HUMAN FACTORS OF SEMI-AUTONOMOUS ROBOTS FOR URBAN SEARCH AND RESCUE

KATHARINA M. GABRECHT, BEng MSc

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

February 2016

Abstract

During major disasters or other emergencies, Urban Search and Rescue (USAR) teams are responsible for extricating casualties safely from collapsed urban structures. The rescue work is dangerous due to possible further collapse, fire, dust or electricity hazards. Sometimes the necessary precautions and checks can last several hours before rescuers are safe to start the search for survivors. Remote controlled rescue robots provide the opportunity to support human rescuers to search the site for trapped casualties while they remain in a safe place.

The research reported in this thesis aimed to understand how robot behaviour and interface design can be applied to utilise the benefits of robot autonomy and how to inform future human-robot collaborative systems. The data was analysed in the context of USAR missions when using semi-autonomous remote controlled robot systems. The research focussed on the influence of robot feedback, robot reliability, task complexity, and transparency. The influence of these factors on trust, workload, and performance was examined. The overall goal of the research was to make the life of rescuers safer and enhance their performance to help others in distress.

Data obtained from the studies conducted for this thesis showed that semi-autonomous robot reliability is still the most dominant factor influencing trust, workload, and team performance. A robot with explanatory feedback was perceived as more competent, more efficient and less malfunctioning. The explanatory feedback was perceived as a clearer type of communication compared to concise robot feedback. Higher levels of robot transparency were perceived as more trustworthy. However, single items on the trust questionnaire were manipulated and further investigation is necessary. However, neither explanatory feedback from the robot nor robot transparency, increased team performance or mediated workload levels.

Task complexity mainly influenced human-robot team performance and the participants' control allocation strategy. Participants allowed the robot to find more targets and missed more robot errors in the high complexity

conditions compared to the low task complexity conditions. Participants found more targets manually in the low complexity tasks.

In addition, the research showed that recording the observed robot performance (the performance of the robot that was witnessed by the participant) can help to identify the cause of contradicting results: participants might not have noticed some of the robots mistakes and therefore they were not able to distinguish between the robot reliability levels.

Furthermore, the research provided a foundation of knowledge regarding the real world application of USAR in the United Kingdom. This included collecting knowledge via an autoethnographic approach about working processes, command structures, currently used technical equipment, and attitudes of rescuers towards robots. Also, recommendations about robot behaviour and interface design were collected throughout the research.

However, recommendations made in the thesis include consideration of the overall outcome (mission performance) and the perceived usefulness of the system in order to support the uptake of the technology in real world applications. In addition, autonomous features might not be appropriate in all USAR applications. When semi-autonomous robot trials were compared to entirely manual operation, only the robot with an average of 97% reliability significantly increased the team performance and reduced the time needed to complete the USAR scenario compared to the manually operated robot. Unfortunately, such high robot success levels do not exist to date.

This research has contributed to our understanding of the factors influencing human-robot collaboration in USAR operations, and provided guidance for the next generation of autonomous robots.

Acknowledgements

I would like to thank my supervisors, Professor Sarah Sharples, Dr. Glyn Lawson, and Dr. Harshada Patel for their expertise, encouragement, support and guidance. Thanks to Professor Sarah Sharples for being a great source of knowledge and high level insight and tirelessly reminding me of keeping my focus and not led astray by so many other interesting research questions. Thanks to Dr. Harshada Patel for her detailed, thorough, and valuable corrections and notes, that made my work so much clearer and better to read. Thanks to Dr. Glyn Lawson for the support and key questions that made me rethink my arguments and structures.

Further, I would like to thank the entire Human Factors Research Group for their support, motivation, and being like a family to me. This PhD would have been a very lonely and sad journey without you. My thanks go to Richard Eastgate for his constantly helpful technical support and to Kirstie Dane and Anne Floyde for their administrative support and bearing with me when a form just didn't provide the fields that I needed.

Acknowledgements go to the University of Nottingham, who awarded me with the Vice Chancellor's Scholarship for Research Excellence (EU) and funded the three years of my PhD generously. Without this support it may not have been possible to conduct this research.

Furthermore, I would like to thank Shaun Yates who enabled me to conduct my research visit at the Fire Service College in Moreton-in-Marsh and who supported me throughout the experiment. And thanks to all USAR course delegates who welcomed me and let me be part of something great.

Particularly thanks go to Dimitrios Darzentas from the Horizon Doctoral Training Centre at the University of Nottingham. He endured my never ending questions and iteration loops and provided substantial help by developing the robot control script for the virtual environment in Unity.

I was particularly happy that Prof. Dr. Peter Hancock was my external examiner. Thank you for your support and guidance. I am also very thankful for the great experiences while staying in the USA. Also, thanks to Gabriella for being such a great person. Furthermore, thanks to Associate Professor

Dr. Sue Cobb for her intense support and utterly valuable insights. Thanks to Dr. Cath Harvey for being my internal examiner and doing such a great job.

Thanks to my friends who listened to my wining and laughing for the entire time of my PhD. In particular, thanks to Laura Lewis for her kind support on everything I asked her for. And thanks to Tang, who has been a loyal and motivating friend.

My thanks also go to my family: „Liebe Mama, lieber Papa, vielen Dank für Eure Fürsorge und Unterstützung. Ich liebe Euch". A very special thanks goes to my partner who encouraged, motivated, and supported me with all I needed to bring this piece of work together. My deepest gratitude to you.

I pay tribute to our rescuers, risking their lives on an everyday basis, to keep us safe.

Table of contents

Abstract ..... **i**

Acknowledgements ..... **iii**

Table of contents ..... **v**

List of figures ..... **xii**

List of tables ..... **xviii**

Acronyms and abbreviations ..... **xx**

1 Introduction **1**

1.1 Chapter overview 1

1.2 Background and motivation 1

1.3 What is a rescue robot? 3

1.4 Scope of PhD 6

1.5 Aims and objectives 7

1.6 Novel contributions 11

1.7 Thesis overview 12

1.8 Chapter summary 15

2 Literature review **17**

2.1 Chapter overview 17

2.2 Chapter introduction 17

2.3 Robotics in USAR 17

2.3.1 Commercially available robots 18

2.3.2 Deployment of robots and lessons learned 22

2.3.3 Robot autonomy levels 30

2.4 Human-robot teams 33

2.4.1 Theoretical Foundations of Teamwork 34

2.4.2 Single robot systems 36

2.4.3 Multi-robot systems 38
2.5 Trust in human-robot teams 39
2.5.1 Why is trust important? 39
2.5.2 What is trust? 40
2.5.3 Models of trust 42
2.6 Conclusion 59
2.7 Chapter summary 61
3 Methodology 63
3.1 Chapter overview 63
3.2 Measuring performance 63
3.2.1 System performance 63
3.2.2 Operator performance 64
3.2.3 Robot performance 65
3.3 Measuring trust 66
3.3.1 Trust in automation (Muir, 1989) 66
3.3.2 Jian trust scale (Jian et al., 1998) 67
3.3.3 HRI trust scale (Yagoda and Gillan, 2012) 67
3.3.4 Real-time trust (Desai, 2012) 68
3.3.5 Human-robot trust scale (Schaefer, 2013) 69
3.4 Experimental methodology 70
3.4.1 Approach to data collection and analysis 70
3.5 Equipment used during experiments 75
3.5.1 Robot system 75
3.5.2 The development of a Virtual USAR scenario in UNITY 76
3.6 Chapter summary 91
4 Study I - Urban Search and Rescue field work 93
4.1 Chapter overview 93
4.2 Introduction 93

4.3 Methodology 96
4.3.1 Participants 96
4.3.2 Materials 96
4.3.3 Experimental design 96
4.3.4 Procedure 97
4.4 Results 97
4.4.1 Collected background knowledge 97
4.4.2 Emerged factors of USAR work 108
4.4.3 Questionnaires and robot attitude 113
4.5 Implications for USAR robots 116
4.6 Discussion 120
4.7 Conclusion 122
4.8 Chapter summary 124
5 Study II - The influence of robot reliability indication and feedback 125
5.1 Chapter overview 125
5.2 Introduction 125
5.3 Methodology 128
5.3.1 Participants 128
5.3.2 Materials 128
5.3.3 Experimental design 130
5.3.4 Procedure 133
5.4 Results 134
5.4.1 General Questionnaire 134
5.4.2 Trial performance and robot perception 134
5.4.3 Visual attention allocation 141
5.4.4 Summary of quantitative results 143
5.4.5 Retrospective verbal protocol analysis 143

5.4.6 Interview analysis 165
5.5 Discussion 172
5.5.1 The influence of additional robot feedback 172
5.5.2 Combined discussion 174
5.5.3 Qualitative data analysis 174
5.5.4 Limitations and future work 175
5.6 Conclusion 176
5.7 Chapter summary 177
6 Study III - The influence of robot reliability and task complexity 179
6.1 Chapter overview 179
6.2 Introduction 179
6.3 Methodology 187
6.3.1 Participants 187
6.3.2 Experimental design 188
6.3.3 Materials 195
6.3.4 Procedure 197
6.3.5 Measures 198
6.4 Results 204
6.4.1 Pre-test: Trust questionnaires comparison (Schaefer/Muir) 204
6.4.2 Mixed mode results 207
6.4.3 Summary of mixed mode results 223
6.4.4 Comparison between manual and mixed mode results 225
6.4.5 Interviews 230
6.5 Discussion 239
6.5.1 The influence of robot reliability on independent variables .. 240
6.5.2 The influence of task complexity on independent variables . 242
6.5.3 Personality scores and rated task difficulty 245

6.5.4 Combined discussion 246
6.5.5 Comparison between manual and mixed mode groups 247
6.5.6 Qualitative data analysis regarding auto and manual mode usage, robot features, and trust 249
6.5.7 Limitations and future work 250
6.6 Conclusion 252
6.7 Chapter summary 253
7 Study IV - The influence of robot transparency and task complexity 255
7.1 Chapter overview 255
7.2 Introduction 255
7.3 Methodology 260
7.3.1 Participants 260
7.3.2 Experimental design 260
7.3.3 Materials 268
7.3.4 Procedure 268
7.4 Results 269
7.4.1 Trust 269
7.4.2 Objective performance 270
7.4.3 Observed robot performance 271
7.4.4 Workload 272
7.4.5 Subjective ratings 273
7.4.6 Event analysis 277
7.4.7 Summary of quantitative results 281
7.4.8 Post-task interview 282
7.5 Discussion 293
7.5.1 Recommendations for future interface designs 295
7.5.2 Combined discussion 296

7.5.3 Limitations and future work 297
7.6 Conclusion 298
7.7 Chapter summary 299
8 General Discussion 300
8.1 Chapter overview 300
8.2 Introduction 300
8.3 Discussion of research findings 302
8.4 Review of aims 306
8.5 Review of novel contributions 314
8.6 Recommendations for robot behaviour and interface design 317
8.7 Personal reflection on trust, human-robot collaboration, and the use of autonomous features 321
8.8 Limitations of research 324
8.9 Chapter summary 326
9 Conclusion and future work 327
9.1 Chapter overview 327
9.2 Concluding statement 327
9.3 Future work 329
9.4 Chapter summary 330
10 References 331
Appendix A - Study I: General Questionnaire 362
Appendix B - Study II: General questionnaire 368
Appendix C - Study II: Post-task questionnaire 371
Appendix D - Study II: RVP analysis; in-between events 374
Appendix E - Study III: General questionnaire 387
Appendix F - Study III: Post-task questionnaire 391
Appendix G - Study III: Programmed robot reliability 396
Appendix H - Study IV: General questionnaire 398

Appendix I - Study IV: Post-task questionnaire 400

Appendix J - Study IV: Analysis of trust questionnaire 403

Appendix K - Digital Appendix 404

List of figures

Figure 1 - CUTLASS Bomb disposal robot used by the British Army ("CUTLASS EOD robot [Image]," 2012) 4
Figure 2 - Pipe crawler Versatrax 150 ("Pipe crawler Versatrax 150 [Image]," 2015) 4
Figure 3 - Overview how studies will address aims and objectives..... 10
Figure 4 - TALON robot from QinetiQ ("TALON robot [Image]," 2015) 18
Figure 5 - Dragon Runner 10 from QinetiQ ("Dragon Runner 10 [Image]," 2015) 19
Figure 6 - 110 First Look from iRobot ("110 First Look [Image]," 2015)..... 20
Figure 7 - Recon ThrowBot LE from Recon Robotics ("Recon Scout - Throwbot LE [Image]," 2015)..... 20
Figure 8 - R2i2 Delta Extreme from RECCE robotics ("R2i2 Extreme [Image]," 2015) 21
Figure 9 - R2i2 Delta Micro from RECCE robotics ("R2i2 Delta Micro [Image]," 2014) 21
Figure 10 - Foster-Miller SOLEM ("Foster-Miller Solem [Image]," 2010)..... 22
Figure 11 - CAESAR robot with flippers ("CAESAR robot [Image]," 2015) 25
Figure 12 - NIFTi UAV (left) and UGV (right) platform ("NIFTi UGV [Image]," 2013) 27
Figure 13 - KOHGA3 ground robot ("KOHGA3 ground robot [Image]," 2011)..... 28
Figure 14 - Level of robot autonomy framework from Beer et al. (2014)..... 32
Figure 15 - "wakamaru" service robot ("Wakamaru [Image]," 2013)..... 37
Figure 16 - McKnight (1998), High level model of initial trust formation..... 43
Figure 17 - Dyadic model of trust in relationships from Simpson (2007). 45
Figure 18 - Technology acceptance model from Davis (1986); picture source: Zaied (2012) 47
Figure 19 - Time series model from Lee and Moray (1994) 49
Figure 20 - Simple qualitative model for trust dynamics based on experiences .. 50
Figure 21 - Trust in automation model from Hoff and Bashir (2014). Dotted arrows represent factors that can change with the course of a single interaction. 51
Figure 22 - Interdependence between system performance, operator trust and reliance strategy 53
Figure 23 - Human-Automation Collaboration Model (Gao et al., 2013) 55
Figure 24 - Factors of trust development in human-robot interaction. Factors included in the correlational analysis are starred (*). Factors included in the group difference analysis are crossed (+). (Hancock, Billings, Schaefer, et al., 2011).. 56

Figure 25 - Updated descriptive human-robot trust model (Schaefer, 2013) 57
Figure 26 - HARRT model from Desai (2012). Blue arrows indicate a positive relationship and orange arrows indicate a negative relationship. 58
Figure 27 - Selected factors that influence human-robot interaction.....60
Figure 28 - Area under the curve real-time trust measure 69
Figure 29 - Required hardware for rescue simulation: Laptop that runs the Unity program (left), a second screen (middle), XboX controller, second keyboard for participants. 78
Figure 30 – Example of a Unity environment (scene) 79
Figure 31 - Waypoint indicator and room label 80
Figure 32 - Rubble, chair, and other objects in the environment.....81
Figure 33 - Smoke particles in the environment 81
Figure 34 - Fire particles in the environment.....82
Figure 35 – Examples of targets in the environment. Top left to bottom right: Victim, hazard sign, bomb, weapons 82
Figure 36 - Collider configured as a trigger. Collider is invisible to the participant. 83
Figure 37 - Script of temperature trigger (collider) 84
Figure 38 - Waypoint with trigger.....85
Figure 39 - Waypoints in the environment with waypoint list.....86
Figure 40 - Script for reaching the next waypoint 87
Figure 41 - Navigation mesh visualised in blue, overlaid with the environment...87
Figure 42 - Script for displaying a message 88
Figure 43 – Rescue robot interface with all elements visible.....89
Figure 44 - Message from the robot displayed on the interface.....91
Figure 45 - SnakeEye monitor (left) and goose neck extension (right) 100
Figure 46 - Search Cam 3000 101
Figure 47 - DELSAR Life Detector LD3 with sensor channel monitor (left) and a sensor on a rubble pile (right).....101
Figure 48 - Gas monitor Impact Series from Honeywell 102
Figure 49 - Handler with search dog during training.....102
Figure 50 - Casualty under a collapsed structure and scent movements 103
Figure 51 - Two post vertical shores 104
Figure 52 - Casualty extrication with crow bars and wedges (left) and Paratech tripod for lifting (right).....105
Figure 53 - Hydraulic concrete chain saw 105
Figure 54 - Concrete breach 106

Figure 55 - Simplified incident organisational structure for Fire Service Operations 107
Figure 56 - A rescuer in a void 108
Figure 57 - Education level of USAR technicians 110
Figure 58 - Tool usage upside down in confined space.....111
Figure 59 – Percentage indication of NARS questionnaire percentage scores 115
Figure 60 - NARS relative median scores for each subset and the overall score 115
Figure 61 - The maze in which the video was recorded 130
Figure 62 – Example targets 131
Figure 63 – Relative mean performance bewteen Parker and Roy. Error bars show 95% confidence intervals.....135
Figure 64 – Relative mean performance between first and second performed task. Error bars show 95% confidence intervals. 135
Figure 65 - Robot effect on task order performance with 95% confidence intervals 136
Figure 66 – Robot communication ratings with confidence intervals (*significant difference).....138
Figure 67 – Robot perception ratings on a scale from 1 = "strongly disagree" to 7 = "strongly agree" with confidence intervals (* significant difference) 140
Figure 68 – Percentage rating of robot contribution to task success with confidence intervals.....141
Figure 69 – Attention allocation towards the robot in low and high reliability phase with confidence intervals 142
Figure 70 – Did participants realise a difference between the robots? 166
Figure 71 – Which robot is preferred over the other? 167
Figure 72 – Which robot would you trust more?.....169
Figure 73 – Which robot is more intelligent? 171
Figure 74 - Qualitative overview of research results of study II; positive influences are indicated with (+), negative influences indicated with (-).....174
Figure 75 - Task-component-factor-dimension framework (P. Liu & Li, 2012)...184
Figure 76 - Overview of reliability profiles.....189
Figure 77 - Screenshot from the low complexity task 191
Figure 78 - Screenshot from the middle complexity task.....191
Figure 79 - Screenshot from the high complexity task 192
Figure 80 - Example of victim (left), hazard sign (middle), and bomb (right) in the environment 193
Figure 81 - Interface of the rescue robot 194
Figure 82 - Experimental setup 196

Figure 83 - Secondary task screenshot.....196
Figure 84 – Comparison of correlations with study variables of the Muir trust score and Schaefer trust score (weak correlations are greyed out).....206
Figure 85 - Schaefer trust scores across robot reliability with 95% confidence intervals.....208
Figure 86 - Schaefer trust scores across task complexity with 95% confidence intervals.....209
Figure 87 – Workload (NASA TLX) across robot reliability; with 95% confidence intervals.....210
Figure 88 - Significantly different workload (NASA TLX) subscales with 95% confidence intervals.....210
Figure 89 - Objective team performance across reliability levels with 95% confidence intervals (bootstrapped).....212
Figure 90 - Objective team performance across complexity levels with 95% confidence intervals (bootstrapped).....213
Figure 91 - Observed robot performance across reliability with 95% confidence intervals (bootstrapped).....214
Figure 92 - Observed robot performance across task complexity with 95% confidence intervals (bootstrapped).....215
Figure 93 - Manual mode times across robot reliability with 95% confidence intervals (bootstrapped).....216
Figure 94 - Trial times across reliability with 95% confidence intervals (bootstrapped).....217
Figure 95 - Rated task difficulty box plots across complexity, whiskers (min/max) 219
Figure 96 - Rated robot performance box plots across reliability; whiskers (min/max).....221
Figure 97 - Rated self-performance box plots across complexity; whiskers (min/max).....222
Figure 98 – Comparison of Objective team performance between manual a with 95% confidence intervals (bootstrapped) 226
Figure 99 - Comparison of secondary task performance between manual and mixed mode group with 95% confidence intervals (bootstrapped).....228
Figure 100 - Comparison of trial times between manual and mixed mode group with 95% confidence intervals (bootstrapped) 229
Figure 101 - Rated task difficulty box plots across reliability levels and manual mode, whiskers (min/max).....230

Figure 102 - TBCA of the question as to why participants used auto mode, with item count in brackets.....231
Figure 103 - TBCA of the question as to why participants used manual mode, with item count in brackets.....233
Figure 104 - Top view map from the robot interface with navigation goal points (orange squares).....235
Figure 105 - TBCA of the question as to how participants used NGPs, with item count in brackets.....235
Figure 106 - TBCA of the question why participants gave different trust ratings, with item count in brackets.....237
Figure 107 - Qualitative overview of research results of study III; positive influences are indicated with (+), negative influences indicated with (-).....246
Figure 108 - Situation awareness-based Agent transparency model (Chen et al., 2014).....262
Figure 109 - Low transparency interface with highlighted display elements.....263
Figure 110 - High transparency interface with highlighted display elements.....265
Figure 111 - Low task complexity (editor view of LT-LC) with waypoints visualised. 266
Figure 112 - Example of victim in a low complexity environment.....266
Figure 113 - High task complexity (editor view of LT-HC) with waypoints visualised. 267
Figure 114 - High task complexity targets (left to right: weapon, hazard sign, victim).....267
Figure 115 - Schaefer (short) trust scores across the two transparency conditions with 95% confidence intervals.....270
Figure 116 - Performance across the two levels of complexity with 95% confidence intervals (bootstrapped).....271
Figure 117 - Significant workload (NASA TLX) subscale analysis of performance with 95% confidence intervals.....273
Figure 118 - Rated self-performance box plots across task complexity; whiskers show minimum and maximum values.....275
Figure 119 - Event distribution between the two task complexity levels.....279
Figure 120 - Event distribution between the two robot transparency levels.....280
Figure 121 - Pie chart of the interface preference of the participants.....283
Figure 122 - Number of comments of low transparency interface elements.....286
Figure 123 - Number of comments of high transparency interface elements.....288
Figure 124 - Situation awareness: percentage deviation across robot transparency levels.....292

