collision Avoidance

Reactive collision avoidance boils down to the computation of motion that is capable of evading obstacles detected by sensors (Minguez et al., 2016). The SC of Soh and Demiris (2012) tackled this problem by projecting robot states forward in time and then validating for safety against an obstacle map constructed from incoming LiDAR scans. A binary model estimated the probability of potential collisions based on these future states. Although simple and effective, this approach assumed like other standard obstacle avoidance algorithms that the mobile robot is both *holonomic* and *disc-shaped* (Minguez et al., 2016; Minguez and Montano, 2009), neither of which apply to the differential-drive rectangular base of the wheelchair. As a result, projected robot states required computationally expensive point-in-polygon checks for colliding obstacle cells (Soh and Demiris, 2012).

In contrast, we employ the ‘gap-based’ collision avoidance method of Mujahed et al. (2018), as it considers the exact robot shape and kinematics through an abstraction layer (Minguez and Montano, 2009). The key idea behind this abstraction layer is to assume that motion trajectories can be approximated by piecewise circular arcs (Fox et al., 1997; Minguez and Montano, 2009). Therefore, the trajectory  $\tau_{1:T}$  produced by Equation (3.4) can be described by a sequence of circular arcs, where the *radius* and *tangent direction* of each circular path is extracted from its endpoint. In turn, the simulated goal point  $g_\tau = (\hat{x}_T, \hat{y}_T)$  allows us to concisely encapsulate  $\tau_{1:T}$  as a single circular path. More details regarding the formal definitions of these trajectory arcs are available in Minguez and Montano (2009); Mujahed et al. (2018).

Gap-based obstacle avoidance approaches generally use rangefinder data (e.g. from LiDAR sensors or depth cameras) to identify ‘gaps’ in the environment, i.e. open navigational spaces for the mobile robot to traverse through (Durham and Bullo, 2008). The “Admissible Gap” strategy of Mujahed et al. (2018) not only accounts for robot shape and kinematics during gap traversal, but is also more computationally efficient, safe, smooth and robust in densely cluttered environments. Inspired by these advantages for scenarios frequently encountered during wheelchair navigation (Minguez and Montano, 2009), we implemented a version of the “Admissible Gap” method tailored to our use case of SC.

As in [Mujahed et al. \(2018\)](#), our collision avoidance objective is to derive collision-free motion commands  $s_t$  that will guide the mobile robot through the gap closest to the user’s intended goal,  $g_\tau$ . A crucial step in this procedure is to test the admissibility of every trajectory traversing a gap by checking if any robot edge  $\mathcal{P}_e$  intersects with obstacle points  $p_o$  along the circular arc. Each intersection point  $p_i$  can then be transformed to the goal frame of reference using:

$$p_i^* = \begin{cases} p_i + g_\tau, & \text{if } \dot{y}_T = 0 \\ \mathcal{R}p_i + g_\tau, & \text{otherwise} \end{cases} \quad (3.5)$$

where rotation matrix  $\mathcal{R}$  (defined in Eq. 31 of [Mujahed et al., 2018](#)) is applied depending on whether motion is purely translational or not. Performing this navigability check yields a vector of potential collision points  $c_o$ , as shown in Algorithm 3.1.

Aside from this collision-checking routine, the remaining steps for deciphering  $s_t$  are contained within [Mujahed et al. \(2018\)](#) and not disclosed here in this thesis. We merely specify how to recover collision vector  $c_o$ , as this vector proves to be a vital AR visualisation in Section 4.3.2.2. However, Appendix A does provide more information on our ROS package of the fused “Admissible Gap” and SC algorithm.

### 3.3.2.2 Arbitration

Given safe velocity commands  $s_t$ , the arbitration stage is then concerned with adjusting user input commands  $u_t$  whenever they are deemed unsafe. [Soh and Demiris \(2012\)](#) mediated user input by selecting the highest-scoring command from a range of discretely sampled velocities in the robot’s control space. The scoring schema for these prospective commands applied a variant of the seminal dynamic window approach ([Fox et al., 1997](#)) in order to compute a velocity command that would optimally align with user intent. We instead adopt a “policy-blending” formalism as an arbitration of the control policies originating from the operator and robot ([Dragan and Srinivasa, 2013](#)), so as to continuously preserve a user’s participation in the SC. In particular, we choose a linear-blending and take precautionary measures to circumvent the known issues of improper wheelchair manoeuvres ([Ezeh et al., 2017](#)). For instance, if the user outputs commands opposing those of the robot (e.g. a forward and reverse motion), then the linear-blended output would be to remain stationary.

The following delineates our linear-blending arbitration scheme. To first avoid unexpected wheelchair manoeuvres resulting from blending (Ezeh et al., 2017), we adapt the motion law of Mujahed et al. (2018) to navigate towards the closest gap based on angular disparity, as opposed to Euclidean distance. Ergo, the robot-generated motion commands  $s_t$  always conform to the user's desired heading for a more anticipatory and sensible outcome in the arbitration (avoiding failure cases like above). The arbitration of control commands  $u_t$  and  $s_t$  is then calculated by:

$$a_t = (1 - \epsilon) * s_t + \epsilon * u_t, \quad (3.6)$$

where  $\epsilon \in [0, 1]$  is typically a measure of confidence in the intention estimation, as in Equation (3.2). Though for our application of smart wheelchairs, we prioritise user-safety by setting  $\epsilon$  according to a perceived indicator of threat (Durham and Bullo, 2008) rather than similarity to user intent:

$$\epsilon = \frac{1}{N} \sum_{i=1}^N \text{sat}_{[0,1]} \left( \frac{D_s + R - l_i}{D_s} \right). \quad (3.7)$$

The aggregate threat score  $\epsilon$  is a saturated function of  $N$  ranger readings  $l_i$ , the robot's radius  $R$  (0.5m in the wheelchair's case), and a safety distance parameter  $D_s$  (set to 0.8m).

Overall, the final assistive command  $a_t$  can be thought of as an amalgam of robot-user input. Intuitively, the robot occupies more control if the user inputs  $u_t$  are collision-prone and vice versa if they are risk-free, hence respecting user authority. In essence, our SC mechanism for smart wheelchairs is designated to exclusively correct the motor commands of operators when in the face of danger.

### 3.4 CURRENT LIMITATIONS

Although our SC methodology ensures safety during wheelchair navigation, it is also liable to bewilder or impede users whenever their actions do not elicit the intended system response. Consequently, a misalignment between the respective internal models of the robot and human may start to form. This is especially problematic for smart wheelchair users due to the steep learning curve associated with accepting navigational assistance on powered wheelchairs (Carlson and Demiris, 2010; Nisbet, 2002). Mismatched internal models may even lead to patient failure in fulfilling the strict eligibility criteria for acquiring ownership of these assistive mobility platforms (Fehr et al.,

2000). This dilemma of model misalignment is primarily rooted in the current lack of *transparency* in our SC (Alonso and de la Puente, 2018; Lyons, 2013; Lyons and Havig, 2014). The building blocks of our SC, namely intention estimation and arbitration, are usually evaluated based on how well the output control commands align with the user’s actual task intent or manage to evade danger. Nevertheless, this does not illustrate whether the user is aware of the underlying robot reasoning and reiterates the need for communication in the SC to expose its inner workings (Losey et al., 2018). We focus on this topic in Chapter 4.

Moreover, the intention estimation technique of our SC derives a representation of human intent that is neither robust nor accurate across different task settings. Despite its straightforward nature, there is no guarantee that a trajectory of low-level wheelchair poses will corroborate a human’s higher-level *plan* of intent (Tomasello et al., 2005). Without conditioning on prior knowledge of the world dynamics or user goals, each recognised intention is only a snapshot representation of a user’s cognitive state. Chapter 5 expands on this representation and improves the robustness of computationally interpreting human intent via a probabilistic inference framework.

### 3.5 CONCLUSIONS

Robotic wheelchairs with built-in assistive features, such as SC, are an emerging means of providing independent mobility to severely disabled individuals. In this chapter, we introduced our architecture for such an assistive robotic wheelchair, as well as its precise methodology of executing SC. The remaining sections of this thesis will refer back to this SC methodology when addressing the limitations mentioned in Section 3.4. Furthermore, later experiments on human subjects will make use of the assistive robot architecture demonstrated in Section 3.2.

The proposed architecture and SC have also served various other research directions. As an example, the hardware and software constituents of our assistive robot architecture supported HRI trials revolving around learning assistance by demonstration through remote interfaces (Schettino and Demiris, 2019). Additionally, the SC algorithm has been extensively used in AR studies where environmental affordances were visually displayed to users operating the robotic wheelchair (Chacón-Quesada and Demiris, 2019, 2020).

Finally, a noteworthy benefit of our SC is its ease of explanation due to its model-based nature. The growing use of deep learning and other “black-box” approaches to intention estimation (Nicolis et al., 2018) and

arbitration (Reddy et al., 2018) are far from explainable to end-users (Fox et al., 2017). Whereas our SC simplifies this endeavour by employing internal mechanisms that are easy to interpret and translate into visual explanations, which becomes apparent in the next chapter.



## TOWARDS EXPLAINABLE SHARED CONTROL USING AUGMENTED REALITY

---

This chapter addresses research questions 2 and 3:

- “How can an Augmented Reality Head-Mounted Display be integrated with a Shared Control system to expose its inner workings?”
- “Can human users of Shared Control have their mental model accuracy improved by an Augmented Reality interface that visually explains the robot’s internal mechanisms?”

Explainable Shared Control (**XSC**) is a paradigm within Shared Control (**SC**) that settles model mismatch during Human-Robot Interaction (**HRI**). Traditional processes of **SC**, i. e. intention estimation and arbitration (Losey et al., 2018), are commonly evaluated based on how well the output control commands align with estimates of the control behaviour intended by a user. In the proposed paradigm, we examine these processes from an additional perspective: *transparency*, where the objective is to best represent the underlying robot reasoning and feed it back to the user. Coinciding with this objective, a suitable medium of communication is required to relay back information to the user. Augmented Reality (**AR**) is presented as an integral component of **XSC** that meets this communication requirement by visually unveiling the robot’s inner workings to human operators.

Whilst Explainable Artificial Intelligence (**XAI**) is a widely explored area of interest, it has only recently garnered similar attention in **SC** (Zolotas and Demiris, 2019; Zolotas et al., 2018). Generating explanations for **SC** poses various new challenges, namely the dependence on a continuous communication channel of physical intent (Losey et al., 2018), the requirement for implicit switching between interaction modes (Goodrich and Olsen, 2003), and the overall lack of consensus on **SC** guidelines (Abbink et al., 2018). The **XSC** paradigm strives to address these challenges and resolve model misalignment by using **AR** visualisations for explanation.

This chapter is structured as follows. In Section 4.1, we clarify our interpretation of commonly used **XAI** terminology. Section 4.2 then integrates an **AR** Head-Mounted Display (**HMD**) onto a robotic wheelchair with built-in

SC, as an early prototype of our explainable system. This section also includes a pilot user study conducted to investigate the influence of different interface design options on the acceptance rate and learning curve of our AR-wheelchair setup. We then outline XSC in Section 4.3 and instantiate it with an updated AR HMD interface for assistive robot navigation, so as to discern the paradigm’s benefits over our initial prototype. This extended work also presents results from a user study that revolves around model misalignment. Section 4.4 summarises the chapter and comments on future directions for XSC.

The majority of research in this chapter has been published in articles Zolotas et al. (2018) and Zolotas and Demiris (2019), as well as the extended abstract Zolotas and Demiris (2020).

#### 4.1 TERMINOLOGY

Many of the key terms in XAI, such as explainability, interpretability or transparency, are constantly evolving and often used interchangeably in different contexts (Clinciu and Hastie, 2019; Gilpin et al., 2018). For instance, transparency in an HRI context has been regarded as a means of creating shared intent and awareness between humans and machines (Lyons, 2013; Lyons and Havig, 2014). Yet in SC, a popular interpretation of transparency relates to the robot’s observability and predictability, i. e. the *what*, *why* and *when* (Abbink et al., 2018; Alonso and de la Puente, 2018). Hence, there are numerous reviews of these key terms (Alonso and de la Puente, 2018; Lyons, 2013; Lyons and Havig, 2014; Theodorou, 2019) and how they relate to one another (Clinciu and Hastie, 2019; Gilpin et al., 2018), e. g. an Artificial Intelligence (AI) becomes transparent by providing explanations or interpretations of its inner workings.

It is therefore imperative for us to first clarify our own usage of these terms. In line with Lyons (2013); Lyons and Havig (2014), we view *transparency* as a communication bridge of intent between human and robot. More specifically, “robot-to-human” transparency (Lyons, 2013; Lyons and Havig, 2014) is about revealing a robot’s intentions and state to interacting humans, which is the topic focus of this chapter. We visually expose this information through *explanations*, i. e. instruments capable of improving a user’s mental model of expected system behaviour (Clinciu and Hastie, 2019). A robot equipped with such instruments is thus deemed “explainable”, as is the SC. The robot’s perspective of human intent is instead viewed as “robot-of-human” transparency (Lyons, 2013; Lyons and Havig, 2014). This matter

will be addressed in Chapter 5 through a human intention inference framework that allows for model *interpretation*, or rather the “extraction of relevant knowledge from domain relationships in the data” (Murdoch et al., 2019).

Throughout this thesis, we will follow the terminology delineated in this section. The way in which constructs of “transparency” and “explainability” are established during HRI will become apparent when we introduce the XSC paradigm in Section 4.3. Nevertheless, the next section emphasises less the role of these constructs and more the system integration of our smart wheelchair with an AR HMD.

## 4.2 EXPLAINABLE ROBOTIC WHEELCHAIR ASSISTANCE

Patients often struggle to build a mental model of their smart wheelchair’s behaviour under different environmental conditions (Nisbet, 2002). Immersive technologies involving HMDs are an emerging solution to help users easily accept the navigational assistance offered by smart wheelchairs. For example, virtual reality HMDs have recently garnered attention as apt training simulators for offline learning of wheelchair control (Alshaer et al., 2017; Devigne et al., 2017; Ktena et al., 2015). However, AR HMDs could serve as an even better mode of communicating assistance to foster transparency, but have yet to be integrated onto physical wheelchairs for online operation.

Motivated by the desire to bridge this gap in transparency, we propose a novel AR system on a robotic wheelchair with built-in SC (see Figure 4.1). A Microsoft HoloLens<sup>1</sup> is incorporated into our real-world setup for the purpose of highlighting to users the inner workings of the SC. Consequently, this section makes two contributions to answer the second research question of this thesis: 1) an AR system in Section 4.2.1 that renders the internal state of a shared controller for powered mobility onto the driver’s view of the world; 2) a pilot study in Section 4.2.2 that evaluates the acceptance rate and learning curve of an immersive training regime for wheelchair control with a variety of tested visualisations. Section 4.2.3 summarises the findings from our pilot study of this initial prototype.

### 4.2.1 Augmented Reality Cueing System

In this section, we describe the core AR system for cueing robot-assisted mobility (refer to Figure 4.1 for an overview). Figure 4.2 summarises its main

<sup>1</sup> <https://www.microsoft.com/en-us/hololens>

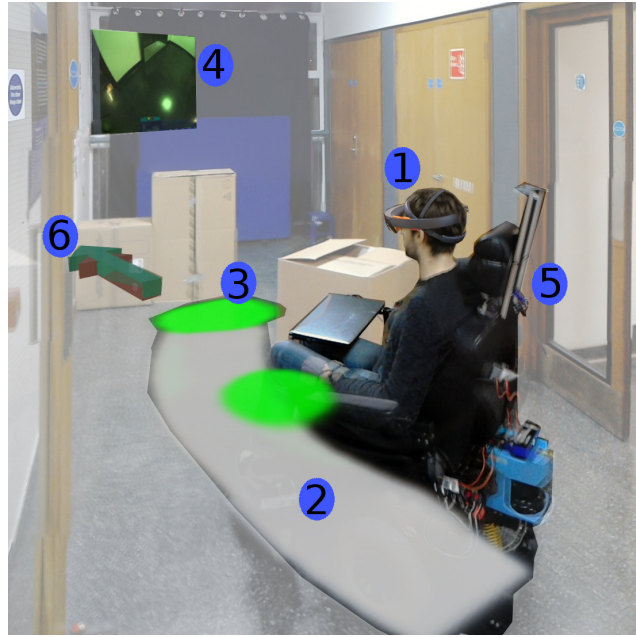


Figure 4.1: Composite image of the visualisations rendered on the user’s view through the AR headset (1). The grey path (2) shows the trajectory generated by the user’s manual input. The green patches (3) highlight objects that pose as potential collisions. The rear-view display (4) captures the camera image mounted on the back of the seat (5), which includes overlaid graphics, such as the path and obstacle cues. The green and red directional arrows (6) represent the user’s raw input and the corrected output, respectively.

components, all of which are coordinated through the Robot Operating System (ROS) (Quigley et al., 2009) and AR graphical cues are designed within the Unity 3D<sup>2</sup> game development environment. The following presents each of these system aspects, except for the shared controller, as it was covered in Section 3.3.

#### 4.2.1.1 Gridmap Processing

Whilst the shared controller captures information relating to a driver’s navigational input, the gridmap processor instead represents environmental context from sensor data. Given incoming rangefinder data, this module identifies dangerous obstacles in the surroundings and constructs an image view of this information to relay back to the user visually via AR (presented in Section 4.2.1.2). All processing steps are entirely local and do not rely on a static map.

<sup>2</sup> <https://unity.com/>



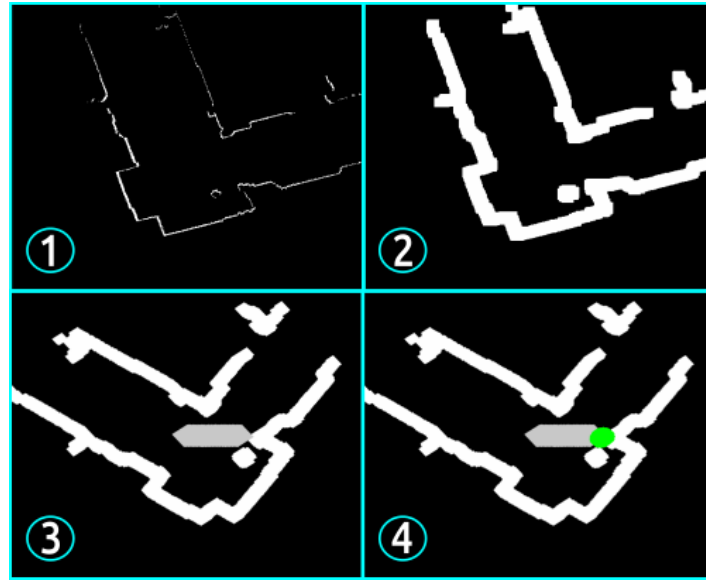


Figure 4.3: Gridmap processing pipeline. The 2D occupancy grid constructed by [LiDAR](#) data is first converted into a binary image (1). Occupied cells are dilated to enlarge potential collisions (2). The image is then rotated to align with the mobile base frame and overlaid with an inflated grey path generated by the user’s input commands (3). Finally, green circles are centred at coordinates where the grey path and obstacles intersect (4).

#### 4.2.1.2 Visualisations for Assistive Feedback

To compensate for the potential misalignment in a user’s interpretation of their wheelchair’s behaviour, we use a Microsoft HoloLens to provide visual feedback on the robot’s dynamics. We envision that an [AR](#) headset will help users form a better mental model of the expected system behaviour with respect to faster improvements in task performance (hypothesis  $H1$  in the Section 4.2.2 study). Furthermore, this approach could reduce the levels of frustration and workload experienced by users of assistive robotic wheelchairs ([Carlson and Demiris, 2012](#)).

Figure 4.1 provides a summary of the four visualisations implemented as [AR](#) feedback<sup>3</sup>. These visualisations are displayed at three different heights relative to the user: floor level, head level and floating above head level. This spatial separation was designed to help limit the likelihood of a user being overloaded with information (hypothesis  $H2$  of Section 4.2.2), or to avoid overlapping visualisations.

<sup>3</sup> Supplementary video material for first-person perspectives of these visualisations is available at: <https://www.youtube.com/watch?v=TJmKZykDudE>

The first visual aid is a rear-view display, which is situated directly above the user's normal viewing direction. From the driver's perspective, this display behaves like a large version of a rear-view mirror, such as those found in road vehicles. The camera display also renders any other graphical effects incorporated into the holistic system. This is achieved by placing a virtual camera's view of the artificial world containing the visualisations in the same position as the real camera. The intrinsic calibration parameters of the real camera are mapped onto the captured virtual image to match the strong fish-eye effect applied in the real camera's display. By applying this fish-eye effect, the user is able to view a very wide angle, which could help navigation in tight manoeuvres typical of indoor wheelchair use.

There are two kinds of visualisations rendered onto the floor. A grey path is projected either forward or backward depending on the direction of travel, which portrays the predicted future state of the wheelchair given the current input commands. If the path intersects with an obstacle then a bright green circle is rendered at that location, which is intended to help drivers identify objects that are likely to make the SC intervene. The construction of this image was described in Section 4.2.1.1.

The last visualisation is a pair of directional arrows that float directly in front of the user. The green arrow corresponds to the user's joystick input and the red arrow is the final command sent to wheelchair after arbitration via SC (see Section 3.3.2.2). The arrows rotate with the direction of the corresponding command velocities and lengthen to represent their magnitude.

These four visualisations fall into two categories of relative placement from a user perspective. The arrows and rear-view display appear fixed to the motion of the wheelchair, behaving similarly to instruments found in an aircraft cockpit or car dashboard. On the contrary, the grey predicted path and green collision markers are perceived as fixed to the environment, not necessarily being locked to the wheelchair as it moves or rotates.

#### 4.2.1.3 *Augmented Reality System Alignment*

All visualisations presented in this work require appropriate alignment with both the world and mobile platform, therefore a correspondence between the HoloLens and wheelchair frames of reference must be determined. The HoloLens maintains its own internal map for the purpose of visual odometry, however by default there is no well-defined origin for the rest of the robotic system to reference. This problem was previously solved in the context of a motion capture arena in Elsdon and Demiriz (2018), however due

to the multi-room nature of indoor wheelchair use, a motion capture system is not a reasonable proposition.

To solve the registration problem, three points were manually marked by placing virtual objects in the [AR](#) environment. The HoloLens has a system known as *spatial anchors*<sup>4</sup>, which use local geometry to latch objects in place despite shifts in the global map. This enables the virtual markers to persist across multiple uses of the HoloLens, whilst also allowing adaptations to be made on-the-fly given any environmental changes. These three points are compared to their equivalent coordinates on the map constructed using the localisation module’s Simultaneous Localization and Mapping ([SLAM](#)) component. Utilising singular value decomposition as outlined in [Ho \(2013\)](#), we obtain the transform between the HoloLens world and the global frame of the mobile base. There are four unknown variables accounted for, three for the position offset between coordinate systems and one representing the rotation in yaw direction. These points should not be collinear to avoid multiple solutions and should span the experimental arena to minimise the effect of placement error.

#### 4.2.2 Experiments

In order to explore the assistive effects of our [AR](#) system on wheelchair control, we conducted a between-subjects experiment that evaluates how different graphical aids affect the user’s experience and learning of our robot’s internal model. The Unity game engine was used to both develop the [AR](#) application and deploy it on the Microsoft HoloLens. Communication with the [HMD](#) was established over a wireless router.

Our hypotheses for this experiment are:

- $H_1$  – [AR](#) will accelerate the learning of our robot’s internal model based on improvements in rate of completion time for three recorded navigation trials.
- $H_2$  – Visualisations in [AR](#) will reduce the physical strain on subjects, according to less variable head motion when performing wheelchair manoeuvres across three trials.

##### 4.2.2.1 Experimental Setup

For this pilot study, we recruited 16 able-bodied participants (13 male, 3 female) aged between 20 and 31. The discrepancy here in gender-balance

<sup>4</sup> <https://azure.microsoft.com/en-gb/services/spatial-anchors/>

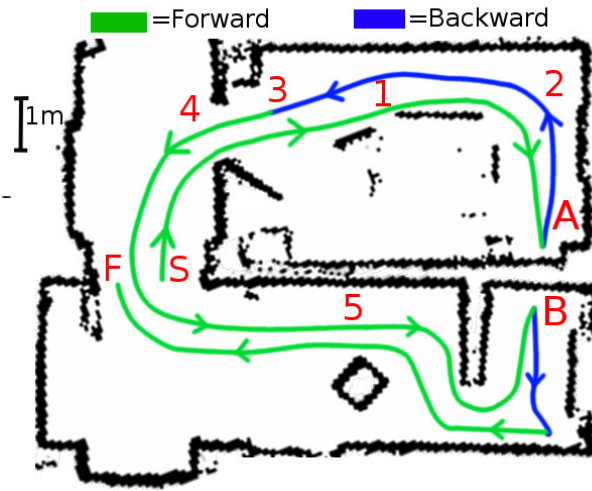


Figure 4.4: An overhead view of the trial route used for the experiment. Overlaid on the map is the path that participants were asked to perform, with green sections requested in forward motion and blue sections in reverse. The small numeric labels denote course sections that required particular manoeuvres, as summarised in Table 4.1. The door opening at location 4 is 90cm wide and the narrow corridor at location 1 is 110cm wide.

correlates with how the subject pool are engineering students and there is a low percentage of women in engineering<sup>5</sup>. Participants were asked to sign a consent form for the collection of data and presentation in this work. Prior and post experiment questionnaires were also handed out for completion. Each subject was requested to complete a navigation route (illustrated in Figure 4.4) four times in sequence, which lasted a total duration of 30 minutes on average. The trial route devised for this experiment includes a subset of evaluation criteria from the Wheelchair Skills Test (WST) manual (version 4.2 Kirby et al., 2002), as shown in Table 4.1.

The purpose of the experimental task was to investigate the effectiveness of different graphical aids and whether the AR accelerated learning of the wheelchair's behaviour. We controlled for this by assigning individuals to one of two groups: with-visualisation and without. People in the control group also wore the HoloLens but without any visualisations displayed, so that head orientation data could still be collected and that the obtrusiveness of the HMD is kept fair for both groups. More importantly, this is done not to add a confounding variable to the post-study statistical analysis.

Participants in the visualisation group differed from the non-visualisation counterpart in two ways. First, they were administered augmented feedback

<sup>5</sup> <https://www.wes.org.uk/content/wesstatistics>

Table 4.1: A summary of the modified WST assessment points. Each of the task-specific positions is numbered correspondingly on the map in Figure 4.4.

Skill	Location
Forward motion in narrow 1m passageway	1
Reverse in narrow 1m passageway	1
Turn while rolling forwards ( $90^\circ$ )	2
Turn while rolling backwards ( $90^\circ$ )	2
Turn in place ( $180^\circ$ )	3
Traverse through open doorway	4
Avoid static obstacles	5
Stop before walls	A & B

and instructed on the meaning of the visualisations, although no advice on how to interpret or make use of them was provided. Second, subjects in the visualisation group were requested to perform the fourth attempt at the course without any graphical aid. This was designed to observe whether a dependency on the AR formed, or if the task-learnt skills were independent of these visual cues. Nonetheless, this also mixes experimental conditions between the groups, and so all significance testing is only run across the first three trials where there is a clear distinction.

#### 4.2.2.2 Empirical Findings

We assessed total time to completion for each trial as a performance indicator of the overall AR feedback. Figure 4.5 indicates that the group without visualisations performed better across all trials and improved consistently in the first three rounds, having plateaued in skill by the third. On the other hand, the group with visualisations demonstrated more variable performance, with a greater decrease in time relative to their first trial, despite taking longer to plateau. Albeit the rate of improvement in timings across the first three trials (i.e. when the two experimental groups are directly comparable) rejects  $H_1$  (two sample t-test,  $p=0.97$ ) and in fact, strongly confirms there is no significant difference between the two conditions. When visualisations were removed on the fourth trial, there was a slight dip in performance, but no strong claim can be made for any dependency forming on the AR assistance.

A possible explanation for this offset in absolute performance between the two groups is suboptimal placement of some of the virtual objects. This is especially true given the narrow Field of View (FoV) of the HoloLens

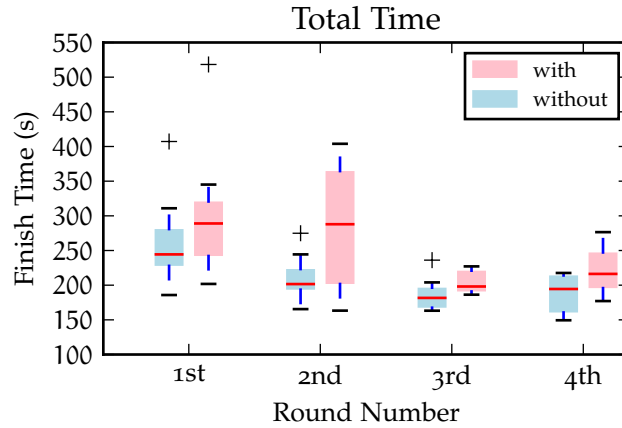


Figure 4.5: Total time to completion for each trial. The group without visualisations were superior in every trial, even in the 4th round where both groups did not have visual aid.