Figure 125 - Qualitative overview of research results of study IV; positive influences are indicated with (+), negative influences indicated with (-).....296
Figure 126 - Overview of objectives and key research findings.....303
Figure 127 - Summary of research findings; black arrows indicate findings from the studies of this PhD; orange arrows indicate verification of research findings by other literature; green arrows indicate other findings from literature; positive influences are indicated with (+), negative influences are indicated with (-). 304
Figure 128 - Overview of the programmed reliability levels of the robot 396
Figure 129 - Overview of programmed robot reliability levels across task complexity 397
Figure 130 - Detail analysis of short trust questionnaire (* indicates the biggest changes between conditions) 403

List of tables

Table 1 -Overview of studies performed and aims/objectives addressed 11
Table 2 - Overview of levels of automation from Endsley and Kaber (1999, pp. 464–465) 31
Table 3 - The differences in feedback given by the two robots. 132
Table 4 - Workload ratings across conditions.....137
Table 5 - Summary of quantitative results 143
Table 6 - TBCA overview of sub-event: Participant waits for robot 145
Table 7 - TBCA overview of sub-event: Robot better 146
Table 8 - TBCA overview of sub-event: Found in low reliability 147
Table 9 - TBCA overview of sub-event: General feelings and general feedback .148
Table 10 - TBCA overview of sub-event: Secondary task 149
Table 11 TBCA overview of sub-event: Participant uncertain/unsure 150
Table 12 - TBCA overview of sub-event: Sub-event 1st target found.....151
Table 13 - TBCA overview of sub-event: General comments.....153
Table 14 - TBCA overview of sub-event: Sub event two mistakes in succession 156
Table 15 - TBCA overview of sub-event: General comments.....157
Table 16 - TBCA overview of sub-event: Low reliability phase 158
Table 17 - TBCA overview of sub-event: High reliability phase 160
Table 18 - Retrospective verbal protocol implications and conclusion.....164
Table 19 - TCBA content theme overview of question 1 166
Table 20 - TCBA content theme overview of question 2 168
Table 21 - TCBA content theme overview of question 3 169
Table 22 - TCBA content theme overview of question 4 171
Table 23 - Independent variable table 188
Table 24 - Task complexity modification overview.....190
Table 25 - Overview of event categories.....199
Table 26 - Example scenario with six events 202
Table 27 - Workload ratings across task complexity 211
Table 28 - Rated task difficulty across complexity 220
Table 29 - Rated robot performance across reliability 221
Table 30 - Rated self-performance across complexity 223
Table 31 - Summary of mixed mode results 225
Table 32 - Rated task difficulty table across reliability and manual mode 230
Table 33 - 2x2 mixed subject design with the variables task complexity and robot transparency.....260
Table 34 - Schaefer (short) trust scores across task complexity 269

Table 35 – Objective performance scores across robot transparency 271
Table 36 - Observed performance across conditions..... 272
Table 37 - Workload ratings across conditions..... 273
Table 38 - Rated task complexity across conditions 274
Table 39 - Rated task difficulty across conditions..... 275
Table 40 – Rated self-performance across complexity 276
Table 41 - Rated self- performance across robot transparency 276
Table 42 - Rated robot performance across conditions 277
Table 43 - Summary of quantitative results 282
Table 44 - TCBA content theme overview of interface preference comments 283
Table 45 - TBCA overview of sub-event: Attention allocation (robot/secondary task) 375
Table 46 - TBCA overview of sub-event: Attention allocation (switching) 378
Table 47 - TBCA overview of sub-event: Robot interface characteristics (positive, neutral, negative) 380
Table 48 - TBCA overview of sub-event: Robot interface characteristics (participant ideas) 383

Acronyms and abbreviations

DARPA	Defense Advanced Research Projects Agency
EOD	Explosive Ordnance Disposal
FSC	Fire Service College
GCSE	General Certificate of Secondary Education
GPS	Global Positioning System
HR	High Reliability
HRC	Human-robot collaboration
HRI	Human-robot interaction
HRTS	Human-robot trust scale (Schaefer, 2013)
IED	Improvised explosive device
JCS	Joint Cognitive Systems
LOA	Level of automation
LR	Low Reliability
NARS	Negative Attitude Toward Robot Scale
NASA TLX	National Aeronautics and Space Administration Task Load Index
NS	Not significant
PN	Participant number
PPE	Personal Protective Equipment
R	Effect size
r_s	Spearman's correlation coefficient
r_{sM}	Spearman's correlation coefficient for Muir trust questionnaire
r_{sS}	Spearman's correlation coefficient for Schaefer trust questionnaire
SAR	Search and Rescue
TBCA	Theme based content analysis
UAV	Unmanned air vehicle

UGV	Unmanned ground vehicle
UKFRS	United Kingdom Fire and Rescue Service
USAR	Urban Search and Rescue
WTC	World Trade Center

1 Introduction

1.1 Chapter overview

This chapter provides an introduction for the thesis and gives a brief background of the topic and explains why this topic is of importance to research and society. The aim of this work is to investigate how robot behaviour and interface design can be applied to utilise the benefits of robot autonomy and inform future human-robot collaborative systems. This is examined in the context of semi-autonomous remote controlled ground robots for Urban Search and Rescue (USAR). Also, this chapter provides an overview of the studies performed to address the individual aims and objectives. Furthermore, a brief overview of the thesis chapters is given.

1.2 Background and motivation

Over the past 30 years the number of natural and technological disasters has risen, as has their impact (Guha-Sapir, Below, & Hoyois, 2015). Earthquakes, storms and floods have the most economic impact (Munich Re, 2015a), whereby storms and earthquakes are the most deadly events (Munich Re, 2015b). The overall damage of natural disasters in USD in 2000 was 61 billion, compared to 371 billion USD in 2011 (Munich Re, 2015a).

The reason for such impacts of natural and technological disasters is mainly due to factors associated with urban occupation. Worldwide, more people live in urban areas and this trend is continuing to rise (United Nations, 2014). Nowadays ca. 54 per cent of the world's population live in urban areas and the United Nations project that by 2050, 66 per cent of the world's population will be urbanised (United Nations, 2014). This is in contrast to 30 per cent of the world's population living in highly populated areas 65 years ago. The more people live in urban areas, the higher will be the numbers of fatalities and economic impact of disasters (Deely et al., 2010). For example, highly populated areas, such as cities, are very vulnerable to the devastating effects of earthquakes. Buildings can collapse, infrastructure can be damaged and consequences such as the nuclear

incident at Fukushima, where the damage of a tsunami caused equipment failure and nuclear meltdowns, can arise. Next to natural disasters also terrorist attacks are posing a threat with increased numbers of attacks, such as explosive devices in crowded urban areas (Institute for Economic and Peace, 2014).

An increase in disasters, terrorist attacks, urbanisation, and their resulting rising impact demands highly qualified Urban Search and Rescue teams with enhanced capacities, flexibility, and appropriate, up-to-date equipment. In general, Urban Search and Rescue (USAR) teams consist of regular firefighters who have had additional USAR training. USAR teams are called when urban structures collapse and people are trapped or buried under rubble or debris. When arriving at an incident site the team has to find and extricate casualties as fast as possible. Their job is very dangerous and they risk their lives to help others (Cowman, Ferrari, & Liao-Troth, 2004).

Fast information collection is key to mission planning and success (The Fire Service College, 2014). Because information at the scene is very limited, often the number of casualties or their location on the incident site is unknown or based on estimates. The more is known about the scene, the better rescuers can plan ahead and aid people quickly and with fewer risks. Therefore, rescuers have to work fast and under constant time pressure in hostile environments. The golden rule of finding people highlights the reward of working fast because the most people alive can be rescued within the first hour (The Fire Service College, 2014). However, sometimes the circumstances at the incident site are too dangerous for rescuers to start their work. These restrictive conditions can include further collapse of buildings, fires, electricity hazards, etc. (The Fire Service College, 2014).

New emerging technologies create a high potential to enhance rescue operations and make the process of rescuing safer, more efficient and more effective. For example during the earthquake in Nepal (2015), rescuers used a specially designed radar, called FINDER, which can detect heart beats under 30 feet of rubble. It was the first time this NASA technology was used in a real world context. FINDER managed to find four people buried under crushed materials (Partnership for Public Service, 2015). In another

example, at the Washington DC mudslide in 2014, Murphy et al. (2015) used an unmanned air vehicle to successfully build 2D and 3D representations of the inaccessible regions of the slide. This allowed geologists and hydrologists to assess the imminent risk to rescuers from further slides and flooding, and provided a better overall understanding of the incident.

Technology holds promises for future rescue operations (Kruijff et al., 2014). Rescue robots can go where humans cannot. They can go into areas such as cavities (voids), which are too dangerous or even inaccessible for the rescue teams. The robot's main task is the collection of information during reconnaissance and mapping missions. This new technology offers the potential to support rescuers in their work and be beneficial for the overall rescue mission. Nevertheless, it is important that the design of these new technologies takes into account the context of their use and the human capabilities of those who will use them. The next paragraphs will clarify what a rescue robot is and where the focus of this PhD lies.

1.3 What is a rescue robot?

A robot in this work is described as a physical entity that is guided by a computer program or an electronic circuitry. A rescue robot is an unmanned, mobile, sensing and physically situated agent (Murphy, 2014). Rescue robots can have different application areas, they can operate on the ground, in the water, in the air or even in space. These mobile rescue robots are particularly challenging to develop. Most robots we hear from or even see, are industrial robots. To date they work very quickly, reliably, and accurately. But, the difference between these and rescue robots is that industrial robots perform pre-programmed repetitive tasks in a fixed environment, and a rescue robot needs to sense, adopt and act in a constantly changing environment. Furthermore, rescue robots need to take care of not destroying forensic evidence, causing rubble to move or bringing casualties further in danger (Murphy, 2014).

Figure 1 - CUTLASS Bomb disposal robot used by the British Army ("CUTLASS EOD robot [Image]," 2012)

Additionally, different robot tasks demand different types of robots. For instance, bomb disposal robots are mostly heavy (armoured) and utilise a robotic arm or a similar manipulator in order to diffuse a bomb remotely (see Figure 1). Another example is pipe inspection robots. As shown in Figure 2, they are designed for a single purpose: to inspect a pipe. They are not designed to drive on any other terrain.

Figure 2 - Pipe crawler Versatrax 150 ("Pipe crawler Versatrax 150 [Image]," 2015)

But rescue robots need to traverse unpredictable, changing, and dangerous terrains. They need to operate in areas where GPS or other signals are very limited.

To the knowledge of the author only one USAR robot is actually part of a rescue team in the American New Jersey Task Force One (Urban Search and Rescue state team). This robot is used for a very distinct purpose (J. Bastan, Task Force Leader on New Jersey Task Force 1, personal communication, September 9, 2014): to find out whether a person is dead or alive. If a person is deep buried in a rubble pile it is an enormously time consuming and demanding task to remove all rubble carefully and gain proper access to that person. Knowing if a person is actually dead or alive is vital to focus workforce on the right tasks. Maybe the person is already dead; therefore the rescuers can concentrate to rescue other people that are still alive.

Rescue robots need to be versatile and adaptable to cope with different tasks and environments (Shah & Choset, 2004). But there is always a trade-off between equipment/sensors and robot weight. More advanced and accurate sensors can provide a better understanding of the remote environment for the human and the robot (Fong, Kaber, Scholtz, & Schultz, 2004; Glas, Kanda, Ishiguro, & Hagita, 2012). They can inform the rescue teams more accurately in order to make better informed decisions. We can produce robots which have all sorts of sensors that will support the operator; however this will increase weight and bulk and the robot might not fit into narrow spaces or will bring rubble to a collapse (cf. Murphy, 2004). For this reason, rescue robots are small and portable (Casper & Murphy, 2003) - small enough to enter voids and light enough not to cause secondary collapses, but big enough to carry important sensors, an appropriate camera, and sufficient battery power. Robot technology is evolving every day and soon technology is able to produce reliable and robust robot systems. With the acceptance and use of rescue robots the life of a rescuer can be made safer and maybe rescue operations can become more efficient (Mioch, Smets, & Neerincx, 2012; Steinbauer, Maurer, & Krajnz, 2014).

The environmental influences and the robot capabilities also influence the human-robot interaction (HRI). Operators have to understand the world through the "eyes" of the robot. This understanding is a very important aspect of HRI and also influences the trust in the robot, acceptance and usage of the robotic system. For operators, remote presence is very demanding because their natural perception is impaired by being detached from the physical environment they have to explore (Chen, Haas, & Barnes, 2007).

With the intention to reduce rescuers' workload, robots are being developed with autonomous features. However, automatic robot features are not used (Larochelle, Kruijff, & Van Diggelen, 2013a), as technology is still prone to error, robots are only slowly being introduced into the area of search and rescue, and robots are far from standard rescue service equipment (Murphy, 2014). There are several issues as to why robots are slowly accepted and barely used. One issue is that standards for rescue robots are missing (Murphy, 2014). Rescue equipment needs to be tested and be of a certain standard to be useful, safe, and reliable (Messina & Jacoff, 2006). Therefore, purchasing a robot for the Fire and Rescue Service, who is responsible for the USAR teams in the U.K., is difficult or even impossible. Most robots used are research robots and they are deployed mostly after the incidents to find dead people. Furthermore, operators do not always place trust in the robotic system and the benefits of such systems cannot be exploited fully. As previously mentioned, technology to date is still prone to error which makes operators quickly lose trust in rescue robots. Why trust seems to be important in human-robot collaboration is part of the research focus of this PhD.

1.4 Scope of PhD

This PhD focusses on human-robot collaboration and interaction in the context of unmanned robots for reconnaissance and mapping on the ground. These are light and agile robots that search inaccessible areas with a camera and other sensors, such as temperature, air quality, 3D scanner etc.

While rescue robots are still prone to error and lacking capabilities in mobility, sensing and appropriate autonomous behaviour, the greatest human related challenge still posed is the limited understanding of human-robot interaction (Murphy, 2004). During human-robot interaction many human factors issues emerge: Is the technology understandable, easy to use, effective, and safe?

Technology is evolving and robots are getting more intelligent and have more degrees of autonomy, which can provide them with abilities such as achieving independently prescribed task objectives, adapting to environmental stages and internal states as well as developing their own objectives (Huang, Messina, & Albus, 2003). Each robot will have its own level of autonomy, but regardless of this level, with new technologies, new interaction styles and new challenges for humans arise (cf. Burghart & Steinfeld, 2008; Sklar et al., 2011). The more complex a system gets, the more human factors needs to be considered. It is hoped that autonomous features can alleviate operators from the workload, mediate error rates and enhance human-robot team performance. However, experiments showed (Lee & Moray, 1992; Lee & See, 2004; Parasuraman & Riley, 1997) that these features are not utilised by operators due to their lack of trust into the technology. The presented work (Chapter 5 to Chapter 7) examines the issue of trust between operators and semi-autonomous robot systems in order to enhance mission performance, reduce workload, optimise robot communication, and establish so called calibrated trust. The goal is not to maximise trust, but rather to ensure an appropriate level of trust. Low levels of trust lead to operators not appropriately using the supporting features; too high levels of trust make operators trust robots when they should not, which leads to further errors (Lee & See, 2004). The detailed aims and objectives of this PhD are presented in the next section.

1.5 Aims and objectives

The overall aim of this PhD is to understand how robot behaviour and interface design can be applied to utilise the benefits of robot autonomy and inform future human-robot collaborative systems. This is examined in the

context of semi-autonomous remote controlled ground robots for Urban Search and Rescue (USAR).

- Aim I: Develop a background understanding of the USAR domain and their work as well as describing the real world application of USAR in order to provide recommendations for the implementations of robots in British USAR teams.

  Objectives:

  1. Gather background knowledge of the USAR domain, especially their technical equipment used to date, as well as investigating the rescue culture and team behaviours within this user group to inform future experiments and robot designs.
  2. Study organisational structures and rescue processes to find an appropriate robot position in the system in order to give recommendations for an implementation of robots in British USAR teams.
  3. Collect data about rescuers' attitudes towards robots.

- Aim II: Improve understanding of underpinning cognitive concepts, thoughts and behaviours of participants while interacting with different autonomous and semi-autonomous robots, in order to inform future robot behaviour and interface design as well as the subsequent studies of this PhD.

  Objectives:

  4. Explore relevant rescue tasks with a retrospective verbal protocol and gather information about thoughts and feelings during human-robot interaction.
  5. Collect interview data regarding robot characteristics and participant preferences.

- Aim III: Investigate how robot and environmental characteristics, influence user cognition, behaviour and performance.

Objectives:

6. Identify the key cognitive concepts that are relevant to USAR.
7. Identify, compare and select appropriate measurements of these key cognitive concepts against each other.
8. Examine the effects of different feedback on trust.
9. Investigate the influence of task complexity and robot reliability on performance, workload and trust. In addition, compare performance levels between semi-autonomous controlled robot and manual controlled robots.
10. Compare, with the aid of the situation awareness transparency model, two different levels of interface transparency across two levels of task complexity.
11. Develop a measurement of performance in semi-autonomous human-robot teams.

How these aims and objectives were addressed in the following experiments is visualised in Figure 3.

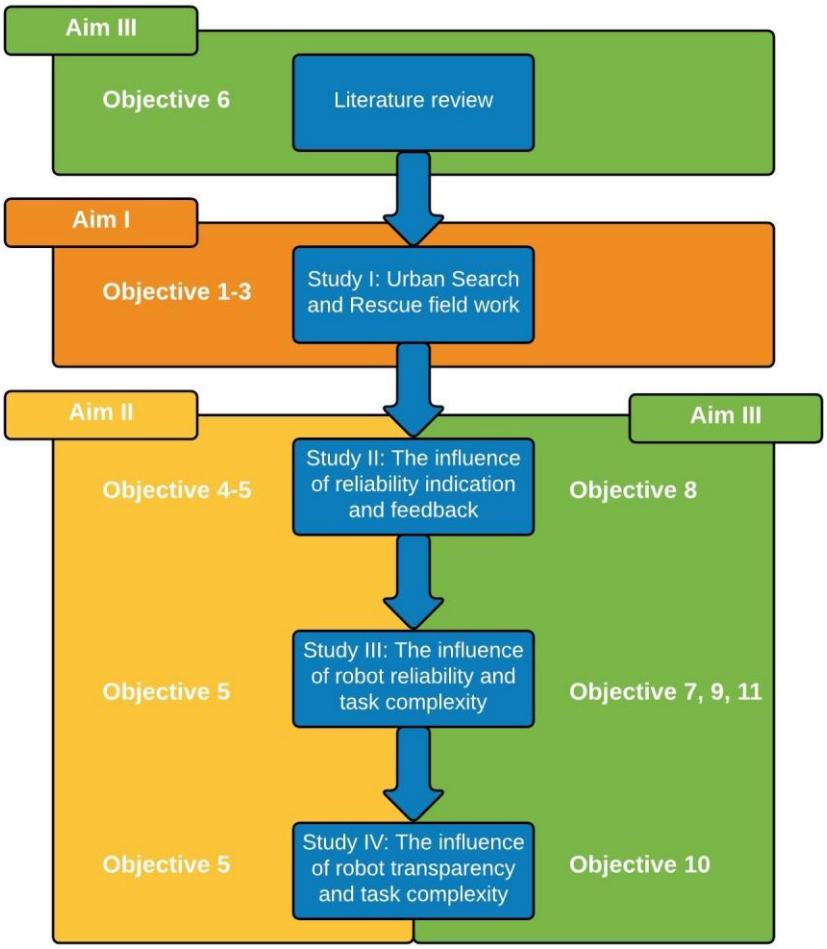


Figure 3 - Overview how studies will address aims and objectives

Overview and details of studies performed and which of the aims and objectives were addressed is shown in Table 1.

Study number	Study title	Study details	Aims & objectives
Study I	Urban Search and Rescue field work	The researcher attended a two weeks USAR course in the U.K. and gathered insight knowledge of the behaviours and thoughts of rescuers. Furthermore, information about tasks, processes, and culture of the rescue domain were collected.	Aim I Objective 1-3
Study II	The influence of robot reliability indication and feedback	This study examined different amounts of robot feedback and reliability justification of the robot. The experiment used the retrospective verbal protocol method to elicit in-depth qualitative information from the participants.	Aim II Objective 4-5 Aim III Objective 8

Study III	The influence of robot reliability and task complexity	This simulation tested the influence of robot reliability and task complexity on trust, workload and performance. Additionally a base line group utilised manual mode only and was compared to the semi-autonomous group regarding trust, workload, and performance.	Aim II Objective 5 Aim III Objective 7, 9, 11
Study IV	The influence of robot transparency and task complexity	The simulated robot showed different levels of transparency and the task changed in complexity. The influence of robot transparency and task complexity on trust, workload and performance was measured.	Aim II Objective 5 Aim III Objective 10

Table 1 -Overview of studies performed and aims/objectives addressed

An overview of the novel contributions to the body of knowledge and a summary of each chapter is provided in the next sections.

1.6 Novel contributions

This PhD addresses a variety of gaps in the literature and aims to add the following novel contributions to the body of research knowledge:

- Although some data about processes, equipment and organisational structures are openly available to the public (HM Government, 2008; "The Personal Qualities and Attributes [Website]," 2014), detailed information about rescue work is still missing:
  - This thesis provides an overview of the work and equipment of USAR personnel in the U.K. from a first-hand perspective.
- So far no research by using a retrospective verbal protocol while interacting with an autonomous remote controlled robot had been done:
  - Using verbalised thought analysis while participants interact with an autonomous robot.
- Many authors have concentrated on autonomous machines and robots (Hoff & Bashir, 2014; Merritt, 2011). However, a complex task such as Urban Search and Rescue still needs the operator in the loop and take over certain aspects of the search tasks (e.g. identifying casualties) (Virk, Gatsoulis, Parack, & Kherada, 2008). In this case semi-autonomous robots are required.

    - Performing a detailed performance analysis with respect to semi-autonomous robot systems and the emerging challenges with it.
    - Developing performance measuring techniques in semi-autonomous robot systems.
- Search and rescue teams encounter unpredictable environments (Y. Liu & Nejat, 2013) that can be highly complex. Investigation with task complexity and other known factors is important to design appropriate robot systems (Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013). Furthermore, transparency is an emerging concept that aims to enhance human-robot team performance and is worth further investigation (Boyce, Chen, Selkowitz, & Lakhmani, 2015; Lyons, 2013).
    - Investigating the effect of task complexity and robot reliability.
    - Investigating the effect of task complexity and robot transparency.
- In human-robot interaction several trust questionnaires exist (Jian, Bisantz, Drury, & Llinas, 2000; Muir, 1989; Schaefer, 2013; Yagoda & Gillan, 2012) but so far literature did not compare these questionnaires with each other regarding their usage in remote controlled semi-autonomous robot systems.
    - Comparing different trust questionnaires and their sensibility.

1.7 Thesis overview

1.7.1 Chapter 1: Introduction

This chapter provides an introduction for the thesis and gives a brief background of the topic and explains why this topic is of importance to research and society. The aim of this work is to investigate how robot behaviour and interface design can be applied to utilise the benefits of robot autonomy and inform future human-robot collaborative systems. This is examined in the context of semi-autonomous remote controlled ground robots for Urban Search and Rescue (USAR). Also, this chapter provides an overview of the studies performed to address the individual aims and objectives. Furthermore, a brief overview of the thesis chapters is given.

1.7.2 Chapter 2: Literature review

This chapter presents a critical review of robots used in USAR and human-robot teams. Furthermore, trust is identified as an important factor of influencing the interaction between operator and robot. Therefore, this chapter reviews the academic literature and reports what trust is, why it is important and what models have been developed in the past.

1.7.3 Chapter 3: Methodology

This chapter provides the methodology used to investigate the research objectives. The chapter starts with a review of the measures used in literature for human-robot performance and trust. Next, a justification for the selection of methods and tools is provided. The chapter concludes with a description of the equipment used. This includes an overview of the development of a simulated robot in a 3D virtual environment. The development and functionalities of the program are explained.

1.7.4 Chapter 4: Study I – Urban Search and Rescue field work

This ethnographic study aimed to gather information about USAR technicians, their training, tasks, working environment, currently used equipment, behaviour and culture. Over a period of two weeks eleven delegates of the USAR Level 1 technician course were observed in order to gather requirements and implications for robots in terms of features, behaviours, interface design, and robot implementation in the U.K. Fire and Rescue Service. Furthermore attitudes and traits of the technicians were collected. The background information was used to inform the subsequent studies. The chapter concludes with recommendations of robot usage in the USAR domain as well as a set of search and rescue scenarios.

1.7.5 Chapter 5: Study II – The influence of reliability indication and feedback

This chapter examines the influence of different amounts of robot feedback on trust, workload, performance, and participant's perception of the robot. Two robots, each providing different amounts of feedback, autonomously

searching an environment for specific targets. Both indicate their reliability level, but one of the robots indicates why it is in a certain reliability level and what type of target it found. This explanatory feedback was perceived as a clearer type of communication and the robot was perceived as more competent, efficient and less malfunctioning. Furthermore, to collect qualitative data about human-robot interactions participants perform retrospective verbal protocols and answer interview questions after the trials.

1.7.6 Chapter 6: Study III – The influence of robot reliability and task complexity

The chapter presents a simulated search and rescue scenario with a semi-autonomous robot system which examines the influence of robot reliability and task complexity on workload, performance, and trust. A post-task questionnaire collects data about trust, subjective workload, robot characteristics, and participant's experiences. This study also informs about possible measurements of performance in semi-autonomous robot systems. In addition, two trust questionnaires and their correlations are compared with each other and recommendations about their application provided.

1.7.7 Chapter 7: Study IV – The influence of robot transparency and task complexity

This chapter's study uses the same simulated environment and robot as the previous study. It examines the influence of robot transparency and task complexity on workload, performance and trust. The quantitative data of the previous study showed that robot transparency is of importance for the operator to understand the robots' states and actions. Transparency levels in this chapter consist of two different interfaces with different levels of feedback and scenario information. Also, the study aims to see if task complexity will have similar effects as in the previous study and that high task complexity leads to lower trust levels and lower performance. Interview data examines and quantifies which elements of the interface were actually used and why, in order to understand the benefit of presenting more or less information for higher transparency levels.

1.7.8 Chapter 8: Discussion

The summary of research findings and the review of aims from the research conducted are discussed in this chapter. Recommendations are made for robot implementation and design. Furthermore, this chapter addresses several discussion points about trust and collaboration in human-robot teams and it highlights the limitations of the research.

1.7.9 Chapter 9: Conclusion and future work

The chapter provides the main conclusions of this thesis in a short concluding statement. Furthermore, possible future work in the area of trust and human-robot collaboration research is outlined.

1.8 Chapter summary

This chapter provided an introduction for the thesis and gave a brief background of the topic and why it is of importance to research and society. The aim of this work is to investigate how robot behaviour and interface design can be applied to utilise the benefits of robot autonomy and inform future human-robot collaborative systems. The chapter explained the field of application, the scope of the thesis, and the novel contributions it will make to the body of knowledge. The structure of the thesis was presented by an overview of the upcoming chapters.

2 Literature review

2.1 Chapter overview

This chapter presents a critical review of robots used in USAR and human-robot teams. Furthermore, trust is identified as an important factor of influencing the interaction between operator and robot. Therefore, this chapter reviews the academic literature and reports what trust is, why it is important and what models have been developed in the past.

2.2 Chapter introduction

Human-robot trust is still a new field with limited but growing research. As with many topics in research, it is very context dependent and advancing technologies produce new challenges and opportunities. Rescue robots can be advanced robot systems with autonomous capabilities and cooperative behaviour that grant opportunities for faster and more effective rescue missions. At the same time new challenges arise: human-robot task distribution and human-robot team coordination, new sources of error, trust issues, information flow, etc. This chapter gives an overview of the literature which provided the basis for the research of this thesis. It is divided into three sections. First, the review discusses rescue robots in the Urban Search and Rescue (USAR) domain by presenting research projects concerned with rescue robots, current commercially available robots and past deployments of rescue robots. In addition, a robot's possible automation capabilities are discussed. Second, theoretical foundations of teamwork and human-robot team structures are explained. The third section is concerned with the issue of trust, why trust is important, what trust is and what models of trust exist. Further review of literature can be found at the beginning of each chapter.

2.3 Robotics in USAR

It is necessary to understand what types of robots have been used to date and the technologies engineers are working on. We can only produce an optimum robot system if development, design, and science work together. For instance, robots have been tested in the field and failed to demonstrate

usefulness to rescuers, due to inadequate design (Casper & Murphy, 2003; Murphy et al., 2015) or usability issues (Matsuno et al., 2014). This section of the PhD aims to look at existing robot systems and current commercially available robot systems, as well as lessons learned from past deployments and projects.

2.3.1 Commercially available robots

The following examples present current commercially available robots that are appropriate or possible to use for search and rescue reconnaissance tasks. Despite this relevance to USAR, most of these robots were not exclusively designed for search and rescue or have fully autonomous capabilities. Most systems are aimed to provide military support (e.g. reconnaissance) or inspection of inaccessible areas (e.g. pipes). However, with these systems it is possible to perform certain tasks in rescue missions that relate to reconnaissance. The review of available robot systems aims to provide the reader with an understanding of currently used robots and their capabilities.

TALON

TALON from QinetiQ (former Foster-Miller) is a 52 kg heavy robot with a gripper arm, obstacle navigation and stairs climbing capabilities. The TALON also provides different cameras such as infrared, thermal, fish eye, and night vision. As shown in Figure 4, this quite large robot, which is 43 cm in height, 57 cm width and a length of 86 cm, is aimed for manipulation tasks such as bomb disposal.

Figure 4 - TALON robot from QinetiQ ("TALON robot [Image]," 2015)

This robot had been purchased for ca. 79,000 GBP in 2005 by the Miami Police Bomb Squad (Martin, 2012). The price includes a camera and other equipment to operate the robot appropriately. Although this robot is not specifically for search and rescue, it can be used in flat terrain for reconaissance missions or manipulation tasks.

Dragon Runner

The little brother of TALON is the Dragon Runner (see Figure 5) and is also produced by QinetiQ. This throwable robot weights 4.5 kg has a microphone, as well as day and night vision. It longest side is 38 cm. This size is much more appropriate for rescue operations. Originally the Dragon Runner was designed for reconnaissance missions; with some add-ons it can also climb stairs. The robot can also be equipped with an additional gripper arm.

Figure 5 - Dragon Runner 10 from QinetiQ ("Dragon Runner 10 [Image]," 2015)

In 2009 QinetiQ had been awarded contracts with the U.K. Ministry of Defence of over 12 million GBP for providing 100 Dragon Runner robots (including spare parts and support) for the military operations in Afghanistan (QinetiQ, 2009).

110 First Look

iRobot produced the reconnaissance robot 110 First Look (Figure 6). It is throwable (e.g. it can also be dropped or thrown towards the area of use), can be equipped with different cameras, and can autonomously self-right itself. The longest side of this robot is 25 cm. The robot is small, rugged and expandable according to iRobot. In 2014 the American Saginaw Police Department purchased a First Look robot for nearly 13,200 GBP (Tower,

2014). The aim of the police department is to drive remote-reconnaissance missions in inaccessible or dangerous areas (e.g. car accidents or taking of hostages).

Figure 6 - 110 First Look from iRobot ("110 First Look [Image]," 2015)

Recon ThrowBot LE

One of the smallest reconnaissance robots is the Recon Scout ThrowBot (see Figure 7). It only weights 500 g and has a length of 19 cm with a diameter of 8 cm. It offers easy transportation and can be deployed immediately ("Recon Scout Throwbot LE [Website]," 2015).

Figure 7 - Recon ThrowBot LE from Recon Robotics ("Recon Scout - Throwbot LE [Image]," 2015)

The operator control unit solely has a screen (showing the video of the robot) and a single joystick. It is also throwable and silent. A working configuration starts at a price of 3,200 GBP (ReconRobotics, 2010).

R2i2 Delta Extreme/Micro

This model was already successfully deployed during the terrorists' attacks on the World Trade Center (WTC) 2001 in New York: the R2i2 Delta Extreme, also known as Inuktun VGTV (Variable Geometry Tracked Vehicle). The robot was initially developed by Inuktun and is now purchasable via RECCE robotics. The tracks are flexible and can traverse very difficult terrain. It incorporates a camera, bi-directional audio and a tether cable. Prices are not readily available but a used robot could be purchased for ca. 5,000 GBP (without sensors) according to Booysen and Mathew (2014). This robot is also utilised as a permanent asset of the USAR team New Jersey Task Force 1 (see Figure 8).

Figure 8 - R2i2 Delta Extreme from RECCE robotics ("R2i2 Extreme [Image]," 2015)

Its system and sensors were adapted by the CRASAR team to fit the needs of the USAR missions. A smaller version of this robot is the 6.2 kg heavy R2i2 Delta Micro (see Figure 9).

Figure 9 - R2i2 Delta Micro from RECCE robotics ("R2i2 Delta Micro [Image]," 2014)

It is equipped with a camera and lighting. Optional are camera tilt feedback and inclinometer.

The next section will emphasise on past deployments of robots in the field and discuss the implications for future robot systems.

2.3.2 Deployment of robots and lessons learned

Terrorist attacks, World Trade Center (2001)

The first real world disaster that led to the use of robotic assistance was the WTC disaster in New York. In 2001 the two towers of the WTC collapsed as a result of being struck by two airplanes, which were hijacked by terrorists. Casper and Murphy (2003) produced a report about this disaster including an extensive set of recommendations for USAR HRI. During the WTC disaster, the Robotic Assisted Search and Rescue Center (CRASAR) were in charge of the robot deployments. CRASAR is located in the US and aims to develop robots and technologies for disaster prevention, response, and recovery. The Center consists of academic researchers, industry partners and rescuers ("CRASAR [Website]," 2015). Their later developed and tested technologies were adopted by official emergency response teams across Europe ("CRASAR [Website]," 2015).

At the WTC disaster robots were used to access small voids and areas that were too dangerous for rescuers to enter in order to search for casualties. Also, robots were used in a later mission on site to record video material for structural engineers. Of the ten robots at the incident site only three were actually used: the models Inuktun MicroTracs, Inuktun MicroVGTV, which can be carried by a single person, and Foster-Miller SOLEM (see Figure 10).

Figure 10 - Foster-Miller SOLEM ("Foster-Miller Solem [Image]," 2010)

The 15 kg heavy SOLEM is not commercially available anymore. The product got replaced by newer versions of the TALON robot and Dragon Runner (see section 2.3.1). The Inuktun models are also replaced by newer versions. The MicroVGTV's new model is the R2i2 Delate Micro and Extreme (see section 2.3.1).

The decision of using the robots stated above was made due to the small entry points of the voids. For example, Casper and Murphy's (2003) report stated that the TALON (described in section 2.3.1) and the Inuktun Pipe crawler were too large to carry to the deployment location or to even enter the voids. The Defense Advanced Research Projects Agency (DARPA) prototype iRobot was too fragile and the operator control unit was not appropriate for field work as well as too difficult to use. Generally, DARPA develops new technologies for the US military ("DARPA: Tactical Technology Office [Website]," 2015). The robots selected for rescue operations were battery powered and could be used between four and seven hours. Further, the report of Casper and Murphy (2003) stated, with respect to personnel, that during deployment a robot needed one operator to be brought to the desired starting location. The bigger Foster-Miller SOLEM needed to be deployed by two people. After bringing the robot to its starting location another two people were usually required to operate the robot: one person had to keep the tether/safety rope from getting caught and the other person remote controlled the robot. Casper and Murphy (2003) explain that in general robot communication was limited to the camera, power, movements, camera tilt, illumination and height change. Also no autonomous features were available. Casper and Murphy (2003) presented seventeen findings in detail; the details concerned with human factors and the control of the robot are outlined below (see Casper & Murphy, 2003, pp. 376–379).

- A major issue was cognitive fatigue of personnel due to lack of sleep. This was a major source of mistakes and decreased performance.
- The skill levels between robot personnel and fully certified rescuers were different and therefore issues of trust towards robots and robot personnel emerged.

- Since video was the only feedback from the robot, it was very difficult for the operators to determine status and location of the robot. Furthermore, the inability to adapt sensor capabilities (e.g. equip, substitute or remove sensors) and the poor video quality inhibited effective use of robots. They also found that there was a need for feedback so that the operator can diagnose problems. Other required information was the position of the robot and the mapping of the environment. Moreover, different viewing angles would help to identify objects faster and with more certainty.
- The acceptance of the robots seems to be determined by the similarity to existing technical search equipment. The USAR specialists selected robots that were small, had simple interfaces, simple control units, and needed minimal numbers of personnel.
- Rescuers still trusted humans, common search tools and search dogs more than the complex robot systems.
- Casper and Murphy (2003) advised that the robot and the interface need to be designed for infrequent use and minimum training time. Proficiency needs to be maintained over longer periods of time. In general, search equipment is used in training for 30 minutes twice a year.
- The visual channel was over-used. Headphones could not be worn due to the protective gear rescuers were required to wear.
- The acceptance of robots and users' confidence diminished when communication failed.

(Casper & Murphy, 2003)

It must be kept in mind that none of the robots was designed for USAR and were not intrinsically safe, which further reduced the deployment of the robots.

Later Stopforth et al. (2010) took up the findings from Casper and Murphy (2003), which are outlined above, and developed a new robot called CAESAR (Figure 11). In 2010 they tested the CAESAR robot on a training site and it could be deployed under three minutes (Stopforth et al., 2010). They especially emphasised the need for good communication between the operator and robot. For better movements the robot had so called flippers.

With the flippers, operators were able to lift and lower the robot as well as push it over obstacles. A composite body construction of the robot kept heat away from critical internal systems. Additionally, the robot tracks were made out of metal and did not soften under high levels of heat. The operator unit displayed colour indication warnings and showed the positions of the flippers.

Figure 11 - CAESAR robot with flippers ("CAESAR robot [Image]," 2015)

Earthquakes, Italy (2012)

After this first usage of robots in 2001, a further 33 robot deployments in real world disasters were reported between 2001 and 2012 (Murphy, 2014). During these disasters different types of robots from different projects and development teams were used. For example, two earthquakes occurred in Italy in May 2012 left widespread damage to urban areas and made thousands of people homeless ("Italy quake homeless in emergency shelters," 2012). The human-robot team from the NIFTi project was called to assess the damage to buildings in order to estimate the risk of further collapse (Kruijff et al., 2012). The EU-funded project NIFTi concentrated their efforts on human-robot teams that perform reconnaissance missions during USAR operations. With their user-centric approach they developed robot requirements, prototypes, and models in autonomous robot behaviour and human-robot collaboration (Kruijff et al., 2012; "NIFTi [Website]," 2014). The deployed unmanned air (UAV) and ground (UGV) vehicles are shown in Figure 12. The NIFTi ground robot incorporated video streaming

from several cameras and a laser-based 3D scanner. Some lessons learned and recommendations regarding the interface and human factors of the NIFTi robots from Kruijff et al. (2012) are outlined below.

UGV

- The interface included a 3D model of the robot and the environment. With this feature the operator could see the position of the flippers in relation to the surface immediately. However operators needed to switch between camera and 3D model, which increased the operator's cognitive load and time to finish the mission.
- Moving the camera separately to the robot's main body (pan and tilt) decreased the situation awareness of the operator. Furthermore, when the camera turned, the operator was not aware of the movement direction of the robot (main body). To counteract this, the team set pre-defined camera positions that were reachable via the interface. Recovery with this feature was easier for the operator.

UAV

- The cognitive load and stress of the operator flying one of the UAVs provoked pilot mistakes. This could be mediated by providing more autonomous features of the robot such as holding a certain position in the air when no command is given.
- The pilots had sub-optimal situation awareness due to poor depth perception via video feed.

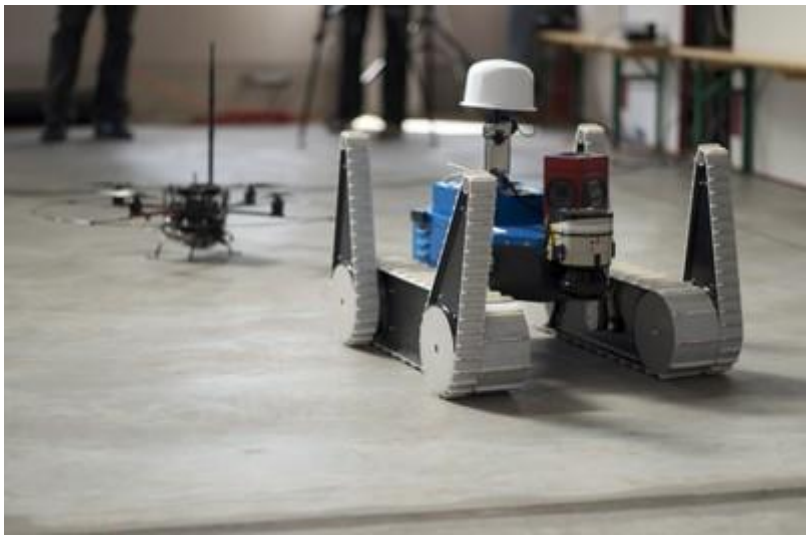


Figure 12 - NIFTi UAV (left) and UGV (right) platform ("NIFTi UGV [Image]," 2013)

In terms of the use of autonomy, Larochelle et al. (2013a) tested the same ground robot as mentioned above, which was developed by the NIFTi project and manufactured by BlueBotics (see Figure 12), in a simulated USAR mission. Although this was not a real incident, real firefighters took part in this very high-fidelity simulation. Their robot had three different autonomy levels: executional (accelerating, observing objects), operational (following a planned route), and tactical (which robot will investigate which areas). The participants used autonomy during scenarios but when their expectations were not met, they switched to manual control. This showed how important initial robot performance is and that the expectations that participants have, has to be calibrated before interaction. Their findings supported previous studies (Beer, Fisk, & Rogers, 2014; Lee, 2008; Yagoda, 2011) in that autonomy was not easily accepted and factors such as reliability, trust and transparency were important influential determinants of operators' control allocation.