(estimated at  $17.5^\circ$  vertically and  $30^\circ$  horizontally). We therefore speculate that subjects could not make proper use of the [AR](#) assistive features outside of their natural [FoV](#).

To further elaborate on how often participants made use of the different visualisations under these restrictive viewing conditions, dwell time was monitored by extending a ray directly forwards from the user’s head and registering intersections with virtual objects. We found that participants in the visualisation group spent a median proportion of 48.4% across the first three trials directed towards the rear-view display and floating arrows. The green obstacle cues were instead oriented towards for a median value of 32.6%. It is worth noting that the viewing direction of the subjects in the non-visualisation group would have also aligned with these obstacle cues for a median of 77.6% had they been rendered. This implies that the floor-based objects adopted a natural orientation angle for wheelchair navigation. Assuming participants maintained a central eye-in-head position, we suspect that floor-plane features occupied a less salient region within the [HMD’s FoV](#) and were thus less effective.

Seeking to explore other aspects of effectiveness that are relevant to the target application, we also evaluated the head orientation data recorded by the HoloLens. Individuals with upper body mobility impairments are prone to colliding with obstacles outside of their viewing capacity during typical wheelchair navigation manoeuvres, such as rotating in place or reversing ([Nisbet, 1996](#); [Simpson, 2008](#)). These day-to-day tasks for wheelchair users, as asserted by the full WST manual ([Kirby et al., 2002](#)), could benefit

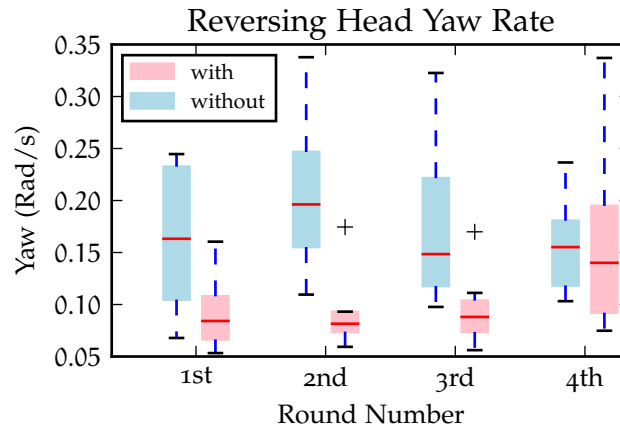


Figure 4.6: Depicts the rate of head rotation in the yaw direction during the reverse passageway section in Figure 4.4. Participants with visual feedback rotate their head significantly less than the control group for the first three trials. When the visualisations are removed in the 4th round, both groups display similar mean rates of yaw rotational movement.

from the inclusion of a rear-view mirror that reduces the necessity for harsh rotational head motion.

Figure 4.6 analyses the effects of the rear-view display on the rate of rotational motion along the yaw axis during the reversal of the narrow passageway. The results demonstrate a significantly lower turning rate across all AR aided trials (two sample t-test,  $p=0.01$ ), and a rate that matches the control group when the visualisations are removed on the last trial. These findings corroborate  $H_2$  and suggest that the rear-view display provided an easily accessible source of information for users, such that they could complete the reversal task with minimal need for strenuous neck movement. This could prove to be particularly beneficial for disabled individuals with limited upper body and neck mobility, such as people suffering from spinal cord injury.

#### 4.2.2.3 Survey Results

In the post-experiment questionnaire, subjects were asked to rate the benefit of each of the provided visualisations on a 5-point scale. A strikingly positive result from this survey was the popularity of the rear-view display. Almost all subjects rated it as either “good” or “very good” ( $4.125 \pm 0.64$ ). Conversely, there was nearly universal disapproval of the grey path and the green obstacle markers ( $2.375 \pm 0.92$  for both). The overall average scores from 1-5 (5 being most positive) are listed in Table 4.2.

Table 4.2: A summary of the user responses to the question: “Rate the following visualisation from 1-5 (1 = very poor, 5 = very good)”.

Visualisation	Mean User Rating	Standard Deviation
Rear-View Display	4.125	0.64
User/ Assistance Arrows	3.125	1.46
Projected Path	2.375	0.92
Highlighted Obstacles	2.375	0.92

The poor ratings associated with the grey path and obstacle cues are informative on floor-based renderings. Although information overlaid on an environment is a fundamental quality of AR, practical considerations should be made for the HMD’s FoV limitations. Some participants provided comments reinforcing this observation by stating that they could rarely notice these floor visualisations, supporting our quantitative analysis on dwell time. The embedding of this environmental information mandates a user to perform a search of their surroundings, which itself could frustrate them. Furthermore, the time and frustration expended on searching for visual aids may have impacted the trial performance reported in Figure 4.5.

Another noteworthy remark is on how intuitive different visualisations appear from a user’s perspective. Many subjects commented on how they misunderstood the purpose of the floating arrows, querying whether they should have aimed to match the corrective red arrow or simply taken both arrows into account as supplementary information. This leads us to believe that low-level cues, such as command indicators, are not necessarily an effective user-centred form of augmented assistance and would require auxiliary instruction to be provided. On the contrary, highlighted obstacles are higher-level and provide more intuitive feedback.

#### 4.2.3 Discussion

To the best of our knowledge, this is the *first* instance of an AR headset being incorporated into a smart wheelchair system. Our findings lead us to believe that there is potential benefit to be gained from the integration of AR headsets with robotic wheelchairs, as long as certain design choices are taken into account. Namely, that virtual objects are placed in easily visible locations that are not within proximity of the mobile base, and preferably do not clutter the natural viewing required for navigation. Moreover, that

graphical cues are high-level and contextual enough for a typical user to garner an augmented experience from the administered aid.

Any [AR](#) cue that fulfils both these requirements, such as a virtual rear-view mirror, could prove to be an attractive component in robotic wheelchair design. The rear-view display yielded enthusiastic participant responses by presenting helpful and intuitive information to users at a comfortable and non-intrusive viewing angle. In the next section, we probe a set of guidelines on how to implement similar [AR](#) cues that are guaranteed to facilitate enhanced information retrieval and transparency in the [HRI](#). As a result, we hope to mitigate the negative consequences of model misalignment on robotic wheelchair navigation.

Regarding the evaluation of model misalignment, we now review a few limitations of the study that may have influenced the outcome. First, the users are not from the actual target population, so any conclusions drawn from this study are discounting the attitudes, needs and preferences of actual wheelchair users ([Viswanathan et al., 2017](#)). Next, the subject numbers are quite low and may not hold sufficient statistical relevance. For instance, if a statistical power of 80% was acceptable for a significance of 0.05 and an effect size of 0.80, then at least 25 subjects would be required. Lastly, we reflect on whether time-to-completion is a fitting metric to test  $H1$ . As underlined in [Section 2.3](#), mental model accuracy can be measured in various ways, such as cognitive load and attentiveness, which are suitable choices for wheelchair control ([Carlson and Demiris, 2009](#)).

A final use-case for our [AR](#) application is to extend beyond first-person wheelchair navigation. Whilst floor-rendered and low-level visualisations did not gather positive participant response, a bystander could benefit from the inclusion of these graphical cues due to their improved ease of interpreting the primary user's intentions. For example, a clinician or therapist wearing a HoloLens can oversee the graphical overlay of the learner's interaction with the robot (similar to the view in [Figure 4.1](#)) and then better understand the reasoning behind any assistive intervention. Communicating the navigational intent of a robotic wheelchair and its driver to passing pedestrians has also been shown to generate smoother interactions between the two parties ([Watanabe et al., 2015](#)). We thus believe that intention communication in multi-[AR](#) headset applications is advantageous for the future of rehabilitation and assistive robotics.

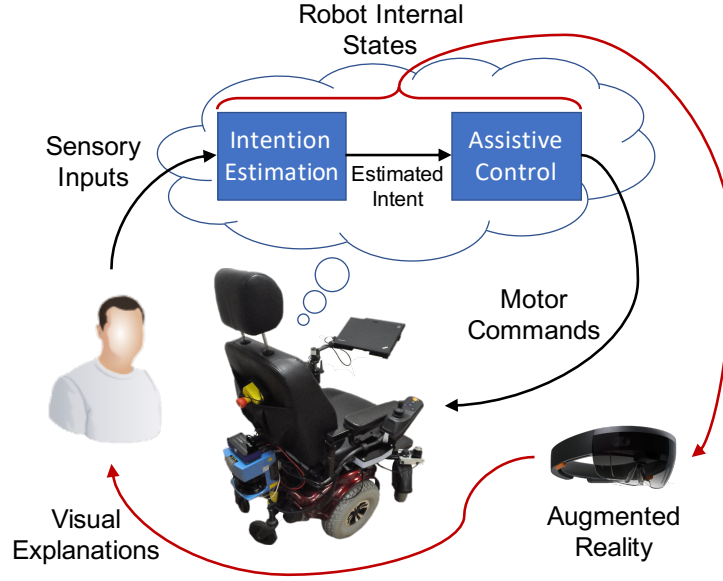


Figure 4.7: Overview of **XSC**. Standard **SC** (indicated by black arrows) will involve an inference of the user’s task intent based on sensory inputs, followed by an assistive control process that outputs motor commands to the robot, e.g. a robotic wheelchair. In the explainable paradigm (indicated by red arrows), the robot’s internal states are also visually represented in **AR**, i.e. making robot “intent” transparent.

#### 4.3 EXPLAINABLE SHARED CONTROL PARADIGM

The smart wheelchair system of the previous section clearly suffered from model misalignment, even with the assistance of **AR** visualisations. Our experimental evaluation of this system exemplified how poorly designed visualisations can harm task performance and even cause greater misunderstanding. These findings suggest that more careful consideration is required when designing visualisations for **SC** due to the active user engagement. Ergo, we introduce **XSC** as a set of **AR** interface guidelines on how to best demystify the **SC**.

**XSC** refers to a novel paradigm where the **SC** simultaneously complements the abilities of a human operator and facilitates rich information exchange. There is a plethora of research on how to complement an operator’s abilities and instil trust in the interaction (Hancock et al., 2011; Soh et al., 2019). Nonetheless, the literature on making processes of **SC** transparent is sparse (Alonso and de la Puente, 2018). As a result, the focus of **XSC** is to expose any internal mechanisms in **AR** using “explanation as visualisation” (Chakraborti, Fadnis, Talamadupula, Dholakia, Srivastava, Kephart

and Bellamy, 2018) and comply with standard axioms of SC (Abbink et al., 2018).

By proposing a means of achieving XSC via an AR interface, we aspire to settle the model mismatch between interacting humans and robots (overview shown in Figure 4.7). Therefore, the main contributions of this section are: 1) to outline XSC in Section 4.3.1 and then instantiate it in Section 4.3.2 with an AR HMD interface for assistive robot navigation; 2) to discern the paradigm’s benefits in Section 4.3.3 through a user study on model misalignment. The setup for this experiment involves our standard robotic wheelchair architecture and a Microsoft HoloLens supplemented with eye tracking capabilities. Section 4.3.4 discusses insights gleaned from this experiment. These contributions aim to further elaborate on research question (2), as well as answer (3).

#### 4.3.1 Guidelines for Transparency

The following are guidelines on how to realise XSC from the perspective of developing internal models, as well as constructing a head-mounted AR interface. By sharing robot intent through an *interface*, both trust and transparency will be injected into the SC (Alonso and de la Puente, 2018; Lyons and Havig, 2014).

##### 4.3.1.1 Causality and Context

First, minimising conflict in the human’s understanding of robot behaviours (Axiom 1 Abbink et al., 2018) should require the SC to exhibit *causality* (Fox et al., 2017). In other words, the internal mechanisms must be able to draw connections between inputs of the world (e.g. robot state, user commands, sensor readings) and the individual stages involved in generating the final output commands. Model-based intention estimation and arbitration algorithms make this an easier endeavour (Fox et al., 2017). Effectively, this helps answer user questions, such as: *why* and *how* did the robot perform that action?

Answering the *why* and *how*, an AR interface guideline of XSC is to represent causality by designing *contextual* visualisations. Contextual visual aids are those that achieve state summarisation of the robot’s perceived environment and any immediate action-effect relationships (e.g. an occupancy grid constructed from LiDAR data or end effector kinematics resulting from actuation). However, the active role of a participant in SC – as opposed to a passive observer – compels their continuous engagement with the task-at-

hand. For such scenarios, we advocate for the visual context to be represented using “embodied” cues (Walker et al., 2018). These cues are generated atop the robot morphology as virtual extensions, e.g. a radar attached to a mobile base that depicts the constructed occupancy grid, or arrow vectors originating from an arm’s end effector to illustrate planned motion (Walker et al., 2018).

#### 4.3.1.2 *Abstraction and Prediction*

Externalising the “brain” of SC is another step towards transparency, which demands that robot reasoning about the task is made explicit (Axiom 2 Ab-bink et al., 2018) through high-level *abstraction* (Chakraborti, Fadnis, Talamadupula, Dholakia, Srivastava, Kephart and Bellamy, 2018). Simply representing raw data streams of the system’s inputs or outputs will not suffice, as it risks overloading users with information that is already observable. Conversely, a trace or trajectory that highlights semantic task-specific characteristics and shows the provenance of information captured in the environment is better for visual portrayal (Chakraborti, Sreedharan, Kulkarni and Kambhampati, 2018). A critical question answered here is: *when* does the robot decide to intervene?

Hence, another AR guideline of XSC is to develop *predictive* visualisations that capture the reasoning behind *when* SC intervenes. Predictive visualisations are those that possess a temporal element and inform the high-level planning of users. A limitation of the aforementioned contextual visual aids is that they display only a snapshot of the current state, which is unlikely to explain when the robot may intervene and can cause even greater misunderstandings (as seen in Section 4.2.2). This is particularly problematic for SC settings, where the active involvement of the user calls for advance planning. By presenting visual traces of the historic or future world states (e.g. the evolution of a virtual robot’s arm motion trajectory), users will be equipped with the necessary information to act preemptively.

#### 4.3.1.3 *Why Headsets?*

Viewing “explanation” as a process of visualisation (Chakraborti et al., 2017), immersive HMDs are an integral component of the paradigm. Whilst most SC methodologies achieve information exchange via force feedback, haptic interfaces are limited at explainability in complex task settings unless combined with a visual modality for a multimodal approach (Losey et al., 2018). Given the rich visual feedback requirements of XSC, we stipulate that embodied interfaces in AR offer a superior medium of unveiling the robot’s

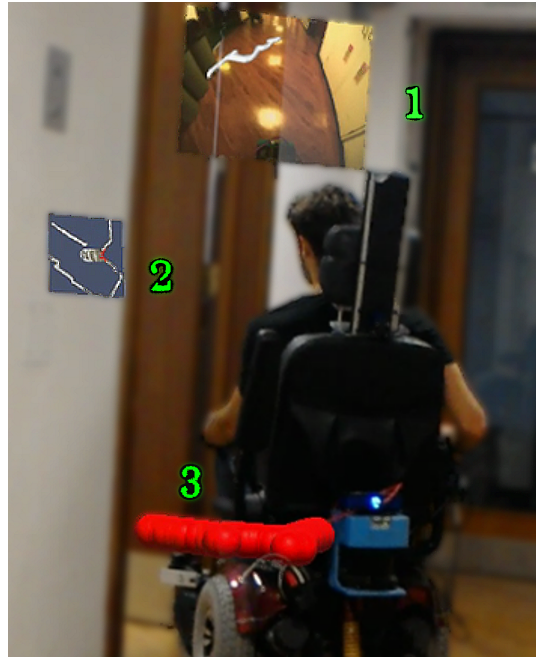


Figure 4.8: Composite image of the visualisations rendered during assistive robot navigation. The rear-view display (1) overlays virtual objects onto an image taken from a camera situated on the back of the seat. The mini-map panel (2) depicts a bird's eye view of the wheelchair configuration and its surroundings. Finally, red spheres (3) are placed atop real-world referents to highlight collisions.

inner workings. In particular, [AR HMDs](#) have the potential to better circumvent model mismatch during an active collaboration like [SC](#), as they provide users with a heightened sense of immersion over monitors or projectors ([Alshaer et al., 2017](#); [Sibirtseva et al., 2018](#)).

#### 4.3.2 *Instantiation for Assistive Robot Navigation*

Acknowledging that the presented outlook on [XSC](#) can be tackled in multiple ways, we now situate the paradigm in an assistive robot navigation setting as a concrete example of its instantiation (final platform shown in [Figure 4.8](#)).

##### 4.3.2.1 *Internal Mechanisms*

The [SC](#) methodology presented in [Section 3.3](#) developed internal mechanisms that are already in harmony with the guidelines of [XSC](#). Projecting trajectories in [Section 3.3.1.2](#) fits the *causality* guideline, as it represents a trace of recognised intent that can also be used to visually explain the wheelchair

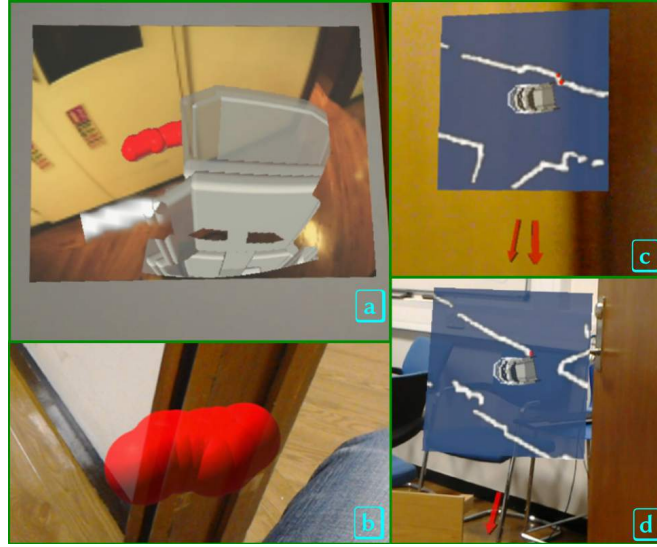


Figure 4.9: First-person perspectives of the AR visualisations. (a) illustrates the rear-view display during reverse motion, where a wheelchair trajectory is projected outwards. (b) shows the red collision spheres overlaid in the environment. (c) & (d) display the mini-map utility and arrow indicators. The mini-map reveals the wheelchair’s future trajectory and can hint at upcoming obstacles, e.g. in (d) where the heading angle is too tight for the doorway and requires adjusting.

dynamics when supplied with input commands. The geometric method of collision-checking in Section 3.3.2.1 also coincides with XSC on *abstraction* by providing a high-level representation of threats based on sensor data. Identifying “gaps” in the surroundings is particularly compelling for XSC, as it adds another semantic layer of *abstraction* around sensory state information. Accordingly, the next phase of XSC is to convey this navigational assistance using a head-mounted AR interface.

#### 4.3.2.2 Augmented Reality Interface

With the interface design guidelines of XSC in mind, three visualisations<sup>6</sup> are devised and categorised as either *environmental* or *embodied*. Any graphical cues that directly overlay the real-world surroundings are considered *environmental*, whilst *embodied* visualisations are fixed to either the robot or headset’s orientation and motion (Kim et al., 2018; Walker et al., 2018). First-person perspectives are demonstrated in Figure 4.9.

<sup>6</sup> Supplementary video material of visualisations available at: <https://www.youtube.com/watch?v=Hja38ghpKN0>

The first virtual aid is a collision sphere paired with a directional arrow. Highlighting collision referents in the physical *environment* with salient red spheres (see Figure 4.9-(b)) augments the *contextual* awareness of users, enabling them to identify why their actions may lead to unexpected behaviour. Directional arrows are also employed as headset-*embodied* cues that signal where these imminent collisions are situated from the operator’s perspective (shown in Figure 4.9-(c) & (d)). These arrows are constrained to always appear within the headset’s FoV.

The second visualisation is a mini-map panel that portrays a birds-eye view of the mobile base and its forward state. In order to explicate the SC methodology discussed in Section 3.3, the mini-map is annotated with laser scan readings and forecasts of the robot’s estimated poses after applying input commands. For instance, Figure 4.9-(d) illustrates a scenario in which the operator has selected a command that leads to a tight angle for door traversal. The *predictive* visual feedback of both the red collision marker and projected robot trajectory suggests that the user should adjust their heading before attempting the doorway.

Lastly, the rear-view display incorporated in our previous AR HMD proposal for wheelchair navigation (see Section 4.2.1.2) is persisted. We extend the display to supplement the rear-view with a virtual wheelchair avatar during backwards motion (translucent grey avatar shown in Figure 4.9-(a)). This aid conforms to XSC by supplying users with a wider *contextual* perspective of the robot’s navigational reasoning and by rendering the *predicted* effects of issuing reversal commands.

Certain design considerations are taken to avoid distracting users throughout operation, which can pose a problem for HMDs (Sibirtseva et al., 2018). The red obstacle spheres are instantiated at the physical targets detected by LiDAR sensors attached to the mobile base, thus providing an accurate depiction of the robot’s collision-checking process. Whenever these spheres fall outside of the egocentric FoV, small arrows appear in the navigator’s periphery as a non-obtrusive indicator. Both the panel and rear-view are robot-*embodied* cues that are rigidly attached to the mobile base. In order to not clutter an operator’s FoV, these two virtual objects are positioned outside the natural viewing angle of wheelchair navigation.

### 4.3.3 Experiments

We conducted a within-subjects experiment to investigate the influence of using XSC with an AR HMD whilst operating our robotic wheelchair platform.

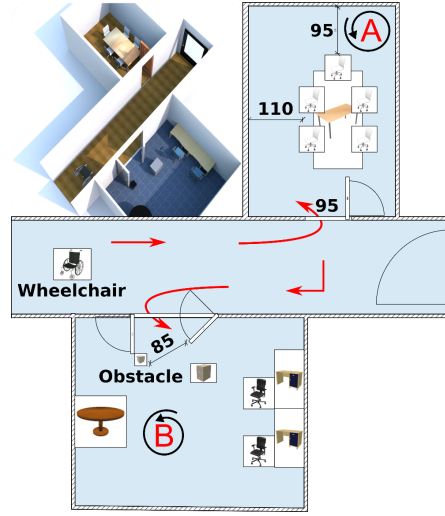


Figure 4.10: A floor plan of the navigation route used for a single experiment trial, with a 3D overhead view in the top-left corner. Each participant is requested to manoeuvre from the wheelchair position, to goal 'A', and then to 'B', before finally returning to the starting wheelchair position. The small numeric labels denote the centimetre widths of spaces along the route.

As before, all navigation processes are implemented atop [ROS](#) ([Quigley et al., 2009](#)) and run using an on-board laptop. The Unity 3D game engine was again utilised for [AR](#) application development and deployed on a HoloLens ( $30^\circ \times 17.5^\circ$  [FoV](#)) that has been supplemented with the Pupil Labs add-on for eye tracking ([Kassner et al., 2014](#)).

Our hypotheses for this indoor navigation study are:

- $H_1$  – [AR](#) will improve the mental model accuracy of participants, as determined by their quicker recovery times from jarring events linked with model misalignment.
- $H_2$  – Visualisations in [AR](#) will reduce the cognitive workload on subjects based on their less variable eye gaze distribution profiles.

#### 4.3.3.1 Experimental Setup & Protocol

We invited a total of 18 able-bodied volunteers (4 female, 14 male) aged 22-65 (median: 25) to take part in the experiment. Once again, the gender discrepancy is owed to the use of engineering students (see Section 4.2.2.1). Participants reported their familiarity with powered mobility, robotic wheelchairs and [AR](#), with the most common grouping reporting no prior experience in any.

A 2D floor plan of the experimental route is illustrated in Figure 4.10. Each subject was requested to complete this route twice, once with and without the aid of the proposed visualisations. To counterbalance the effects of trial order, even-numbered participants performed their first trial with visual feedback before proceeding onto their second trial without, and vice versa for the odd-numbered. Participants wore the headset across both trials for the purpose of data collection and fairness in comfort.

An entire experiment run took approximately 30 minutes and consisted of the following five phases: (1) preliminaries, (2) training if necessary, (3) eye gaze calibration, (4) navigation task and (5) post-experiment questionnaire. Phase (1) asked volunteers to fill out an introductory questionnaire, sign a consent form and watch video demonstrations of the AR assistance. Phase (2) was an optional 5-minute training of wheelchair control, especially for those with no prior experience. Step (3) involved fitting the HoloLens on the subject and calibrating the eye tracker using the plugin for HMDs (Kassner et al., 2014). Phase (4) consisted of the two navigation trials, followed by a post-experiment questionnaire in stage (5).

There are two unique locations along the navigation route that test the benefits of XSC guidelines on AR interface design. First, the passageway to goal 'A' includes a chair in the top-left corner of the room, which leads to a tight bend around the table that is challenging to manoeuvre (*contextual*). Second, the obstacle box located in the office is small and requires advance notice for smooth circumvention without regular downward glances (*predictive* – see Figure 4.11).

#### 4.3.3.2 Evaluation Metrics

A variety of task-specific metrics are examined to evaluate the XSC system. Aside from standard performance measures for mobile robotics, such as time-to-completion, the focus of evaluation is also directed towards human factors associated with model misalignment in SC.

One notable human factor that has previously been investigated for its impact on SC is cognitive workload (Carlson and Demiris, 2012). Self-reported questionnaires are often utilised to assess this metric, but eye gaze is known to be correlated with heightened workload or difficulty in manoeuvring a wheelchair (Carlson and Demiris, 2012; Simpson, 2008). More specifically, prior literature has drawn connections between gaze-based attention and the accuracy of mental models (Goodrich and Olsen, 2003). Higher degrees of eye movement are linked to cognitive overload, which coincides with an inappropriate mental model on the basis of a less efficient interaction (Carlson



Figure 4.11: Illustrates two participants navigating the office doorway entrance (refer to map in Figure 4.10) mid-trial. The volunteer on the left had to regularly attend to the obstacle box without any graphical aid. Contrarily, the volunteer on the right could manoeuvre around the box with little need to perform downward glances due to the graphical assistance.

and Demiris, 2009, 2012). As a result, we report eye gaze patterns as a physiological measure on the mental models of participants.

Additionally, we record the time to traverse specific navigation events that are relevant to XSC and any task load incurred due to model mismatch. Bypassing doorways and avoiding incidents where the wheelchair gets stuck are both prominent issues for powered mobility (Simpson, 2008). Given the role of transparent assistance in these situations, we identified events where participants encountered doors and “stucks”, so as to record the time it takes to overcome such events. Door positions are set by their midpoints and tracked as events whenever they fall within the wheelchair’s footprint. Likewise, “stucks” occur whenever the wheelchair does not escape its own clearance for a duration of 10 seconds.

Lastly, a post-experiment survey was handed out to volunteers for subjective feedback. The survey asked users to rate the benefit of the different visualisations and their general perceptions of the overall system (5-point Likert scale).

#### 4.3.3.3 Empirical Findings

Fixating on the evaluation of transparency attributed to XSC, Figure 4.12 demonstrates traversal times for the events described in Section 4.3.3.2. The results indicate that volunteers recovered faster from “stucks” (related t-test,

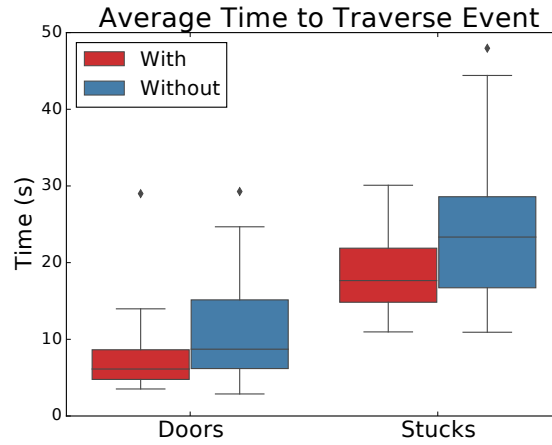


Figure 4.12: Average timing results across commonly occurring events that can result from model misalignment in SC for robot-assisted navigation. For both doorways and “stucks”, participants overcame these circumstances faster in their visualisation-aided trial, highlighting the benefit of predictive cueing at inciting quicker recovery times.

$p=0.03$ ) when guided with AR (median: 18.61s, IQR=22.7-15.17s) than when not (median: 26.23s, IQR=44.41-16.88s). Similarly, doorways had quicker traversal times (related t-test,  $p=0.093$ ) on trials with visualisations (median: 6.13s, IQR=8.64-4.77s) in comparison to without (median: 8.71s, IQR=15.14-6.19s). These findings support  $H1$  in that subjects overcame hazardous and otherwise jarring incidents more effectively, hinting at higher mental model accuracy.

In the post-survey responses, participants made various comments that reinforce the quantitative findings on event traversal times. Some claimed that the aids “helped with understanding *where* the problem was”, as well as “explaining *why* the safety algorithm was changing the way the wheelchair behaved”. These observations echo the XSC guidelines discussed in Section 4.3.1 and fortify the value of explicating SC, such that human operators can quickly recover from model misalignment.

Time-to-completion per participant is shown in Figure 4.13. We report average relative improvement in timings, with median values of 16.05% (IQR=24.91-12.86%) when switching to visualisations on the second trial and 13.75% (IQR=21.36-8.49%) when visualisations were instead removed. Despite the positive trend in improvements, the experimental scenario only involves two trials and thus cannot discount the possibility of a larger effect due to natural learning rates between successive trials. As stated in Section 4.2.3, the issue of statistical power persists in this study as well, meaning

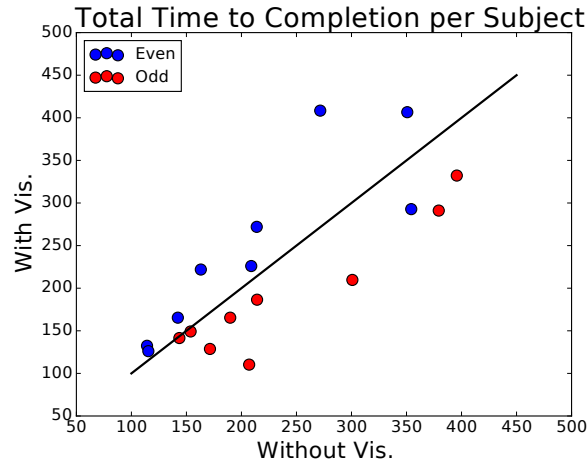


Figure 4.13: Time-to-completion per participant, with and without visualisations. Even-numbered subjects had graphical aid for their first trial, and odd-numbered subjects on their second. Points below the main diagonal show subjects that attained lower times with visual aid in comparison to without.

if more subjects were available then an investigation into the learning effects could prove fruitful.

Joint angular gaze distribution plots of eye tracking results are shown in Figure 4.14. Adhering to the manufacturer’s guidelines for our selected eye tracker, we filtered all gaze points below a specified confidence threshold before generating these plots. For the angular coordinates with visualisations, Figure 4.14-(a) presents mean angles of  $-1.74^\circ$  (SD  $9.63^\circ$ ) and  $-7.57^\circ$  (SD  $9.88^\circ$ ) in the horizontal and vertical directions, respectively. Contrarily, Figure 4.14-(b) presents for non-visualisation trials mean angles of  $-3.97^\circ$  (SD  $11.39^\circ$ ) and  $-12.54^\circ$  (SD  $9.97^\circ$ ) in the horizontal and vertical directions, respectively.

Although no strong claims can yet be made for the relationship between these gaze patterns and mental models (related t-test,  $p > 0.1$ , rejecting  $H_2$ ), there are still a few important observations. First, subjects occupied the negative vertical region less frequently in the visually-aided trial, signifying that they could successfully complete the task without repeatedly glancing downwards for obstacles (exemplified in Figure 4.11). This implies that the AR HMD interface provided an easily accessible source of contextual information regarding the SC. Furthermore, volunteers maintained a more centrally-oriented gaze when operating the wheelchair with visualisations than without. A reduction in variability could be indicative of lower mental workload during the prescribed navigation task (Carlson and Demiris, 2012). Lastly, greater care may need to be taken to not divert user attention

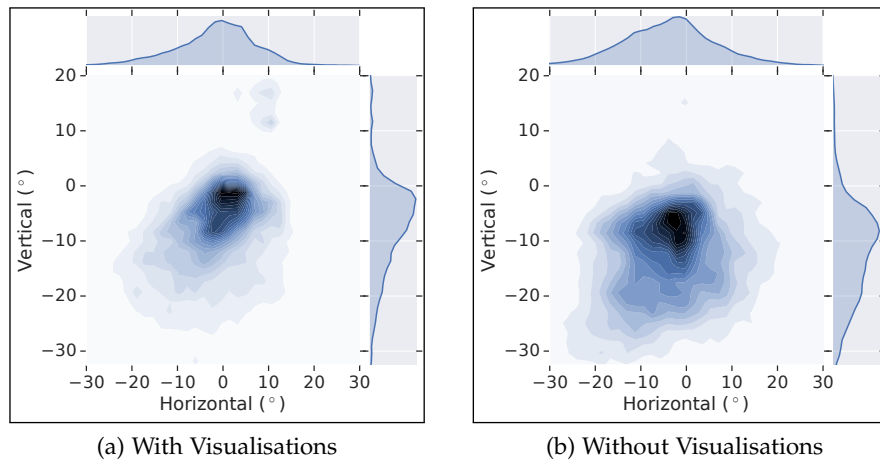


Figure 4.14: Joint angular gaze distribution across all participants along the horizontal and vertical axes. When operating the wheelchair with visualisations **(a)**, volunteers maintain a more centrally-oriented gaze angle and perform less downward glances in comparison to without **(b)**.

through visual explanations, as there are more instances of looking upwards in visualisation trials.

#### 4.3.3.4 Survey Results

An uplifting result from the survey responses was the positive user appraisal of all AR visualisations. Table 4.3 lists the average scores from 1-5 (5 being most positive) for the graphical aids. Most notably, the red collision indicators garnered nearly universal approval ( $4.39 \pm 0.92$ ) and even surpassed the popularity of the rear-view display ( $4.06 \pm 0.94$ ). Although the mini-map was less appreciated ( $3.22 \pm 1.17$ ), the result was still in favour of its inclusion.

Table 4.4 presents general opinions on the overall system, which promote the efficacy of conforming to XSC through AR HMDs. The general disagreement with the notion of feeling *distracted* ( $2.17 \pm 0.99$ ) asserts that our visualisations did not have any inherently misleading effects. In line with the aim to administer transparent assistance, subjects responded with high levels of *clarity* on the purpose of each virtual aid ( $4.11 \pm 0.68$ ). Finally, the positive

Table 4.3: Summary of responses to: “Rate the benefit of the following visualisation from 1-5 (1 = very poor, 5 = very good)”.

Visualisation	Mean User Rating	Standard Deviation	Mode
Red Collision Indicators	4.39	0.92	5
Mini-map Display	3.22	1.17	3
Rear-view Display	4.06	0.94	5

Table 4.4: Summary of subjective responses to general perceptions from 1-5 (1 = strongly disagree, 5 = strongly agree).

Question	Mean Rating $\pm$ Std. Dev.	Mode
I felt <i>clarity</i> on the visualisations	$4.11 \pm 0.68$	4
I found the visualisations <i>effective</i>	$3.67 \pm 0.84$	4
I felt <i>distracted</i> by the visualisations	$2.17 \pm 0.99$	2

tendency for users to find the visualisations *effective* at predicting the robot’s behaviour ( $3.67 \pm 0.84$ ) bolsters the prospects of predictive cueing in AR.

#### 4.3.4 Discussion

By applying a new interface design approach, the benefits of integrating AR HMDs onto “smart” mobility platforms have begun to emerge. The results from our user study demonstrate a reduction in the traversal times of doorways and trapped situations, as well as diminished head movement, all of which are cognitively straining everyday challenges for wheelchair users (validating  $H1$ ). Moreover, the recorded eye gaze data has provided tentative insight into how the introduced AR setup may expose users to less mentally demanding distribution profiles, albeit the patterns are not statistically significant yet (rejecting  $H2$ ). As with our previous study, this may partially be attributed to the lack of statistical power and choice of non-regular wheelchair users as participants (see Section 4.2.3).

A couple of drawbacks that persist in this updated architecture are related to the hardware. First, the LiDAR sensors only take a horizontal slice of the 3D space and thereby fail to capture a holistic perspective of what the wheelchair operator sees. This issue transfers into the AR visualisations by only presenting obstacle indicators at the height of the planar scans and constraining map portrayal to 2D grids. Second, the restricted FoV and heftiness of current HMDs could hinder the assistance of the actual SC, especially over longer durations, e. g. due to user fatigue (Sibirtseva et al., 2018). Nevertheless, sufficient technological advancements are expected in this area, rendering the problem obsolete.

In contrast to our earlier study presented in Section 4.2.2, deploying *contextual* and *predictive* AR visual aids have greatly enhanced the likelihood of potential users adopting our system. Unlike before, our updated AR HMD interface has neither distracted nor negatively impacted subjects during the indoor navigation task. In fact, this follow-up study revealed that subjects exhibited quicker recovery times from adverse events that are typically en-

countered during model misalignment in wheelchair navigation. Overall, these findings advocate the positive societal impact of AR headsets for assistive wheelchair navigation and act as a stepping-stone in paving forth this field of research.

#### 4.4 CONCLUSIONS

In this chapter, XSC was introduced to resolve the model misalignment problem that frequents many HRIs. Answering research question (2), a novel AR system was presented in Section 4.2, where a Microsoft HoloLens acted as a head-mounted aid for smart wheelchair navigation. Through an evaluation of this system, we gleaned preliminary insights into the beneficial and adverse nature of different AR cues for assistive navigation. In particular, we asserted that care should be taken in the presentation of information, with effort-reducing cues for augmented information acquisition (e. g. a rear-view display) being the most appreciated. Section 4.3 then delineated XSC and instantiated the paradigm for assistive navigation, where an AR headset played the integral role of visually demystifying the SC. Experimental results on the effectiveness of XSC demonstrated quicker user recovery times from adverse situations commonly encountered during model misalignment, resolving question (3).

There are many ways of instantiating XSC, however the guidelines applied in this chapter set a precedent for multiple application domains. In medical applications, such as surgical navigation, the weight of impeding performance through inappropriately placed virtual objects can have life-threatening consequences (Dixon et al., 2013). *Predictive* aids that augment the surgeon’s trajectory planning and *contextual* awareness are an exemplar use-case of XSC. Likewise, we anticipate that designing SC mechanisms to exhibit *causality* and *abstraction* could help disambiguate robot intentions in other HMD-based human-robot collaborations, such as aerial navigation (Walker et al., 2018) and shared workspace manipulation (Sibirtseva et al., 2018).

Future work could explore how to generalise the XSC paradigm to be robot-agnostic and less task-dependent. The resounding appreciation for the red collision indicators of Section 4.3.2.2 suggests that visualisations augmenting the *environment* are perhaps more tightly coupled with XSC, warranting further investigation in other task domains. Another worthwhile avenue of research is to formalise the XSC guidelines and trace exactly how they reconcile model misalignment over a continuous interaction instead of solely on a session basis. One way of testing this could be to purposefully

inject incorrect robot behaviours into the SC and observe how user misconceptions are corrected with different visualisations.

This chapter has been dedicated to the human’s perspective of robot intent without addressing the robot’s perspective of human intent, i. e. “robot-of-human” transparency (Lyons, 2013; Lyons and Havig, 2014). In the next chapter, we complete our XSC objective for transparency by contributing a probabilistic inference framework for robots to learn human intentions from observed behaviour. Crucially, the model is *interpretable* and thereby enables explanations to be administered to both system designers, as well as end-users.



## DISENTANGLED SEQUENCE CLUSTERING FOR HUMAN INTENTION INFERENCE

---

This chapter answers our last research question:

“How can an interpretable model be developed for robots to infer human intentions without making any assumptions about specific task constraints?”

Equipping robots with the ability to estimate human intent satisfies the robot’s viewpoint of transparency in Explainable Shared Control (XSC). Many computational approaches towards this objective employ probabilistic reasoning to recover a distribution of “intent” conditioned on the robot’s perceived sensory state. When adopting such a probabilistic stance, the intention estimation of Shared Control (SC) can be regarded as an *inference* problem. However, most approaches to this problem assume task-specific representations of human intent (e. g. labelled goals) are known *a priori*.

In this chapter, we overcome such task-oriented constraints by proposing an original clustering framework – the Disentangled Sequence Clustering Variational Autoencoder (DiSCVAE) – to learn a distribution of human intent in an *unsupervised* manner. The DiSCVAE is a subset of the Variational Autoencoder (VAE) (Kingma and Welling, 2013; Rezende et al., 2014), a widely used generative model for learning complex distributions over large, high-dimensional datasets. Moreover, the DiSCVAE incorporates ideas from sequence learning networks, e. g. Recurrent Neural Networks (RNNs), combined with VAEs to efficiently infer latent variables over sequential data (Chung et al., 2015; Fraccaro et al., 2016; Goyal et al., 2017; Krishnan et al., 2017). Though unlike previous sequence-based frameworks, the proposed variant exploits the notion of *disentanglement* in representation learning (Bengio et al., 2013) to infer a discrete (or categorical) variable for the purpose of clustering. The role of disentanglement in modelling a distribution of intent will be made apparent later in the chapter.

Two sets of experiments are conducted to evaluate the DiSCVAE. First, we validate its general capacity to discover high-level classes over sequential data by testing on an unlabelled video dataset of bouncing digits, known as Moving MNIST (Srivastava et al., 2015). For this dataset, the classes that

must be extracted from the aforementioned categorical variable are the individual digit identities. Importantly, we also report findings from a real-world Human-Robot Interaction (HRI) experiment involving our smart wheelchair platform. The primary ambition here is to glean insights into how the inferred categorical variable coincides with human intent in this setting (see Figure 5.1 for an overview of the experiment setup).

The chapter is organised as follows. Section 5.2 motivates why enabling robots to infer human intent in unconstrained task scenarios is vital for HRI. In Section 5.2, we describe the preliminary material necessary to define our DiSCVAE in Section 5.3. Experimental results on the Moving MNIST and robotic wheelchair domains are presented in Sections 5.4 and 5.5. Section 5.6 discusses techniques that bear close ties with the DiSCVAE and its application to intention inference. Finally, Section 5.7 concludes with the implications of this work and its future extensions. Note that we maintain the same terminology for intent throughout this chapter as in Section 2.4.2.

Research from this chapter has been submitted to a peer-reviewed journal.

## 5.1 MOTIVATION

Humans are remarkably proficient at accurately and rapidly inferring the implicit intentions of others from their overt behaviour (Blakemore and Decety, 2001; Tomasello et al., 2005). Consequently, they are adept at planning their own actions when collaborating with one another in shared physical environments. It therefore stands to reason that intention inference may be equally imperative in creating fluid and effective HRIs. Robots endowed with this ability have been extensively explored in collaborative robotics (Demiris, 2007; Jain and Argall, 2019; Losey et al., 2018), yet their migration into real-world settings remains an open research problem.

One major impediment to real-world instances of human intention inference is the assumption that a known representation of intent exists. For example, most prevalent frameworks in collaborative robotics assume a discrete set of task goals is known *a priori*. Under this assumption, the robot can infer a distribution of human intent by applying Bayesian reasoning over the entire goal space (Hu et al., 2018; Jain and Argall, 2019; Javdani et al., 2015). Whilst such a distribution offers a versatile and practical representation of intent, the need for predefined labels to acquire it is not always feasible or realistic unless restricted to a specific task scope (Locatello et al., 2019).

Another challenge in estimating human intent is that many diverse actions often fulfil the same intention (Jordan and Rumelhart, 1992). A prominent

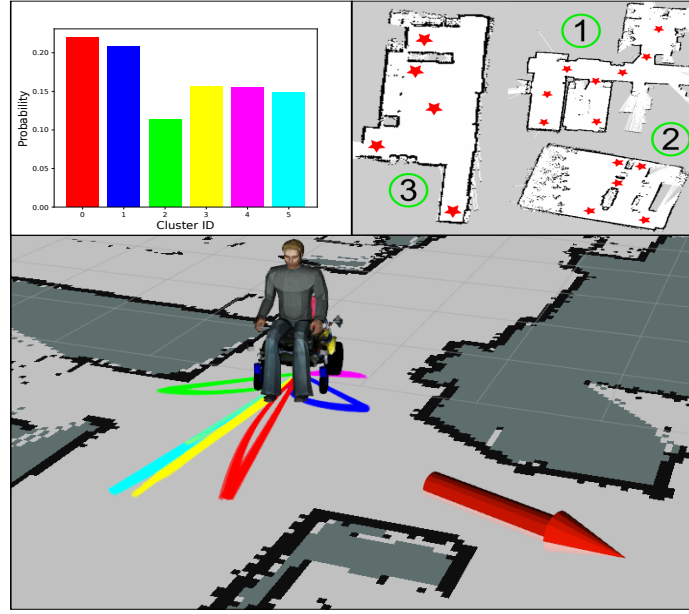


Figure 5.1: Overview of the intention inference experiment on a robotic wheelchair (see Figure 3.1 for the real robot platform). **Bottom:** Recorded output of an actual human subject navigating towards a goal (red arrow), which is visualised here as an animated simulation. **Top Right:** Maps of the three experiment settings, with red stars denoting target locations. **Top Left:** Probability histogram of the categorical variable modelling “intentions” at this particular snapshot of the data for  $K=6$  clusters. The bars are coloured to align with the wheelchair trajectories generated by sampling from the corresponding clusters. Multiple diverse trajectories can be sampled from the same cluster and each trajectory’s length is dependent on the velocity commands drawn from the generative model.

class of probabilistic algorithms that aptly tackle this challenge are generative models, which derive a distribution of observations by introducing latent random variables to capture any hidden underlying structure. Within the confines of intention inference, the modelled latent space can then be presumed to represent all possible causal relations between intentions and observed human behaviour (Hu et al., 2019; Tanwani and Calinon, 2017; Wang et al., 2013). The advent of deep generative models, such as VAEs (Kingma and Welling, 2013; Rezende et al., 2014), has also made it possible to efficiently infer this latent space from abundant sources of highly complex data.

Inspired by the prospects of not only extracting hidden “intent” variables but also interpreting their meaning, we frame the intention inference problem as a process of *disentangling* the latent space. Disentanglement is a core research direction in representation learning, and refers to the recovery of abstract concepts from independent factors of variation that are assumed to

be responsible for generating the observed data (Bengio et al., 2013; Locatello et al., 2019; Tschannen et al., 2018). These independent factors could be the handwriting style of digits in the MNIST dataset (Kingma et al., 2014), or the orientation and motion of objects in videos (Hsieh et al., 2018; Yingzhen and Mandt, 2018), or even the speaker identity in audio signals (Hsu et al., 2017, 2018; van den Oord et al., 2017). The interpretable structure of such disentangled representations is exceedingly desirable for human-in-the-loop scenarios (Fortuin et al., 2019), like robotic wheelchair assistance. Despite how this desirable quality has spurred on considerable advances in representation learning algorithms, very few have transferred over to the robotics domain (Hu et al., 2019).

As a result, we strive to bridge this gap by proposing an *unsupervised* clustering framework for human intention inference that circumvents the barriers to utility under unconstrained task conditions. Capitalising on prior disentanglement techniques for sequence modelling, we learn a latent representation of sequential observations (e.g. of human behaviour) that divides into a local (time-varying) and global (time-preserving) part (Hsieh et al., 2018; Hsu et al., 2017; Yingzhen and Mandt, 2018). Though unlike previous approaches, our variant simultaneously infers a categorical variable to construct a mixture model with the continuous global variable. Each cluster thereby grants an *interpretable* way of inferring discrete high-level features, e.g. the navigation intentions of a wheelchair user. The overall framework is generally suited for class discovery in sequences.

## 5.2 PRELIMINARIES

Before defining our clustering framework for intention inference, the following describes principles from representation learning that underpin its operation. As the VAE acts as the basis of the DiSCVAE, we begin with a brief overview of its foundations. We then examine how to tailor VAEs to sequential data, as they are not directly suitable for time-series analysis in their original form.

### 5.2.1 Variational Autoencoders

Deep generative models are density estimators of data that rely on neural networks to predict probability distribution parameters. In the context of VAEs (Kingma and Welling, 2013; Rezende et al., 2014), the generative process to acquire the joint distribution  $p_\theta(\mathbf{x}, \mathbf{z})$  follows two steps. A latent vari-

able  $\mathbf{z}$  is first drawn from a prior  $p_\theta(\mathbf{z})$  (often a multivariate Gaussian), and then observations  $\mathbf{x}$  are reproduced from a conditional distribution  $p_\theta(\mathbf{x}|\mathbf{z})$ . The ‘deep’ aspect here relates to how the parameters  $\theta$  of these distributions are learnt by a neural network with non-linear activation functions, sometimes termed the *generative network*.

Of central interest to VAEs and Bayesian inference in general is the sought-after posterior  $p_\theta(\mathbf{z}|\mathbf{x})$ . However, the integrals required to evaluate this posterior are often intractable, especially with complicated likelihood functions, like non-linear neural networks. To bypass this problem, an approximation  $q_\phi(\mathbf{z}|\mathbf{x})$  of the true posterior is instead computed, where parameters  $\phi$  are learnt via a *recognition network* (Kingma and Welling, 2013; Rezende et al., 2014). VAEs specifically perform a mean-field approximation of  $q_\phi(\mathbf{z}|\mathbf{x})$ , meaning the resulting variational distribution is fully factorised by assuming independence across latent variables. This entire methodology of approximating the inference process for scalable Bayesian modelling falls under the umbrella of variational inference (Zhang et al., 2019).

The VAE is thus a deep latent variable model (displayed in Figure 5.2a) that consists of both a generative and recognition network (Kingma and Welling, 2013; Rezende et al., 2014). Training this model can be achieved by maximising the marginal log-likelihood of the data  $\log p_\theta(\mathbf{x})$ , or equivalently, maximising the Evidence Lower Bound (ELBO)  $\mathcal{L}(\mathbf{x}; \theta, \phi)$  as a surrogate objective function:

$$\begin{aligned} \log p_\theta(\mathbf{x}) &\geq \mathcal{L}(\mathbf{x}; \theta, \phi) \\ &\equiv \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[ \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \\ &\equiv \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_\theta(\mathbf{z})), \end{aligned} \tag{5.1}$$

where the first term of the last line can be viewed as a reconstruction error and the second Kullback-Leibler (KL) divergence term as a regulariser that encourages the variational posterior  $q_\phi(\mathbf{z}|\mathbf{x})$  to be close to the prior  $p_\theta(\mathbf{z})$ .

Key to the work of Kingma and Welling (2013); Rezende et al. (2014) is an efficient training scheme to jointly learn the network parameters  $\theta$  and  $\phi$  that optimise the ELBO. Whilst it would be attractive to use stochastic gradient descent for this optimisation (as is typical in neural networks), differentiating the ELBO expectations is either impractical or impossible. Seeking to overcome such impracticality, Kingma and Welling (2013); Rezende et al. (2014) applied the *reparameterisation trick* to estimate expectation gradients through Monte Carlo sampling. In essence, the trick expresses continuous

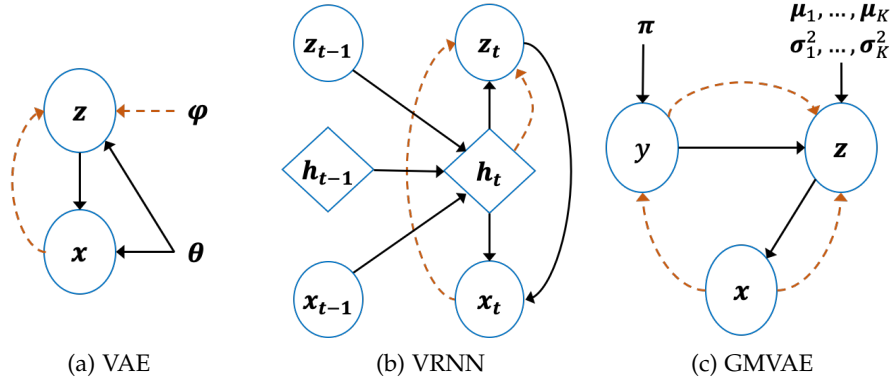


Figure 5.2: Deep generative models for: (a) variational inference (Kingma and Welling, 2013; Rezende et al., 2014); (b) a sequential VAE that conditions on the deterministic hidden states of an RNN at each timestep (VRNN Chung et al., 2015); (c) a VAE with a Gaussian mixture prior (GMVAE). Dashed lines denote inference and bold lines indicate generation.

latent variables in a deterministic form that is differentiable for stochastic gradient descent, thereby facilitating variational inference over large quantities of data.

We have presented the seminal VAE, but it does not account for time-series analysis in its original form, and so the next section examines variational inference for sequences. For notational simplicity, parameters  $\phi$  and  $\theta$  learnt by recognition and generative networks will be omitted hereafter.

### 5.2.2 Variational Inference for Sequences

Most real-world data are characterised by time-varying attributes, motivating the development of deep generative models for sequences. Nevertheless, the multi-layered structure of sequential data is a challenging modelling task where variables at different timesteps are highly correlated and cannot be assumed independent as in the factorised variational distribution of a VAE. In order to explicitly capture these correlations, a recurrence relationship must be established between the internal states of the latent variable model.

State Space Models (SSMs), such as Hidden Markov Models (HMMs), are probabilistic graphical models that accomplish this feat using structured variational inference. The structured qualifier denotes how the variational approximation directly models dependencies between *stochastic* variables, rather than factor them out (Zhang et al., 2019). Despite the rich history associated with SSMs, they have only recently been merged with neural networks

to leverage the expressive power of [RNNs](#) for sequence learning ([Fraccaro et al., 2016](#); [Krishnan et al., 2017](#)). Deep [SSMs](#) are therefore graphical models formed of structured inference (or recognition) networks that are parameterised by [RNNs](#) and optimised according to the [VAE](#) learning principle.

Alternatively, [RNNs](#) can be augmented to include latent variables and solely model connections between *deterministic* states. By persisting the recurrent connections of hidden states with themselves, the [RNN](#) retains its autoregressive nature. A notable example is the [VRNN](#) ([Chung et al., 2015](#)) (shown in Figure 5.2b), which differs from [SSMs](#) by *indirectly* conditioning on random variables and observations from previous timesteps through the deterministic hidden state,  $\mathbf{h}_t(\mathbf{x}_{t-1}, \mathbf{z}_{t-1}, \mathbf{h}_{t-1})$ . This leads to a joint distribution over the observation sequence and latent states:

$$\begin{aligned} p(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T}) &= \prod_{t=1}^T p(\mathbf{x}_t | \mathbf{z}_{\leq t}, \mathbf{x}_{< t}) p(\mathbf{z}_t | \mathbf{x}_{< t}, \mathbf{z}_{< t}) \\ &= \prod_{t=1}^T p(\mathbf{x}_t | \mathbf{z}_t, \mathbf{h}_t) p(\mathbf{z}_t | \mathbf{h}_t), \end{aligned} \quad (5.2)$$

where the true posterior  $p(\mathbf{z}_t | \mathbf{h}_t)$  is conditioned on information pertaining to previous observations  $\mathbf{x}_{< t}$  and latent states  $\mathbf{z}_{< t}$ , hence accounting for temporal dependencies. The [VRNN](#) state  $\mathbf{h}_t$  is also shared with the inference procedure to yield the following variational posterior distribution:

$$\begin{aligned} q(\mathbf{z}_{\leq T} | \mathbf{x}_{\leq T}) &= \prod_{t=1}^T q(\mathbf{z}_t | \mathbf{x}_{\leq t}, \mathbf{z}_{< t}) \\ &= \prod_{t=1}^T q(\mathbf{z}_t | \mathbf{x}_t, \mathbf{h}_t). \end{aligned} \quad (5.3)$$

Deciding how to incorporate stochastic variables into a recurrent model is an architecture choice ([Goyal et al., 2017](#); [Zhang et al., 2019](#)). Whilst deep [SSMs](#) ([Fraccaro et al., 2016](#); [Krishnan et al., 2017](#)) provide a tighter [ELBO](#) than that of a [VRNN](#), they lose autoregressive structure by not conditioning on the deterministic hidden state ([Goyal et al., 2017](#)). A recent framework aiming to unify several of these architecture choices emphasised the benefit of indirectly conditioning on stochastic latent variables to encode a “plan” about future states during inference ([Goyal et al., 2017](#)). As our preliminary results resonated with these findings, the [DiSCVAE](#) specified in the next section also elects an approach akin to a [VRNN](#) ([Chung et al., 2015](#)).

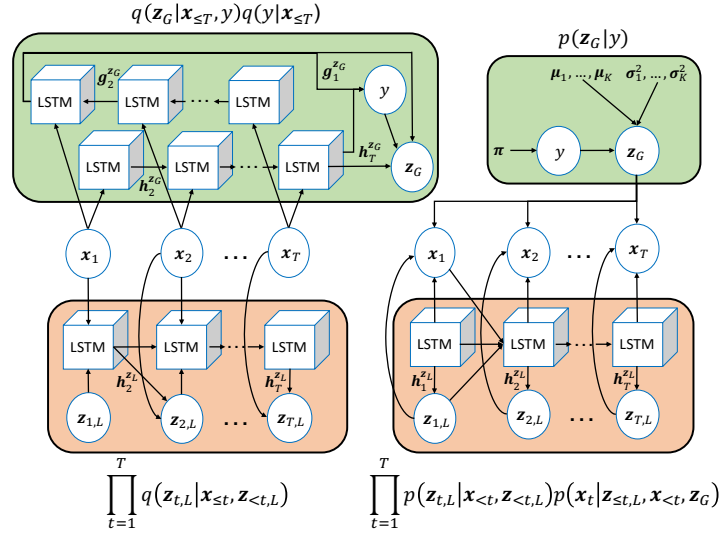


Figure 5.3: Computation graph of the inference  $q(\cdot)$  and generative  $p(\cdot)$  networks. **Green** blocks contain the global variables  $y$  and  $z_G$ , with a bidirectional **LSTM** used to condition over the input sequence  $x_{\leq T}$ . Hidden states  $h_t^{z_G}$  and  $g_1^{z_G}$  then compute the  $q(\cdot)$  distribution parameters. **Orange** blocks encompass the local sequence variable  $z_{t,L}$ , where a regular **LSTM** updates its internal states  $h_t^{z_L}$  at each timestep and is combined with current inputs  $x_t$  during inference of  $z_{t,L}$ . Generating  $x_t$  requires both  $z_G$  and  $z_{t,L}$ .