Earthquake, Japan (2014)

Another robot was used in the aftermath of the Great Eastern Japan earthquake which involved the Tsunami and after the Fukushima power plant accident (Matsuno et al., 2014). Overall, Matsuno and colleagues (2014) experiences showed that robots are useful for victim and economic recovery missions. They deployed the research robot KOHGA3 robot (Figure 13). The pan-tilt-zoom camera was very valuable because zooming could

give a clear picture of distant objects without the need to move the robot to a certain location and the camera could be tilted to get different viewing angles. Additionally, the sensor arm made it possible to look into rooms or behind obstacles, even though they were not accessible for the robot's main body. Matsuno et al. (2014) further found that a battery and signal strength indicator was missing on the user interface and that the system should provide easy exchange of sensors to fit the mission needs.

Figure 13 - KOHGA3 ground robot ("KOHGA3 ground robot [Image]," 2011)

With respect to autonomous features, Matsuno et al. (2014) reported that the robot should autonomously be able to indicate whether a terrain is traversable or not. To reduce operator's workload they desired autonomous/semi-autonomous functions such as show possible routes or the ability to return to the start point.

The robots needed the capability to record pictures and video to be able to report states and show other stakeholders the situation. Matsuno et al. (2014) also found that the robots needed to be easily portable, fast to deploy, rugged and be able to record evidence. Interestingly, the human-robot interaction was still very challenging; loss of situation awareness and lack of feedback were the main concerns that emerged during field deployment. Matsuno's team cooperated with the International Rescue System Institute and CRASAR.

Mudslide, Washington State USA (2014)

Recently, CRASAR deployed commercially available UAVs during the Washington State mudslides in 2014 (Murphy et al., 2015). The rain soaked site and logging in the area fostered a landslide that killed 44 people and destroyed the riverside community (Cornwall, 2014).

The use of small air vehicles provided some advantages over helicopters which are too dangerous to fly near to the ground and they are ten times more expensive than deploying UAVs (Murphy et al., 2015). The UAVs were tasked with helping to identify the geological and hydrological state of the mudslide with pictures, 2D, and 3D reconstructions of the area. Due to limited time and manual flying, data sets were incomplete because the area was not covered appropriately. Murphy et al. (2015) emphasised, again, the fact that the system needs to be easily portable (e.g. over rough terrain). Engineers expected to have a remote view, but autonomous data collection without a live stream from the camera compromised the acceptance and utility of the system. Although the robot flew autonomously, due to the Federal Aviation Administration requirements a constant line-of-sight needed to be maintained. This posed high demands on the operator as it was possible to lose sight of the UAV when checking the operator control unit at the same time.

Lessons learned summary

These insights from selected real world deployments repeatedly suggest that robotic rescue systems must be rugged in terms of hardware and reliability otherwise they will not be used. If a flimsy connector breaks in the field and has to be repaired it will cost time and resources, especially when the equipment needed for the repair is not immediately available due to the remote position of the rescue teams (e.g. Matsuno et al., 2014).

In general, for rescue operations, robots need to be small and man-portable. Existing commercial systems do not fit the needs of the rescuers yet. More development with the collaboration of rescuers, academia and industry is required. Many issues in human-robot interaction still exist in real world contexts and need to be investigated, preferably in the field.

Much of the literature discusses a demand for more modular sensors and interfaces, so that a robot's capabilities can be customised and the interface accordingly to cope with the huge variety of possible incidents, whether they have to inspect buildings in danger of secondary collapse or to search for casualties in voids (Casper & Murphy, 2003; Driewer, Schilling, & Baier, 2005; Peschel, 2012; Rule & Forlizzi, 2012). However, every new feature will have implications for human-robot interaction and can create the potential for new possible types of errors (Hoff & Bashir, 2014).

2.3.3 Robot autonomy levels

Team organisation and robot capabilities are dependent on the autonomy level of the robot. Some robots do not have any autonomous capabilities and are used as a remote viewing tool. Other robots might be capable of driving and locating targets autonomously. In order to capture the differences between these systems this section discusses autonomy levels and other types of autonomy.

Sheridan and Verplank (1978) introduced a ten level of automation taxonomy. The levels are described by who is doing what and where information is held and are applicable mainly to cognitive tasks. This taxonomy is very much focussed on pure automation aids. Therefore Endsley and Kaber (1999) developed a new taxonomy of levels of automation (LOA) based on Sheridan and Verplank's (1978) work to have a wider scope, including teleoperation. They incorporated domains with multiple goals and tasks, and high demands as well as limited resources. They further identified general functions that needed to be incorporated: human monitoring, option/strategy generation, option/strategy selection, and the implementation of these options/strategies. The stated functions were either assigned to the human or the system (Endsley & Kaber, 1999, pp. 464–465). An overview of these function allocations is illustrated in Table 2.

LOA		Moni-toring	Gener-ating	Selec-ting	Imple-menting	Comment
1	Manual control	H	H	H	H	Human is still able to intervene in certain task steps
2	Action support	H	H	H	H/S	
3	Batch processing	H	H	H	S	
4	Shared control	H	H/S	H	H/S	
5	Decision support	H	H/S	H	S	
6	Blended decision making	H	H/S	H/S	S	
7	Rigid system	H	S	H	S	
8	Automated decision making	H	H/S	S	S	
9	Supervisory control	H	S	S	S	
10	Full automation	S	S	S	S	Human out of the loop
H - human; S - system						

Table 2 - Overview of levels of automation from Endsley and Kaber (1999, pp. 464–465)

Endsley and Kaber's (1999) experimental studies found that the level of automation influenced automated system performance. Automation levels that used human option/strategy generation and robot implementation were superior to pure manual control or full automation. Endsley and Kaber (1999) pointed out that automated system guidance (system provides option guidance) or option selection (system generates options that can be selected) was counter-productive, since participants had to gauge their choices against the robot's choices.

With the expansion of the body of literature in human-robot interaction, Beer et al. (2014) proposed a framework that would help to choose the required levels of autonomy for certain applications in human-robot interaction. Although the previous models (Endsley & Kaber, 1999; Sheridan & Verplank, 1978) are a good guide for a human-robot interaction framework, they did not take into account autonomous robots as a physical entity with mobility, capable of environmental manipulation or social interaction (Beer et al., 2014). The decision aid framework is embedded into five guidelines as shown in Figure 14. Beer et al. (2014) considered this framework for a search and rescue operation. The following USAR example

was made by Beer et al. (2014) to clarify their guidelines. Please refer to the guideline questions in Figure 14.

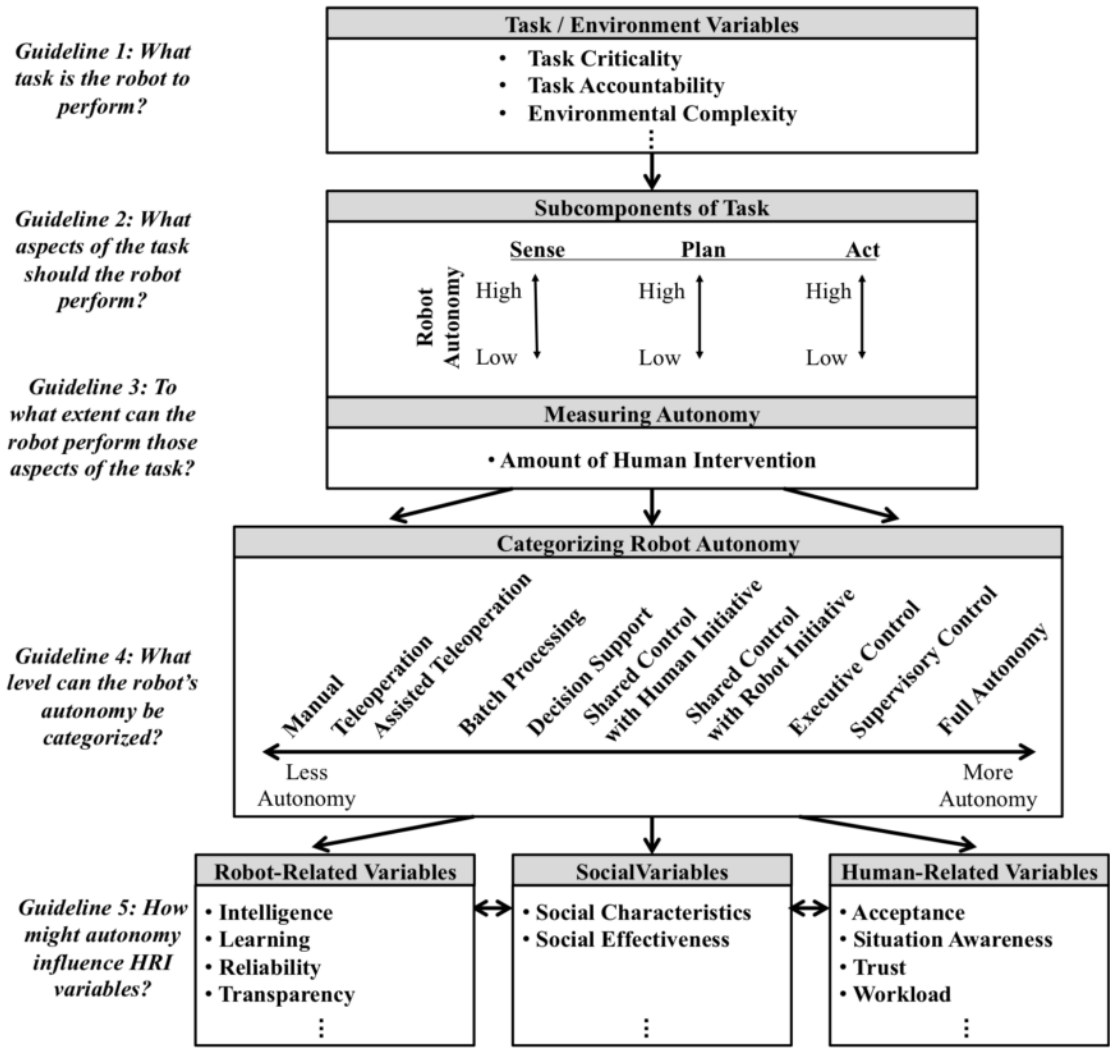


Figure 14 - Level of robot autonomy framework from Beer et al. (2014).

Guidelines according to Beet et al. (2014). Guideline 1: High task criticality and high environmental complexity is present in USAR scenarios. This suggests that it might be challenging for a robot to perform all tasks reliably. Therefore, an initial thought might be to design the robot with low autonomy levels. Guideline 2: Workload of operators is very high during teleoperation and situation awareness poses another great challenge. Guideline 3: literature shows that the unpredictable environments and new situations a robot encounters are quite challenging and that the human often needs to intervene or help the robot. Guideline 4: previous guideline answers suggest that for a rescue robot the autonomy category teleoperation or supervisory

control, depending on task complexity and the operator's situation awareness, can be chosen. Guideline 5: If a robot autonomy is chosen, its effects on robot-related variables, social variables and human-related variables need to be considered. These are task and robot specific factors that vary due to application or specific task.

The final consideration of Beer et al. (2014) was to use sliding autonomy for USAR robots in order to adapt to different task complexities and situation awareness levels. Sliding autonomy presents the possibility for the operator or the robot to adjust the level of autonomy as needed (Desai & Yanco, 2005). Beer et al.'s (2014) suggestion to use sliding autonomy for USAR robots is in agreement with other literature (Sellner, Heger, Hiatt, Simmons, & Singh, 2006). Furthermore sliding autonomy can also be beneficial for multi-robot control (Heger & Singh, 2006).

What is important in human-robot teams and how they can be organised, is documented in the next section.

2.4 Human-robot teams

Robotic technology is evolving rapidly and robots are changing from merely being tools to becoming team mates. They are capable of making their own decisions and collaborating with humans. Advances in technology create new questions as to how machines and human can work together. This section will discuss theoretical foundations of teamwork, by examining the different underlying concepts/models/frameworks that researchers use to describe human-robot team organisation, followed by an overview of single- and multi-robot systems.

According to Salas et al. (2005), teams are more adaptable, and provide higher productivity and creativity than a single person. Collaboration will also increase the flexibility of robot automation (Bradshaw, Dignum, Jonker, & Sierhuis, 2012; Brogårdh, 2009). In teams, approaches to solve organisational problems can be more complex, more innovative and comprehensive than for individuals. Salas et al. (2005) also defined the "Big Five" factors of teamwork and coordinating mechanisms: team leadership, mutual performance monitoring, backup behaviour, adaptability, and team

orientation (goals). Coordinating mechanisms within "Big Five" factors are shared mental models, mutual trust and closed-loop communication.

Gao et al. (2012) noted that in teams where humans collaborate with autonomous agents, the complexity increases and the methods of communication and interaction may change. For instance, human-human teams communicate in a natural way (e.g. using speech or gesturing), but a robot may not be able to understand natural communication and is dependent on specific commands. Furthermore, Groom and Nass (2007) stated that the acceptance of robots by human team members will determine if the robots' benefits are fully utilised. In order to understand how human-robot teams can work together, the theoretical foundations of teamwork are discussed next.

2.4.1 Theoretical Foundations of Teamwork

Teamwork models can inform computer science on how to program a robot to execute collaborative behaviour; the behaviour must produce human-like interaction to be understood intuitively by the human partner and to be suitable for the given situation (Kim et al. 2007). The theoretical foundations of teamwork are very well described in Hoffman & Breazeal (2004). Similarly to Salas et al. (2005), they claim that collaboration consists of shared activity, joint intention, common ground and goals:

- *Shared activity*: According to Bratman (1992) a shared activity consists of mutual responsiveness, commitment to the joint activity and commitment to mutual support. Bratman further explains that in order to create a joint activity, "meshing up" of individual plans is of importance.
- *Joint intention*: The Joint Intention Theory from Cohen (cf. P. R. Cohen & Levesque, 1991) describes the importance of communication in a team. Team members have to communicate with each other for maintaining mutual beliefs about the state of the goal. Scheutz et al. (2006) used the Joint Intention Theory to inform their robot architecture. The robot is able to show positive and negative affect in order to contribute to a mutual belief of the joint intention

in the team. Showing affect enhanced human-robot team performance.

- *Common ground*: Common ground is the sum of the shared knowledge, beliefs or assumptions of the team (Clark, 1996). According to Klein et al. (2005) common ground is important for inter-predictability, and interdependent actions of joint activities.
- *Goals*: A goal-centric view is of importance to teams, allowing team members to interpret other team members' actions on the basis of intended goals and not on the performed physical activities (Hoffman & Breazeal 2004) as there are different approaches to reaching a goal. Goals must be mutual and coordinated for successful teamwork.

Understanding the underlying concepts of teams can be useful when modelling robot behaviour. For example, Bicho et al. (2010) used the above described elements of teamwork to inform a control architecture for a collaborative robot, which was able to anticipate human needs, and detect and respond to unexpected events (cf. Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005). A robot needs to be predictable and give sufficient feedback in order to establish joint intention or common ground.

Other models and frameworks address specific problems within human-robot collaboration (HRC), for example Fiore et al. (2005) proposed a theoretical framework for understanding memory failures in distributed human-robot teams and they argued that it would be productive to use agent technology to augment team cognition. Cuevas et al. (2007) used elements from organisational and cognitive science to inform a framework with the goal to enhance human-agent team cognition and coordination. The framework describes the influence of information exchange and updating on the different levels of cognition. Also incorporated are task-related stressors, such as workload or time pressure. In contrast to other team models, it explicitly includes automation technology.

Due to the rising complexity of human-robot teams, they can also be seen as Joint Cognitive Systems (JCS). JCS were introduced by Hollnagel & Woods (1999) and aim to describe a human-robot team as a system capable

of modifying behaviour, on the basis of experience, by mutual interaction (Kim et al. 2007; Neerincx & Grant 2010).

Some common team organisations are described in the next section. It is important to differentiate between co-located and teleoperated robots because the nature of interaction is different. For instance, co-located robots engage with humans in close physical proximity and communication happens face-to-(robot)face. Here, social behaviour and safety are critical concerns (Bradshaw et al., 2004). With teleoperated robots, the interaction is limited to displays and controls. In remote control situations focus lies on the interface's usability (e.g. screens, control units, etc.) (Eliav, Lavie, Parmet, Stern, & Edan, 2011). The interface needs to convey the information from the remote environment to the operator and generate an accurate shared mental model for the human-robot team (Oleson, Billings, Kocsis, Chen, & Hancock, 2011).

2.4.2 Single robot systems

Single robot systems are teams consisting of a single robot and one or more humans. These robots are common in social robotics (e.g. medication or entertainment robots) and in search and rescue missions (Murphy, 2014). If a robot has much more autonomy it can be used in multi-robot systems (H. Wang, Chien, et al., 2009), which are described in the subsequent section.

2.4.2.1 **Co-located Robots**

Examples of co-located teams are robots in our homes (e.g. service robots or health-care robots), where individual users engage with single robots on a social level. For instance the personal service robot "wakamaru", as shown in Figure 15, can undertake conversations with humans, can give wake-up calls, inform about the appointments of the day, detect burglars and provide desired information from the internet (Shiotani et al., 2006).

Figure 15 - "wakamaru" service robot ("Wakamaru [Image]," 2013)

In such social robotics, the goal is to enhance communication and mutual understanding. The autonomy level of such robots is very high and mostly there is very little monitoring or support required. There are also single robots which interact with a lot of different people, for instance mobile assistive robots, which can support an assembly line in logistic tasks (e.g. bringing goods from person A to person B) (see Angerer, Strassmair, Rootenbacher, & Robertson, 2012). Regardless of the domain, health-care robots or rescue robots, robots will need social interfaces (Fincannon, Barnes, Murphy, & Riddle, 2004). For instance, Murphy et al. (2011) developed the Survivor Buddy; this robot acts as a medium between a trapped victim and the outside world.

2.4.2.2 Teleoperated Robots

Teleoperated robots are used in health care (teleoperated nurse robots/telepresence robots), in the military, and in the search and rescue domain. In search and rescue missions it is necessary to separate the operator from a possibly dangerous environment (e.g. collapsing houses, chemicals, fire etc.). The communication between the human and robot is via displays and controls. Social interaction is not the first priority in these performance-driven and time-pressured teams. The aim is to enhance the operator's situation awareness and reduce the workload to increase overall team performance (Crossman, Marinier, & Olson, 2012). A single robot is used by one operator, if the robot's autonomy level is low and mainly

manual control of the user is required. However, with increasing robot autonomy the human operator can experience less workload from the task. Researchers have proposed that operators could make use of that capacity by handling multiple robots (Crandall & Goodrich, 2005).

2.4.3 Multi-robot systems

Increasing robot autonomy enables humans to control multiple teleoperated robots simultaneously. Although, operators are not required to manually control each robot, they still have a high workload caused by monitoring, communicating, coordinating and complex decision-making across multiple robots (Gao et al., 2012). The number of team members and the ratio of humans to robots is mostly dependent on the type of task, characteristics of the robot and the setting.

The goal is to maximise the performance in human multi-robot teams. Unfortunately this performance is very much dependent on the autonomy level of the robot and the workload of the task. Crandall & Goodrich (2005) developed a model to determine the fan-out of human-robot systems. Fan-out is the maximum number of robots a single operator can handle. They describe a measurement methodology to obtain necessary values for their fan-out algorithm, which includes, among other variables, the ratio of neglect time and the interaction time of the human with each robot. A similar approach can be found in the work of Olsen & Wood (2004). In the search and rescue domain, Lewis et al. (2011) recommended on the basis of their experiments, that a single operator should monitor a maximum of 10 robots.

Team complexity increases if a team consists of several humans interacting with each other and with multiple robots. In the search and rescue setting, two organisation structures of teams are possible (Lewis et al., 2011): the first structure is to allocate certain assignments of a robot to one operator, called *assigned robots*. The second possibility is that all operators share all robots and look after a robot if it requires attention. This structure is called *shared pool*. Overall *assigned robots* reduce the number of robots for each operator. Additionally, Lewis et al. (2011) claimed that if the robots are grouped in a defined area the situation awareness of a single operator is

higher. In contrast to that, in a *shared pool,* one operator (overlapping assignment) may observe things a second operator had missed.

Chien & Lewis's (2012) experiment examined team structure (*assigned robots* and *shared pool*) and how automated robot navigation or operator assigned way-point navigation interacts with team performance. The results of the search and rescue simulations showed that in tasks performed with the *assigned robots* structure and with automated path planning, more victims were found, more territory was explored and victims were marked more accurately. In general, teams with automated navigation had more cognitive resources available for other tasks. Increased automation enhanced performance in both *assigned robots* and *shared pool* conditions. However, if the robot took over navigation, the situation awareness of the operator decreased.

Although multi-robot teams sounds a promising endeavour, the current developmental stage of rescue robots is still challenged by designing performance enhancing interaction with one robot with little autonomy. Literature showed that trust plays an important role in human-robot collaboration. Therefore, the issue of trust, why trust is important and what trust is, is the object of investigation of the next section. In addition, an overview of relevant trust models provides an understanding of the factors that influence trust and how these factors are connected to each other.

2.5 Trust in human-robot teams

2.5.1 Why is trust important?

When people interact with automated or semi-autonomous systems their subjective trust in these systems can predict and influence the allocation of functions (e.g. letting the robot navigate autonomously or navigate manually) within the human-automation system (Muir & Moray, 1996). An appropriate level of trust is key to the usage of automated systems (Lee & See, 2004). An inappropriate level of trust can lead to misuse or disuse of the system, for example, that errors can occur due to over-trusting or under-trusting the system and eventually the potential benefits of the system can be lost (Parasuraman & Riley, 1997). This appropriate level of

trust is also the key to improving safety and productivity (Hoff & Bashir, 2014).

If we accept the notion that robots can be team mates capable of proper communication and independent task performance, then trust is an important issue because of its impact on effectiveness of collaboration. Generally, if team members do not trust each other they will waste time and resources checking and inspecting other team members' work and consequently team participation, team contribution, cycle times and task quality can be negatively affected (Salas et al., 2005). In addition, humans tend to interpret a person's behaviour differently according to the level of trust in that person (Simons & Peterson, 2000). What is valid in human-human teams can also partly be projected onto human-robot teams. If the users mistrust a robot, their willingness to accept robot-generated information decreases (Freedy & de Visser, 2007); they will constantly be checking the autonomous work of the robot, or they will perform the tasks on their own, to the detriment of the system's performance. It could also be that the user will interpret robot behaviour negatively. For instance, if the robot tells the operator to take regular breaks, the human can negatively interpret this as being controlled by a machine. The next section will discuss what trust is and what types of trust exist.

2.5.2 What is trust?

We experience trust every day, sometimes unconsciously, sometimes consciously. Trust is a very fuzzy, interconnected construct, and is difficult to define and measure. Trust often refers to a variety of constructs and is used in many different disciplines with different definitions (for a comprehensive overview see Mcknight & Chervany, 1996). For instance the personality psychologist Rotter (1967, p. 651) defined interpersonal trust as, "an expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon". The sociologist Shapiro (1987) saw trust as a structural phenomenon (e.g. trust is controlled by norms, constraints, restrictions, policies, etc.) and economic-oriented researchers such as Castaldo et al. (2010, p. 666) defined trust as, "an expectation (or a belief, a reliance, a confidence, and

synonyms/aliases) that a subject distinguished by specific characteristics (honesty, benevolence, competencies, and other antecedents) will perform future actions aimed at producing positive results for the trustor in situations of consistent perceived risk and vulnerability". Bhattacharya et al. (1998) captured key trust elements across several disciplines and proposed a mathematical trust model, which they simplistically described as, "Trust is an expectancy of positive (or nonnegative) outcomes that one can receive based on the expected action of another party in an interaction characterised by uncertainty".

In terms of trust in automation, Lee and See (2004, p. 50) defined trust as, "the attitude [of the trustor] that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability", whereby the agent can be an automated system or another person. McKnight and Chervany (1996) argued that most definitions are too narrow to define trust properly, such as that put forward by Wagner (2015, p. 485), who operationalised the definition from Lee and See (2004) to the following: Trust is "a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor's risk in a situation in which the trustor has put its outcomes at risk". This definition is less appropriate in human-robot trust regarding rescue robots because it neglects the fact that trust is something you are willing to give or not. In human-robot interaction, Hancock et al. (2011) defined trust as follows: "[…] as the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others".

All of these definitions describe trust and they have certain factors in common. First, from most of the trust definitions one can infer that there is an expectation towards the trustee, which includes a certain amount of predictability or faith. Thus it can be argued that an operator has certain expectations towards a robot and that predictability is an important characteristic of that robot. Second, trust is always required in situations where uncertainty exists. Remotely operating a robot in complex environments, such as Urban Search and Rescue, includes various uncertainties. Third, having trust means we are willing to depend on someone or something. For example an operator is willing to depend on a robot to find casualties at an incident site. To bring together the three

common factors from previous definitions, in this PhD a very broad definition from Bhattacharya et al. (1998, p. 462) was adapted and modified to fit the needs of this work:

Trust is the willingness to rely on a system with the expectancy of positive outcomes that one can receive based on the expected action of that system in an interaction characterised by uncertainty.

The following section introduces different models and types of trust and will illustrate similarities and differences between them.

2.5.3 Models of trust

Models of trust include factors that influence trust and can show how they are connected or related. Trust changes over time and can be influenced by a variety of discipline-specific factors, therefore each discipline, such as management, psychology or human-robot interaction, has its own trust models. This section considers a variety of trust types and concepts to compare their underlying trust constructs and the implications for this work and human-robot trust in the rescue domain.

2.5.3.1 Model of initial trust

An important dimension of trust is time. Trust needs to be established over time, for example through positive experiences (Wiethoff & Lewick, 2000) or in the case of human-robot interaction, through initial exposure to the technology (Oriz, Fiorella, & Vogel-Walcutt, 2010). A countermeasure against lack of trust is training. Training can be employed to reduce initial biases, for example teaching people about the capabilities of the robot (Freedy & de Visser, 2007) and the underlying assumptions of the software. The operator needs to get to know the robot and find out the robots' strengths and weaknesses. Therefore many models assume a low initial level of trust (e.g. Williamson, 1993).

However, there is also evidence that high levels of early trust exist (McKnight, Cummings, & Chervany, 1998). For example Kramer (1994) studied participants who did not know each other. It was expected that they would show low levels of initial trust but surprisingly some participants showed high initial trust levels. McKnight et al. (1998) therefore developed

a model of initial trust with the aim to define factors and processes that happen when initial trust establishes. Initial trust cannot be based on any type of experience instead it will be based on one's disposition to trust. McKnight et al. (1998) combined several theories of trust research because in this model they argued that each of the trust theories (calculative-based trust, knowledge-based trust, trusting intentions, trusting beliefs) is necessary to understand the big picture. This model, shown in Figure 16, depicts an individual's general *disposition to trust* or their tendency to trust in general.

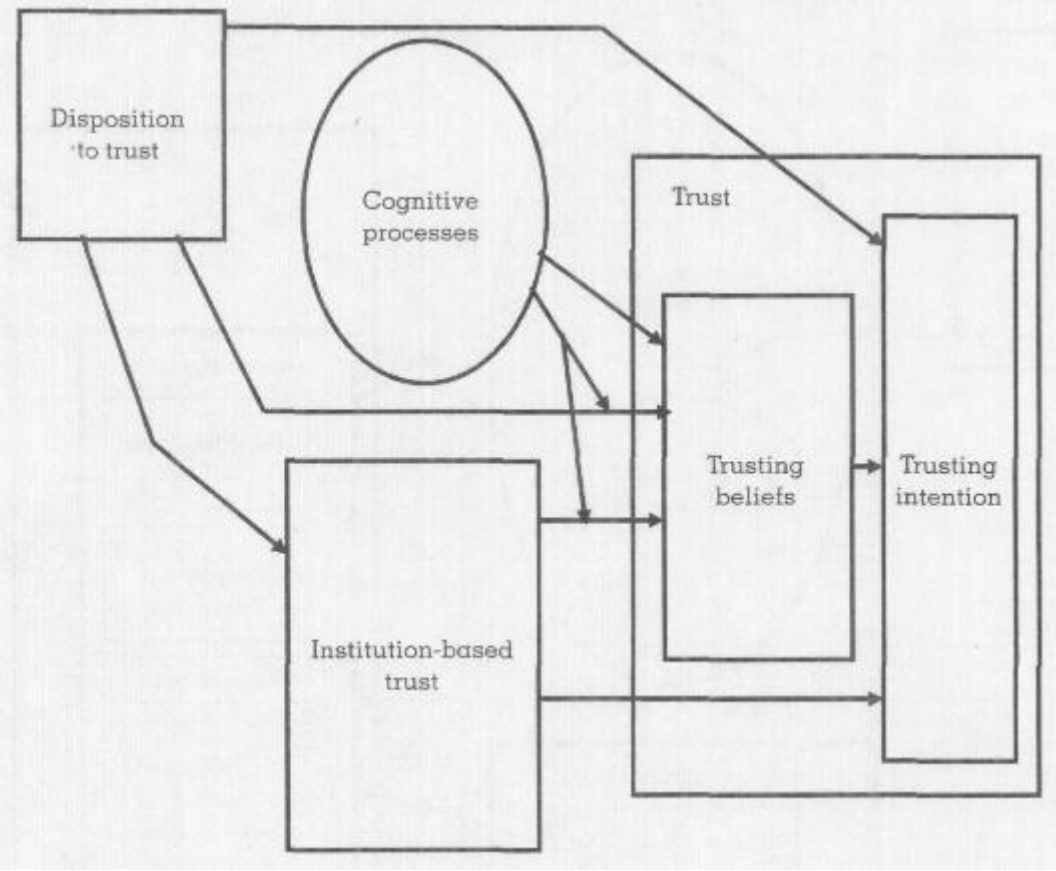


Figure 16 - McKnight (1998), High level model of initial trust formation

It consists of the trusting stance, which describes to what extent people think (or believe) dealing with other people is beneficial and how much they have faith in humanity, which is the extent people believe others are typically reliable and well-meaning. Although this model derives from the economic discipline, the factor *institution-based trust*, can be seen as the belief that the circumstances and the environment (e.g. regulations, guarantees, contracts, etc.) are appropriate for successful interaction. This

belief influences the general *trusting intention* and the *trusting beliefs*. *Trusting beliefs* are the individual's beliefs that the other person is benevolent, competent, honest and predictable in a certain situation. Trusting beliefs are mainly influenced by *cognitive processes* such as categorisation (due to reputation, stereotyping, etc.) or how much a person thinks they are in control of the situation. Overall, the *disposition to trust*, *trusting beliefs* and *institution-based trust* influence the *trust intention* someone has in an initial encounter.

The model explains how trusting intentions can be very high in new relationships and how trust can be robust or even fragile, due to the different influencing constructs. This model suggests that in terms of rescue robots, the attitude towards robots or technology is an influencing factor that should be captured prior to experiments or prior to interaction and that positive trust experiences can foster higher levels of trust between the human and the robot. However, also the regulations (e.g. only use the robot under a certain temperature) and guarantees (e.g. the robot can dive up to 1 meter) have to be clearly communicated to support successful interactions.

2.5.3.2 Interpersonal trust

Interpersonal trust is the expectancy that a person holds, that the other person can be relied upon (cf. Rotter, 1967). Therefore, interpersonal trust is trust humans have among one another. In the trust in automation literature Muir and Moray (1996) mention that interpersonal trust can capture some important aspects of human-machine trust. They based their experiments on a model of "dynamics of trust" from Rempel et al. (1985). Rempel and colleagues (1985) considered three attributional components in their model of trust: predictability, dependability and faith. The predictability of a person is dependent on consistency and the stability of the social environment. In close relationships it is believed that predictability is learned by experiences. Strongly related to predictability is dependability. It is important how much a partner depends on the other and how trustworthy they are. Dependability is based on past experiences and evidence. Sometimes unpredictable things happen and in these situations the role of faith is eminent. With faith people assume things that go beyond

their previous evidence. In a close relationship where we have faith, we trust our partner to be responsive and caring even in unpredictable situations. Rempel et al. (1985) mention that past predictability and dependability provides a foundation for faith and therefore these three components are interrelated.

Simpson (2007) formulated six fundamental principles of interpersonal trust in relationships: the partners dispositions to trust, test situations, joint decisions, expectancies, perceptions of trust, and perceptions of felt security. With these principles he developed a dyadic model of trust to highlight the most important situational and psychological processes in a pair of individuals, as well as the development and maintenance of close relationships. The squared boxes in Figure 17 depict the constructs and the two circles entail the individual differences of the partners. This model assumes that the initial trust and intention to trust lies within the *partners' dispositions*.

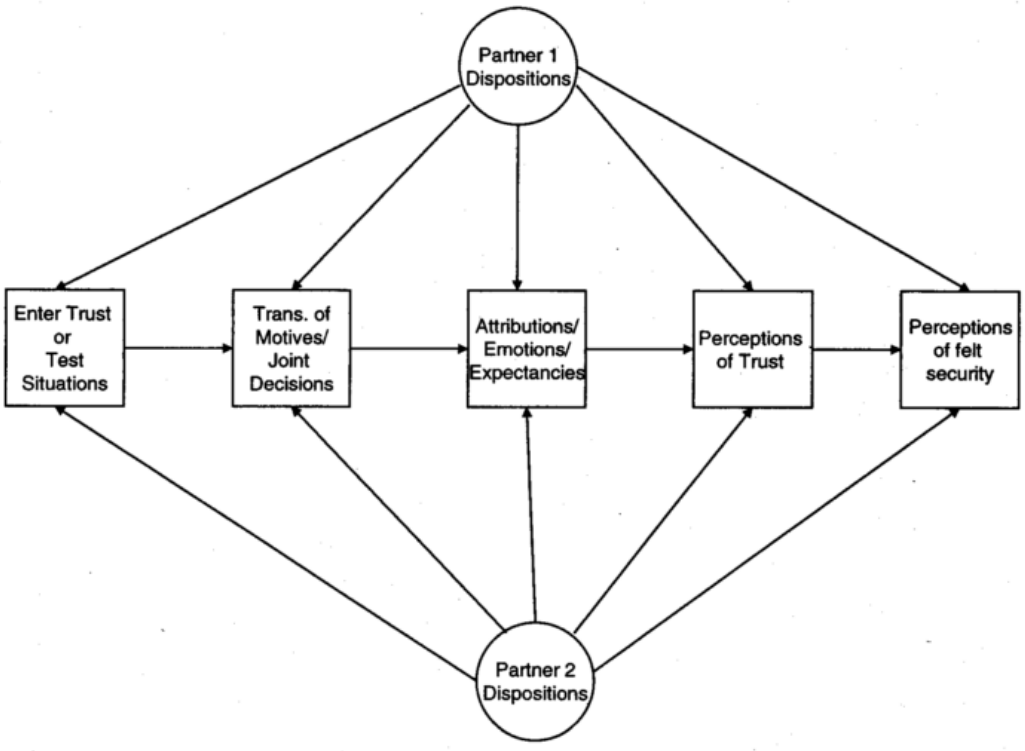


Figure 17 - Dyadic model of trust in relationships from Simpson (2007).

Both partners enter into "test situations", which are situations where the other could fail or win in showing how trustworthy they are. This is similar

to the model of Rempel et al. (Rempel et al., 1985), where experiences play a key role in developing trust. If a partner succeeds and wins those trust situations and shows they care for each other, it may positively change the motives of the relationship and the attributes designated to partners, thus the resulting expectations can change. With time the perception of trust develops and positive trust situations give the partners a perception of security in the relationship. The model assumes that each step has feedback loops. Additionally, the authors raise the issue that early in a relationship many trust transformations are possible but over time partners integrate the other person into their self-concepts which leads to less transformation of motivation later on in the relationship.

According to interpersonal trust concepts, the human and the robot need to have many positive experiences in order to build and maintain trust over time. However, this might be difficult since it is usually only possible to have limited training time with a robot and real-world deployments (e.g. earthquakes, terrorist attacks) are relatively rare compared to everyday firefighting.

2.5.3.3 **Human-animal/human-robot analogy**

Human-animal trust has been used as an analogy to human-robot trust (Billings et al., 2012; Coeckelbergh, 2010; Schaerer, Kelley, & Nicolescu, 2009). On the one hand, in both concepts, predictability, performance, and anthropomorphic characteristics seem to influence the development of trust. In addition, the concepts have in common that the experience and amount of training of the human with the animal is associated with trust. Other factors, such as communication or risk, influence the relationship between human and animal as well as between human and robot (Billings et al., 2012). On the other hand, animals possess self-preservation and other instincts that allow them to act differently to their trained behaviour (Billings et al., 2012). Robots do not have instincts and will act as intended by the designer/programmer, even if that means to destroy themselves. Furthermore humans are familiar with animals and tend to forgive them more easily but their expectations towards machines/robots are very high and the level of forgiveness is quite low (Billings et al., 2012).

The human-animal model also emphasises that the experience in teams plays a major role in the development of trust. Interestingly expectations and forgiveness are different between animals and machines. An intriguing question arises from here: Can we design a robot form and behaviour so that people forgive it more easily, like R2-D2 from Star Wars?

2.5.3.4 Technology acceptance model

The technology acceptance model might shed light on the mechanisms and behaviours humans display when interacting with robots. Although this is not a trust model, it needs to be considered that a robot is still technology. Davis (1986) modified the social psychology grounded theory of reasoned action, which was used as an intention model across several domains towards the technology acceptance model (TAM). TAM aimed to address computing technology usage behaviour (see Figure 18). It is a tool to identify why a certain technology is not accepted and determine the impact factors.

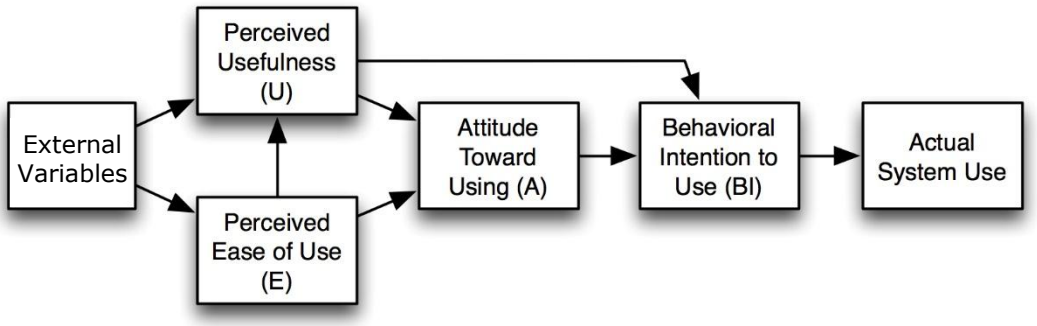


Figure 18 - Technology acceptance model from Davis (1986); picture source: Zaied (2012)

The core of this model is built on the assumption that *perceived usefulness* and *perceived ease of use* are the main factors of technology acceptance and usage. *Perceived usefulness* is the belief of a person that using the technology will increase their task performance. *Perceived ease of use* is the expected level, of the technology, of being free of effort. In some domains the perceived usefulness is one of the most critical factors that determines the usage of technology (K. Chen & Chan, 2011; Hancock, Billings, Schaefer, et al., 2011; Wilkowska & Ziefle, 2009). Some authors further

mention that if the benefits are valuable enough, they can overcome the barriers of low perceived usability (Wilkowska & Ziefle, 2009).

Further, an attitude towards a system is built on the beliefs of the *perceived usefulness* and the *perceived ease of use*. Ease of use is also claimed to influence the *perceived usefulness* of a system (Holden & Rada, 2011). Davis et al. (1989) distinguish two mechanisms of the *perceived ease of use*: self-efficacy and instrumentality. They claim that the easier it is to use a system, the higher the level of the subjective efficacy and perceived control. "Instrumentality" refers to the perception of performing better with less effort. In addition, the model assumes that a person's intention to use a technology is jointly influenced by the attitude towards the technology and the *perceived usefulness*. In particular, people intend to use a system they have positive affect for (Davis et al., 1989). The main focus of people in using a technology is the improvement of their task performance; this is incorporated in the model with a direct relationship between *perceived usefulness* and *behavioural intention to use* the technology. According to Davis et al. (1989), people would use a technology that is beneficial to their performance, whether they have a positive or negative attitude towards it.

To date only some evidence is available that rescue robots are beneficial in real world emergency scenarios (Matsuno et al., 2014; Steinbauer et al., 2014). It seems that in order to create an intention to use robots, the robot primarily needs to demonstrate usefulness in the field and has to be perceived as easy to use.

2.5.3.5 Trust in automation

Most of the research in robot trust was done in the area of automation (see Oriz et al., 2010). When a robot is solely remote controlled the focus of trust is not on its autonomy, instead it is focussed on safety, physical/technological reliability, and the level of performance compared to a human (Oriz et al., 2010). Trust in automation and trust in a remote controlled robot are two different trust concepts. Since this PhD focusses on semi-autonomous robot systems, trust in automation is an important consideration. Moray and Inagaki (1999) classified trust models between humans and machines into five categories: regression models, time series

models, qualitative models, argument-based probabilistic trust, and neutral net models.

Regression models are based on multiple regressions in order to identify independent variables that influence trust or capture the time humans spend in automatic control or willingly relied on automatic control (Lee & Moray, 1992; Muir, 1989). The downside of the regression models is that they do not capture the dynamics of trust. Based on Lee and Moray (1992) findings showed that control allocation is based on someone's own confidence in their ability to perform a task correctly. In a later study Lee and Moray (1994) found that control allocation is influenced by both, trust and self-confidence. They developed a model that is classified as a time series model, which captures the dynamics of trust over time.

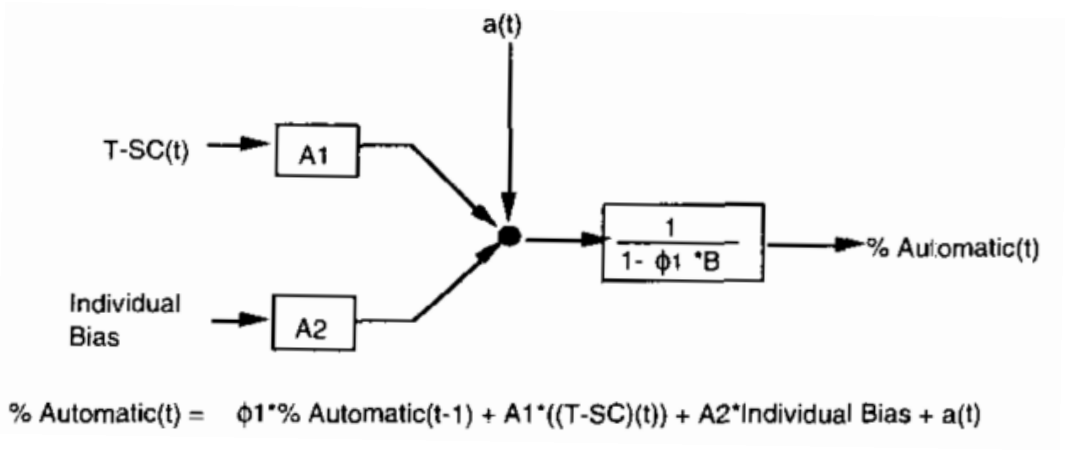


Figure 19 - Time series model from Lee and Moray (1994)

The model helps to understand the association between trust, self-confidence and the use of automatic controllers. As shown in Figure 19, the use of automatic features over time [% Automatic(t)] is influenced by the difference between trust and self-confidence [T-SC(t)]. Furthermore, previous experience [Φ] and individual biases can play a role. The dot [a(t)] represents normally distributed independent fluctuations.