### 5.3 THE DISENTANGLED SEQUENCE CLUSTERING VARIATIONAL AUTOENCODER

In this section, we introduce the **DiSCVAE** (graphically shown in Figure 5.3), a probabilistic clustering framework suited for human intention inference. Our method of clustering is initially presented as an adaptation of the standard **VAE** to incorporate a Gaussian mixture prior and a categorical variable. We then formulate the **DiSCVAE** by combining this clustering **VAE** with a sequential latent variable model capable of *disentanglement*. Finally, we discuss how everything fits into the scope of intention inference.

#### 5.3.1 Clustering with Variational Autoencoders

A key notion of learning “good” representations through a generative model is to express a prior capable of naturally clustering the data space (Bengio et al., 2013). Previous research on **VAEs** has pursued this objective and segmented the latent space into distinct classes by applying a Gaussian mixture

prior instead of the standard unimodal Gaussian (Dilokthanakul et al., 2016; Jiang et al., 2017), sometimes referred to as a **GMVAE**.

Our approach is similar, with the exception of two modifications. First, we leverage the categorical reparameterisation trick (Jang et al., 2016; Maddison et al., 2016) to obtain differentiable samples of *discrete* variables and thereby enable stochastic gradient descent for model optimisation. This differs from the standard reparameterisation trick that exclusively operates on continuous variables (Kingma and Welling, 2013; Rezende et al., 2014). Second, we alter the **ELBO** objective to mitigate the precarious issues of posterior collapse and cluster degeneracy (or mode collapse). Posterior collapse refers to the phenomenon of latent variables being ignored or overpowered by highly expressive decoders during training, such that the posterior mimics the prior, i.e. the KL divergence term in Equation (5.1) falls to zero (Higgins et al., 2017; Hsieh et al., 2018; van den Oord et al., 2017). On the other hand, when multiple modes of the prior have collapsed into one (e.g. a single cluster component), then this indicates mode collapse or cluster degeneracy in the case of mixture models (Dilokthanakul et al., 2016; Hsu et al., 2018; Shi et al., 2020).

The **GMVAE** used for this work (see Figure 5.2c) is outlined below. Assuming observations  $\mathbf{x}$  are generated according to some stochastic process with discrete latent variable  $y$  and continuous latent variable  $\mathbf{z}$ , and that the aim is to divulge  $K$  clusters, then we can write the joint probability as:

$$p(\mathbf{x}, \mathbf{z}, y) = p(\mathbf{x} | \mathbf{z})p(\mathbf{z} | y)p(y), \quad (5.4)$$

where

$$\begin{aligned} y &\sim \text{Cat}(\boldsymbol{\pi}) \\ \mathbf{z} &\sim \mathcal{N}(\boldsymbol{\mu}_z(y), \text{diag}(\boldsymbol{\sigma}_z^2(y))) \\ \mathbf{x} &\sim \mathcal{N}(\boldsymbol{\mu}_x(\mathbf{z}), \mathbf{I}) \text{ or } \mathcal{B}(\boldsymbol{\mu}_x(\mathbf{z})), \end{aligned}$$

and functions  $\boldsymbol{\mu}_z$ ,  $\boldsymbol{\sigma}_z^2$  and  $\boldsymbol{\mu}_x$  are neural networks whose outputs parameterise the distributions of  $\mathbf{z}$  and  $\mathbf{x}$ , respectively. More specifically, the generative process involves three steps: (1) sampling  $y$  from a categorical distribution  $\text{Cat}(y | \boldsymbol{\pi})$  parameterised by probability vector  $\boldsymbol{\pi}$ , with  $\pi_k$  set to  $K^{-1}$  in favour of an uninformative uniform prior; (2) sampling  $\mathbf{z}$  from the marginal prior  $p(\mathbf{z} | y)$ , which results in a Gaussian Mixture Model (**GMM**) with a diagonal covariance matrix and uniform mixture weights; and (3) generating data  $\mathbf{x}$  from a likelihood function  $p(\mathbf{x} | \mathbf{z})$ , e.g. a fixed unit variance Gaussian if real-valued ( $\mathbf{I}$  denoting the identity matrix) or a Bernoulli if binary.

A mean-field approximation  $q(\mathbf{z}, \mathbf{y} | \mathbf{x})$  of the true posteriors on  $\mathbf{z}$  and  $\mathbf{y}$  is introduced in its factorised form as:

$$q(\mathbf{z}, \mathbf{y} | \mathbf{x}) = q(\mathbf{z} | \mathbf{x}, \mathbf{y})q(\mathbf{y} | \mathbf{x}), \quad (5.5)$$

where both the normally distributed  $q(\mathbf{z} | \mathbf{x}, \mathbf{y})$  and categorical  $q(\mathbf{y} | \mathbf{x})$  are also parameterised by neural networks. Nonetheless, the reparameterisation trick does not directly apply to non-differentiable *discrete* samples (Kingma and Welling, 2013; Rezende et al., 2014). Instead, we employ a continuous relaxation of  $q(\mathbf{y} | \mathbf{x})$  during training, coined as the Concrete (Maddison et al., 2016) or Gumbel-Softmax (Jang et al., 2016) distribution<sup>1</sup>.

The ELBO objective for this clustering model is:

$$\begin{aligned} \mathcal{L}(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{z}, \mathbf{y} | \mathbf{x})} \left[ \log \frac{p(\mathbf{x}, \mathbf{z}, \mathbf{y})}{q(\mathbf{z}, \mathbf{y} | \mathbf{x})} \right] \\ &= \mathbb{E}_{q(\mathbf{z} | \mathbf{x}, \mathbf{y})} [\log p(\mathbf{x} | \mathbf{z})] \\ &\quad - \mathbb{E}_{q(\mathbf{y} | \mathbf{x})} [\text{KL}(q(\mathbf{z} | \mathbf{x}, \mathbf{y}) \| p(\mathbf{z} | \mathbf{y}))] \\ &\quad - \text{KL}(q(\mathbf{y} | \mathbf{x}) \| p(\mathbf{y})), \end{aligned} \quad (5.6)$$

where the first term acts as a reconstruction loss on observations  $\mathbf{x}$ , and the latter two terms push the variational posteriors to be close to their corresponding priors.

Optimising the ELBO for our GMVAE with a powerful decoder is prone to posterior collapse and mode collapse. Cluster degeneracy is particularly problematic for  $p(\mathbf{z} | \mathbf{y})$ , where the KL divergence term on  $\mathbf{y}$  opts to use the same mean and variance for each mixture component during training (Dilokthanakul et al., 2016; Hsu et al., 2018). To prevent both posterior collapse and cluster degeneracy, we exploit the relationship with index-code mutual information  $I(\mathbf{y}; \mathbf{x})$  between  $\mathbf{y}$  and  $\mathbf{x}$  (Hoffman and Johnson, 2016):

$$\begin{aligned} \mathbb{E}_{p(\mathbf{x})} [\text{KL}(q(\mathbf{y} | \mathbf{x}) \| p(\mathbf{y}))] &= I(\mathbf{y}; \mathbf{x}) \\ &= \mathcal{H}(\mathbf{y}) - \mathcal{H}(\mathbf{y} | \mathbf{x}) \\ &= \log K - \mathcal{H}(\mathbf{y} | \mathbf{x}). \end{aligned} \quad (5.7)$$

Provided that  $p(\mathbf{y})$  is uniform, directly maximising the entropy on  $q(\mathbf{y} | \mathbf{x})$  is equivalent to minimising  $\text{KL}(q(\mathbf{y} | \mathbf{x}) \| p(\mathbf{y}))$ . As with previous VAE-based clustering models (Hsu et al., 2018; Shi et al., 2020), we make this replacement in Equation (5.6) to alleviate posterior collapse and impose a constraint

<sup>1</sup> In practice, a continuous relaxation of a one-hot categorical distribution is utilised, such that samples from  $q(\mathbf{y} | \mathbf{x})$  are one-hot vectors.

that specialises clusters to individual observations. Empirical evidence supporting this argument is supplied in Section 5.4.

### 5.3.2 Model Specification

Having established a means of categorising the latent space learnt using VAEs for static data, we now derive the DiSCVAE as a sequential extension (graphically shown in Figure 5.3). The motivation behind our proposed framework is to automatically disentangle representations and cluster them into meaningful classes of information. In doing so, we aim to preserve the benefits of autoregressive latent variable models for prediction, as well as boost model *interpretability* via controlled generation.

Disentangling latent variables is an emerging area of research that has recently gained traction in deep generative models for sequences (Hsieh et al., 2018; Hsu et al., 2017, 2018; Yingzhen and Mandt, 2018). A common choice of disentangled representation amongst these models involves segregating into *sequence-level* and *segment-level* parts (Hsu et al., 2017). This has also been regarded as a split into static and dynamic, or time-invariant and time-dependent attributes (Hsieh et al., 2018; Yingzhen and Mandt, 2018). Inspired by these approaches, we denote our disentangled representation at timestep  $t$  as  $\mathbf{z}_t = [\mathbf{z}_G, \mathbf{z}_{t,L}]$ , where  $\mathbf{z}_G$  and  $\mathbf{z}_{t,L}$  encode *global* (time-invariant) and *local* (time-dependent) sequence characteristics, respectively.

The novelty of our approach lies in how we solely cluster the global (time-invariant) variable  $\mathbf{z}_G$  extracted from sequences. Related temporal clustering models have either mapped the entire sequence  $\mathbf{x}_{\leq T}$  to a discrete latent manifold (Fortuin et al., 2019) or inferred a categorical factor of variation  $y$  to cluster over an *entangled* continuous latent representation (Hsu et al., 2018; Shi et al., 2020). Whereas the DiSCVAE clusters high-level attributes  $\mathbf{z}_G$  in isolation from lower-level dynamics  $\mathbf{z}_{t,L}$ . The DiSCVAE model itself is an amalgamation of prior works, e.g. the VRNN (Chung et al., 2015), GMVAEs (Dilokthanakul et al., 2016; Jiang et al., 2017) and sequence disentanglement (Yingzhen and Mandt, 2018), however its formulation plays a symbolic role in our interpretation of intention inference, as is made apparent in Section 5.3.4.

Using the clustering scheme described in Section 5.3.1, we define the generative model  $p(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T,L}, \mathbf{z}_G, y)$  as:

$$p(\mathbf{z}_G | y)p(y) \prod_{t=1}^T p(\mathbf{x}_t | \mathbf{z}_{t,L}, \mathbf{z}_G, \mathbf{h}_t^{\mathbf{z}_L})p(\mathbf{z}_{t,L} | \mathbf{h}_t^{\mathbf{z}_L}). \quad (5.8)$$

---

**Algorithm 5.1:** Sampling procedure to produce diverse predictions of novel states from the inferred cluster

---

**Inputs:** Sequence  $\mathbf{x}_{\leq t}$ ; sample length  $L$ ;  
**Output:** Predictions  $\tilde{\mathbf{x}}_{t+1}, \dots, \tilde{\mathbf{x}}_{t+L}$

- 1 Feed prefix  $\mathbf{x}_{\leq t}$  into inference model via Equation (5.9)
- 2 Assign cluster  $c$  using Equation (5.11)
- 3 Draw fixed sample from  $p(\mathbf{z}_G | y = c)$
- 4 **for**  $i \in \{t+1, \dots, t+L\}$  **do**
- 5      $\mathbf{h}_i \leftarrow \text{RNN}(\mathbf{z}_{i-1}, \mathbf{x}_{i-1}, \mathbf{h}_{i-1})$
- 6     Sample dynamics from  $p(\mathbf{z}_{i,L} | \mathbf{h}_i)$
- 7     Predict  $\tilde{\mathbf{x}}_i \sim p(\mathbf{x}_i | \mathbf{z}_{i,L}, \mathbf{h}_i, \mathbf{z}_G)$
- 8 **end**

---

The GMM prior  $p(\mathbf{z}_G | y)$  encourages mixture components (indexed by  $y$ ) to develop in the latent space of variable  $\mathbf{z}_G$ . Akin to a VRNN (Chung et al., 2015), the posterior of  $\mathbf{z}_{t,L}$  is parameterised through deterministic state  $\mathbf{h}_t^{\mathbf{z}_L}$ . It is also important to highlight the dependency on both  $\mathbf{z}_{t,L}$  and  $\mathbf{z}_G$  upon generating  $\mathbf{x}_t$ .

To perform posterior approximation, we adopt the variational distribution  $q(\mathbf{z}_{\leq T, L}, \mathbf{z}_G, y | \mathbf{x}_{\leq T})$  and factorise it as:

$$q(\mathbf{z}_G | \mathbf{x}_{\leq T}, y) q(y | \mathbf{x}_{\leq T}) \prod_{t=1}^T q(\mathbf{z}_{t,L} | \mathbf{x}_t, \mathbf{h}_t^{\mathbf{z}_L}). \quad (5.9)$$

As before, categorical  $y$  is injected into the inference process and relaxed to acquire Monte Carlo sample estimates of gradients during training (Jang et al., 2016; Maddison et al., 2016). An alternative variational distribution  $q(\mathbf{z}_{\leq T, L}, \mathbf{z}_G, y | \mathbf{x}_{\leq T})$  could use a “full” structure that conditions on  $\mathbf{z}_G$  and  $y$  as well (Yingzhen and Mandt, 2018), but we find the chosen “factorised”  $q(\cdot)$  to be more effective for our experimental domains. By conforming to a factorised structure, we assume that global features are independent of local dynamics, e.g. a human’s high-level intention to grasp an object does not correlate with the precise intricacies of their grasping behaviour. The variational posterior over  $\mathbf{z}_{t,L}$  could also summarise information from the future  $\mathbf{x}_{\leq T}$ , but Fraccaro et al. (2016) points out that this is unlikely to render improvements given the shared deterministic state  $\mathbf{h}_t^{\mathbf{z}_L}$ .

Under the VAE framework, the DiSCVAE is maximised according to the time-wise learning objective:

$$\begin{aligned}
\mathcal{L}(\mathbf{x}_{\leq T}) &= \mathbb{E}_{q(\cdot)} \left[ \log \left[ \frac{p(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T, L}, \mathbf{z}_G, \mathbf{y})}{q(\mathbf{z}_{\leq T, L}, \mathbf{z}_G, \mathbf{y} | \mathbf{x}_{\leq T})} \right] \right] \\
&= \mathbb{E}_{q(\cdot)} \left[ \sum_{t=1}^T \left( \log [p(\mathbf{x}_t | \mathbf{z}_{t, L}, \mathbf{z}_G, \mathbf{h}_t^{\mathbf{z}_L})] \right. \right. \\
&\quad \left. \left. - \text{KL}(q(\mathbf{z}_{t, L} | \mathbf{x}_t, \mathbf{h}_t^{\mathbf{z}_L}) \| p(\mathbf{z}_{t, L} | \mathbf{h}_t^{\mathbf{z}_L})) \right) \right. \\
&\quad \left. \left. - \text{KL}(q(\mathbf{z}_G | \mathbf{x}_{\leq T}, \mathbf{y}) \| p(\mathbf{z}_G | \mathbf{y})) + \mathcal{H}(q(\mathbf{y} | \mathbf{x}_{\leq T})) \right] \right].
\end{aligned} \tag{5.10}$$

This summation of lower bounds across timesteps,  $\mathcal{L}(\mathbf{x}_{\leq T})$ , is decomposed into: (1) the expected log-likelihood of input sequences; (2) KL divergences for variables  $\mathbf{z}_{t, L}$  and  $\mathbf{z}_G$ ; and (3) a measure of conditional entropy based on our preceding argument in Equation (5.7).

### 5.3.3 Network Architecture

The complete DiSCVAE network architecture is demonstrated in Figure 5.3. An RNN is used to parameterise the posteriors over  $\mathbf{z}_{t, L}$ , with the shared hidden state  $\mathbf{h}_t^{\mathbf{z}_L}$  allowing  $\mathbf{x}_{< t}$  and  $\mathbf{z}_{< t, L}$  to be indirectly conditioned on in Equations (5.8) and (5.9). For time-invariant variables  $\mathbf{y}$  and  $\mathbf{z}_G$ , a bidirectional RNN (Graves and Schmidhuber, 2005) is applied to extract feature representations over the entire sequence  $\mathbf{x}_{\leq T}$ , analogous to prior architectures (Fracaro et al., 2016; Krishnan et al., 2017; Yingzhen and Mandt, 2018). All RNNs have LSTM cells (Hochreiter and Schmidhuber, 1997), as these gated units excel at handling very long segments of spatio-temporal data (Sutskever et al., 2014). One-hidden layer Multilayer Perceptrons (MLPs) are also dispersed throughout to output the mean and variance of any Gaussian distributions, as per the VAE ideology. Decoded inputs  $\mathbf{x}_t$  at each timestep depend on the concatenated disentangled representation,  $\mathbf{z}_t$ .

### 5.3.4 Intention Inference

Let us now relate back to the problem of intention inference and how we deem intent as *both* a goal and a plan of action (Tomasello et al., 2005). Under such a unified representation, we claim that the latent class attribute  $\mathbf{y}$  models a K-dimensional repertoire of action plans for any specific task.

From this perspective, intention inference is a matter of assigning clusters to observations  $\mathbf{x}_{\leq T}$  of human behaviour and their environment (e.g. joystick commands and Light Detection And Ranging (LiDAR) sensor readings). Ultimately, human intent is computed as the most probable element of the component posterior:

$$c = \arg \max_k q(y_k | \mathbf{x}_{\leq T}), \quad (5.11)$$

where  $c$  is the assigned cluster identity, i.e. the inferred intention label  $x$  of the tuple  $i_x$  in Section 2.4.2. The *goal*  $g_x$  associated with this cluster is then modelled by  $\mathbf{z}_G$ , and local variable  $\mathbf{z}_{t,L}$  captures the various behaviours capable of accomplishing the inferred action *plan*  $u_x$  (see Section 2.4.2 for a reminder on the representation).

Aside from classifying observations, another major benefit of DiSCVAE is its capacity for controlled generation of sequences. Given an observation sequence  $\mathbf{x}_{\leq T}$ , we can infer  $c$  and then fix it to generate new sequences by conditioning on the model priors. Repeatedly sampling  $\mathbf{z}_{t,L}$  also allows for diversity in how the predicted trajectories  $\tilde{\mathbf{x}}_t$  pan out according to the global plan. The procedure of generating novel states (e.g. intent-driven behaviours) is summarised in Algorithm 5.1.

#### 5.4 VALIDATION SETTING: MOVING MNIST

As a means of validating our clustering framework, this section presents results on Moving MNIST (Srivastava et al., 2015), a dataset for video representation learning. Code for the DiSCVAE and experimentation is publicly available online<sup>2</sup>.

##### 5.4.1 Dataset and Implementation

Moving MNIST is a video dataset (Srivastava et al., 2015) comprised of multiple digits bouncing off their surrounding frame edges at random velocities. It has recently become a popular domain for investigating disentanglement amongst sequential latent variable models (Hsieh et al., 2018; Kosiorsek et al., 2018) due to its intuitive separation of dynamic and static sequence attributes, namely the motion and identity of each digit. The dataset also poses as a pertinent validation testbed for our work, as clustering on MNIST digits is a common ground for evaluation (Dilokthanakul et al., 2016; Jiang et al., 2017).

<sup>2</sup> See Appendix A for information regarding this open-source software.

Although the dataset is flexible in size given how the moving digits can be rendered on-the-fly, we decide to synthesise a fixed total of only 10000 video sequences. Not relying on an endless amount of training data offers better judgement on how our model might perform in more limited data settings, e.g. in [HRI](#). From the 10000 sequences, we split 8000 into training and 1000 for both validation and testing. Every video is of length  $T=20$  frames, with each frame occupying a  $64 \times 64$  patch. Unlike the original dataset ([Srivastava et al., 2015](#)), our version contains a single  $28 \times 28$  digit moving randomly within each frame so as to not overcomplicate the clustering evaluation by needing to try out much higher values than  $K=10$ . Digit classes are evenly distributed across the dataset.

The DiSCVAE architecture adheres to the graph visualised in Figure 5.3. Wrapped around this probabilistic architecture is a four-layer Convolutional Neural Network (CNN) to encode and decode the raw video sequences, acting as a feature extractor. The entire network is then configured as follows: convolutional layers have  $3 \times 3$  kernels and strides of two, MLP layers (leaky ReLU activations) and the bidirectional LSTM state have 512 hidden units, and the  $\mathbf{h}_t^{\mathbf{z}_L}$  state shared between the inference and generative processes has 128 units. Moreover, each pixel is modelled as a Bernoulli variable represented by latent variables of dimensionality  $\dim(\mathbf{z}_{t,L})=32$  and  $\dim(\mathbf{z}_G)=128$ . The hyperparameter tuning process is exceptionally sensitive when learning disentangled representations ([Locatello et al., 2019](#)), and so rigorous experimental evaluations took place to settle on these values.

With regard to model training, less rigorous testing was required as there are common choices in the relevant literature. For instance, we use the popular Adam optimiser ([Kingma and Ba, 2014](#)) to maximise the ELBO, with a learning rate of  $3 \times 10^{-4}$  and batch size of 16 (relatively typical values). The temperature parameter defining how approximately discrete  $q(\mathbf{y} | \mathbf{x}_{\leq T})$  should be is also set to 1.0, as recommended in [Jang et al. \(2016\)](#); [Maddison et al. \(2016\)](#). All functionality is implemented using TensorFlow ([Abadi et al., 2016](#)) and its Probability library ([Dillon et al., 2017](#)).

#### 5.4.2 Evaluation Protocol

Clustering performance is determined by a frequently applied metric for unsupervised classification accuracy ([Jiang et al., 2017](#)):

$$\text{ACC} = \max_{m \in \mathcal{M}} \frac{\sum_{i=1}^N \mathbf{1}\{l_i = m(c_i)\}}{N}, \quad (5.12)$$

where  $l_i$  and  $c_i$  are the actual label and cluster assignment associated with observation  $x_i$ , respectively. The set  $M$  covers all possible one-to-one mappings between labels and learnt clusters, for which the best permutation is computed via the Hungarian linear assignment algorithm (Kuhn, 1955). To evaluate prediction performance on Moving MNIST, the standard protocol is to sample 10 frames into the future given 10 preceding input frames and then compare these samples with the ground truth sequence (Srivastava et al., 2015). Sticking to this protocol, we report on metrics of binary cross-entropy (BCE) and mean squared error (MSE).

The following methods are considered for this experiment, with each one possessing the same general network structure as specified above:

- **SeqVAE-GMM:** A sequential VAE where the variables at each timestep are treated as independent of one another, plus a GMM separately trained on the learnt latent space;
- **SeqGMVAE:** A sequential GMVAE based on the clustering schema presented in Section 5.3.1, but again not handling temporal dependencies between variables;
- **VRNN-GMM:** A VRNN (Chung et al., 2015) with a GMM fit to its latent space during an isolated optimisation phase;
- **DDPAE-GMM:** The Decompositional Disentangled Predictive Auto-Encoder (Hsieh et al., 2018) (coupled with a GMM for classification), a model that attained state-of-the-art results on Moving MNIST by disentangling and decomposing sequence representations;
- **DiSCVAE:** The proposed model of Section 5.3.2;
- **DiSCVAE-KLY:** A model variation where the ELBO  $\mathcal{L}(\mathbf{x}_{\leq T})$  has the entropy measure in Equation (5.10) exchanged for its respective KL divergence, i. e. like in Equation (5.6).

In addition to these unsupervised approaches, we train a *supervised* bidirectional LSTM (BiLSTM) using the aforementioned CNN encoder and a softmax classifier. All models are optimised over 10 training runs at different random seeds until validation-based early stopping.

### 5.4.3 Results

Table 5.1 summarises classification results on the test set of Moving MNIST. Excluding the DiSCVAE, all other algorithms achieve unsatisfactory per-

Table 5.1: Performance on Moving MNIST test set (10 random seeds), with missing values for non-temporal models that do not output sequence predictions

Model	ACC (%) $\uparrow$	BCE $\downarrow$	MSE $\downarrow$
SeqVAE-GMM	$22.16 \pm 2.03$	-	-
SeqGMVAE	$18.59 \pm 1.45$	-	-
VRNN-GMM	$15.04 \pm 0.5$	$319.61 \pm 26.24$	$73.03 \pm 0.78$
DDPAE-GMM	$27.96 \pm 3.68$	<b><math>246.52 \pm 4.36</math></b>	$72.9 \pm 1.17$
DiSCVAE-KLY	$28.54 \pm 3.28$	$286.9 \pm 8.08$	$70.9 \pm 0.95$
DiSCVAE	<b><math>77.04 \pm 6.76</math></b>	$279.4 \pm 8.87$	<b><math>68.9 \pm 1.28</math></b>
Superv. BiLSTM	$92.59 \pm 1.64$	-	-

formance at discerning digit identities from the video sequences. The VRNN-GMM combination is notably poor on account of its entangled latent representation. On the other hand, the models viewing each frame independently are marginally better due to their focus on time-invariant characteristics for reconstruction. Further improvements in the DiSCVAE-KLY and DDPAE (Hsieh et al., 2018) with a GMM fit to its learnt “content” vector (conceptually similar to our global variable) reinforce the value of both *disentanglement* and *temporal correlations*. Nonetheless, the DiSCVAE obtains a far superior classification accuracy that is also impressively comparable to the supervised BiLSTM.

Predictive performance is also exhibited in Table 5.1, with only the autoregressive VRNN and DDPAE models serving as baselines. The VRNN is outperformed by the DiSCVAE and DDPAE, suggesting that disentangling latent attributes can even occasionally aid in the prediction of entangled future states. Between the DiSCVAE and DDPAE, the former produces better MSE scores and worse BCE measurements, indicating that predictions are on average more accurate but less robust in the presence of uncertainty. The higher certainty surrounding the DDPAE estimates is possibly attributed to how this model learns decomposed representations of input sequences (e.g. individual digits within the video), acting as an attention mechanism towards salient regions for prediction (Hsieh et al., 2018; Kosiorrek et al., 2018). Any discrepancy in DDPAE performance from the original work (Hsieh et al., 2018) is by virtue of a much smaller dataset encompassing single-digit sequences.

Qualitative results of the DiSCVAE applied to the test set are demonstrated in Figure 5.4. The bottom row shows the entire ground truth sequence and on the right-hand side are 10 forward sampled states from each

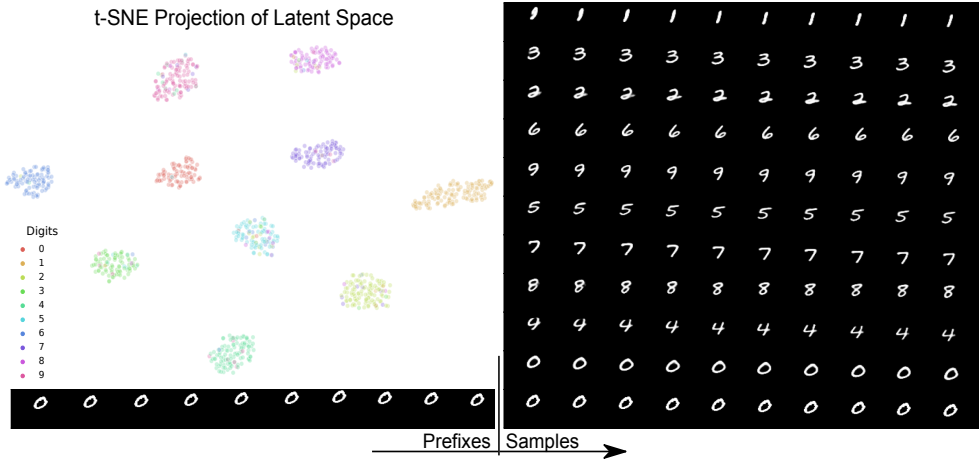


Figure 5.4: Bouncing digits generated by drawing samples from different mixture components. **Bottom:** Shows the ground truth sequence, where the first 10 frames prefix the sampling procedure disclosed in Algorithm 5.1. **Right:** Illustrates how predictions sampled from every cluster correctly persist the bouncing profile of the digit but alter its identity. The sequence just above ground truth is generated from the assigned cluster  $c$  and correctly matches the “zero” class. **Top Left:** A visualisation of the t-SNE projections for digit sequences held in the test set (each point colour-coded by its label).

component of the mixture prior  $p(\mathbf{z}_G | \mathbf{y})$ . For every cluster injected into the sampling procedure of Algorithm 5.1, the velocity characteristics of the bouncing digit are maintained yet the identity distinctly changes to match the corresponding component. Note that the trajectory just above ground truth is drawn from the inferred cluster  $c$ , i.e. the one that maximises posterior probability as in Equation (5.11). Furthermore, the top left plot in Figure 5.4 depicts t-SNE embeddings of global latent  $\mathbf{z}_G$  (Maaten and Hinton, 2008), where coloured data points expose the true digit labels. This plot is generated for the best run (90.2% accuracy), hence the strong coherence.