Qualitative models are useful to guide research and describe factors that are known to influence trust. An example of this is the simple qualitative model for trust dynamics based on experiences from Jonker and Treur (1999), as shown in Figure 20. However, compared to the time series model depicted above, they lack the ability to make quantitative predictions. Other

qualitative models were described by Muir (1994) and Desai (2012). Muir's model showed that in order to develop trust, the user needs to interact with the system and experience faults and other non-nominal situations (e.g. misunderstanding the system's intent). In a later work, Muir and Moray (1996) also emphasise that to know what the system is doing next (predictability) and to be able to rely on the constant behaviour of a system (dependability) is important for the development of trust.

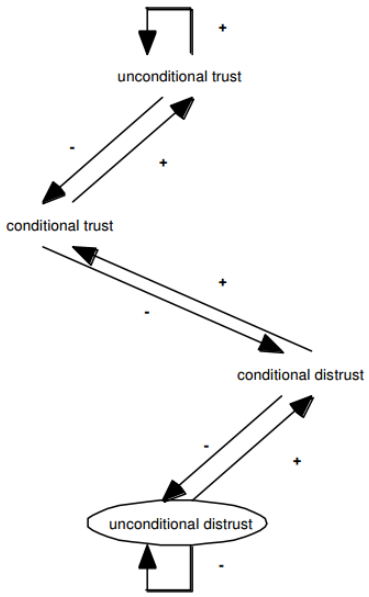


Figure 20 - Simple qualitative model for trust dynamics based on experiences

An example for an argument-based probabilistic trust model is the qualitative work by Cohen et al. (1998). They proposed that the issue of accepting automation is not under-trust or over-trust, but inappropriate trust. Their model is based on the information value theory and uses available evidence from the task, the human, and the system to reduce uncertainty. Basically this model measures an automated decision aid's performance by the likelihood (probability) of the appropriateness of the system's actions. This model is able to show specific conditions under which an automated aid will not perform properly. It is especially useful to identify the conditions that affect good and bad system performance.

The findings of these previous studies point towards the important fact that trust is a major influencing factor that may predict whether a person uses automatic or manual control. Ultimately, control allocation influences the

human-machine performance. In addition, the development of trust is shaped by prior experiences with the system and is dependent on dispositional trust. With the aim of predicting the use of automation, later researchers specified factors and system characteristics that influence trust (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Jian et al., 2000; Lee & See, 2004; Merritt, Lee, Unnerstall, & Huber, 2014; Merritt, 2011; Schaefer et al., 2013). An overview of these factors with respect to HRI are very well summarised in the work of Schaefer (2013). The corresponding trust model will be explained in a later section with the aid of Figure 25.

Recently, Hoff and Bashir (2014) followed up on a systematic review, previously undertaken by Lee and See (2004), and reviewed 127 empirical research studies on trust in automation between 2002 and 2013. They developed a comprehensive model of trust in automation with three main layers of trust: dispositional trust, situational trust, and learned trust (see Figure 21).

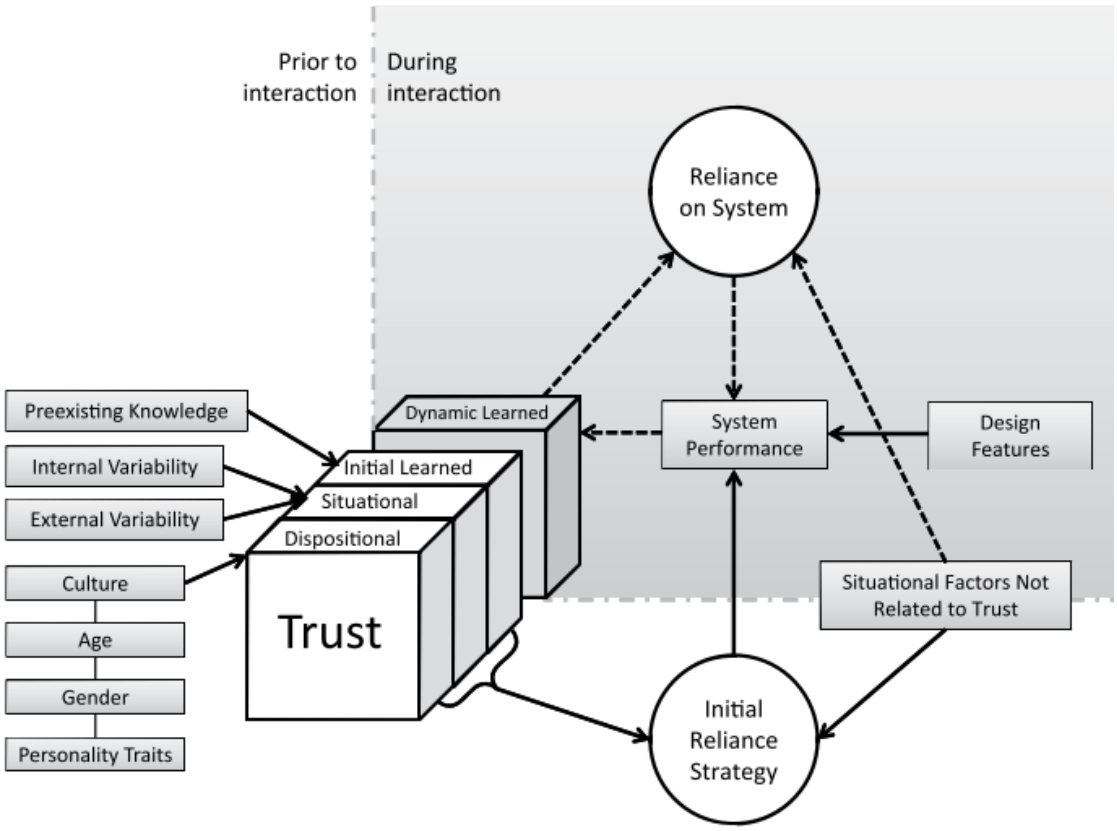


Figure 21 - Trust in automation model from Hoff and Bashir (2014). Dotted arrows represent factors that can change with the course of a single interaction.

As mentioned previously, *dispositional trust* is based on the intrinsic traits of a person (e.g. *culture, gender, age, and personality traits*). Therefore it is the tendency of someone to trust automation. The second layer consists of *situational trust*, representing the factors influencing trust in a given situation. These influences can create *external variability* (e.g. type of automation, task difficulty, perceived benefits, etc.) or *internal variability* (e.g. self-confidence, mood, attentional capacity, etc.). The last layer, *initial learned trust*, represents the developed trust that builds on past experiences with the system or similar systems and the current interaction.

The three layers described above build the *initial reliance strategy* of the human. However, an interaction is of a dynamic nature. *Dynamic learned trust* tends to vary due to the current *system performance* (Hoff & Bashir, 2014). *System performance* can be affected by certain *design features*. Overall, the *reliance on the system* is determines the current *system performance* and is influenced by *situational factors* (not related to trust) and the *dynamic learned trust*. This model also shows that different concepts of trust are not exclusive but can also complement each other.

Hoff and Bashir (2014) state that it is important to see the interdependence of system performance, dynamic learned trust and the operator's reliance on the system (see Figure 22): The performance of the system influences the trust of the operator and the trust level determines the reliance strategy, which again can affect the system's performance.

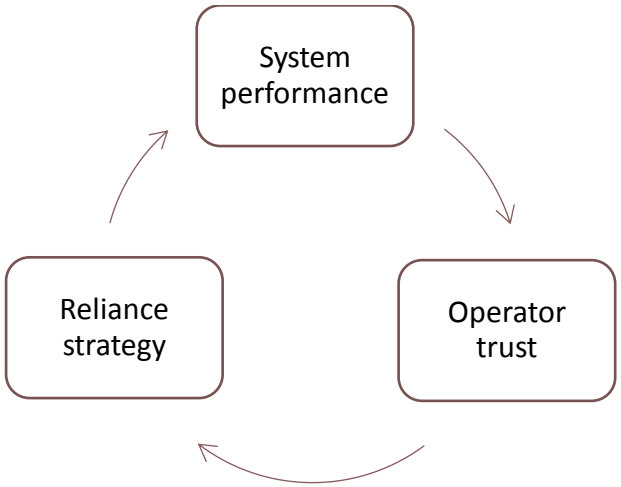


Figure 22 - Interdependence between system performance, operator trust and reliance strategy

Further, Hoff and Bashir (2014) found that the higher the complexity of the automation, the more situations with uncertainty occur, and the more opportunities operators have to compare automated performance to manual. In addition, the higher the level of the human's freedom to make decisions, the stronger is the relationship between reliance and trust.

Interpersonal trust and trust in automation have many factors in common. The factors predictability and dependability are as valid in interpersonal trust as in trust in automation. Additional factors in trust in automation are the operator's self-confidence and the level of complexity of an automated system. The higher the complexity, the more uncertainty is involved in the interaction. This can be interpreted as a warning not to over-complicate automated systems. Complicated and complex automated systems can be challenging to predict and eventually the ironies of automation will lead to the failure of a human-automation system (Bainbridge, 1983).

One of the main advantages of automation is that a system has the ability to perform complex and repetitive tasks quickly and without errors.

However, search and rescue is not a repetitive task. There is still a difference between pure automation and autonomous/semi-autonomous rescue robots because robots may be mobile, encounter situations with great uncertainty, possess different levels of autonomy and/or may be designed like living creatures (e.g. humans, dogs, cats, etc.). There is insufficient evidence that findings from automation studies can be transferred "one-for-one" to human-robot interactions (Desai, 2012; cf. Hancock, Billings, Schaefer, et al., 2011). Trust in automation provides a solid base for research in trust in human-robot interaction.

2.5.3.6 Trust in human-robot interaction

Trust is especially important in the context of the military and emergency situations where the safety and survival of people depend on good human-robot system performance (Hancock, Billings, Schaefer, et al., 2011). Building a universal trust model for human-robot interaction is quite challenging. There are different characteristics of users, unpredictable and changing environments in varying fields of application, and many different robot systems with different tasks and purposes. Therefore only models of trust that are general or aimed to address remote controlled robots are considered in this review.

Computational and mathematical models focus mainly on the entire human-robot system. Figure 23 shows a computational systems approach to modelling collaboration between a human and automation during the control of multiple robots in a search and rescue scenario (Gao, Clare, Macbeth, & Cummings, 2013).

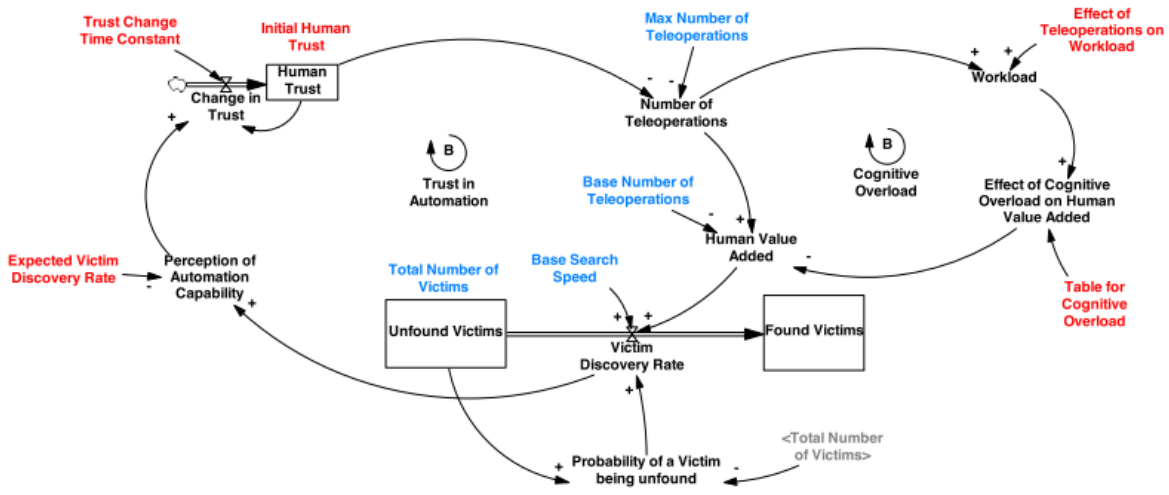


Figure 23 - Human-Automation Collaboration Model (Gao et al., 2013)

The model can be used in computer simulations to predict human behaviour and performance. It consists of three major parts: system performance, trust in automation, and cognitive overload. Factors which influence victim discovery rate (performance) are the total number of victims, the speed of the robot, and the human influence, which is an additive factor to the victim discovery rate. For the trust in automation loop the initial trust, expectations, and general fluctuation/dynamic trust behaviour (Trust Change Time Constant) of the operator are considered. The third part consists of the cognitive overload loop. This part acknowledges that rising workload levels can improve performance, but when workload levels are too high it can cause cognitive overload leading to a decline of performance. In a validation study the model accurately predicted the performance of the system but gave mixed results in terms of the frequency of teleoperation (Gao et al., 2013). There are also other computational models that are concerned with single robot control and which feature similar trust elements as well as mixed control modes and real time/dynamic trust concepts (Y. Wang, Shi, Wang, & Zhang, 2014; Xu & Dudek, 2013).

These computational models can be very accurate in the situation they were developed for. However, these kind of simulations use a variety of simplifications, such as a simple search process or that operators correctly perceive the capabilities of the system (Gao et al., 2013). USAR cannot be simplified in such ways and entails still a great deal of uncertainties. Still, the influence of trust factors and workload have been shown to influence human control allocation and performance.

Hancock et al. (2011) conducted a fundamental and extensive meta-analysis of factors affecting trust in general human-robot interaction. Twenty-nine empirical studies measuring trust towards a robot and used human participants were selected to calculate effect sizes. Figure 24 shows possible factors influencing human-robot trust. The factors were collected via a literature review and subject matter expert guidance. In this overview, three major elements influence human-robot trust: the human, the robot and the environment. *Human-related factors* include *ability-based factors* such as *expertise* or level of *competence*, and other *characteristics* such as

personality traits, *self-confidence* or *propensity to trust*. Many of these factors are similar to those in trust in automation and interpersonal trust models.

Robot-related determinants are classified into *performance-based* and *attribute-based* types. *Environmental* factors are grouped into *team collaboration* and *task characteristics*.

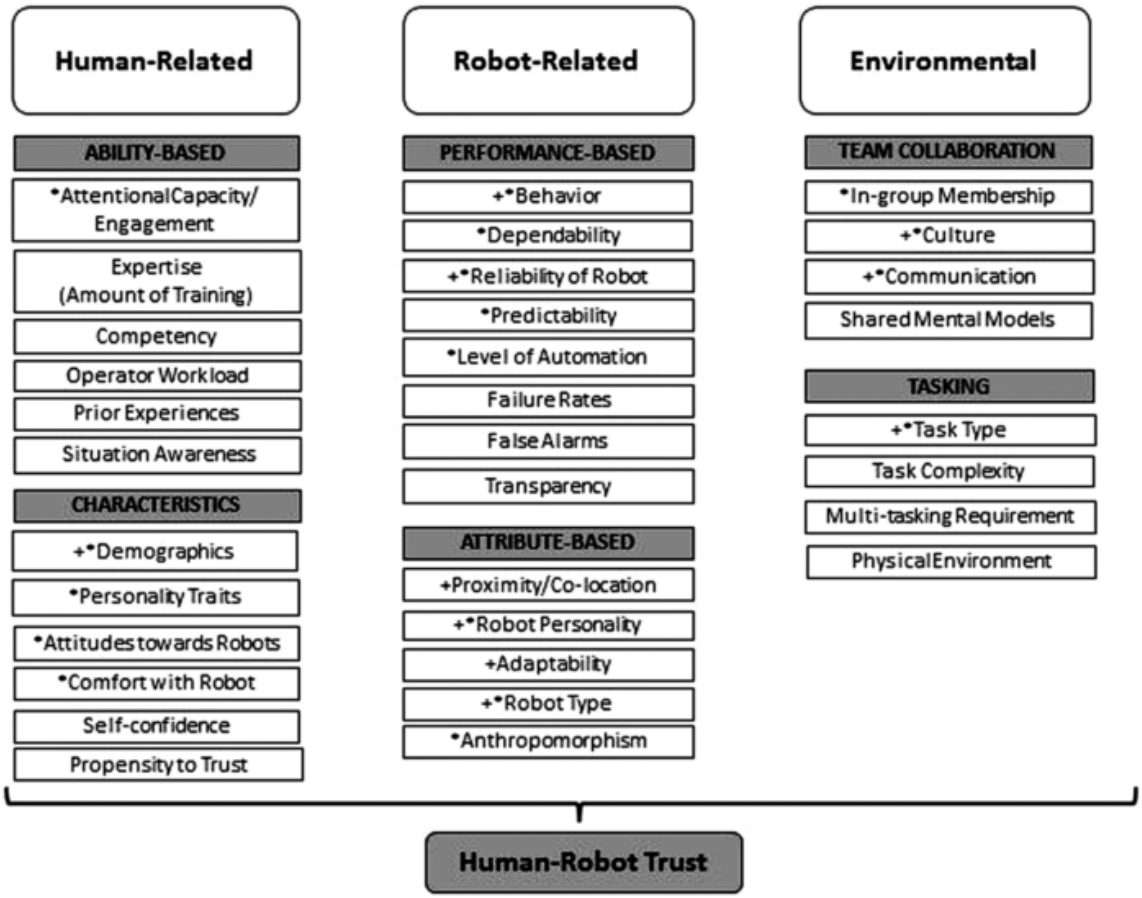


Figure 24 - Factors of trust development in human-robot interaction. Factors included in the correlational analysis are starred (*). Factors included in the group difference analysis are crossed (+). (Hancock, Billings, Schaefer, et al., 2011)

Ten studies were included in a correlational analysis (see Figure 24, * items); the data showed that there was a moderate global effect between trust and all starred factors (Figure 24). An analysis for each of the categories revealed that robot-related factors were mostly associated with trust, followed by the environmental characteristics and the human dimensions. Overall, the largest influencing factors in their analysis were

performance-based characteristics of robots. However the performance based factors rely on only two studies.

Still, there is agreement in the level of reliability associated with the level of trust: the more reliable the system, the higher the level of trust (Desai, 2012; Hancock, Billings, Schaefer, et al., 2011). Studies where group differences were tested (see Figure 24, + items) showed that there was a large global effect regarding trust. Again, the largest effects were associated with robot characteristics, followed by a moderate effect of environmental dimensions and small effects of human influences.

Based on this model and another meta-analysis (Oleson et al., 2011) a new descriptive model was developed (Schaefer et al., 2013), keeping the robot, human and environmental factors at its core (see Figure 25).

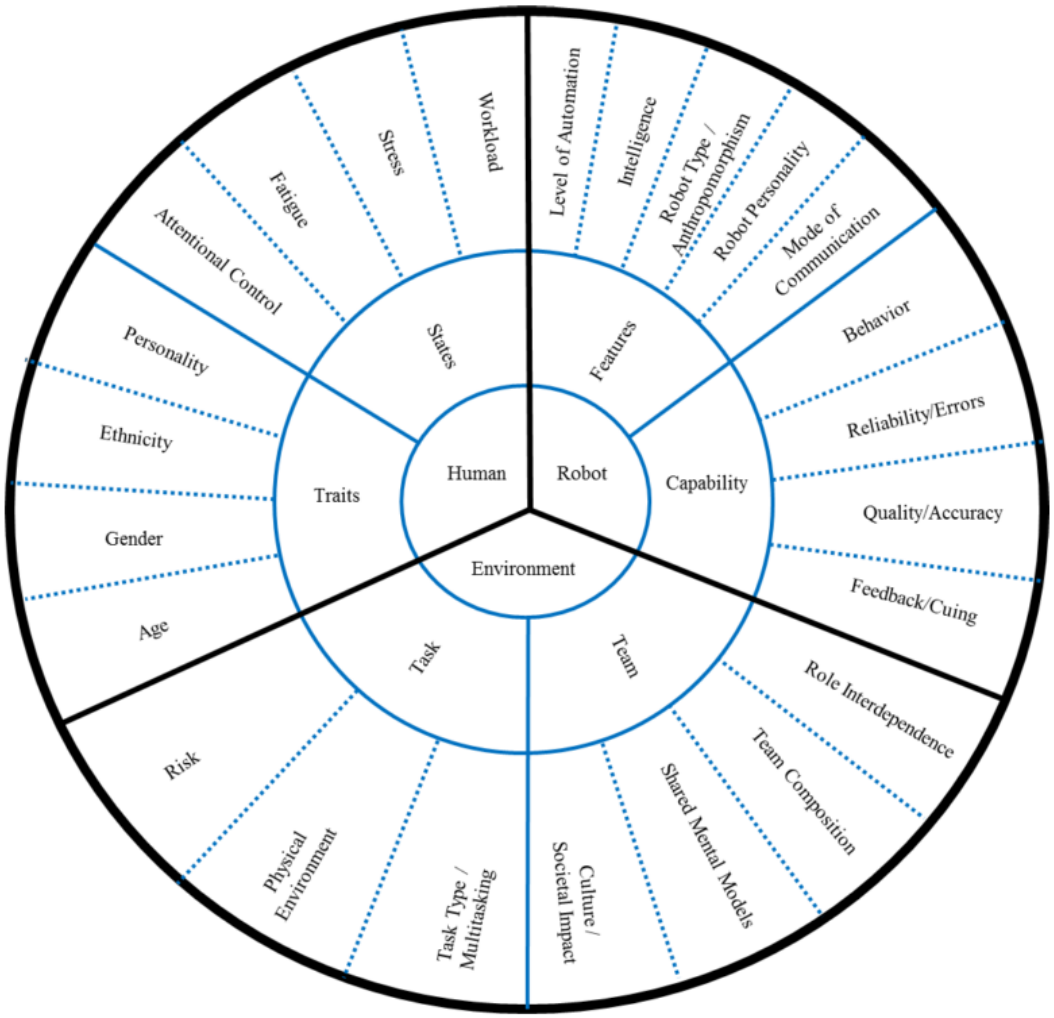


Figure 25 - Updated descriptive human-robot trust model (Schaefer, 2013)

The human factors were divided into traits and states and the robot factors into features and capabilities. In particular, due to the influence of trust in automation (Schaefer et al., 2013), mode of communication and feedback was added to the model because these variables were shown to have a small or moderate effect on trust. An extensive review of the factors in this model can be found in the work of Schaefer et al. (2013). This is the first model to show the "big picture" of factors influencing trust in HRI.