#### 5.4.4 Ablation Study

Lastly, we explore the issues of posterior collapse and cluster degeneracy when modelling discrete variable  $\mathbf{y}$ . Many techniques have been applied to evade the occurrence of a zero KL divergence term, such as KL cost annealing (Higgins et al., 2017) or vector-quantisation (van den Oord et al., 2017), yet we find maximising the entropy of  $q(\mathbf{y} | \mathbf{x}_{\leq T})$  to be a simple and sufficient solution. This has also been corroborated by other recent works (Hsu et al., 2018; Shi et al., 2020). To justify the claim, we examined how different models and regularisation terms for  $\mathcal{L}(\mathbf{x}_{\leq T})$  affect the posterior when learning with

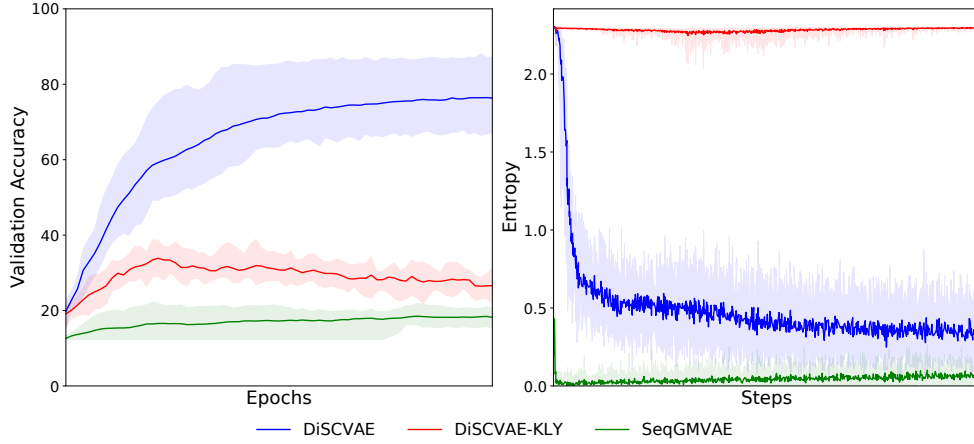


Figure 5.5: Validation set classification accuracy and conditional entropy of  $q(y | \mathbf{x}_{\leq T})$  monitored during training runs on Moving MNIST. The model trained to minimise the KL divergence over  $y$  (DiSCVAE-KLY) exhibits cluster degeneracy, with the conditional entropy plot indicating that every observation is equally likely to be assigned to the same component. At the other extreme, the sequential GMVAE (SeqGMVAE) clearly suffers from posterior collapse. Only the entropy formulation for training a DiSCVAE yields reasonable clustering performance and a meaningful variational posterior.

a powerful CNN decoder. In particular, we tracked the clustering accuracy metric in Equation (5.12) and the conditional entropy during training runs of the DiSCVAE, DiSCVAE-KLY and SeqGMVAE models.

Figure 5.5 illustrates the evolution of these measured quantities on the validation set. When including the entropy term without modelling dependencies between timesteps (SeqGMVAE), the reconstruction loss is prioritised and leads to a uniform posterior distribution that ignores latent variable  $y$  altogether (posterior collapse). On the contrary, if temporal correlations are captured but the model is optimised under KL divergence regularisation (DiSCVAE-KLY), every sequence will project to the same component and suffer from mode collapse. In turn, we appoint an entropy constraint (DiSCVAE) that encourages cluster specialisation to particular input observations and avoids degeneracy despite using expressive non-linear neural networks. Table 5.1 additionally shows how the DiSCVAE surpasses its KL divergence competitor in both classification and prediction.

## 5.5 INTENTION INFERENCE ON ROBOTIC WHEELCHAIRS

After validating the framework’s capacity to disclose classes from synthetic video sequences without supervision, we now explore intention inference in the scope of assistive wheelchair navigation. The problem statement here is to infer the action plans of users from observations of their joystick commands and surroundings, as perceived using laser rangefinders. Equipping robots with this capability is a fundamental ambition of assistive robotics (Demiris, 2007).

### 5.5.1 Dataset

Ten healthy subjects (aged 25-33, all male) with prior experience using a robotic wheelchair were recruited to navigate three mapped environments (top right of Figure 5.1). Each subject was requested to manually control the wheelchair using its joystick and follow a random route designated by goal arrows appearing on a graphical interface, like the one shown in Figure 5.1. In keeping with our terminology of “intent”, we highlight that these goals are incomplete representations of human intent, as they do not reflect the local plans of subjects, i. e. *how* they act in pursuit of a goal.

Experiment data collected during trials was recorded at a frequency of 10Hz, with sequences of length  $T=20$ . All signals perceived by the robot are constrained to this specific frequency rate as the LiDAR sensors act as a bottleneck. As for  $T$ , the length is inspired from related work on estimating the short-term “local” intentions of robotic wheelchair operators (Poon et al., 2017). Every sequence was composed of user joystick commands  $\mathbf{a}_t \in \mathbb{R}^2$  (linear and angular velocities), as well as LiDAR readings  $\mathbf{I}_t \in \mathbb{R}^{360}$  ( $1^\circ$  angular resolution). The resulting dataset amounted to a total of 8823 sequences.

Experimental evaluation on this dataset occurs in two ways. First, the generalisability of our intention inference framework is assessed according to different mapped environments. As a result, any trials that took place in Map 3 (see Figure 5.1) are excluded from the training and validation sets, leaving splits of 5881/1580/1422 for training/testing/validation. The rationale for dividing the dataset in this way is to investigate performance under variations in task context, and verify whether our *interpretable* DiSCVAE can elucidate human intent irrespective of such change. Second, user-specific models are evaluated on a subset of the dataset. This subset adheres to the same map-based split, but consists of only a single subject (3885/1035/882 for training/testing/validation).

### 5.5.2 Post-Processing

Post-processing steps include synchronising and labelling the gathered data. Synchronisation of incoming LiDAR and joystick sequences to a common frequency is necessary, as each signal is recorded at varying rates (even between different LiDAR sensors). Even though the mapped goal poses are incompatible with our definition of intent, we still desire labels for post-analysis. Hence, we appoint ground truth labels to the *implicit* intent of subjects based on the local manoeuvres they make while pursuing task goals. More precisely, the following automated labelling routine is utilised.

Each sequence is initially categorised as either *narrow* or *wide* depending on the measure of threat we previously applied in our SC methodology, refer to Equation (3.7). In essence, this score captures the danger of imminent obstacles per timestep and designates a narrow sequence whenever the averaged measure exceeds a certain threshold.

Next, we discern the intended navigation manoeuvres of participants from the wheelchair’s odometry information. After empirically testing different thresholds for translational and angular velocity, we determined six manoeuvres: in-place rotations (left/right), forward and reverse motion, as well as forward turns (left/right). Overall, this results in 12 classes that account for the influence of both environment state and actions. The majority class across the training and validation sets is the wide in-place rotation (left and right), whilst in the test set it is the narrow reverse. This switch in label frequency between training and testing highlights the task diversity resulting from different maps. Note that odometry data is not supplied as input to the DiSCVAE network.

### 5.5.3 Implementation

The mobile platform is a powered wheelchair integrated with an on-board computer and three LiDAR sensors, as described in Section 3.2. All software processes and inter-device communication are handled within Robot Operating System (ROS) (Quigley et al., 2009).

The DiSCVAE network is portrayed in Figure 5.6 and mimics the graphical model visualised in Figure 5.3, excluding the extra steps to deal with two input modalities. Both the raw LiDAR vector  $\mathbf{l}_{\leq T}$  and 2D user control commands  $\mathbf{a}_{\leq T}$  are passed through separate single-layer MLP neural networks with 512 units (ReLU activations) for feature extraction. The derived code vectors are then concatenated together to yield  $\mathbf{x}_{\leq T}$ , which is fed into

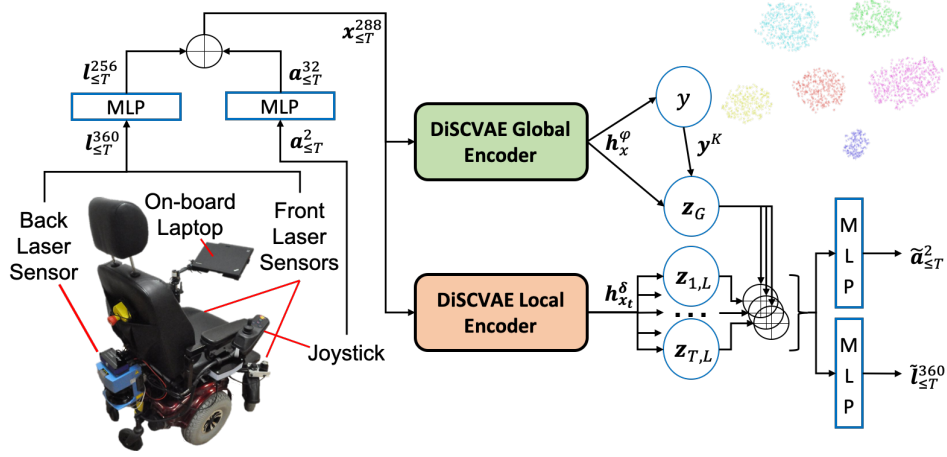


Figure 5.6: Complete robot and network architecture for the robotic wheelchair experiment. Joystick and LiDAR data are fed into independent MLP layers to produce a non-linear transformation of the concatenated input sequence,  $\mathbf{x}_t \in \mathbb{R}^{288}$ . This sequence then feeds into the DiSCVAE encoder graph (see Figure 5.3) to produce the bi-directional LSTM state representation  $\mathbf{h}_x^\phi$  and hidden states  $\mathbf{h}_{x_t}^\delta$  responsible for inferring variables  $\mathbf{z}_G$  and  $\mathbf{z}_{t,L}$ , respectively. These variables are then concatenated and passed onto another set of MLP layers, which decode the joystick commands  $\tilde{\mathbf{a}}_t \in \mathbb{R}^2$  and range values  $\tilde{\mathbf{l}}_t \in \mathbb{R}^{360}$ .

the DiSCVAE to infer latent variables  $\mathbf{z}_G$  and  $\mathbf{z}_{\leq T,L}$ . Upon generation, two individual decoders are conditioned on these variables and trained to reconstruct the original input modalities.

All other training details match our Moving MNIST implementation, except for the following. Sensory observations are modelled as Gaussian variables with fixed variance instead of Bernoulli variables. Sequences are also normalised per modality before entering the network using the mean and standard deviation of the training set, as in previous sequential latent variable models (Fraccaro et al., 2016; Maddison et al., 2017). Likewise, the latent variables representing observations have dimensions  $\dim(\mathbf{z}_G) = \phi = 32$  and  $\dim(\mathbf{z}_{t,L}) = \delta = 32$ . The Adam optimiser (Kingma and Ba, 2014) is also configured with a learning rate of  $1 \times 10^{-3}$ . Again, hyperparameter selection mostly relied on experimental evaluation.

#### 5.5.4 Choosing $K$

A crucial design choice of the DiSCVAE is to select  $K$  for the action plan repertoire size. Although this is straightforward for Moving MNIST, the

Table 5.2: Test set metrics to determine number of clusters

No. Clusters	Whole Dataset		User-Specific	
	ELBO $\uparrow$	NMI $\uparrow$	ELBO $\uparrow$	NMI $\uparrow$
4	-585.5	0.13	-574.0	0.19
6	-585.8	0.18	-574.1	0.21
10	-585.2	0.19	<b>-571.9</b>	<b>0.27</b>
13	<b>-583.1</b>	<b>0.22</b>	-573.4	0.25
16	-583.5	0.21	-573.8	0.26

decision is less apparent when there are many potential interpretations of classes (e.g. between a forward-left turn or an in-place rotation left). The lack of access to ground truth factors of variation also complicates the matter, possibly suggesting that an unsupervised metric would prove useful in diagnosing clustering performance (Locatello et al., 2019). Therefore, two metrics are utilised: the unsupervised log-likelihood per timestep (ELBO) and the Normalised Mutual Information (NMI) across our assigned labels. The NMI measure occupies a range of  $[0, 1]$  and is extensively used to assess the quality of clustering, even amongst similar VAE-based algorithms for interpretable discrete representation learning (Fortuin et al., 2019).

Table 5.2 provides the ELBO and NMI for different  $K$  in the range 4-16. Whilst there is no compelling difference between the number of clusters, we settled on  $K = 13$  and  $K = 10$  for the holistic dataset and its user-specific subset, respectively. This decision was partially based on the slightly superior performance, as well as the close proximity to the number of predefined labels, aiding the linear cluster assignment process in Equation (5.12). It might also be foreseeable that the user-specific subset would occupy a smaller repertoire size, given the more regular patterns of a single subject’s actions.

#### 5.5.5 Evaluation

Complementary to the procedure for Moving MNIST, we report on clustering performance through the accuracy metric defined in Equation (5.12). Note that the linear assignment for this metric optimises a one-to-one mapping, so the leftover clusters or labels (depending on whichever is less) will be assigned arbitrarily. Hence, we also train a classifier (k-nearest neighbour) over the learnt latent representation,  $\mathbf{z}_G$ , to digest the prospects for semi-supervised classification and obtain a more absolute assessment of dis-

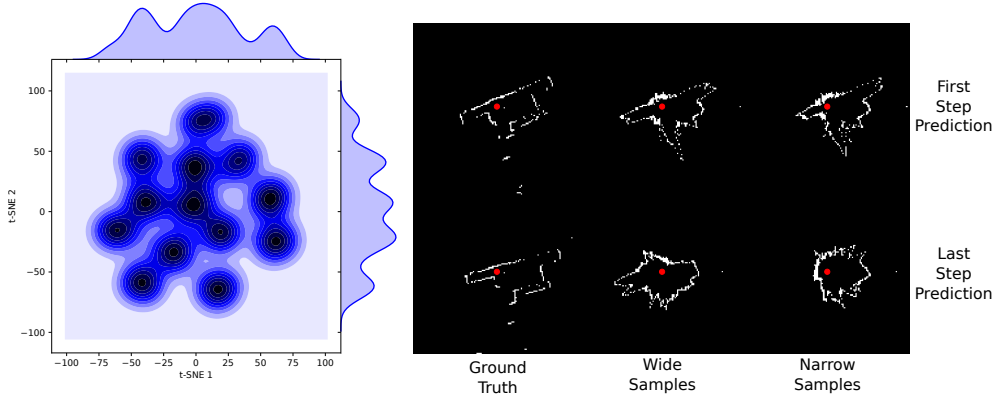


Figure 5.7: **Left:** t-SNE embeddings of component samples drawn from a  $K = 13$  mixture prior. The joint density plot illustrates how a multimodal latent space is learnt from the robotic wheelchair data, with separable clusters representing intent. **Right:** Predicted laser scans (converted to a 2D grid) on the test set when sampling from a “wide” and “narrow” type cluster. Wide samples create spacious proximity around the wheelchair (red dot) and preserve the corridor (left of dot), whilst narrow samples gradually sever this corridor gap.

criminative performance. For this classifier, we compute the mean average precision (mAP) individually across all 12 classes before taking their averaged result as an indicator of the effectiveness of this semi-supervised approach. Trajectory predictions of user actions  $\tilde{\mathbf{a}}_t$  and LiDAR readings  $\tilde{\mathbf{I}}_t$  are also judged by comparing “ground truth” with 10 forward sampled states. The metric for comparison is MSE, a standard error measure in intention estimation (Hu et al., 2018; Tanwani and Calinon, 2017).

For this experiment, only a VRNN (Chung et al., 2015), the DiSCVAE and a supervised BiLSTM are considered, with each trained across both datasets (“Whole” and “User”). The BiLSTM classifier is trained to learn mappings between inputs and the labels identified in Section 5.5.2, whilst the VRNN is only optimised for the regression of trajectories. All methods maintain the same network structure as in Figure 5.6 to encode observations.

### 5.5.6 Results

With respect to qualitative analysis, we demonstrate how action and state samples emerge from the model’s prior latent structure. Figure 5.1 portrays forecasted trajectories by sampling from each mixture component during a recorded interaction with a subject. There is clear variability in the trajectory outcomes predicted for a specific wheelchair configuration ( $K = 6$  to ease visualisation). Importantly, the plotted categorical probability histogram (top

Table 5.3: Performance on Wheelchair test set (10 random seeds)

Model	ACC (%) $\uparrow$	mAP (%) $\uparrow$	MSE $\downarrow$
VRNN-Whole	$21.7 \pm 1.5$	$62.8 \pm 1.2$	$3.50 \pm 0.1$
VRNN-User	$21.9 \pm 1.3$	$63.0 \pm 2.8$	$3.60 \pm 0.2$
DiSCVAE-Whole	$33.3 \pm 3.1$	<b><math>75.9 \pm 1.8</math></b>	<b><math>3.49 \pm 0.1</math></b>
DiSCVAE-User	$37.6 \pm 3.2$	<b><math>80.2 \pm 2.2</math></b>	<b><math>3.49 \pm 0.1</math></b>
BiLSTM-Whole	<b><math>44.9 \pm 2.4</math></b>	$38.1 \pm 1.0$	-
BiLSTM-User	<b><math>46.6 \pm 1.9</math></b>	$44.9 \pm 2.4$	-

left of Figure 5.1) indicates that the most probable trajectory aligns with the wheelchair user’s current goal (red arrow), i. e. the correct “intention”. Meanwhile, Figure 5.7-Right exemplifies how future environment states manifest when sampling from clusters categorised as “wide” or “narrow”. Figure 5.7-Left also depicts how random component samples from a  $K = 13$  mixture prior form distinguishable clusters and reveal a multimodal latent space.

Table 5.3 presents the quantitative results for this experiment. As anticipated, highly variable wheelchair control in an unconstrained navigation task makes classifying intent extremely challenging. The DiSCVAE attains a low error rate of 33.3% on the “Whole” dataset and even the supervised BiLSTM obtains a classification rate of merely 44.9% on the unseen test environment. Nevertheless, learning representations of intent can clearly garner benefits in inference, as the mAP is significantly improved (approximately doubled) by training a classifier over the labelled latent attributes of the DiSCVAE and VRNN. MSE scores for trajectory prediction of joystick and ranger values show that the DiSCVAE and VRNN are alike in their error estimates, which is consistent with how *entangled* representations often yield better predictions at the expense of interpretability (Higgins et al., 2017; Tschannen et al., 2018).

A few conclusions can be drawn from these results. First and foremost, relying on labelled goals for evaluation is not necessarily informative on the underlying distribution of intent. Instead, interpretability is germane to the evaluation of disentangled representations (Locatello et al., 2019), in which the DiSCVAE substantially gains as a transparent clustering model of human intent (elaborated on in Section 5.5.7). Second, the elevated mAP hints that semi-supervised learning may pose a worthwhile avenue to explore in future applications, especially in user modelling on larger interaction datasets. Lastly, the time-invariant and time-varying elements of wheelchair naviga-

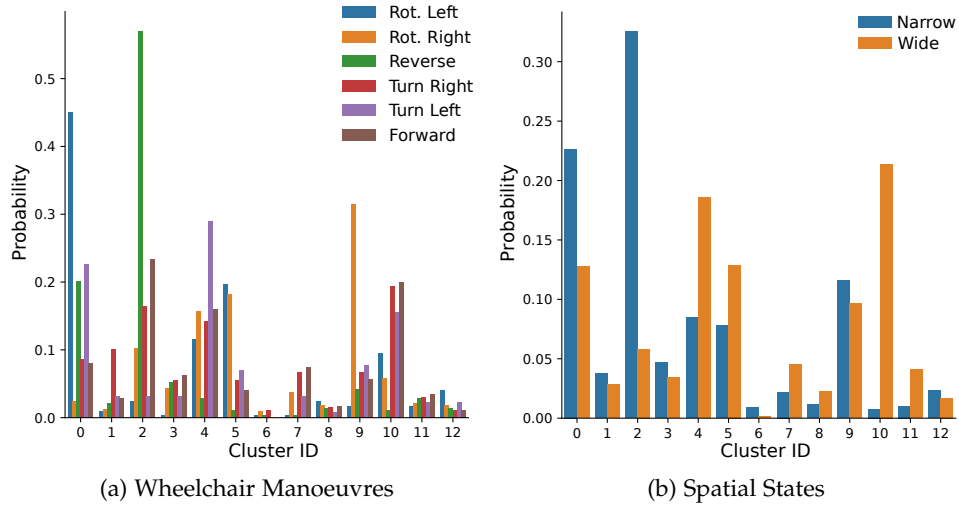


Figure 5.8: Assignment distribution of  $y$  for  $K=13$  with post-processed labels for (a) wheelchair manoeuvres and (b) perceived spatial context. The plot illuminates how various clusters are associated with user intent under different environmental conditions. For example, most backward motion and “narrow” state samples fall within cluster 2. Similar patterns are noticeable for in-place rotations (0 and 9), as well as for “wide” forward motion (4 and 10).

tion are possibly interdependent, as reinforced by the balanced dimensionality of local and global features.

#### 5.5.7 Illuminating the Clusters

Straying away from the purely discriminative task of classifying intent, we now use our framework to decipher the plans intended by users at the “local” scope of wheelchair navigation. In particular, we plot the assignment distributions of  $y$  for each test set example in the “Whole” dataset to understand the underlying meaning of our clustered latent space. The labels from Section 5.5.2 primarily serve to assist in this endeavour. We also point out that identical assignment distributions are found in the “User” data.

Figure 5.8a provides further clarity on how certain clusters have learnt isolated wheelchair manoeuvres. For instance, cluster 2 is distinctly linked with motion trajectories of the wheelchair going in reverse. Likewise, clusters 0 and 9 are affiliated with left and right in-place rotations. Furthermore, the spatial state assignments exhibited in Figure 5.8b delineate how these listed clusters are most often categorised as *narrow*. This result is to be expected of evasive actions that habitually take place in cluttered spaces. In contrast, predominantly forward-oriented manoeuvres fall into *wide* clusters (e.g. 4

and 10). These findings substantiate that an action plan repertoire has been aptly inferred using our proposed framework.

## 5.6 RELATED WORK

Our work builds on the VAE paradigm (Kingma and Welling, 2013; Rezende et al., 2014) and its sequential extensions (Chung et al., 2015; Fraccaro et al., 2016; Goyal et al., 2017; Krishnan et al., 2017) by uncovering meaning behind their learnt latent representations (Bengio et al., 2013). Recent sequential latent variable models have advanced towards this objective by learning to disentangle representations into sequence- and segment-level attributes of data (Hsu et al., 2017; Yingzhen and Mandt, 2018). As a result, these models can distinguish and manipulate features present in the sequence at both a local and global scale, e.g. separating pose information from content in video data (Hsieh et al., 2018). Our model adheres to this scheme of disentanglement but embraces an extra level of *discrete* attributes for the purpose of clustering.

Incorporating discrete variables into deep generative models plays an integral role in understanding latent spaces, especially by drawing connections with abstract concepts, such as human intent. Some sequential models complementary to ours have disentangled latent variables using decomposition (Hsieh et al., 2018) or attention (Kosiorrek et al., 2018) to enable abstract reasoning (e.g. to compartmentalise or count the bouncing digits in Moving MNIST), yet few have additionally considered clustering. The most comparable work is a hierarchical generative model for text-to-speech synthesis (Hsu et al., 2018) that also employed a GMM prior with categorical attributes to cluster different speakers. However, a notable difference is that the DiSCVAE parameterises the latent GMM with a recognition network and does not presume categorical labels are visible *a priori*.

Intention understanding essentially revolves around the problem of an observing agent deriving a model to match an acting agent’s behaviour (Demiris, 2007). In HRI, this problem is typically addressed by equipping an observing robot with a probabilistic model that infers intent from human actions (Hu et al., 2018; Jain and Argall, 2019; Javdani et al., 2015). The growing interest in scalable learning techniques for modelling agent intent has also spurred on applications in robotics for purposes like SC (Losey et al., 2019; Reddy et al., 2018) and multi-agent governance (Xie et al., 2020). However, disentangled representation learning remains sparse in the literature, with the only known comparable work to ours being a conditional VAE that dis-

entangled latent variables in a multi-agent driving setting (Hu et al., 2019). Albeit similar in principle, we stipulate that our approach is the first to *infer* a discrete “intent” variable from human behaviour and subsequently *cluster* their action plans.

## 5.7 CONCLUSIONS

In this chapter, we embraced a novel outlook on human intention inference and introduced a deep generative model to simultaneously disentangle and cluster sequence representations. The overall framework is broadly applicable to sequential data, as proven by revealing classes from synthetic video sequences of bouncing digits, whilst correctly preserving motion characteristics on trajectory generation. A real-world experiment on intention inference involving robotic wheelchairs also gleaned insights into how our model could discern primitive action plans from observations, e. g. rotating in-place or reversing. We believe the contributions of this chapter can equally serve the machine learning and robotics communities.

The implications of an unsupervised, interpretable means of inferring intent are promising for numerous research avenues in HRI. For instance, the task-agnostic prior could be exploited in downstream tasks, such as user modelling, to augment the wider adoption of collaborative robotics in unconstrained environments. Our findings on semi-supervised learning from the robotic wheelchair experiment fortify this idea. The interpretable latent structure could also prove fruitful in assistive robots that warrant explanation by visually relaying inferred intentions back to end-users (as in Chapter 4). A final course of inquiry could be to incorporate these explanations into an interactive learning procedure (Locatello et al., 2019), e. g. for user personalisation.

By contributing a means of creating “robot-of-human” transparency (Lyons, 2013; Lyons and Havig, 2014), we have now fulfilled all the prerequisites for effective XSC. Despite the “black-box” nature of our DiSCVAE for intention inference, the capability to attach abstract concepts to its latent structure enables us to reap the expressive power of deep learning (Bengio and Delalleau, 2011) in applications that benefit from explanation, e. g. SC for robotic wheelchair assistance.

## CONCLUSIONS AND FUTURE DIRECTIONS

---

The final chapter of this thesis serves three purposes. First, to compile together our contributions to various subject domains. Second, to remark on any limitations of the presented work. Lastly, to review future research directions that may address these limitations and enhance the pervasiveness of assistive robotics.

### 6.1 OVERVIEW OF THESIS CONTRIBUTIONS

Our primary objective in this thesis has been to resolve the model misalignment that frequents Shared Control (SC) by exposing robot and human intentions to one another. In pursuing this objective, we have made contributions to a wide selection of research areas, including Human-Robot Interaction (HRI), Augmented Reality (AR) and representation learning. A focal point across all these contributions has been to advance the prevalence of assistive robots, hence our chosen platform: a “smart” wheelchair.

Our first contribution answered research question (2) with a novel architecture integrating an AR Head-Mounted Display (HMD) onto a smart wheelchair. From the visualisations devised to accelerate mental model accuracy of the SC, only the rear-view display garnered positive user ratings. Smart wheelchair manufacturers could shed light from this result, as the large rear-view is especially advantageous for disabled individuals with poorer eyesight or limited upper body and neck mobility. Moreover, we discovered that users could not exploit all the AR aids to their benefit because some were either too low-level and thereby puzzling (e.g. command vectors), or poorly positioned and thus inaccessible (e.g. floor-level objects). The latter issue is partly affected by the headset’s sturdiness and narrow Field of View (FoV), however with sufficient technological advancements, many of these concerns about how to spatially situate AR aids will become obsolete. For the former issue, we found the need for a more principled interface design.

Striving to guide AR HMD interface design on how to avoid these outcomes and successfully “explain” any internal mechanisms of robots employing SC, we introduced the Explainable Shared Control (XSC) paradigm. This paradigm concerns both the development of internal SC processes, as

well as their visualisation in [AR](#). In particular, we described an [SC](#) methodology where intention estimation and arbitration exhibited *causality* and *abstraction*, allowing corresponding *contextual* and *predictive AR* cues to be generated. These two traits in the [SC](#) methodology helped answer our first research question. Experiments with the updated smart wheelchair system showed that subjects engaging in [XSC](#) had less strenuous eye gaze patterns and were quicker to overcome adverse events than subjects utilising standard [SC](#). These findings corroborated the capacity for [XSC](#) to combat model misalignment in assistive navigation, answering question (3).

Having addressed the human’s perception of robot intent through visual “explanations”, our last question regarding [XSC](#) was to supply robots with an *interpretable* framework for human intention inference. This inference problem often relies on assumptions about the task-at-hand in order to operate under constrained conditions. However, the proposed Disentangled Sequence Clustering Variational Autoencoder ([DiSCVAE](#)) made no such assumptions and enabled an assistive robot to instead learn how to represent intentions directly from observed human behaviour. Unlike many “black-box” methods that suffer from lack of model interpretability, the [DiSCVAE](#) latent space consisted of clusters that could illuminate developers and users alike on what human intentions were deciphered from observations.

Each of these contributions sought to fulfil the core requirements of [XSC](#) and illustrate the beneficial impact on assistive robots, such as the smart wheelchair. Our commitment to publish and make publicly available all software derived from this thesis is also likely to pave the way towards more explainable robot behaviour in the fields of [SC](#) and [HRI](#). Appendix [A](#) summarises the details of this open-source software.

## 6.2 OUTSTANDING ISSUES

Before outlining future avenues for the [XSC](#) paradigm, the issues associated with its current form must be expressed. This section is devoted to these outstanding issues.

### 6.2.1 Addressing the Target Population

In spite of an encouraging forecast for [XSC](#), we have yet to conduct user trials with the actual target population. Testing on able-bodied volunteers is favourable when it comes to engineering an assistive robot prototype, as rapidly testing on users with minimal health risks is a mandatory and

pragmatic first step. However, if the prototype is to eventually migrate outside of controlled lab settings, then a larger scale study where a small number of disabled patients are introduced is vital (Carlson and Demiris, 2012; Viswanathan et al., 2017). Otherwise, healthy subjects may yield biased results and there is no guarantee that the main study conclusions will transfer over to target end-users. Once we have evaluated XSC in a case study including disabled individuals, only then can we draw concrete conclusions about its benefits for assistive robotics.

### 6.2.2 General Applicability of Explainable Shared Control

As with any paradigm in robotics, validating its applicability over diverse robot architectures is a non-trivial challenge. Even the traditional SC paradigm adopts numerous definitions and cannot easily find a common ground in the literature (Abbink et al., 2018). This is primarily due to the variations in physical characteristics and roles of robots assisting people via SC.

For instance, our AR HMD interface for SC on smart wheelchairs may not prove suitable when considering aerial robots or robotic arms. Aside from obvious differences in hardware and functionality between these robots, there is also the fact that smart wheelchairs are unique in how humans embody them as operators. This embodiment largely affects the choice of AR visualisations and requires extra care regarding the HMD's FoV, as we identified from our pilot study in Section 4.2. A fresh set of AR aids may then be necessary in scenarios where humans are remotely sharing control with robots, e. g. for telerobotic control (Milgram et al., 1995) or multi-agent teaming (Dias et al., 2008).

Nevertheless, the recommended guidelines of XSC are intended to be general enough to inform the design of AR HMD interfaces in varied HRI settings. In Appendix C, we refer to these guidelines in order to conceive an AR HMD interface for dual-arm collaborative robots. For this application, the arm manipulator's intent is projected through *contextual* and *predictive* AR cues. Although SC was not applied in this setup, elements of explainability from XSC are still utilised effectively. In turn, we suspect that the same concepts introduced in this thesis would prove effective on platforms other than the robotic wheelchair.

### 6.2.3 *Tracing Model Misalignment*

Given the *active* nature of SC, tracing model mismatch should take place over a continuous interaction, as opposed to a session basis. Whilst the user study in Section 4.3.3 recognised the value of this trait and detected events associated with model misalignment during navigation trials, it did not explore how mental models are reconciled per event. By aggregating the results across entire trials, it is unclear how exactly the AR headsets helped users recover from “stucks” or other jarring incidents. In a deeply collaborative setting, like SC, it would be prudent to actively track the reconciliation process throughout each event.

Tracking reconciliation is broadly part of a wider debate on how to evaluate SC systems. Abbink et al. (2018) reflect on this dilemma and introduce an evaluation corollary where experimental conditions of SC should “include *static* and *dynamic* conditions that fall *within* and *beyond* the boundaries of the task domain”. At present, our experimental trials have remained within the boundaries of assistive navigation and are examined only in terms of static conditions. It is therefore necessary to expand the scope of XSC beyond the prescribed navigation tasks and establish new criteria for dynamically delimiting model misalignment and reconciliation. Probing the role of cognitive load using saccadic eye movement could act as such a dynamic criterion on the learning gauge of mental models.

### 6.2.4 *Communicating the Human Intention Inference Model*

Humans often struggle to comprehend the internal motives surrounding robot behaviour (Jain and Argall, 2019). There are many reasons for this struggle, including the correspondence problem (i. e. observers not possessing the same internal mechanics as demonstrators Nehaniv and Dautenhahn, 2001), appearance constraints (e. g. robots lacking anthropomorphic features Walker et al., 2018), illegible robot actions (Dragan et al., 2013), and so forth. Regardless of the reason, it is of paramount importance to *transparency* that the intentions or objectives of robots are conveyed back to users (Alonso and de la Puente, 2018; Huang et al., 2019).

As a result, a key expectation of XSC is for humans and robots to share intent in a *bi-directional* manner. In other words, the robot must infer human intent in parallel with communicating its internal model for inference. Despite the AR interface of Chapter 4 unveiling our smart wheelchair’s intention estimation mechanism, this internal process was not an accurate inference of

human intent, as pointed out in Section 3.4. On the other hand, the framework in Chapter 5 learnt an action plan repertoire for wheelchair navigation that depicts a more complete representation of human intent. Though we have yet delved into how the interpretable nature of the DiSCVAE could be relayed back to *end-users* visually.

#### 6.2.5 Closing the Explainable Shared Control Loop

Possibly the most fundamental issue remaining is to close the XSC loop and define a holistic system. Our endeavour to bring transparency to the human and robot perspectives of SC was fulfilled in isolation without suggesting a method of integration. Integration would require the DiSCVAE latent space to be encapsulated into the AR HMD interface (as stated in Section 6.2.4), and for its generated samples to be administered as assistive robot commands during the SC. Given how the DiSCVAE was trained over navigation trials involving healthy subjects, any generated actions could be supplied to adjust noisy inputs of actual patients (akin to learning assistance by demonstration Kucukyilmaz and Demiris, 2018; Soh and Demiris, 2015).

### 6.3 FUTURE RESEARCH

This thesis only provides an initial step forward in XSC, with many of its qualities demanding further investigation. In the following, we describe a few research directions for XSC and their significance to assistive robotics.

#### 6.3.1 Mutual Model Adaptation

One interesting path of research is to analyse the interplay between internal models of humans and robots as they actively participate in closed-loop XSC. This idea coincides with what is known as mutual model adaptation in SC (Nikolaidis et al., 2017), where the internal models of *both* the human and robot are regularly undergoing adaptation. Situated in XSC, the intention inference model of the robot would have to be capable of online updates and the AR interface must be dynamically configurable according to these updates. Provided with real-time visual feedback on the robot's internal states, users would then evoke responsive behaviour that triggers additional changes in the learnt action repertoire, e.g. modifying or adding action plans.

There is a myriad of lucrative aspects to studying this interplay, so we will only name a few. First, model misalignment could be properly traced as the interaction is unfolding and assessed based on the frequency of mutual model updates. It could be hypothesised that a higher update frequency correlates with worse misalignment. Next, the *XSC* decision-making could exploit incoming information surrounding the interplay for purposes like fine-tuning the arbitration policy, or disabling *AR* visualisations that are no longer useful. Inference models on how humans make sense of the *AR* interface could then also be created to discern which cues are most informative. Huang et al. (2019) recently tackled a similar problem, where a robot inferred which behaviours would best teach end-users about its objective function. The work did not aim to *explicate* any internal robot reasoning and so *AR* could complement this line of research.

### 6.3.2 Hierarchical Intention Prediction

Another pertinent research question is whether long-term intentions can be estimated from the short-term action plans extracted in Chapter 5. As mentioned in Section 2.4.2, we frame the recognition of future-directed intentions as *prediction*, rather than estimation or inference. A source of biological inspiration on how to effectively approach this prediction problem are *internal models* (see Figure 2.2). These models are neural functions that enable humans to foresee the resulting behaviour of perceived actions according to internal simulations of their own sensory-motor repertoire of actions (Wolpert et al., 2003, 1998). Given how we acquired such a repertoire in Chapter 5, it bears considering how our learnt clusters, or alternatively internal models, could be extrapolated to predict future intentions.

Many intention prediction architectures based on internal models rely on two fundamental concepts: multiplicity and hierarchy (Demiris, 2007; Demiris and Khadhour, 2006; Wolpert et al., 2003, 1998). *Multiplicity* refers to the idea that the motor system must handle multiple contexts with multiple possible responses or behaviours, probing the idea that the motor apparatus follows a distributed layout (Wolpert et al., 1998). *Hierarchy* instead expresses how our motor system mediates and reasons about low-level actions from higher-level internal representations (Wolpert et al., 2003).

Viewing the clustered latent space of our *DiSCVAE* as an already decomposed representation of multiple actions and contexts, the missing quality for prediction is hierarchy. The initiative here could be to map all the local actions of the repertoire into a hierarchical structure of intent, where the

highest levels represent human deliberation and planning, whilst the lowest layers reflect motor cognition (Hamilton and Grafton, 2007; Pacherie, 2008). For example, a “drive forward” action may combine with a situational-dependent cluster that has learnt to “avoid obstacles”, and this may then translate into the higher intention of “drive safely through doorway”.

One strategy for assembling this hierarchical structure in our intention inference framework could involve learning a layer of high-level internal models above our clustered latent space. Lee et al. (2013) successfully used Stochastic Context-Free Grammars (SCFGs) for similar motives by parsing imitated action sequences and capturing the underlying structure in complicated tasks composed of multi-layered behaviours. These SCFGs offer numerous compelling traits for hierarchical learning, such as robustness to noise, compactness in representation and the ability to handle recursive symbol sequences. Another invaluable asset of SCFGs for XSC lies in the output’s human-readability, meaning the resulting grammar could relay back a group of interpretable action symbols to visually depict the robot’s intent. As a result, symbol parsing for intention prediction using SCFGs could define a novel way of mediating control in XSC.

### 6.3.3 *Multisensory Modalities*

A third course of inquiry will be to glean further insights into XSC from multisensory modalities, such as eye gaze. Section 4.3.3 briefly probed this notion by tracking user eye movements during wheelchair navigation to draw connections with model misalignment. Our enlightening findings from this sensory signal motivates a deeper look into other modalities, e.g. auditory stimuli. In particular, a biological characteristic of humans that plays an immense role in their perceptual capabilities is *multisensory integration*. Multisensory integration refers to our brain’s simultaneous processing of an array of sensory inputs, such as visual stimuli, for the purpose of constructing a robust multimodal percept (Driver and Spence, 2000; Stein and Stanford, 2008). The synthesis of multiple data sources is known to ameliorate uncertainty in coherency of the physical surroundings, leading to a more robust form of percept (Ernst and Bühlhoff, 2004; Stein and Stanford, 2008).

Merging multiple sensory sources into the XSC loop could enhance the information throughput of the AR interface. The AR literature is rich in interfaces that combine multimodal cues (e.g. touch, speech, gaze) into the augmentation (Azuma, 1997; Carmigniani et al., 2011; Sibirtseva et al., 2019). Hence, multimodal communication has been increasingly noted for its poten-

tial to elevate *transparency* in HRI (Lakhmani et al., 2016) and disambiguate internal robot functions (Huang et al., 2019; Sibirtseva et al., 2019). We thus believe that generating multimodal “explanations” is a promising avenue for XSC in fostering an immersive and transparent user experience.

Likewise, fusing multisensory inputs in the intention prediction of XSC could improve its robustness and accuracy. With the onset of pervasive sensing in robotics, multimodal representation learning has spurred on auspicious results (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012). For example, Noda et al. (2014) utilised deep autoencoders to integrate multiple representations of sensory information into a noise-robust behaviour prediction network. In vehicle manoeuvre prediction, Jain et al. (2016) and Lee et al. (2017) both exploited a Fusion-Recurrent Neural Network (RNN) layer to combine high-level representations of separate sensory sources and anticipate future trajectories with state-of-the-art accuracy, despite only possessing a partial temporal context of the multimodal input. Lee et al. (2017) notably used a Variational Autoencoder (VAE)-based framework that could generate multiple hypotheses about the future predictions, posing an attractive choice for the *multiplicity* desired in intention prediction.

#### 6.3.4 User Personalisation

A final ambition for XSC is to fulfil the challenging demands of providing proper conditional assistance by developing a personalised user model that can accommodate each specific patient’s characteristics. SC for smart mobility is well-known for being highly dependent on each individual operator and their preferences (Erdogan and Argall, 2017; Viswanathan et al., 2017). Vanhooydonck et al. (2010) undertook this endeavour by using a neural network to learn an *implicit* model of users’ driving behaviour for wheelchair navigation. Conversely, Jain and Argall (2019) relied on an *explicit* parameter to define the level of assistance required by each user. We envision that a combination of learning implicit user characteristics and adjusting an explicit SC arbitration parameter is an appropriate tactic for personalisation. Furthermore, the resulting user model could be incorporated into the AR interface, e. g. to adjust the scale of assistance.

### 6.4 EPILOGUE

The overarching aspiration of XSC has been to establish a seamless collaboration between assistive robots and humans. By documenting ways of mani-

festing transparency from the human perspective via [AR](#), as well as the robot angle through intention inference, we hope that this thesis serves as an early step towards realising explainability in [SC](#). As [XSC](#) is only in its primitive form, there are many exciting opportunities across interdisciplinary research domains that could extend this paradigm, and thereby propel forward the widespread use of assistive robots.



## BIBLIOGRAPHY

---

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. and Zheng, X. (2016), Tensorflow: A system for large-scale machine learning, in 'USENIX Symposium on Operating Systems Design and Implementation', USENIX Association, pp. 265–283. (pages [95](#), [142](#)).
- Abbink, D. A., Carlson, T., Mulder, M., de Winter, J. C. F., Aminravan, F., Gibo, T. L. and Boer, E. R. (2018), 'A Topology of Shared Control Systems—Finding Common Ground in Diversity', *IEEE Transactions on Human-Machine Systems* **48**(5), 509–525. doi: [10.1109/THMS.2018.2791570](#) (pages [15](#), [21](#), [22](#), [24](#), [25](#), [51](#), [52](#), [66](#), [67](#), [111](#), [112](#)).
- Abbott, W. W. and Faisal, A. A. (2012), 'Ultra-low-cost 3D gaze estimation: an intuitive high information throughput compliment to direct brain-machine interfaces', *Journal of Neural Engineering* **9**(4), 1–11. doi: [10.1088/1741-2560/9/4/046016](#) (page [143](#)).
- Abramson, L. Y., Seligman, M. E. and Teasdale, J. D. (1978), 'Learned helplessness in humans: Critique and reformulation', *Journal of Abnormal Psychology* **87**(1), 49–74. doi: [10.1037/0021-843X.87.1.49](#) (page [22](#)).
- Agree, E. M. (2014), 'The potential for technology to enhance independence for those aging with a disability', *Disability and Health Journal* **7**, S33–S39. doi: [10.1016/j.dhjo.2013.09.004](#) (pages [15](#), [22](#), [39](#)).
- Alonso, V. and de la Puente, P. (2018), 'System Transparency in Shared Autonomy: A Mini Review', *Frontiers in Neurorobotics* **12**, 83. doi: [10.3389/fnbot.2018.00083](#) (pages [28](#), [36](#), [48](#), [52](#), [65](#), [66](#), [112](#)).
- Alshaer, A., Regenbrecht, H. and O'Hare, D. (2017), 'Immersion factors affecting perception and behaviour in a virtual reality power wheelchair simulator', *Applied Ergonomics* **58**, 1–12. doi: [https://doi.org/10.1016/j.apergo.2016.05.003](#) (pages [26](#), [53](#), [68](#)).

- Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S. and MacIntyre, B. (2001), 'Recent advances in augmented reality', *IEEE Computer Graphics and Applications* **21**(6), 34–47. doi: [10.1109/38.963459](https://doi.org/10.1109/38.963459) (page 26).
- Azuma, R. T. (1997), 'A Survey of Augmented Reality', *Presence: Teleoperators and Virtual Environments* **6**(4), 355–385. doi: [10.1162/pres.1997.6.4.355](https://doi.org/10.1162/pres.1997.6.4.355) (pages 26, 115).
- Bengio, Y., Courville, A. and Vincent, P. (2013), 'Representation learning: A review and new perspectives', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**(8), 1798–1828. doi: [10.1109/TPAMI.2013.50](https://doi.org/10.1109/TPAMI.2013.50) (pages 34, 35, 81, 84, 88, 107).
- Bengio, Y. and Delalleau, O. (2011), On the Expressive Power of Deep Architectures, in J. Kivinen, C. Szepesvári, E. Ukkonen and T. Zeugmann, eds, 'Algorithmic Learning Theory', Springer Berlin Heidelberg, pp. 18–36. doi: [10.1007/978-3-642-24412-4\\_3](https://doi.org/10.1007/978-3-642-24412-4_3) (pages 34, 108).
- Blakemore, S. J. and Decety, J. (2001), 'From the perception of action to the understanding of intention', *Nature Reviews Neuroscience* **2**(8), 561–567. doi: [10.1038/35086023](https://doi.org/10.1038/35086023) (pages 29, 32, 82).
- Blaylock, N. and Allen, J. (2006), Fast Hierarchical Goal Schema Recognition, in 'National Conference on Artificial Intelligence', AAAI'06, AAAI Press, pp. 796–801. (page 33).
- Borji, A., Sihite, D. N. and Itti, L. (2014), 'What/where to look next? Modeling top-down visual attention in complex interactive environments', *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **44**(5), 523–538. doi: [10.1109/TSMC.2013.2279715](https://doi.org/10.1109/TSMC.2013.2279715) (page 145).
- Bratman, M. E. (1990), 'What is intention', *Intentions in communication* pp. 15–32. (pages 16, 29, 30).
- Brose, S. W., Weber, D. J., Salatin, B. A., Grindle, G. G., Wang, H., Vazquez, J. J. and Cooper, R. A. (2010), 'The role of assistive robotics in the lives of persons with disability', *American Journal of Physical Medicine and Rehabilitation* **89**(6), 509–521. doi: [10.1097/PHM.0b013e3181cf569b](https://doi.org/10.1097/PHM.0b013e3181cf569b) (pages 15, 22).
- Bryson, J. and Winfield, A. (2017), 'Standardizing ethical design for artificial intelligence and autonomous systems', *Computer* **50**(5), 116–119. doi: [10.1109/MC.2017.154](https://doi.org/10.1109/MC.2017.154) (page 27).

- Bütepage, J., Kjellström, H. and Kragic, D. (2017), 'Anticipating many futures: Online human motion prediction and synthesis for human-robot collaboration', *arXiv preprint arXiv:1702.08212* . (page 35).
- Buttussi, F. and Chittaro, L. (2018), 'Effects of Different Types of Virtual Reality Display on Presence and Learning in a Safety Training Scenario', *IEEE Transactions on Visualization and Computer Graphics* **24**(2), 1063–1076. doi: [10.1109/TVCG.2017.2653117](https://doi.org/10.1109/TVCG.2017.2653117) (page 26).
- Carlson, T. and Del R. Millan, J. (2013), 'Brain-controlled wheelchairs: A robotic architecture', *IEEE Robotics and Automation Magazine* **20**(1), 65–73. doi: [10.1109/MRA.2012.2229936](https://doi.org/10.1109/MRA.2012.2229936) (pages 40, 143).
- Carlson, T. and Demiris, Y. (2009), 'Using Visual Attention to Evaluate Collaborative Control Architectures for Human Robot Interaction', *Proceedings of New Frontiers in Human-Robot Interaction* pp. 38–43. (pages 27, 64, 72).
- Carlson, T. and Demiris, Y. (2010), Increasing robotic wheelchair safety with collaborative control: Evidence from secondary task experiments, in 'IEEE International Conference on Robotics and Automation', pp. 5582–5587. doi: [10.1109/ROBOT.2010.5509257](https://doi.org/10.1109/ROBOT.2010.5509257) (page 47).
- Carlson, T. and Demiris, Y. (2012), 'Collaborative Control for a Robotic Wheelchair: Evaluation of Performance, Attention, and Workload', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **42**(3), 876–888. doi: [10.1109/TSMCB.2011.2181833](https://doi.org/10.1109/TSMCB.2011.2181833) (pages 23, 24, 25, 31, 56, 72, 73, 75, 111).
- Carmigniani, J., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E. and Ivkovic, M. (2011), 'Augmented Reality Technologies, Systems and Applications', *Multimedia Tools Appl.* **51**(1), 341–377. doi: [10.1007/s11042-010-0660-6](https://doi.org/10.1007/s11042-010-0660-6) (pages 26, 115).
- Chacón-Quesada, R. and Demiris, Y. (2019), Augmented reality controlled smart wheelchair using dynamic signifiers for affordance representation, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 4812–4818. (page 48).
- Chacón-Quesada, R. and Demiris, Y. (2020), Augmented reality user interfaces for heterogeneous multirobot control, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems'. accepted for publication. (page 48).

- Chakraborti, T., Fadnis, K. P., Talamadupula, K., Dholakia, M., Srivastava, B., Kephart, J. O. and Bellamy, R. K. E. (2018), Visualizations for an Explainable Planning Agent, in 'International Joint Conference on Artificial Intelligence', AAAI Press, pp. 5820–5822. (pages 16, 28, 65, 67).
- Chakraborti, T., Sreedharan, S., Kulkarni, A. and Kambhampati, S. (2018), Projection-Aware Task Planning and Execution for Human-in-the-Loop Operation of Robots in a Mixed-Reality Workspace, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 4476–4482. doi: [10.1109/IROS.2018.8593830](https://doi.org/10.1109/IROS.2018.8593830) (pages 16, 28, 67).
- Chakraborti, T., Sreedharan, S., Zhang, Y. and Kambhampati, S. (2017), Plan Explanations As Model Reconciliation: Moving Beyond Explanation As Soliloquy, in 'International Joint Conference on Artificial Intelligence', pp. 156–163. (pages 16, 27, 67).
- Chang, M. L., Gutierrez, R. A., Khante, P., Short, E. S. and Thomaz, A. L. (2018), Effects of Integrated Intent Recognition and Communication on Human-Robot Collaboration, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 3381–3386. doi: [10.1109/IROS.2018.8593359](https://doi.org/10.1109/IROS.2018.8593359) (page 36).
- Chatzopoulos, D., Bermejo, C., Huang, Z. and Hui, P. (2017), 'Mobile Augmented Reality Survey: From Where We Are to Where We Go', *IEEE Access* 5, 6917–6950. doi: [10.1109/ACCESS.2017.2698164](https://doi.org/10.1109/ACCESS.2017.2698164) (page 26).
- Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C. and Bengio, Y. (2015), A Recurrent Latent Variable Model for Sequential Data, in C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds, 'Advances in Neural Information Processing Systems 28', Curran Associates, Inc., pp. 2980–2988. (pages 35, 81, 86, 87, 91, 92, 96, 104, 107).
- Clinciu, M.-A. and Hastie, H. (2019), A Survey of Explainable AI Terminology, in 'Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence', pp. 8–13. (pages 27, 52).
- Cohen, P. R. and Levesque, H. J. (1990), 'Intention is choice with commitment', *Artificial Intelligence* 42(2), 213–261. doi: [https://doi.org/10.1016/0004-3702\(90\)90055-5](https://doi.org/10.1016/0004-3702(90)90055-5) (pages 29, 30).
- Cowan, R. E., Fregly, B. J., Boninger, M. L., Chan, L., Rodgers, M. M. and Reinkensmeyer, D. J. (2012), 'Recent trends in assistive technology for mobility', *Journal of NeuroEngineering and Rehabilitation* 9(1), 1–8. doi: [10.1186/1743-0003-9-20](https://doi.org/10.1186/1743-0003-9-20) (pages 21, 22).

- Demiris, Y. (2007), 'Prediction of intent in robotics and multi-agent systems', *Cognitive Processing* 8(3), 151–158. doi: [10.1007/s10339-007-0168-9](https://doi.org/10.1007/s10339-007-0168-9) (pages [16](#), [29](#), [31](#), [32](#), [42](#), [82](#), [100](#), [107](#), [114](#)).
- Demiris, Y. (2009), Knowing when to assist: Developmental issues in lifelong assistive robotics, in 'Annual International Conference of the IEEE Engineering in Medicine and Biology Society', pp. 3357–3360. doi: [10.1109/IEMBS.2009.5333182](https://doi.org/10.1109/IEMBS.2009.5333182) (pages [15](#), [22](#), [40](#)).
- Demiris, Y. and Khadhour, B. (2006), 'Hierarchical attentive multiple models for execution and recognition of actions', *Robotics and Autonomous Systems* 54(5), 361–369. doi: [10.1016/j.robot.2006.02.003](https://doi.org/10.1016/j.robot.2006.02.003) (pages [33](#), [34](#), [114](#)).
- Desai, M. and Yanco, H. A. (2005), Blending human and robot inputs for sliding scale autonomy, in 'IEEE International Workshop on Robot and Human Interactive Communication', pp. 537–542. doi: [10.1109/ROMAN.2005.1513835](https://doi.org/10.1109/ROMAN.2005.1513835) (page [21](#)).
- Devigne, L., Babel, M., Nouviale, F., Narayanan, V. K., Pasteau, F. and Gallien, P. (2017), Design of an immersive simulator for assisted power wheelchair driving, in 'IEEE International Conference on Rehabilitation Robotics', pp. 995–1000. doi: [10.1109/ICORR.2017.8009379](https://doi.org/10.1109/ICORR.2017.8009379) (page [53](#)).
- Dias, M. B., Kannan, B., Browning, B., Jones, E., Argall, B., Dias, M. F., Zinck, M., Veloso, M. and Stentz, A. (2008), Sliding autonomy for peer-to-peer human-robot teams, in 'International Conference on Intelligent Autonomous Systems', pp. 332–341. (pages [21](#), [111](#)).
- Diaz, C., Walker, M., Szafir, D. A. and Szafir, D. (2017), Designing for Depth Perceptions in Augmented Reality, in 'IEEE International Symposium on Mixed and Augmented Reality', pp. 111–122. doi: [10.1109/ISMAR.2017.28](https://doi.org/10.1109/ISMAR.2017.28) (page [28](#)).
- Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M. and Saurous, R. A. (2017), 'Tensorflow distributions', *arXiv preprint arXiv:1711.10604* . (pages [95](#), [142](#)).
- Dilokthanakul, N., Mediano, P. A. M., Garnelo, M., Lee, M. C. H., Salimbeni, H., Arulkumaran, K. and Shanahan, M. (2016), 'Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders', *arXiv preprint arXiv:1611.02648* . (pages [35](#), [89](#), [90](#), [91](#), [94](#)).
- Dixon, B. J., Daly, M. J., Chan, H., Vescan, A. D., Witterick, I. J. and Irish, J. C. (2013), 'Surgeons blinded by enhanced navigation: the effect

- of augmented reality on attention', *Surgical Endoscopy* **27**(2), 454–461. doi: [10.1007/s00464-012-2457-3](https://doi.org/10.1007/s00464-012-2457-3) (page 78).
- Doshi, A., Morris, B. and Trivedi, M. (2011), 'On-road prediction of driver's intent with multimodal sensory cues', *IEEE Pervasive Computing* **10**(3), 22–34. doi: [10.1109/MPRV.2011.38](https://doi.org/10.1109/MPRV.2011.38) (page 32).
- Doshi, A. and Trivedi, M. M. (2012), 'Head and eye gaze dynamics during visual attention shifts in complex environments', *Journal of Vision* **12**(2), 9–9. doi: [10.1167/12.2.9](https://doi.org/10.1167/12.2.9) (page 25).
- Dragan, A. D., Lee, K. C. T. and Srinivasa, S. S. (2013), Legibility and predictability of robot motion, in 'ACM/IEEE International Conference on Human-Robot Interaction', pp. 301–308. (page 112).
- Dragan, A. D. and Srinivasa, S. S. (2013), 'A policy-blending formalism for shared control', *The International Journal of Robotics Research* **32**(7), 790–805. doi: [10.1177/0278364913490324](https://doi.org/10.1177/0278364913490324) (pages 23, 46).
- Driver, J. and Spence, C. (2000), 'Multisensory perception: Beyond modularity and convergence', *Current Biology* **10**(20), R731 – R735. doi: [10.1016/S0960-9822\(00\)00740-5](https://doi.org/10.1016/S0960-9822(00)00740-5) (page 115).
- Durham, J. W. and Bullo, F. (2008), Smooth Nearness-Diagram Navigation, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 690–695. doi: [10.1109/IROS.2008.4651071](https://doi.org/10.1109/IROS.2008.4651071) (pages 45, 47).
- Elsdon, J. and Demiris, Y. (2018), Augmented Reality for Feedback in a Shared Control Spraying Task, in 'IEEE International Conference on Robotics and Automation', pp. 1939–1946. doi: [10.1109/ICRA.2018.8461179](https://doi.org/10.1109/ICRA.2018.8461179) (page 57).
- Erdogan, A. and Argall, B. D. (2017), 'The effect of robotic wheelchair control paradigm and interface on user performance, effort and preference: An experimental assessment', *Robotics and Autonomous Systems* **94**, 282–297. doi: [10.1016/j.robot.2017.04.013](https://doi.org/10.1016/j.robot.2017.04.013) (pages 24, 25, 116).
- Ernst, M. O. and Bühlhoff, H. H. (2004), 'Merging the senses into a robust percept', *Trends in Cognitive Sciences* **8**(4), 162–169. doi: [10.1016/j.tics.2004.02.002](https://doi.org/10.1016/j.tics.2004.02.002) (page 115).
- Ezeh, C., Trautman, P., Devigne, L., Bureau, V., Babel, M. and Carlson, T. (2017), Probabilistic vs linear blending approaches to shared control for wheelchair driving, in 'International Conference on Rehabilitation Robotics', pp. 835–840. (pages 24, 25, 46, 47).