As discussed above Hancock et al. (2011) and Schafer et al. (2013) found that robot performance characteristics are the most influencing factor on trust. This is also in agreement with Desai et al. (2012) who created a model for human-robot interaction regarding remote controlled robots (see Figure 26).

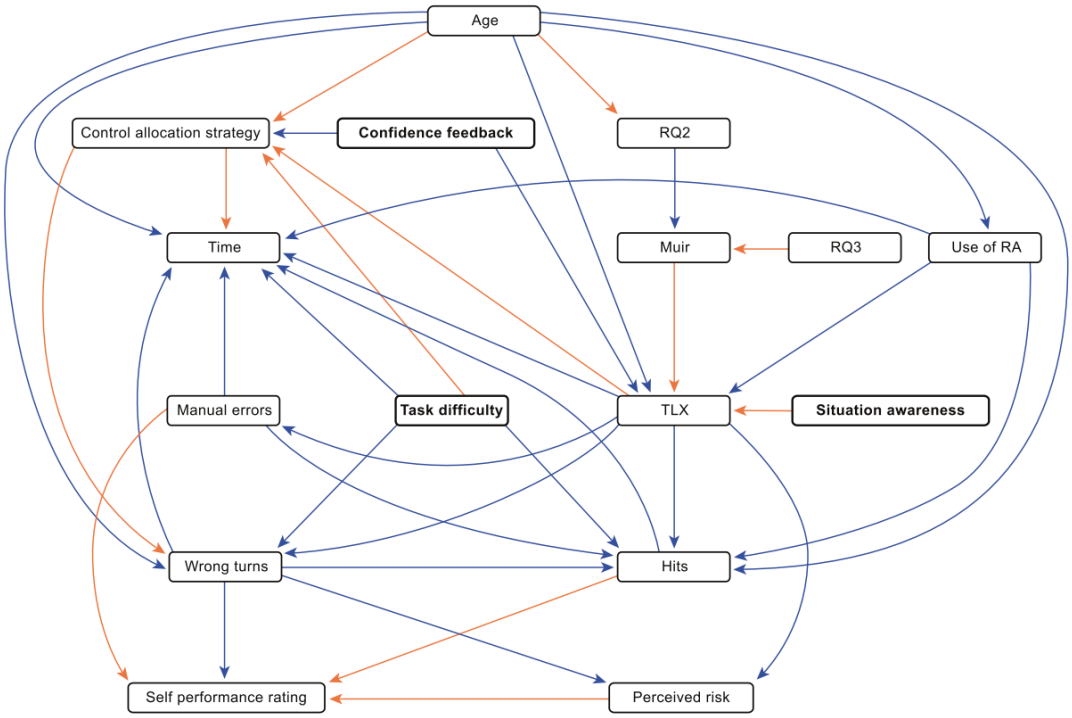


Figure 26 - HARRT model from Desai (2012). Blue arrows indicate a positive relationship and orange arrows indicate a negative relationship.

The data was collected across several studies. The model is a qualitative visualisation of correlations and does not consider the magnitudes of association (see Figure 26). Trust was measured using the Muir (1989) trust questionnaire (Desai, 2012). *RA* is the abbreviation for robot assisted mode, which indicates a lower level of autonomy (similar to manual mode)

compared to the full autonomy mode. In this model trust (*Muir*) was mainly influenced by the risk attitudes of the participants (*RQ2* and *RQ3*). On the one hand, this model does not incorporate the performance of the robot, which was the main influencing factor in previous literature (Hancock, Billings, Schaefer, et al., 2011). On the other hand, the overall human-robot team performance was incorporated by the items *Time*, *Manual errors*, *Wrong turns* and *Hits*. Interestingly, there is no direct relationship between control allocation and trust (*Muir*). Desai (2012) claims that this does not mean that there is no relationship, just that there are stronger relationships that influence the control allocation strategy.

This section has provided an overview of the main aspects and the models that will inform the design of the studies in this PhD. Most relevant for the subsequent studies will be the models of trust in HRI. They provide relevant factors and the relationships between factors in order to select appropriate independent variables to examine and which dependent variables to measure.

2.6 Conclusion

The chapter has reviewed literature about the three main topics: Robotics in USAR, human-robot teams and trust. Although the body of knowledge that is directly concerned with rescue robots and human factors is rare, some literature could be found and discussed.

From the literature emerged that trust in human-robot teams is of importance and requires further investigation. Trust determines the use of autonomous/semi-autonomous robot systems and can possibly influence the mission performance. Especially, the more autonomy is introduced by a system, the more complex and challenging gets the design of the robot/interface, and it is more difficult to foster successful usage and collaboration. The lessons learned from the deployment of the robots showed that especially in terms of robot behaviour, feedback and transparency, research and development is still needed. Further, literature does not provide sufficient information about U.K. rescue teams, their behaviour, training, working processes, and current equipment in order to

develop an appropriate and beneficial system for the U.K. Fire and Rescue Service.

The following factors selected from literature, as shown in Figure 27, are of relevance to human-robot interaction and will be further investigated in this PhD.

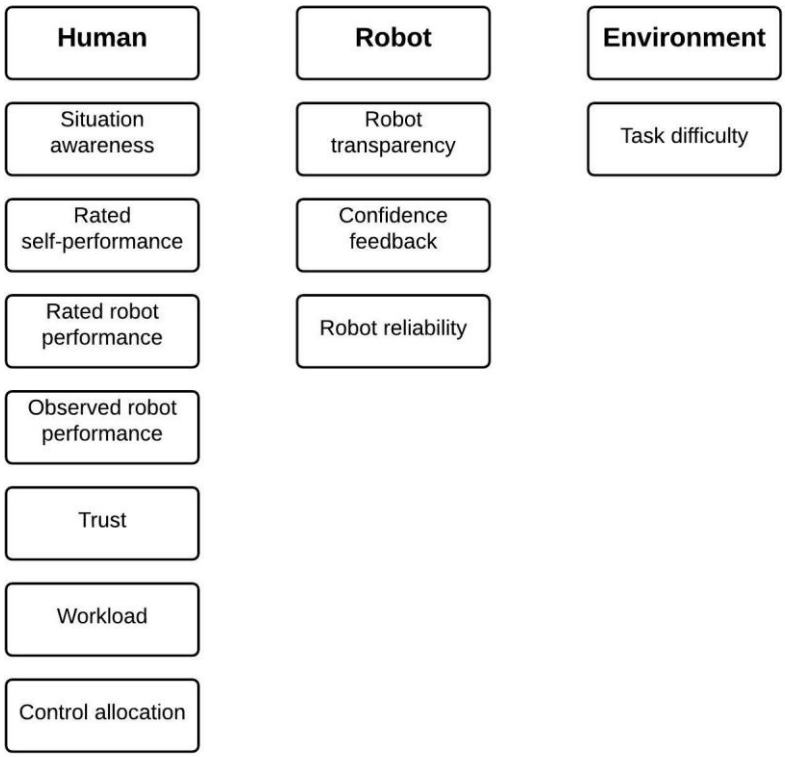


Figure 27 - Selected factors that influence human-robot interaction

The USAR environment delivers a testbed for performance oriented, time critical and complex tasks to investigate. Therefore this PhD will investigate trust in conjunction with robot feedback, transparency, reliability and task complexity with the aid of USAR scenarios. The goal is to design robot systems that are easily accepted, require minimum amounts of training, provide situation awareness and support the rescue team in an efficient and effective manner.

The next chapter discusses the methodology of the overall PhD, explains different measures and methods used and justifies the experimental approaches of each of the studies.

2.7 Chapter summary

This chapter presented a literature review that gave an overview of robots used in USAR, their deployments and autonomy levels. It further identified trust to be an important factor affecting the collaboration between operators and robots. Subjective trust can predict the allocation of functions within the human-robot team, misuse and disuse of a system, and establishing appropriate levels of trust can optimise performance and reveal the potential benefits of using an automated robotic system. The research reported in this thesis is informed by this literature review and carried out with the focus on trust, workload and performance.

3 Methodology

3.1 Chapter overview

This chapter provides the methodology used to investigate the research objectives. The chapter starts with a review of the measures used in literature for human-robot performance and trust. Next, a justification for the selection of methods and tools is provided. The chapter concludes with a description of the equipment used. This includes an overview of the development of a simulated robot in a 3D virtual environment. The development and functionalities of the program are explained.

3.2 Measuring performance

The biggest challenges in measuring performance are the many different types of robots, robot autonomy levels, tasks, and areas of application which need to be considered. Most frameworks and sets of metrics are therefore only valid in certain domains or solely applicable under strict task constraints (Fong et al., 2004; Singer & Akin, 2011).

Common metrics to measure the performance of humans and robots in human-robot task-oriented interaction were investigated by Fong et al. (2004). They categorised the performance into system performance, operator performance and robot performance. This is the same taxonomy that Murphy and Schreckenghost (2013) adopted.

3.2.1 System performance

System performance can be quantitative or subjective. The quantitative performance measures judge the effectiveness and efficiency of the human-robot team (Fong et al., 2004). Efficiency can be the time required to complete the task. Effectiveness is how well the mission was accomplished (e.g. 80%). According to Fong et al. (2004) effectiveness measures need to consider the design of the autonomy of the robot. For example, a robot that is designed to support only in certain situations autonomously (e.g. 50% of the time) but the human needs to intervene 25% of the time the robot acts, the system has an effectiveness of 37.5 %.

Quantitative measures often used in human-robot teams are the number and mental costs of human interventions (Fong et al., 2004), task completion time and average time between failures (Singer & Akin, 2011). Specific to search and rescue the measures victims found, area explored, and workload are of importance (Hamp et al., 2013; H. Wang, Lewis, Velagapudi, Scerri, & Sycara, 2009). In order to assess how the quantitative performance is perceived, subjective ratings should be collected from all involved team members (Fong et al., 2004).

Another category of system performance presents the issue of appropriately regulating the control allocation (Desai et al., 2013). Control allocation can be measured by calculating the ratio of performance benefit to resource allocation or by measuring the effort of the human to work as a team (Fong et al., 2004). The latter is, according to Fong et al. (2004), most appropriate when both competencies, from the human and the robot, are required. Desai (2012) measured performance in semi-autonomous teleoperation objectively by hits (hitting an obstacle with the robot), time needed and wrong turns. Additionally, he used subjective measures by asking the participants to rate the robot's performance and their own performance. He distinguished between the human's and robot's performance by determining the source of the error (e.g. errors made by automation or errors made during manual operation). Problems emerge when participants do not perceive the mistakes of the system, then the perceived or observed performance of the system is different to the actual performance (Chien & Lewis, 2012).

3.2.2 Operator performance

The human's performance is influenced by their capabilities, situation awareness, trust and workload (Murphy & Schreckenghost, 2013). One method to measure subjective workload is the NASA Task Load index (Hart & Staveland, 1988). This method is widely used in HRI (Chien & Lewis, 2012; de Visser & Parasuraman, 2011; Desai, 2012; Helldin, 2014; Sanders, Wixon, Schaefer, Chen, & Hancock, 2014; Selkowitz, Lakhmani, Chen, & Boyce, 2015). In order to assess situation awareness often the SAGAT, short for "Situation Awareness Global Assessment Technique"

(Endsley, 1988), has been used in HRI (T. B. Chen, Campbell, Gonzalez, & Coppin, 2014; Desai, 2012; Selkowitz et al., 2015). Nevertheless, Singer and Akin (2011) mention that situational awareness can have different perspectives: The human's awareness of the overall missions and tasks, or the human awareness of the robot's current state and the environment it is in. Therefore, it should be clear which perspective needs to be measured.

Furthermore the accuracy of mental models needs to be considered (Fong et al., 2004; Murphy & Schreckenghost, 2013). This includes the appropriate design of affordances, operator expectations, and matching the interface and controls to the human mental model (Fong et al., 2004). Singer and Akin (2011) mentioned that also the information type and variety delivered by the robot are influencing factors on human performance.

3.2.3 Robot performance

In manual teleoperation without autonomous robot support, only the robot's physical capabilities and technical reliability are influencing factors of the robot performance. With increasing robot autonomy additional metrics emerge. These metric are robot autonomy level, self-awareness and human-awareness (Fong et al., 2004). Self-awareness of a robot is the ability of the robot to know its current state and be aware of possible errors. For example the indication of the robot about its current reliability level shows a high self-awareness. Human-awareness often refers to co-located robots and to the extent the robot is able to sense, recognise and interpret human behaviour. If human and robot share a task, such as a hand-over task, the robot needs to be aware of the human's positon and trajectory (e.g. Dehais, Sisbot, Alami, & Causse, 2011). For remote controlled robot systems the human-awareness would be the correct implementation of human commands or the adaption of the system to the current state of the human. For instance, the heart rate and pupil diameter of a human is changing and indicating higher levels of workload (Kramer, 1991): a robot, being human-aware, could adapt its autonomy or the amount of feedback to support the human.

As mentioned by Murphy and Schreckenghost (2013) literature identified many metrics that can be captured but does not give many suggestions as to how to measure these. This PhD will investigate this issue and propose measures that can be applied in semi-autonomous remote controlled human-robot teams. The next section gives an overview of existing trust questionnaires for HRI.

3.3 Measuring trust

As established in the literature review (section 2.5.3 Models of trust) there are different types of trust. But most questionnaires only measure one type of trust. For example, trust can be measured by a single item question such as, "Do you trust this system?" with an binary answer: yes or no (e.g. Robinette, Wagner, & Howard, 2015). However, trust is a far more complex construct and influenced by many different factors. In order to investigate trust issues and their cause, more detailed questionnaires are necessary, but the details and trust factors are domain dependent (trust in management, trust in people, trust in robots, etc.). Most questionnaires and trust models for HRI were developed from interpersonal trust and trust in automation questionnaires because these domains are most similar to trust in human-robot teams. The sections below provide an overview of the trust scales used in HRI to investigate which questionnaire might be appropriate to use in semi-autonomous remote controlled robot systems.

3.3.1 Trust in automation (Muir, 1989)

The first questionnaire developed for trust in automation, and later used for HRI, was developed by Muir (1989). Early on Muir suggested that the amount of trust will determine the use of automatic controllers. Therefore he developed a fast and easy to complete questionnaire with four questions regarding the system in question, which could be answered on a scale from 1 (Not at all) to 10 (Completely). The items included the system's predictability, effectiveness, faith in future performance and a direct trust question. This scale was used for measuring trust in HRI by Desai et al. (2012, 2013). They found that the Muir scale is not sensitive to changes in trust when participants were presented with different reliability profiles of robots (Desai, 2012). In a later study Desai et al. (2013) stated that the

trust scale, since it is a post-task questionnaire, seems to be biased by a primacy-recency effect. However, the biggest advantage is the short length of the questionnaire.

3.3.2 Jian trust scale (Jian et al., 1998)

The Jian trust scale was also developed for measuring trust in automation (Jian et al., 2000). This questionnaire did not distinguish between the different trust domains such as human-human or human-machine trust. By using, among other methods, a word elicitation study and paired comparison they developed a 12-item questionnaire that consisted of statements about the system (e.g. the system is reliable). The statements can be answered on a 7-point scale (1 = not at all and 7 = extremely). In parts, the questionnaire is negatively phrased with questions such as, "The system is deceptive" or "The system's actions will have a harmful or injurious outcome". Modified versions of this questionnaire were used by Chen and Terrence (2009) in a military context involving robotics tasks and Chien and Lewis (2012) in the context of USAR with multiple robots. In both cases, no significant difference in trust could be found.

Desai (2012) compared Jian et al.'s (2000) questionnaire to Muir's (1989) questionnaire and obtained nearly identical results for both questionnaires in user trials with a semi-autonomous robot system and suggested that using just one of these questionnaires is sufficient.

3.3.3 HRI trust scale (Yagoda and Gillan, 2012)

Yagoda and Gillan (2012) developed a HRI trust scale by creating a list of factors that influence trust from literature and subject matter experts. The main HRI topics that the questionnaire contains are team configuration, team process, context, task, and system. All 37 items were tested for inter-correlation and an exploratory factor analysis was performed. Each item is scored on a 7-point Likert rating scale ranging from "strongly disagree" to "strongly agree". In addition a field with "N/A" provides the possibility to omit certain questions if they are not appropriate or applicable in the given context.

Yagoda and Gillan (2012) emphasise on the additional use of the Interpersonal trust scale from Rotter (1967). For a comprehensive picture of trust, the incorporation of individual differences is important. What distinguished this questionnaire from typical automation questionnaires was the use of questions that asked for team configurations and the physical environment. However, this questionnaire was not further used or validated in research literature.

3.3.4 Real-time trust (Desai, 2012)

Desai (2012) developed a real-time trust measure in order to capture the dynamics of trust over time and show immediate changes of trust after robot failures. Participants were asked to state every 25 seconds if their trust increased, decreased or had not changed by pressing an upward, downward or horizontal arrow. The data points were drawn up on a graph over time and the area under the curve was calculated as a trust value. This method was also used to prove that early failures of the robot have a higher impact on trust than later failures.

However, using the area under the curve produces a mathematical problem. Since the graph starts at zero, which already assumes initial trust is zero, negative effects accumulate mathematically. For example, as shown in Figure 28, if an early mistake is made and trust declines, but then rises continuously because of high reliability until the end, the area under the curve is lower than a later drop in reliability. The early drop curve can never be bigger than the later drop curve, because the participant can only indicate increase (+1) or decrease (-1) during fixed intervals. This means that from a mathematical point of view an earlier drop in trust will always produce less trust under the curve than a later drop. It might be useful to use absolute values and then calculate the area under the curve for an overall trust score. This can also overcome the shortfall of assuming that trust starts at zero because initial trust can also be high (Kramer, 1994).

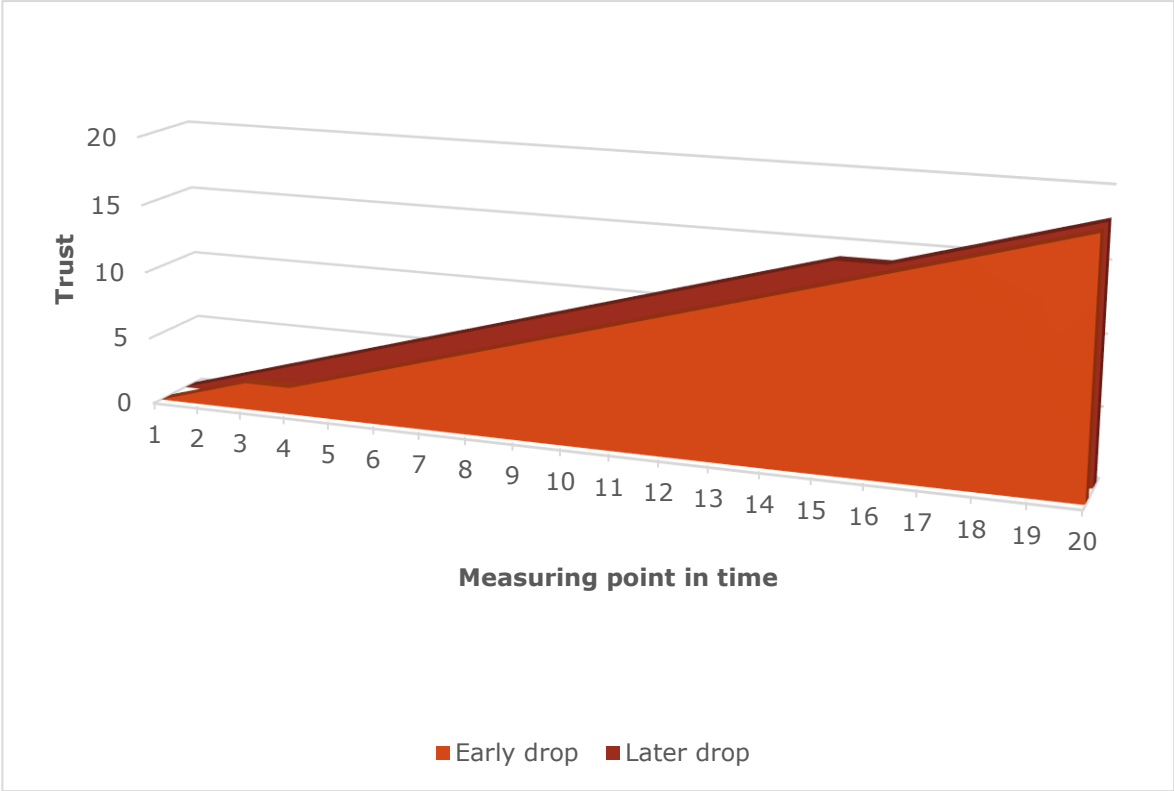


Figure 28 - Area under the curve real-time trust measure

Therefore taking the area under the curve as evidence that early failures of the robot have a more negative impact on trust than later failures, might not be sufficient.