- Fehr, L., Langbein, W. E. and Skaar, S. B. (2000), 'Adequacy of power wheelchair control interfaces for persons with severe disabilities: a clinical survey.', *Journal of Rehabilitation Research and Development* 37(3), 353–360. (pages 40, 47, 143).
- Fischer, T., Chang, H. J. and Demiris, Y. (2018), Rt-gene: Real-time eye gaze estimation in natural environments, in 'European Conference on Computer Vision'. (page 144).
- Fortuin, V., Hüser, M., Locatello, F., Strathmann, H. and Rätsch, G. (2019), Deep Self-Organization: Interpretable Discrete Representation Learning on Time Series, in 'International Conference on Learning Representations'. (pages 36, 84, 91, 103).
- Fox, D., Burgard, W. and Thrun, S. (1997), 'The dynamic window approach to collision avoidance', *IEEE Robotics and Automation Magazine* 4(1), 23–33. doi: 10.1109/100.580977 (pages 45, 46).
- Fox, M., Long, D. and Magazzeni, D. (2017), Explainable Planning, in 'International Joint Conference on Artificial Intelligence Workshop on Explainable AI'. (pages 16, 27, 35, 49, 66).
- Fraccaro, M., Sønderby, S. K., Paquet, U. and Winther, O. (2016), Sequential Neural Models with Stochastic Layers, in 'Advances in Neural Information Processing Systems', Curran Associates, Inc., pp. 2199–2207. (pages 81, 87, 92, 93, 102, 107).
- Gallese, V. and Goldman, A. (1998), 'Mirror neurons and the simulation theory of mind-reading', *Trends in Cognitive Sciences* 2(12), 493–501. doi: 10.1016/S1364-6613(98)01262-5 (page 29).
- Gallese, V., Keysers, C. and Rizzolatti, G. (2004), 'A unifying view of the basis of social cognition', *Trends in Cognitive Sciences* 8(9), 396–403. doi: <https://doi.org/10.1016/j.tics.2004.07.002> (page 29).
- Ghorbel, M., Pineau, J., Gourdeau, R., Javdani, S. and Srinivasa, S. (2018), 'A Decision-Theoretic Approach for the Collaborative Control of a Smart Wheelchair', *International Journal of Social Robotics* 10(1), 131–145. doi: 10.1007/s12369-017-0434-7 (pages 24, 25).
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. and Kagal, L. (2018), Explaining explanations: An overview of interpretability of machine learning, in 'IEEE International Conference on Data Science and Advanced Analytics', pp. 80–89. doi: 10.1109/DSAA.2018.00018 (pages 27, 52).

- Goodrich, M. A. and Olsen, D. R. (2003), Seven principles of efficient human robot interaction, in 'IEEE International Conference on Systems, Man and Cybernetics', Vol. 4, pp. 3942–3948. doi: [10.1109/ICSMC.2003.1244504](https://doi.org/10.1109/ICSMC.2003.1244504) (pages [15](#), [16](#), [27](#), [51](#), [72](#)).
- Goyal, A., Sordoni, A., Côté, M.-A., Ke, N. R. and Bengio, Y. (2017), Z-Forcing: Training Stochastic Recurrent Networks, in 'Advances in Neural Information Processing Systems', Curran Associates, Inc., pp. 6713–6723. (pages [81](#), [87](#), [107](#)).
- Grafton, S. T. (2009), 'Embodied cognition and the simulation of action to understand others', *Annals of the New York Academy of Sciences* **1156**, 97–117. doi: [10.1111/j.1749-6632.2009.04425.x](https://doi.org/10.1111/j.1749-6632.2009.04425.x) (pages [29](#), [30](#), [32](#)).
- Graves, A. and Schmidhuber, J. (2005), 'Framewise phoneme classification with bidirectional LSTM and other neural network architectures', *Neural Networks* **18**(5), 602–610. (page [93](#)).
- Grèzes, J., Armony, J. L., Rowe, J. and Passingham, R. E. (2003), 'Activations related to “mirror” and “canonical” neurones in the human brain: an fMRI study', *NeuroImage* **18**(4), 928–937. doi: [https://doi.org/10.1016/S1053-8119\(03\)00042-9](https://doi.org/10.1016/S1053-8119(03)00042-9) (page [30](#)).
- Hamilton, A. F. and Grafton, S. T. (2007), 'The motor hierarchy: from kinematics to goals and intentions', *Sensorimotor foundations of higher cognition* **22**, 381–408. (pages [30](#), [34](#), [115](#)).
- Han, J.-S., Bien, Z. Z., Kim, D.-J., Lee, H.-E. and Kim, J.-S. (2003), Human-machine interface for wheelchair control with EMG and its evaluation, in 'Annual International Conference of the IEEE Engineering in Medicine and Biology Society', Vol. 2, pp. 1602–1605. doi: [10.1109/IEMBS.2003.1279672](https://doi.org/10.1109/IEMBS.2003.1279672) (page [143](#)).
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J. and Parasuraman, R. (2011), 'A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction', *Human Factors* **53**(5), 517–527. doi: [10.1177/0018720811417254](https://doi.org/10.1177/0018720811417254) (pages [28](#), [65](#)).
- Hart, S. G. and Staveland, L. E. (1988), Development of nasa-tlx (task load index): Results of empirical and theoretical research, in P. A. Hancock and N. Meshkati, eds, 'Human Mental Workload', Vol. 52 of *Advances in Psychology*, North-Holland, pp. 139 – 183. doi: [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9) (page [25](#)).

- Hesslow, G. (2002), 'Conscious thought as simulation of behaviour and perception', *Trends in Cognitive Sciences* 6(6), 242–247. doi: [10.1016/S1364-6613\(02\)01913-7](https://doi.org/10.1016/S1364-6613(02)01913-7) (pages [29](#), [33](#)).
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M. M., Mohamed, S. and Lerchner, A. (2017), beta-vae: Learning basic visual concepts with a constrained variational framework, in 'International Conference on Learning Representations'. (pages [89](#), [98](#), [105](#)).
- Ho, N. (2013), 'Finding Optimal Rotation and Translation Between Corresponding 3D points'. (page [58](#)).
- Hochreiter, S. and Schmidhuber, J. (1997), 'Long Short-Term Memory', *Neural Computation* 9(8), 1735–1780. doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735) (page [93](#)).
- Hoffman, M. D. and Johnson, M. J. (2016), Elbo surgery: yet another way to carve up the variational evidence lower bound, in 'Workshop in Advances in Approximate Bayesian Inference, NIPS', Vol. 1, p. 2. (page [90](#)).
- Hsieh, J.-T., Liu, B., Huang, D.-A., Fei-Fei, L. F. and Niebles, J. C. (2018), Learning to Decompose and Disentangle Representations for Video Prediction, in 'Advances in Neural Information Processing Systems', pp. 517–526. (pages [36](#), [84](#), [89](#), [91](#), [94](#), [96](#), [97](#), [107](#)).
- Hsu, W.-N., Zhang, Y. and Glass, J. (2017), Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data, in 'Advances in Neural Information Processing Systems', pp. 1878–1889. (pages [36](#), [84](#), [91](#), [107](#)).
- Hsu, W.-N., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Wang, Y., Cao, Y., Jia, Y., Chen, Z., Shen, J. and others (2018), 'Hierarchical generative modeling for controllable speech synthesis', *arXiv preprint arXiv:1810.07217*. (pages [36](#), [84](#), [89](#), [90](#), [91](#), [98](#), [107](#)).
- Hu, Y., Zhan, W., Sun, L. and Tomizuka, M. (2019), Multi-modal Probabilistic Prediction of Interactive Behavior via an Interpretable Model, in 'IEEE Intelligent Vehicles Symposium', pp. 557–563. doi: [10.1109/IVS.2019.8813796](https://doi.org/10.1109/IVS.2019.8813796) (pages [33](#), [36](#), [83](#), [84](#), [108](#)).
- Hu, Y., Zhan, W. and Tomizuka, M. (2018), Probabilistic Prediction of Vehicle Semantic Intention and Motion, in 'IEEE Intelligent Vehicles Symposium', pp. 307–313. doi: [10.1109/IVS.2018.8500419](https://doi.org/10.1109/IVS.2018.8500419) (pages [33](#), [82](#), [104](#), [107](#)).

- Huang, C.-M., Andrist, S., Sauppe, A. and Mutlu, B. (2015), 'Using gaze patterns to predict task intent in collaboration', *Frontiers in Psychology* 6, 1049. doi: [10.3389/fpsyg.2015.01049](https://doi.org/10.3389/fpsyg.2015.01049) (page 32).
- Huang, C.-M. and Mutlu, B. (2016), Anticipatory Robot Control for Efficient Human-Robot Collaboration, in 'ACM/IEEE International Conference on Human Robot Interaction', HRI '16, IEEE Press, pp. 83–90. (pages 23, 32).
- Huang, S. H., Held, D., Abbeel, P. and Dragan, A. D. (2019), 'Enabling Robots to Communicate Their Objectives', *Autonomous Robots* 43(2), 309–326. doi: [10.1007/s10514-018-9771-0](https://doi.org/10.1007/s10514-018-9771-0) (pages 112, 114, 116).
- Iezzoni, L. I., McCarthy, E. P., Davis, R. B. and Siebens, H. (2001), 'Mobility difficulties are not only a problem of old age', *Journal of General Internal Medicine* 16(4), 235–243. doi: [10.1046/j.1525-1497.2001.016004235.x](https://doi.org/10.1046/j.1525-1497.2001.016004235.x) (page 39).
- Itti, L. and Koch, C. (2001), 'Computational modelling of visual attention.', *Nature Reviews Neuroscience* 2(3), 194–203. doi: [10.1038/35058500](https://doi.org/10.1038/35058500) (page 145).
- Ivanovic, B., Schmerling, E., Leung, K. and Pavone, M. (2018), Generative Modeling of Multimodal Multi-Human Behavior, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 3088–3095. doi: [10.1109/IROS.2018.8594393](https://doi.org/10.1109/IROS.2018.8594393) (page 35).
- Jacob, R. J. (1995), 'Eye tracking in advanced interface design', *Virtual environments and advanced interface design* 258, 288. (page 145).
- Jain, A., Singh, A., Koppula, H. S., Soh, S. and Saxena, A. (2016), Recurrent Neural Networks for driver activity anticipation via sensory-fusion architecture, in 'IEEE International Conference on Robotics and Automation', pp. 3118–3125. doi: [10.1109/ICRA.2016.7487478](https://doi.org/10.1109/ICRA.2016.7487478) (page 116).
- Jain, S. and Argall, B. (2018), Recursive Bayesian Human Intent Recognition in Shared-Control Robotics, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 3905–3912. doi: [10.1109/IROS.2018.8593766](https://doi.org/10.1109/IROS.2018.8593766) (pages 23, 32).
- Jain, S. and Argall, B. (2019), 'Probabilistic Human Intent Recognition for Shared Autonomy in Assistive Robotics', *Journal of Human-Robot Interaction* 9(1). doi: [10.1145/3359614](https://doi.org/10.1145/3359614) (pages 23, 26, 29, 32, 82, 107, 112, 116).
- Jang, E., Gu, S. and Poole, B. (2016), 'Categorical reparameterization with gumbel-softmax', *arXiv preprint arXiv:1611.01144* . (pages 89, 90, 92, 95).

- Javdani, S., Srinivasa, S. S. and Bagnell, J. A. (2015), 'Shared Autonomy via Hindsight Optimization', *Robotics science and systems: online proceedings* . doi: [10.15607/RSS.2015.XI.032](https://doi.org/10.15607/RSS.2015.XI.032) (pages [23](#), [26](#), [32](#), [82](#), [107](#)).
- Jeannerod, M. (2001), 'Neural simulation of action: A unifying mechanism for motor cognition', *NeuroImage* **14**(1), S103–S109. doi: [10.1006/nimg.2001.0832](https://doi.org/10.1006/nimg.2001.0832) (pages [29](#), [32](#)).
- Jiang, Z., Zheng, Y., Tan, H., Tang, B. and Zhou, H. (2017), Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering, in 'International Joint Conference on Artificial Intelligence', AAAI Press, pp. 1965–1972. (pages [35](#), [89](#), [91](#), [94](#), [95](#)).
- Jordan, M. I. and Rumelhart, D. E. (1992), 'Forward models: Supervised learning with a distal teacher', *Cognitive Science* **16**(3), 307–354. doi: [10.1016/0364-0213\(92\)90036-T](https://doi.org/10.1016/0364-0213(92)90036-T) (pages [34](#), [42](#), [82](#)).
- Karniel, A. (2002), 'Three creatures named "forward model"', *Neural Networks* **15**(3), 305–307. doi: [10.1016/S0893-6080\(02\)00020-5](https://doi.org/10.1016/S0893-6080(02)00020-5) (pages [33](#), [42](#), [43](#)).
- Kassner, M., Patera, W. and Bulling, A. (2014), Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction, in 'ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication', UbiComp '14 Adjunct, ACM, pp. 1151–1160. doi: [10.1145/2638728.2641695](https://doi.org/10.1145/2638728.2641695) (pages [41](#), [71](#), [72](#), [144](#)).
- Kim, H., Gabbard, J. L., Anon, A. M. and Misu, T. (2018), 'Driver Behavior and Performance with Augmented Reality Pedestrian Collision Warning: An Outdoor User Study', *IEEE Transactions on Visualization and Computer Graphics* **24**(4), 1515–1524. doi: [10.1109/TVCG.2018.2793680](https://doi.org/10.1109/TVCG.2018.2793680) (pages [28](#), [69](#)).
- Kingma, D. P. and Ba, J. (2014), 'Adam: A Method for Stochastic Optimization', *arXiv preprint arXiv:1412.6980* . (pages [95](#), [102](#)).
- Kingma, D. P., Mohamed, S., Jimenez Rezende, D. and Welling, M. (2014), Semi-supervised Learning with Deep Generative Models, in 'Advances in Neural Information Processing Systems', pp. 3581–3589. (pages [35](#), [84](#)).
- Kingma, D. P. and Welling, M. (2013), 'Auto-Encoding Variational Bayes', *arXiv preprint arXiv:1312.6114* . (pages [18](#), [35](#), [81](#), [83](#), [84](#), [85](#), [86](#), [89](#), [90](#), [107](#)).
- Kirby, R., Swuste, J., Dupuis, D. J., MacLeod, D. A. and Monroe, R. (2002), 'The Wheelchair Skills Test: A pilot study of a new outcome measure', *Archives of Physical Medicine and Rehabilitation* **83**(1), 10–18. doi: [10.1053/APMR.2002.26823](https://doi.org/10.1053/APMR.2002.26823) (pages [25](#), [59](#), [61](#)).

- Koppula, H. S. and Saxena, A. (2016), 'Anticipating Human Activities Using Object Affordances for Reactive Robotic Response', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(1), 14–29. doi: [10.1109/TPAMI.2015.2430335](https://doi.org/10.1109/TPAMI.2015.2430335) (page 32).
- Kosiorrek, A., Kim, H., Teh, Y. W. and Posner, I. (2018), Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects, in S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds, 'Advances in Neural Information Processing Systems', Curran Associates, Inc., pp. 8606–8616. (pages 36, 94, 97, 107).
- Krishnan, R. G., Shalit, U. and Sontag, D. (2017), Structured Inference Networks for Nonlinear State Space Models, in 'AAAI Conference on Artificial Intelligence', AAAI'17, AAAI Press, pp. 2101–2109. (pages 81, 87, 93, 107).
- Ktena, S. I., Abbott, W. and Faisal, A. A. (2015), A virtual reality platform for safe evaluation and training of natural gaze-based wheelchair driving, in 'IEEE/EMBS Conference on Neural Engineering', pp. 236–239. doi: [10.1109/NER.2015.7146603](https://doi.org/10.1109/NER.2015.7146603) (pages 40, 53, 143, 144, 145).
- Kucukyilmaz, A. and Demiris, Y. (2018), 'Learning Shared Control by Demonstration for Personalized Wheelchair Assistance', *IEEE Transactions on Haptics* 11(3), 431–442. doi: [10.1109/TOH.2018.2804911](https://doi.org/10.1109/TOH.2018.2804911) (pages 24, 26, 113).
- Kuhn, H. W. (1955), 'The hungarian method for the assignment problem', *Naval research logistics quarterly* 2(1-2), 83–97. (page 96).
- Lakhmani, S., Abich, J., Barber, D. and Chen, J. (2016), A proposed approach for determining the influence of multimodal robot-of-human transparency information on human-agent teams, in D. D. Schmorrow and C. M. Fidopiastis, eds, 'Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience', Springer International Publishing, pp. 296–307. (page 116).
- Leaman, J. and La, H. M. (2017), 'A Comprehensive Review of Smart Wheelchairs: Past, Present, and Future', *IEEE Transactions on Human-Machine Systems* 47(4), 486–489. doi: [10.1109/THMS.2017.2706727](https://doi.org/10.1109/THMS.2017.2706727) (pages 39, 40).
- Lee, K., Su, Y., Kim, T. K. and Demiris, Y. (2013), 'A syntactic approach to robot imitation learning using probabilistic activity grammars', *Robotics and Autonomous Systems* 61(12), 1323–1334. doi: [10.1016/j.robot.2013.08.003](https://doi.org/10.1016/j.robot.2013.08.003) (pages 32, 115).

- Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H. S. and Chandraker, M. (2017), DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents, *in* 'IEEE Conference on Computer Vision and Pattern Recognition', pp. 336–345. doi: [10.1109/CVPR.2017.233](https://doi.org/10.1109/CVPR.2017.233) (pages [35](#), [116](#)).
- Li, H., Kutbi, M., Li, X., Cai, C., Mordohai, P. and Hua, G. (2016), An ego-centric computer vision based co-robot wheelchair, *in* 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 1829–1836. doi: [10.1109/IROS.2016.7759291](https://doi.org/10.1109/IROS.2016.7759291) (pages [40](#), [143](#), [145](#)).
- Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B. and Bachem, O. (2019), Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, *in* 'International Conference on Machine Learning', Vol. 97, pp. 4114–4124. (pages [35](#), [36](#), [82](#), [84](#), [95](#), [103](#), [105](#), [108](#)).
- Losey, D. P., McDonald, C. G., Battaglia, E. and O'Malley, M. K. (2018), 'A Review of Intent Detection, Arbitration, and Communication Aspects of Shared Control for Physical Human-Robot Interaction', *Applied Mechanics Reviews* **70**(1), 10804–10819. doi: [10.1115/1.4039145](https://doi.org/10.1115/1.4039145) (pages [22](#), [25](#), [26](#), [29](#), [32](#), [42](#), [48](#), [51](#), [67](#), [82](#)).
- Losey, D. P., Srinivasan, K., Mandlekar, A., Garg, A. and Sadigh, D. (2019), 'Controlling Assistive Robots with Learned Latent Actions', *arXiv preprint arXiv:1909.09674*. (pages [24](#), [25](#), [35](#), [107](#)).
- Lyons, J. B. (2013), Being Transparent about Transparency: A Model for Human-Robot Interaction, *in* 'AAAI Spring Symposium: Trust and Autonomous Systems'. (pages [16](#), [27](#), [28](#), [36](#), [48](#), [52](#), [79](#), [108](#)).
- Lyons, J. B. and Havig, P. R. (2014), Transparency in a Human-Machine Context: Approaches for Fostering Shared Awareness/Intent, *in* 'Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments', Springer International Publishing, pp. 181–190. (pages [16](#), [27](#), [28](#), [36](#), [48](#), [52](#), [66](#), [79](#), [108](#)).
- Maaten, L. v. d. and Hinton, G. (2008), 'Visualizing data using t-sne', *Journal of machine learning research* **9**, 2579–2605. (page [98](#)).
- Maddison, C. J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A. and Teh, Y. (2017), Filtering Variational Objectives, *in* 'Advances in Neural Information Processing Systems', Curran Associates, Inc., pp. 6573–6583. (page [102](#)).

- Maddison, C. J., Mnih, A. and Teh, Y. W. (2016), 'The concrete distribution: A continuous relaxation of discrete random variables', *arXiv preprint arXiv:1611.00712* . (pages 89, 90, 92, 95).
- Matsubara, T., Miro, J. V., Tanaka, D., Poon, J. and Sugimoto, K. (2015), Sequential intention estimation of a mobility aid user for intelligent navigational assistance, in 'IEEE International Symposium on Robot and Human Interactive Communication', pp. 444–449. doi: 10.1109/ROMAN.2015.7333580 (pages 31, 33).
- Matsumoto, Y., Ino, T. and Ogsawara, T. (2001), Development of intelligent wheelchair system with face and gaze based interface, in 'IEEE International Workshop on Robot and Human Interactive Communication', pp. 262–267. doi: 10.1109/ROMAN.2001.981912 (page 145).
- Metz, D. H. (2000), 'Mobility of older people and their quality of life', *Transport Policy* 7(2), 149–152. doi: [https://doi.org/10.1016/S0967-070X\(00\)00004-4](https://doi.org/10.1016/S0967-070X(00)00004-4) (page 39).
- Milgram, P., Rastogi, A. and Grodski, J. (1995), Telerobotic control using augmented reality, in 'IEEE International Workshop on Robot and Human Communication', pp. 21–29. doi: 10.1109/ROMAN.1995.531930 (pages 26, 111).
- Milgram, P., Zhai, S., Drascic, D. and Grodski, J. (1993), Applications of augmented reality for human-robot communication, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', Vol. 3, pp. 1467–1472. doi: 10.1109/IROS.1993.583833 (page 26).
- Minguez, J., Lamiriaux, F. and Laumond, J.-P. (2016), Motion Planning and Obstacle Avoidance, in 'Springer Handbook of Robotics', Springer International Publishing, pp. 1177–1201. doi: 10.1007/978-3-319-32552-1\_47 (pages 44, 45).
- Minguez, J. and Montano, L. (2009), 'Extending Collision Avoidance Methods to Consider the Vehicle Shape, Kinematics, and Dynamics of a Mobile Robot', *IEEE Transactions on Robotics* 25(2), 367–381. doi: 10.1109/TRO.2009.2011526 (page 45).
- Mujahed, M., Fischer, D. and Mertsching, B. (2018), 'Admissible gap navigation: A new collision avoidance approach', *Robotics and Autonomous Systems* 103, 93–110. doi: 10.1016/J.ROBOT.2018.02.008 (pages 43, 45, 46, 47, 142).

- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. (2019), 'Interpretable machine learning: definitions, methods, and applications', *arXiv preprint arXiv:1901.04592* . (pages 27, 53).
- Murphy, K. P. and Paskin, M. A. (2002), Linear-time inference in Hierarchical HMMs, in T. G. Dietterich, S. Becker and Z. Ghahramani, eds, 'Advances in Neural Information Processing Systems', MIT Press, pp. 833–840. (page 33).
- Narayanan, V. K., Spalanzani, A. and Babel, M. (2016), A semi-autonomous framework for human-aware and user intention driven wheelchair mobility assistance, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 4700–4707. doi: 10.1109/IROS.2016.7759691 (page 31).
- Nehaniv, C. L. and Dautenhahn, K. (2001), 'Like me?- measures of correspondence and imitation', *Cybernetics and Systems* 32(1-2), 11–51. doi: 10.1080/019697201300001803 (pages 34, 112).
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A. Y. (2011), Multimodal Deep Learning, in 'International Conference on Machine Learning', ICML'11, Omnipress, pp. 689–696. (page 116).
- Nicolis, D., Zanchettin, A. M. and Rocco, P. (2018), Human Intention Estimation based on Neural Networks for Enhanced Collaboration with Robots, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 1326–1333. doi: 10.1109/IROS.2018.8594415 (pages 32, 48).
- Nikolaidis, S., Zhu, Y. X., Hsu, D. and Srinivasa, S. (2017), Human-robot mutual adaptation in shared autonomy, in 'ACM/IEEE International Conference on Human-Robot Interaction', pp. 294–302. (page 113).
- Nisbet, C. (1996), 'Smart' Wheelchairs for Mobility Training', *Technology and Disability* 5(1), 49–62. doi: 10.3233/TAD-1996-5107 (page 61).
- Nisbet, P. D. (2002), Who's intelligent? Wheelchair, driver or both?, in 'Proceedings of the International Conference on Control Applications', Vol. 2, pp. 760–765. doi: 10.1109/CCA.2002.1038696 (pages 15, 22, 47, 53).
- Noda, K., Arie, H., Suga, Y. and Ogata, T. (2014), 'Multimodal integration learning of robot behavior using deep neural networks', *Robotics and Autonomous Systems* 62(6), 721–736. doi: 10.1016/j.robot.2014.03.003 (page 116).

- Ognibene, D. and Demiris, Y. (2013), Towards Active Event Recognition, in 'International Joint Conference on Artificial Intelligence', IJCAI '13, AAAI Press, pp. 2495–2501. (page 32).
- Pacherie, E. (2008), 'The phenomenology of action: A conceptual framework', *Cognition* **107**(1), 179–217. doi: [10.1016/j.cognition.2007.09.003](https://doi.org/10.1016/j.cognition.2007.09.003) (pages 24, 30, 31, 34, 115).
- Pellegrinelli, S., Admoni, H., Javdani, S. and Srinivasa, S. (2016), Human-robot shared workspace collaboration via hindsight optimization, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 831–838. doi: [10.1109/IROS.2016.7759147](https://doi.org/10.1109/IROS.2016.7759147) (pages 23, 32).
- Pentland, A. and Liu, A. (1999), 'Modeling and Prediction of Human Behavior', *Neural Computation* **11**(1), 229–242. doi: [10.1162/089976699300016890](https://doi.org/10.1162/089976699300016890) (page 33).
- Pineau, J., West, R., Atrash, A., Villemure, J. and Routhier, F. (2011), 'On the feasibility of using a standardized test for evaluating a speech-controlled smart wheelchair', *International Journal of Intelligent Control and Systems* **16**(2), 124–131. (page 25).
- Poon, J., Cui, Y., Miro, J. V., Matsubara, T. and Sugimoto, K. (2017), Local driving assistance from demonstration for mobility aids, in 'IEEE International Conference on Robotics and Automation', pp. 5935–5941. doi: [10.1109/ICRA.2017.7989699](https://doi.org/10.1109/ICRA.2017.7989699) (pages 31, 44, 100).
- Quigley, M., Conley, K., Gerkey, B. P., Faust, J., Foote, T., Leibs, J., Berger, E., Wheeler, R. and Ng, A. Y. (2009), ROS: an open-source Robot Operating System, in 'ICRA Workshop on Open Source Software'. (pages 41, 54, 71, 101, 141).
- Reddy, S., Dragan, A. and Levine, S. (2018), Shared Autonomy via Deep Reinforcement Learning, in 'Proceedings of Robotics: Science and Systems'. doi: [10.15607/RSS.2018.XIV.005](https://doi.org/10.15607/RSS.2018.XIV.005) (pages 24, 25, 49, 107).
- Reimer, B. and Mehler, B. (2011), 'The impact of cognitive workload on physiological arousal in young adult drivers: A field study and simulation validation', *Ergonomics* **54**(10), 932–942. doi: [10.1080/00140139.2011.604431](https://doi.org/10.1080/00140139.2011.604431) (page 25).
- Rezende, D. J., Mohamed, S. and Wierstra, D. (2014), Stochastic Backpropagation and Approximate Inference in Deep Generative Models, in 'Internation-

- tional Conference on Machine Learning', *Proceedings of Machine Learning Research*, pp. 1278–1286. (pages [18](#), [35](#), [81](#), [83](#), [84](#), [85](#), [86](#), [89](#), [90](#), [107](#)).
- Rizzolatti, G., Fadiga, L., Gallese, V. and Fogassi, L. (1996), 'Premotor cortex and the recognition of motor actions', *Cognitive Brain Research* **3**(2), 131–141. doi: [https://doi.org/10.1016/0926-6410\(95\)00038-0](https://doi.org/10.1016/0926-6410(95)00038-0) (page [30](#)).
- Sarabia, M. and Demiris, Y. (2013), A humanoid robot companion for wheelchair users, in G. Herrmann, M. J. Pearson, A. Lenz, P. Bremner, A. Spiers and U. Leonards, eds, 'International Conference on Social Robotics', ICSR 2013, Springer International Publishing, pp. 432–441. doi: [10.1007/978-3-319-02675-6\\_43](https://doi.org/10.1007/978-3-319-02675-6_43) (page [40](#)).
- Schettino, V. and Demiris, Y. (2019), Inference of user-intention in remote robot wheelchair assistance using multimodal interfaces, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 4600–4606. (pages [24](#), [48](#)).
- Seligman, M. E. (1972), 'Learned helplessness', *Annual Review of Medicine* **23**, 407–412. doi: [10.1146/annurev.me.23.020172.002203](https://doi.org/10.1146/annurev.me.23.020172.002203) (page [22](#)).
- Shi, W., Zhou, H., Miao, N. and Li, L. (2020), 'Dispersed exponential family mixture vaes for interpretable text generation'. (pages [89](#), [90](#), [91](#), [98](#)).
- Sibirtseva, E., Ghadirzadeh, A., Leite, I., Björkman, M. and Kragic, D. (2019), Exploring temporal dependencies in multimodal referring expressions with mixed reality, in 'International Conference on Human-Computer Interaction', Springer, pp. 108–123. (pages [115](#), [116](#)).
- Sibirtseva, E., Kontogiorgos, D., Nykvist, O., Karaoguz, H., Leite, I., Gustafson, J. and Kragic, D. (2018), A comparison of visualisation methods for disambiguating verbal requests in human-robot interaction, in 'IEEE International Symposium on Robot and Human Interactive Communication', pp. 43–50. (pages [26](#), [28](#), [68](#), [70](#), [77](#), [78](#)).
- Simpson, R. C. (2005), 'Smart wheelchairs: A literature review', *Journal of Rehabilitation Research and Development* **42**(4), 423. doi: [10.1682/JRRD.2004.08.0101](https://doi.org/10.1682/JRRD.2004.08.0101) (page [39](#)).
- Simpson, R. C. (2008), 'How many people would benefit from a smart wheelchair?', *Journal of Rehabilitation Research and Development* **45**(1), 53–72. doi: [10.1682/JRRD.2007.01.0015](https://doi.org/10.1682/JRRD.2007.01.0015) (pages [39](#), [61](#), [72](#), [73](#), [143](#)).

- Simpson, R. C. and Levine, S. P. (1997), Adaptive shared control of a smart wheelchair operated by voice control, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', Vol. 2, pp. 622–626. doi: [10.1109/IROS.1997.655076](https://doi.org/10.1109/IROS.1997.655076) (page 143).
- Simpson, R., LoPresti, E., Hayashi, S., Nourbakhsh, I. and Miller, D. (2004), 'The Smart Wheelchair Component System', *Journal of Rehabilitation Research and Development* 41(3b), 429. doi: [10.1682/JRRD.2003.03.0032](https://doi.org/10.1682/JRRD.2003.03.0032) (pages 16, 22, 39).
- Soh, H. and Demiris, Y. (2012), Towards Early Mobility Independence : An Intelligent Paediatric Wheelchair with Case Studies, in 'IROS Workshop on Progress, Challenges and Future Perspectives in Navigation and Manipulation Assistance for Robotic Wheelchairs'. (pages 40, 45, 46, 141).
- Soh, H. and Demiris, Y. (2013), When and how to help: An iterative probabilistic model for learning assistance by demonstration, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 3230–3236. doi: [10.1109/IROS.2013.6696815](https://doi.org/10.1109/IROS.2013.6696815) (page 23).
- Soh, H. and Demiris, Y. (2015), 'Learning Assistance by Demonstration: Smart Mobility With Shared Control and Paired Haptic Controllers', *Journal of Human-Robot Interaction* 4(3), 76–100. doi: [10.5898/JHRI.4.3.Soh](https://doi.org/10.5898/JHRI.4.3.Soh) (pages 23, 113).
- Soh, H., Pan, S., Chen, M. and Hsu, D. (2019), Trust Dynamics and Transfer Across Human-robot Interaction Tasks: Bayesian and Neural Computational Models, in 'International Joint Conference on Artificial Intelligence', AAAI Press, pp. 6226–6230. (pages 28, 65).
- Sohn, K., Lee, H. and Yan, X. (2015), Learning Structured Output Representation using Deep Conditional Generative Models, in C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds, 'Advances in Neural Information Processing Systems', Curran Associates, Inc., pp. 3483–3491. (page 35).
- Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B. and Mehler, B. (2014), Classifying driver workload using physiological and driving performance data, in 'Annual ACM conference on Human factors in computing systems', CHI '14, ACM, pp. 4057–4066. doi: [10.1145/2556288.2557068](https://doi.org/10.1145/2556288.2557068) (page 25).

- Srivastava, N., Mansimov, E. and Salakhutdinov, R. (2015), Unsupervised Learning of Video Representations Using LSTMs, in 'International Conference on International Conference on Machine Learning', pp. 843–852. (pages [18](#), [81](#), [94](#), [95](#), [96](#), [142](#)).
- Srivastava, N. and Salakhutdinov, R. R. (2012), Multimodal Learning with Deep Boltzmann Machines, in F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems', Curran Associates, Inc., pp. 2222–2230. (page [116](#)).
- Stein, B. E. and Stanford, T. R. (2008), 'Multisensory integration: current issues from the perspective of the single neuron', *Nature Reviews Neuroscience* **9**, 255–266. doi: [10.1038/nrn2331](#) (page [115](#)).
- Sutskever, I., Vinyals, O. and Le, Q. V. (2014), Sequence to Sequence Learning with Neural Networks, in 'Advances in Neural Information Processing Systems', Curran Associates, Inc., pp. 3104–3112. (page [93](#)).
- Tanwani, A. K. and Calinon, S. (2017), A generative model for intention recognition and manipulation assistance in teleoperation, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 43–50. doi: [10.1109/IROS.2017.8202136](#) (pages [33](#), [83](#), [104](#)).
- Theodorou, A. (2019), AI governance through a transparency lens, PhD thesis, University of Bath. (pages [27](#), [52](#)).
- Tomasello, M., Carpenter, M., Call, J., Behne, T. and Moll, H. (2005), 'Understanding and sharing intentions: The origins of cultural cognition', *Behavioral and Brain Sciences* **28**(5), 675–691. doi: [10.1017/S0140525X05000129](#) (pages [16](#), [24](#), [29](#), [30](#), [48](#), [82](#), [93](#)).
- Trick, S., Koert, D., Peters, J. and Rothkopf, C. A. (2019), Multimodal uncertainty reduction for intention recognition in human-robot interaction, in 'IEEE/RSJ International Conference on Intelligent Robots and Systems', pp. 7009–7016. (page [32](#)).
- Tschannen, M., Bachem, O. and Lucic, M. (2018), 'Recent advances in autoencoder-based representation learning', *arXiv preprint arXiv:1812.05069*. (pages [35](#), [84](#), [105](#)).
- Tsotsos, J. K. (2001), 'Motion understanding: Task-directed attention and representations that link perception with action', *International Journal of Computer Vision* **45**(3), 265–280. doi: [10.1023/A:1013666302043](#) (page [145](#)).

- van den Oord, A., Vinyals, O. and Kavukcuoglu, K. (2017), Neural Discrete Representation Learning, in 'Advances in Neural Information Processing Systems', pp. 6306–6315. (pages 35, 36, 84, 89, 98).
- Vanhooydonck, D., Demeester, E., Hüntemann, A., Philips, J., Vanacker, G., Van Brussel, H. and Nuttin, M. (2010), 'Adaptable navigational assistance for intelligent wheelchairs by means of an implicit personalized user model', *Robotics and Autonomous Systems* 58(8), 963–977. doi: 10.1016/j.robot.2010.04.002 (page 116).
- Vasquez, D., Fraichard, T., Aycard, O. and Laugier, C. (2008), 'Intentional motion on-line learning and prediction', *Machine Vision and Applications* 19(5), 411–425. doi: 10.1007/s00138-007-0116-9 (page 33).
- Vasquez, D., Fraichard, T. and Laugier, C. (2009), 'Growing Hidden Markov Models: An Incremental Tool for Learning and Predicting Human and Vehicle Motion', *The International Journal of Robotics Research* 28(11–12), 1486–1506. doi: 10.1177/0278364909342118 (page 33).
- Viswanathan, P., Zambalde, E. P., Foley, G., Graham, J. L., Wang, R. H., Adhikari, B., Mackworth, A. K., Mihailidis, A., Miller, W. C. and Mitchell, I. M. (2017), 'Intelligent wheelchair control strategies for older adults with cognitive impairment: user attitudes, needs, and preferences', *Autonomous Robots* 41(3), 539–554. doi: 10.1007/s10514-016-9568-y (pages 24, 25, 40, 64, 111, 116, 143).
- Walker, M., Hedayati, H., Lee, J. and Szafir, D. (2018), Communicating Robot Motion Intent with Augmented Reality, in 'ACM/IEEE International Conference on Human-Robot Interaction', ACM, pp. 316–324. doi: 10.1145/3171221.3171253 (pages 28, 67, 69, 78, 112).
- Wang, J. M., Fleet, D. J. and Hertzmann, A. (2008), 'Gaussian Process Dynamical Models for Human Motion', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(2), 283–298. doi: 10.1109/TPAMI.2007.1167 (page 33).
- Wang, Z., Merel, J. S., Reed, S. E., de Freitas, N., Wayne, G. and Heess, N. (2017), Robust Imitation of Diverse Behaviors, in 'Advances in Neural Information Processing Systems', Curran Associates, Inc., pp. 5320–5329. (pages 32, 35).
- Wang, Z., Mülling, K., Deisenroth, M. P., Amor, H. B., Vogt, D., Schölkopf, B. and Peters, J. (2013), 'Probabilistic movement modeling for intention

- inference in human–robot interaction’, *The International Journal of Robotics Research* **32**(7), 841–858. doi: [10.1177/0278364913478447](https://doi.org/10.1177/0278364913478447) (pages [32](#), [33](#), [83](#)).
- Watanabe, A., Ikeda, T., Morales, Y., Shinozawa, K., Miyashita, T. and Hagita, N. (2015), Communicating robotic navigational intentions, in ‘IEEE/RSJ International Conference on Intelligent Robots and Systems’, pp. 5763–5769. doi: [10.1109/IROS.2015.7354195](https://doi.org/10.1109/IROS.2015.7354195) (page [64](#)).
- Wolpert, D. M., Doya, K. and Kawato, M. (2003), ‘A unifying computational framework for motor control and social interaction’, *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **358**(1431), 593–602. doi: [10.1098/rstb.2002.1238](https://doi.org/10.1098/rstb.2002.1238) (pages [29](#), [30](#), [31](#), [33](#), [114](#)).
- Wolpert, D. M., Miall, R. and Kawato, M. (1998), ‘Internal models in the cerebellum’, *Trends in Cognitive Sciences* **2**(9), 338–347. doi: [10.1016/S1364-6613\(98\)01221-2](https://doi.org/10.1016/S1364-6613(98)01221-2) (pages [29](#), [33](#), [43](#), [114](#)).
- Wortham, R. H., Theodorou, A. and Bryson, J. J. (2017), Improving robot transparency: Real-time visualisation of robot ai substantially improves understanding in naive observers, in ‘IEEE International Symposium on Robot and Human Interactive Communication’, pp. 1424–1431. doi: [10.1109/ROMAN.2017.8172491](https://doi.org/10.1109/ROMAN.2017.8172491) (page [28](#)).
- Xie, A., Losey, D. P., Tolsma, R., Finn, C. and Sadigh, D. (2020), ‘Learning Latent Representations to Influence Multi-Agent Interaction’, *arXiv preprint arXiv:2011.06619* . (page [107](#)).
- Yingzhen, L. and Mandt, S. (2018), Disentangled Sequential Autoencoder, in ‘International Conference on Machine Learning’, pp. 5670–5679. (pages [36](#), [84](#), [91](#), [92](#), [93](#), [107](#)).
- Zhang, C., Bütepage, J., Kjellström, H. and Mandt, S. (2019), ‘Advances in Variational Inference’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(8), 2008–2026. doi: [10.1109/TPAMI.2018.2889774](https://doi.org/10.1109/TPAMI.2018.2889774) (pages [85](#), [86](#), [87](#)).
- Zhu, C., Cheng, Q. and Sheng, W. (2008), Human intention recognition in Smart Assisted Living Systems using a Hierarchical Hidden Markov Model, in ‘2008 IEEE International Conference on Automation Science and Engineering’, pp. 253–258. doi: [10.1109/COASE.2008.4626513](https://doi.org/10.1109/COASE.2008.4626513) (page [33](#)).
- Zolotas, M. and Demiris, Y. (2019), Towards Explainable Shared Control using Augmented Reality, in ‘IEEE/RSJ International

- Conference on Intelligent Robots and Systems', pp. 3020–3026. doi: [10.1109/IROS40897.2019.8968117](https://doi.org/10.1109/IROS40897.2019.8968117) (pages [39](#), [51](#), [52](#)).
- Zolotas, M.** and Demiris, Y. (2020), 'Transparent Intent for Explainable Shared Control in Assistive Robotics', *International Joint Conference on Artificial Intelligence* pp. 5184–5185. doi: [10.24963/ijcai.2020/732](https://doi.org/10.24963/ijcai.2020/732) (page [52](#)).
- Zolotas, M.**, Elsdon, J. and Demiris, Y. (2018), Head-Mounted Augmented Reality for Explainable Robotic Wheelchair Assistance, in 'IEEE International Conference on Intelligent Robots and Systems', pp. 1823–1829. doi: [10.1109/IROS.2018.8594002](https://doi.org/10.1109/IROS.2018.8594002) (pages [51](#), [52](#)).

## SOFTWARE PACKAGES

---

The following outlines several software packages and open-source contributions that were developed on during this thesis.

### A.1 LOCALISATION AND NAVIGATION PACKAGES FOR MOBILE ROBOTS

Localisation and navigation are integral capabilities required of the robotic wheelchair architecture described in Section 3.2. The Robot Operating System (ROS) is a software framework that offers a collection of libraries and tools that can accommodate these capabilities (Quigley et al., 2009). For example, autonomous navigation can be handled using the ROS navigation stack<sup>1</sup>, which is composed of modules for path planning, localisation, and so forth. However, our wheelchair scenario demanded for a human user's active engagement with the robot, and so we applied the Shared Control (SC) implementation delineated in Section 3.3 for all navigation purposes. The localisation module on the other hand was composed of many ROS packages available online. In particular, we utilised *gmapping*<sup>2</sup> and *hector\_mapping*<sup>3</sup> for Simultaneous Localization and Mapping (SLAM), as well as *amcl*<sup>4</sup> for pure localisation on pre-constructed maps.

It is worth highlighting that we established a unified software repository for the laboratory to easily configure localisation and navigation packages, as well as our SC method. As a result, many generic localisation and navigation routines are now available to other laboratory mobile platforms (e.g. our paediatric smart wheelchair Soh and Demiris, 2012) with minimal parameter configuration.

### A.2 OPEN-SOURCE CONTRIBUTIONS

Three open-source code repositories are by-products of this thesis:

---

<sup>1</sup> <http://wiki.ros.org/navigation>  
<sup>2</sup> <http://wiki.ros.org/gmapping>  
<sup>3</sup> [http://wiki.ros.org/hector\\_mapping](http://wiki.ros.org/hector_mapping)  
<sup>4</sup> <http://wiki.ros.org/amcl>

- **Reactive Assistance:** A ROS package of our SC method presented in Section 3.3. Contains a C++ implementation of the admissible gap navigation algorithm (Mujahed et al., 2018) adapted for the purpose of SC. Currently operates on differential-drive rectangular mobile bases with full Field of View (FoV) 2D planar Light Detection And Ranging (LiDAR) data.
- **Scan-Image Converter:** A ROS package containing a C++ node that takes as input LiDAR messages and converts them into image representations (polar to Cartesian space translation). Node was utilised in the gridmap processing phase of Section 4.2.1.1.
- **Disentangled Sequence Clustering Variational Autoencoder:** Code repository for a deep generative model that simultaneously clusters *and* disentangles latent representations of sequences. The model has been built using the TensorFlow (Abadi et al., 2016) and TensorFlow Probability (Dillon et al., 2017) libraries for deep learning. A set of scripts are also provided in this repository to evaluate the clustering framework on the Moving MNIST dataset (Srivastava et al., 2015). A similar but separate repository is available for the robotic wheelchair experiment of Section 5.5. This code is not yet publicly available, as the work is still under review.

## EYE-GAZE WHEELCHAIR

---

In this appendix, a non-invasive, eye-gaze controlled wheelchair is introduced as a hands-free solution for power mobility users.

### B.1 MOTIVATION

A traditional input method for controlling electric powered wheelchairs is a joystick, however not all patients possess the cognitive or motor capacity to safely navigate an environment via this input device (Fehr et al., 2000). Alternative hands-free solutions include using voice recognition (Simpson and Levine, 1997), electromyography (Han et al., 2003) and head gestures (Li et al., 2016) for wheelchair navigation. Yet there remains a small group of individuals with severe motor-disabilities (e.g. amyotrophic lateral sclerosis or spinal cord injury) who still cannot comfortably or easily employ these control interfaces (Fehr et al., 2000; Simpson, 2008; Viswanathan et al., 2017). For this target population, brain-computer interfaces (BCIs) and eye movement have become increasingly popular options (Carlson and Del R. Millan, 2013; Ktena et al., 2015).

Despite the exotic appeal of BCIs, there are a couple of advantages in favour of eye-directed wheelchairs. First, BCIs require immense levels of concentration and impose intense demands on mental workload (Carlson and Del R. Millan, 2013). These burdens are far less prominent in gaze-based controllers. Second, brain signals are noisy by nature and thus result in diminished information rates (Carlson and Del R. Millan, 2013). In contrast, gaze signals can yield high-information throughput for tasks such as wheelchair navigation (Abbott and Faisal, 2012; Ktena et al., 2015). Given these advantages, we settled on eye gaze as a non-invasive means of controlling smart wheelchairs.

### B.2 SYSTEM DESIGN

In the following, we describe the operation of our proposed eye-gaze controlled wheelchair (high-level system diagram shown in Figure B.1). First, the front-facing webcam of the on-board laptop records RGB images of

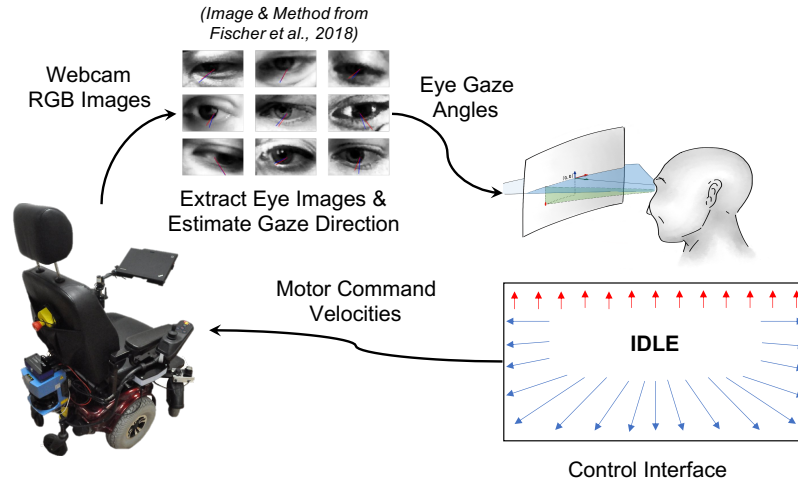


Figure B.1: System diagram of the eye-gaze controlled smart wheelchair platform. The user-facing webcam of the laptop records RGB images and feeds them into the gaze estimation network of Fischer et al. (2018). This network then outputs a tuple of gaze angles in real-time, which are subsequently converted into appropriate motor command velocities through a threshold-based control interface. Red arrows indicate reverse motion (looking upwards), blue arrows indicate linear (looking down) and rotational movement (looking left/right), and 'IDLE' is to remain stationary.

the user and feeds them into a neural network algorithm for gaze estimation (Fischer et al., 2018). The method of Fischer et al. (2018) is accurate, versatile in natural environments, and capable of processing images in real-time, making it particularly suitable for our application. Once this gaze estimation method extracts eye gaze angles from the user's face images, these angles are then converted into motor command velocities. Our controller for this conversion is inspired by the "natural free-view" interface for wheelchairs presented in Ktena et al. (2015), where a *continuous* control field determines linear and rotational command velocities according to user gaze angles.

A key benefit of our eye-controlled wheelchair setup is its non-intrusive functionality. Without requiring any headsets or eye trackers (Kassner et al., 2014), users can comfortably sit in the wheelchair and issue motor commands. Moreover, safety is guaranteed, as all the input velocities still undergo adjustment using the Shared Control (SC) methodology of Section 3.3. Videos of the overall system are available online<sup>1</sup>.

Although we have developed a promising prototype for wheelchair users with severe motor-disability, there are still numerous challenges to overcome.

<sup>1</sup> Supplementary video material of the eye-gaze controlled wheelchair operating *indoors*: <https://www.youtube.com/watch?v=Ey2G2HUYG6Y>  
As well as *outdoors*: <https://www.youtube.com/watch?v=deBISM4cRI>

One of the most pressing challenges is to distinguish goal-directed eye behaviour from “attentive” saccades, which is known as the “Midas touch problem” (Jacob, 1995). Prior systems have mitigated this problem by either relying on explicit screen-interfaces to issue wheelchair commands or restricting gaze-prediction models to only output a discrete set of states (Ktena et al., 2015; Li et al., 2016; Matsumoto et al., 2001). Nevertheless, a more fitting approach may instead be to computationally model visual attention based on eye gaze behaviour (Itti and Koch, 2001). The aim of this model would be to discriminate between stimulus-driven influences (e.g. salient regions in the scene) and goal-directed contributions relating to the task-at-hand (Borji et al., 2014; Tsotsos, 2001).



AUGMENTED REALITY FOR DUAL-ARM ROBOTS

---

This appendix presents an Augmented Reality (AR) Head-Mounted Display (HMD) interface for an application involving a dual-arm collaborative robot, namely the ABB YuMi. Whilst the thesis has focused on robotic wheelchairs and Shared Control (SC), the following demonstrates how AR headsets can improve *transparency* in Human-Robot Interactions (HRIs) other than assistive navigation.

## C.1 EXPLAINING AFFORDABLE ROBOT BEHAVIOURS

The objective of this application is to visually explain the YuMi's affordable behaviours through an AR HMD interface. In turn, the user will better understand the capabilities of the robot and anticipate how different behaviours might unfold. Moreover, users are able to explicitly initiate robot actions via hand gestures, and so the AR interface also acts as a controller.

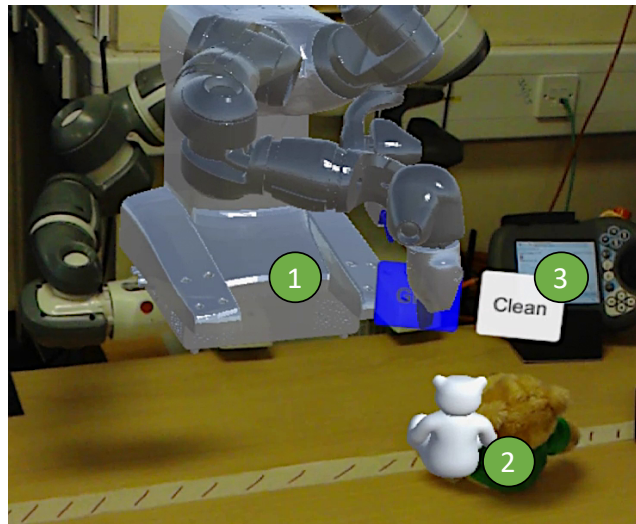


Figure C.1: AR view of: (1) A dual-arm collaborative robot (the ABB YuMi) and its overlaid 3D model to project intent; (2) An interactable object (a teddy bear); (3) An action (“clean”) that the user can select for the robot to perform.

Figure C.1 illustrates a wearer's view of the interaction with YuMi captured in AR, where affordable robot behaviours are projected as indicators

of intent. To create this user perspective, a depth camera attached on top of the YuMi first extracts the identities and locations of objects on the shared workspace, such that arm motion trajectories can be planned and executed to manipulate these objects. We then broadcast the detected object poses and robot arm joint poses for different manipulation trajectories into a concurrent [AR](#) application. The resulting 3D visual overlay<sup>1</sup> shows users how different grasping actions (e.g. “give” or “clean” the teddy bear) pan out when activated.

## C.2 LINKS TO EXPLAINABLE SHARED CONTROL

Although users can initiate collaborative actions through the [AR](#) interface and do not need to physically interact or share control with the robot, the guidelines of Explainable Shared Control ([XSC](#)) still benefit visualisation design. For instance, projecting the arm manipulator’s intentions is clearly an example of *predictive* [AR](#) feedback. Note that these 3D arm trajectories are portrayed even when a user has not yet selected a specific behaviour and is instead contemplating options by hovering their head gaze over different action panels (shown in Figure [C.1](#)). Furthermore, *contextual* [AR](#) cues are also utilised in the depiction of affordances, i. e. objects that the robot can interact with. Despite not yet conducting a user study with this [AR](#)-YuMi setup, we envision that by adhering to the [XSC](#) guidelines in creating the proposed interface, there would be corresponding improvements in transparency.

---

<sup>1</sup> Supplementary video material of the interaction is provided in: <https://www.youtube.com/watch?v=n7dFFJBrMbA>

AUTHOR'S PUBLICATIONS

---

The following is a compilation of all peer-reviewed publications related to this thesis, supplemented with a summary of their relevance to the thesis.

**Zolotas, M.,** Elsdon, J. and Demiris, Y. (2018), Head-Mounted Augmented Reality for Explainable Robotic Wheelchair Assistance, *in* 'IEEE International Conference on Intelligent Robots and Systems', pp. 1823–1829. doi: [10.1109/IROS.2018.8594002](https://doi.org/10.1109/IROS.2018.8594002).

- Presents the first instance of an augmented reality headset being incorporated into a robotic wheelchair system and investigates the influence of different interface design options through a pilot user study. Results from this study demonstrate that care should be taken in the presentation of information, with effort-reducing cues for augmented information acquisition (for example, a rear-view display) being the most appreciated.
- Section [4.2](#) is based on this conference paper.

**Zolotas, M.** and Demiris, Y. (2019), Towards Explainable Shared Control using Augmented Reality, *in* 'IEEE International Conference on Intelligent Robots and Systems', pp. 3020–3026. doi: [10.1109/IROS40897.2019.8968117](https://doi.org/10.1109/IROS40897.2019.8968117).

- Introduces the paradigm of Explainable Shared Control and provides guidelines on how to best visualise the internal state of shared control in order to combat model misalignment. Findings from an assistive navigation experiment with users indicate that the paradigm facilitates transparent assistance by improving recovery times from adverse events associated with model misalignment.
- Section [3.3](#) and Section [4.3](#) are based on this conference paper.

**Zolotas, M.** and Demiris, Y. (2020), Transparent Intent for Explainable Shared Control in Assistive Robotics, *in* 'International Joint Conference on Artificial Intelligence', pp. 5184–5185. doi: [10.24963/ijcai.2020/732](https://doi.org/10.24963/ijcai.2020/732).

- Summarises our research on establishing transparency in shared control by enabling both the robot and human to understand each other's underlying "intent".
- Parts of Chapter [4](#) are based on this extended abstract.

In addition to the above, the following article has also been submitted to a peer-reviewed journal.

**Zolotas, M.** and Demiris, Y. (2020), Disentangled Sequence Clustering for Human Intention Inference, *submitted*.

- Contributes a clustering algorithm involving a latent variable model to infer human intent from robot sensory observations. The proposed algorithm is generally applicable to sequential data and was thus evaluated on both a video dataset for unsupervised classification, as well as data collected during a robotic wheelchair experiment for intention inference. Experimental results indicate that human intent can be interpreted from the model's learnt latent space, without requiring any supervision.
- Chapter 5 is based on this journal submission.

This last publication is unrelated to the thesis, but is a by-product of my teaching supervision in robotics during this Doctor of Philosophy.

Bagga, S., Maurer, B., Miller, T., Quinlan, L., Silvestri, L., Wells, D., Winqvist, R., **Zolotas, M.** and Demiris, Y. (2019), instruMentor: An Interactive Robot for Musical Instrument Tutoring, *in 'Towards Autonomous Robotic Systems' (Oral Presentation)*, pp. 303-315. doi: [10.1007/978-3-030-23807-0\\_25](https://doi.org/10.1007/978-3-030-23807-0_25).

- Master's degree group project on a musical instrument tutor robot for students learning the recorder.
- Project was transformed under my supervision into a paper accepted at the leading annual UK Robotics conference.
- Edited and presented the paper, which was also awarded the conference Prize for Innovation, sponsored by the Institution of Engineering and Technology (IET) Robotics.