3.3.5 Human-robot trust scale (Schaefer, 2013)

With the aid of the human-robot trust model (see Figure 25, p. 57) Schaefer (2013) developed a human-robot trust scale with 40 items, as well as a short version with 14 items. An extensive review of the literature on trust led to the collection of 172 items that were proposed to be influencing factors of trust. These items were reduced by conducting six experiments, deduction techniques, and with the support of subject matter experts. Furthermore, two of the six experiments validated the scale. The scale produces a trust percentage score between 0% and 100%. This scale was later used by Schaefer's research group (Sanders, Harpold, Kessler, & Hancock, 2015; Sanders et al., 2014). However, the scale was not specifically developed for semi-autonomous robot systems.

In conclusion, the only trust measures that have been used in conjunction with semi-autonomous robot systems are the Jian et al. (2000) (Ross, Szalma, & Hancock, 2007) and Muir (1989) questionnaire (Desai, 2012). And since both show similar results the Muir (1989) questionnaire is favoured due to its short length. However, a very detailed and promising questionnaire is the Schaefer (2013) trust measure. In subsequent studies the Muir (1989) and Schaefer (2013) questionnaires will be tested. The next section will discuss the selection of the experimental methods of this thesis.

3.4 Experimental methodology

This section discusses and justifies the selection of the different experimental methodologies used. The specific application of the methods is explained in the methods chapter of the individual study chapters.

3.4.1 Approach to data collection and analysis

Overall, four studies were performed by using a variety of methods that include qualitative and quantitative data. This mixed methods approach mostly collected quantitative data via questionnaires and qualitative data with the aid of interviews. Qualitative data collected during the studies was analysed with the theme-based content analysis (TBCA) from Neale & Nichols (2001). This method was selected due to the fact that it retains the raw quotes and orders the qualitative data into quantifiable themes.

An appropriate level of trust is key to the usage of automated systems (Lee & See, 2004) and exploiting the potential benefits of an automated system (Parasuraman & Riley, 1997). In this thesis trust was generally measured with a single item rating question and the Schaefer (2013) human-robot trust questionnaire. The Schaefer (2013) trust questionnaire was chosen, because it proved to be sensitive to trust changes and is specific to human-robot interaction. Added to the questionnaires were ‘not applicable’ options for each item, because some questions were aimed at social or co-located robots.

Workload can influence misuse and disuse of automation (Parasuraman & Riley, 1997) as well as the task performance (Prewett, Johnson, Saboe, Elliott, & Coovert, 2010). After each task participants were asked to

complete a NASA-Task Load Index (NASA-TLX) questionnaire (Hart & Staveland, 1988). The NASA TLX is a multi-dimensional scale to obtain subjective workload and was developed by the Human Performance Group at NASA's Ames Research Center. This tool was used because of the easy utilisation and its common use (Fong et al., 2004).

According to a meta-analysis from Hancock et al. (2011) personality traits are also likely to influence trust ratings. The shortened version of the personality questionnaire of Goldberg's Big Five factor structure (Goldberg, 1992) was used. This shortened 50-item 'IPIP' personality questionnaire (Gow, Whiteman, Pattie, & Deary, 2005) was applied because it showed good internal consistency and related strongly with the known dimensions of personality (Gow et al., 2005).

The specific approaches are grouped into the aims and objectives established in the introduction of this thesis. An overview of the studies is visible in Table 1 (p. 11).

3.4.1.1 **Aim I**

Aim I: Develop a background understanding of the USAR domain and their work as well as describing the real world application of USAR in order to provide recommendations for the implementations of robots in British USAR teams. This aim includes the objectives:

1. Gather background knowledge of the USAR domain, especially their technical equipment used to date, as well as investigating the rescue culture and team behaviours within this user group to inform future experiments and robot designs.
2. Study organisational structures and rescue processes to find an appropriate robot position in the system in order to give recommendations for an implementation of robots in British USAR teams.
3. Collect data about rescuers' attitudes towards robots.

In order to gather background knowledge and address objective 1 an autoethnographic approach was used. The author took part of a USAR

Technician Course and was involved in all training aspects a firefighter would get in order to be qualified for a USAR mission (see Chapter 4). Autoethnography comes from "auto" (self), "ethno" (culture) and "graphy" (writing) (Munro, 2011) and is a qualitative method that combines autobiography and ethnography (Ellis, Adams, & Bochner, 2011), therefore a combination of personal experience and observation. Studies can be composed of a pure autoethnographic 'story' (e.g. "There are survivors" from Ellis (1993)) or they can combine autoethnographic elements with other observations. Autoethnography is a method to describe and analyse personal experience with the aim to understand cultural experience (Ellis et al., 2011). To the knowledge of the author no studies in the USAR domain used an autoethnographic approach yet. With respect to human factors it is used in the investigation of workplaces, their culture, conflicts and performance. For example Sobre-Denton (2012) investigated workplace bullying, gender discrimination and white privilege with an autoethnographic approach. She chose this approach as a method of sense-making through her own personal identity situated in her workplace and experiencing the social activity around her. Also, a very amusing piece of autoethnographic work with invaluable insight into job satisfaction and informal interaction was done by Roy (1959). The title "banana-time" derived from the fact that within the group, where the author became part of, most working breaks got names such as, coffee time, peach time, or banana time. From his autoethnographic work he could derive a variety of practical and theoretical considerations regarding job satisfaction. With the background knowledge gathered, future experiments could be adequately informed and designed to reproduce a real-world like scenario.

For objective 2 informal interviews were used to gather relevant quotes that support inferences of the author. With the study of organisational structures and rescue processes possible implementation recommendations for robotic aids could be established.

Objective 3: In order to collect the attitude of rescuers towards robot the "Negative Attitude Toward Robots Scale" (NARS) was used (Tsui et al., 2010). This attitudinal questionnaire was chosen because it is not a general attitude questionnaire towards technology, but is directly aimed at robots.

The NARS has been applied in the area of autonomous and telepresence robots (Tsui, Desai, Yanco, Cramer, & Kemper, 2010). The scale provides a baseline of the general attitude of the participant towards rescue robots in general. The scale is divided into three subscales which ask for the different areas of attitude: negative attitudes toward situations and interactions with robots, negative attitudes toward social influence of robots, and negative attitudes toward emotions in interaction with robots.

3.4.1.2 **Aim II**

Aim II: Improve understanding of underpinning cognitive concepts, thoughts and behaviours of participants while interacting with different autonomous and semi-autonomous robots, in order to inform future robot behaviour and interface design as well as the subsequent studies of this PhD. The aim comprises the following objectives:

4. Explore relevant rescue tasks with a retrospective verbal protocol and gather information about thoughts and feelings during human-robot interaction.
5. Collect interview data regarding robot characteristics and participant preferences.

In order to inform the objectives above, Study II (can be found in Chapter 5) used a video stream of an autonomous rescue robot. Participants had to interact with the autonomous robot and were filmed during that task. After the task they watched their own video and performed a retrospective verbal protocol (RVP). The pilot study showed that a concurrent verbal protocol interfered too much with the main task and secondary task. The RVP data was divided into certain events that happened during the task (e.g. robot successfully identified target). In conjunction with the theme-based content analysis (TBCA) the events were further grouped into emerging themes. With this method, feelings and thoughts in form of quotes could be appropriately grouped and quantified. In addition, in study II (Chapter 5), III (Chapter 6), and IV (Chapter 7), interviews provided information about robot characteristics (e.g. robot speed, map visualisation, etc.) and participant's preferences of these characteristics.

3.4.1.3 **Aim III**

Aim III: Investigate how robot and environmental characteristics, influence user cognition, behaviour and performance. The aim contains the objectives listed below.

6. Identify the key cognitive concepts that are relevant to USAR.
7. Identify, compare and select appropriate measurements of these key cognitive concepts against each other.
8. Examine the effects of different feedback on trust.
9. Investigate the influence of task complexity and robot reliability on performance, workload and trust. In addition, compare performance levels between semi-autonomous controlled robot and manual controlled robots.
10. Compare, with the aid of the situation awareness transparency model, two different levels of interface transparency across two levels of task complexity.
11. Develop a measurement of performance in semi-autonomous human-robot teams.

Objective 6 was addressed with a critical literature review (Chapter 2). The review gave insight into the factors that were relevant to remote controlled robot systems and human-robot collaborative systems. The subsequent studies were designed on the basis of this review.

The methodology, in section 3.2 and 3.3, also informed about existing measures for performance and trust, which responds to objective 7. In addition, study III (Chapter 6) compared two trust questionnaires with each other: Muir (1989) and Schaefer (2013). The comparison aimed to determine how sensitive these questionnaires were and which was appropriate to use in the subsequent studies.

For objective 8 the second study (Chapter 5) compared two robots with different amounts of feedback. Qualitative data was gathered with a retrospective verbal protocol.

Study III (Chapter 6) informs objective 9 by comparing the performance scores and other collected measures between participants who used a mixed

control mode (using automatic and manual teleoperation) and participants who exclusively used manual mode. This also aimed to quantify the benefits of robot automation on rescue performance.

Objective 10 is addressed with Study IV (Chapter 7). The study used the situation awareness transparency model to create two different levels of interface transparency. They were tested across two different task complexities.

In study III (Chapter 6) different performance measures were developed in order to approach objective 11. So far there were no standard performance measures for human-robot teams. This is a starting point for discriminating between observed robot performance and objective robot performance. Because mistakes from the robot can be overlooked by the participant and alter their perception of the robot performance and consequently their trust in the robot. Therefore the measure "observed performance" was proposed.

3.5 Equipment used during experiments

3.5.1 Robot system

The robot used in Study II was a LEGO Mindstorms NXT 2.0 robot with a 32-bit microprocessor and 4 output ports. Connected to the ports were two servo motors for moving the robot, an ultrasonic sensor for measuring proximity to obstacles, and a colour sensor for guided navigation. Furthermore a wireless camera with TV signal was paired with a PC screen that was visible to the participant. The NXT could drive automatically or be controlled via a PC keyboard. The robot was programmed with LabVIEW. With the use of the colour sensor it was able to follow a line on the floor and an operator could take over manual control if necessary. Due to technical faults and unreliable signal strength it was decided to use a video stream from the robot rather than an uncontrollable robot system. An unreliable system would have introduced unintended and uncontrollable system faults that would have influenced the measures of the study significantly. For later studies a new virtual rescue scenario was developed, tested and utilised. The development of the scenario is described in the next section.

3.5.2 The development of a Virtual USAR scenario in UNITY

To overcome technical issues experienced in Study II, a new search and rescue scenario was developed in UNITY. The scenario includes a simulated robot in a simulated 3D environment. Participants would interact with the program via a desktop PC. The developed program was used in study III and IV. The development and functionalities of the program are explained in the following paragraphs.

Many research areas use virtual reality based experimental setups to save resources, time and/or not bringing their participants in danger (e.g. driving simulators) or discomfort them (e.g. Lewis, 2014). In this PhD a virtual reality approach was used because of limited resources, the possibility to create laboratory-like conditions, and not bringing participants in dangerous emergency situations. Furthermore, the remote position and the interface interaction are very similar to the real world application, where the operator is only able to interact with the robot via a screen and a controller.

Previous rescue robot studies were performed either, with real robot systems, the Wizard of Oz method, or in USARSim (Chien, Lewis, Mehrotra, Brooks, & Sycara, 2012; Gao et al., 2012; Horsch, Smets, Neerincx, & Cuijpers, 2013). USARSim is a high-fidelity simulation software of USAR robots and environments. It is used as a research tool for robot development and human-robot interaction. The program has the capability to accurately display user interfaces, robot automation, and the remote environment. The unreal engine based simulation is very coding intensive and requires a skilled programmer to develop high-fidelity simulations.

Craighead, Burke and Murphy (2008) discussed the use of the Unity game engine for a search and rescue gaming environment. They stated that Unity can present a high-fidelity simulation environment for search and rescue. Unity is a game development engine for creating multiplatform (Windows, Mac, Smartphones, etc.) games in 3D and 2D. It has an easy to learn and user-friendly environment (Menard, 2011). Unity can be used free for private and academic developers. Although Unity is a game engine it has the flexibility and capability to also support researchers in their work

(Robinette et al., 2015). For example it can be used for stimulus presentation in psychology experiments and it can collect data in text files that can be used in other programs (e.g. SPSS).

For Chapter 6 and Chapter 7 virtual USAR scenarios were developed by using the Unity game engine. The main reason were the limited time for development and the limited skills of programming of the author. The software has an object oriented work-flow that allows easy development of environments via drag and drop. In comparison to other software a minimum of coding is required. Furthermore, Unity has an asset store where developers can download ready-made 3D models, animations, scripts, and shaders. Models can be imported and assigned physical properties and control scripts.

The robot control script was developed with the substantial help of Dimitrios Darzentas from the Horizon Doctoral Training Centre at the University of Nottingham. The design of the virtual environment, robot behaviour, and tasks were informed by the experiences and documentation of study I (Chapter 4).

3.5.2.1 System overview

The simulation requires a laptop or computer that is able to run the Unity Engine. Two screens were used, one for the researcher to observe the participant's performance and the simulation for errors, and a screen for the Game View, which represents the robot interface for the participant/operator. In order to steer the robot through the environment an XboX controller was used, in some cases the keyboard needed to be used (e. g. secondary task). The XboX controller was also used by Desai (2012). Figure 29 shows the experimental setup from the perspective of the participant.

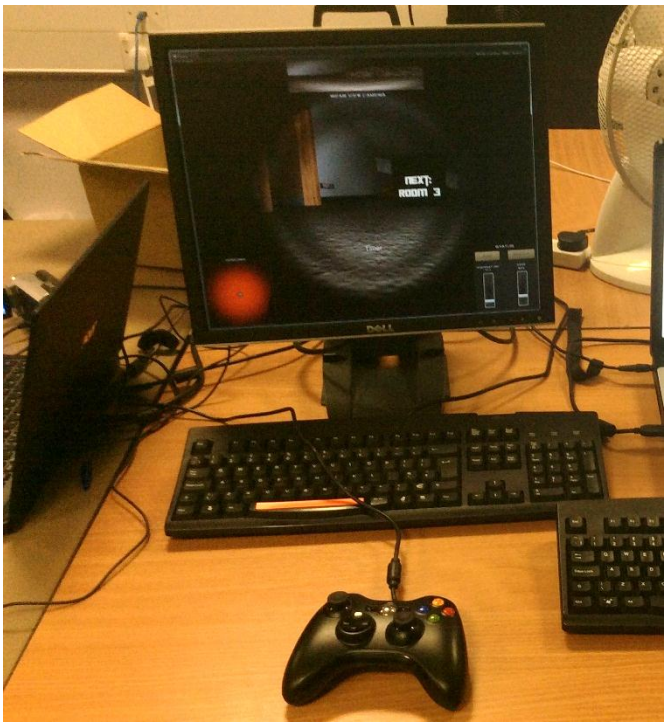


Figure 29 - Required hardware for rescue simulation: Laptop that runs the Unity program (left), a second screen (middle), XboX controller, second keyboard for participants.

A Unity project includes different environments (in specialist jargon called scenes). Environments are the simulated surroundings that participants see on the screen. For the different studies different environments were used to avoid a learning effects of routes and targets' positions. An example of such an environment is shown in Figure 30 from the perspective of the developer (bird's eye view).

Figure 30 – Example of a Unity environment (scene)

The next paragraphs will explain the different simulated Unity components, such as the environments, the robot, and the interface (only visible component to the participant). Screenshots and example scripts are provided for a better understanding of the simulation. The scripts are mostly coded in C# and Java.

3.5.2.2 **Simulated components**

3.5.2.2.1 Environments

Each environment consisted of a variety of objects in the environment, so called "3D models". Models could be rubble piles, doors, canisters, chairs, and other objects. They were mostly imported from the asset store and occasionally modified. Each environment visualised a partly collapsed office complex, with corridors and rooms. The objects in the environment are explained below.

Waypoints - For the purpose of comparability waypoint indicators showed participants where they had to go next, depending from which direction they approached the cube. Furthermore the doors had labels that corresponded

to the waypoint indicators (see Figure 31). This ensured that all participants used a similar route and therefore experienced similar viewing angles of the environment. There were several groups of objects which are outlines below.

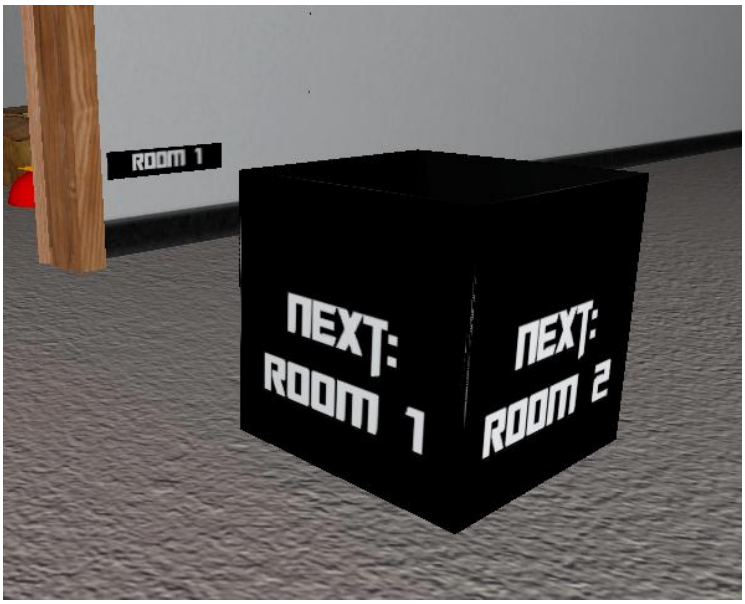


Figure 31 - Waypoint indicator and room label

Rubble - Most of the environment consisted of rubble, barrels, chairs, tables etc. Each object possess physical characteristics and a collider. A collider is an invisible box around an object, with the purpose that other objects, such as the robot, cannot penetrate it, and therefore they would collide with each other.

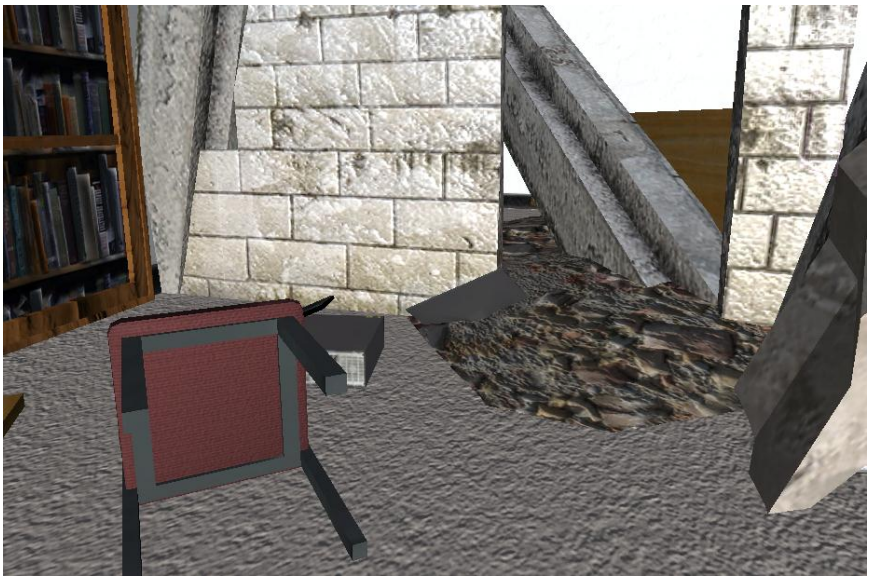


Figure 32 - Rubble, chair, and other objects in the environment

Figure 32 shows an example of the rubble and other objects that were used to clutter the environment. During the experimental conditions the amount of rubble and objects changed depending on the task complexity level.

Smoke – Smoke particles were used to make sensors inaccurate and the visibility more difficult. Smoke could also produce lag in the signal between the robot and the participant. Smoke is shown in Figure 33.

Figure 33 - Smoke particles in the environment

Fire – Fire produced the same issues as smoke but did additionally trigger the temperature gauge of the interface to rise. If participants drove too near to a fire the robot would be incapacitated and the mission aborted. An example of a fire in the environment is visualised in Figure 34.

Figure 34 - Fire particles in the environment

Targets – The participant was required to find specific targets, as shown in Figure 35. These targets could be trapped humans (victims), hazard signs,

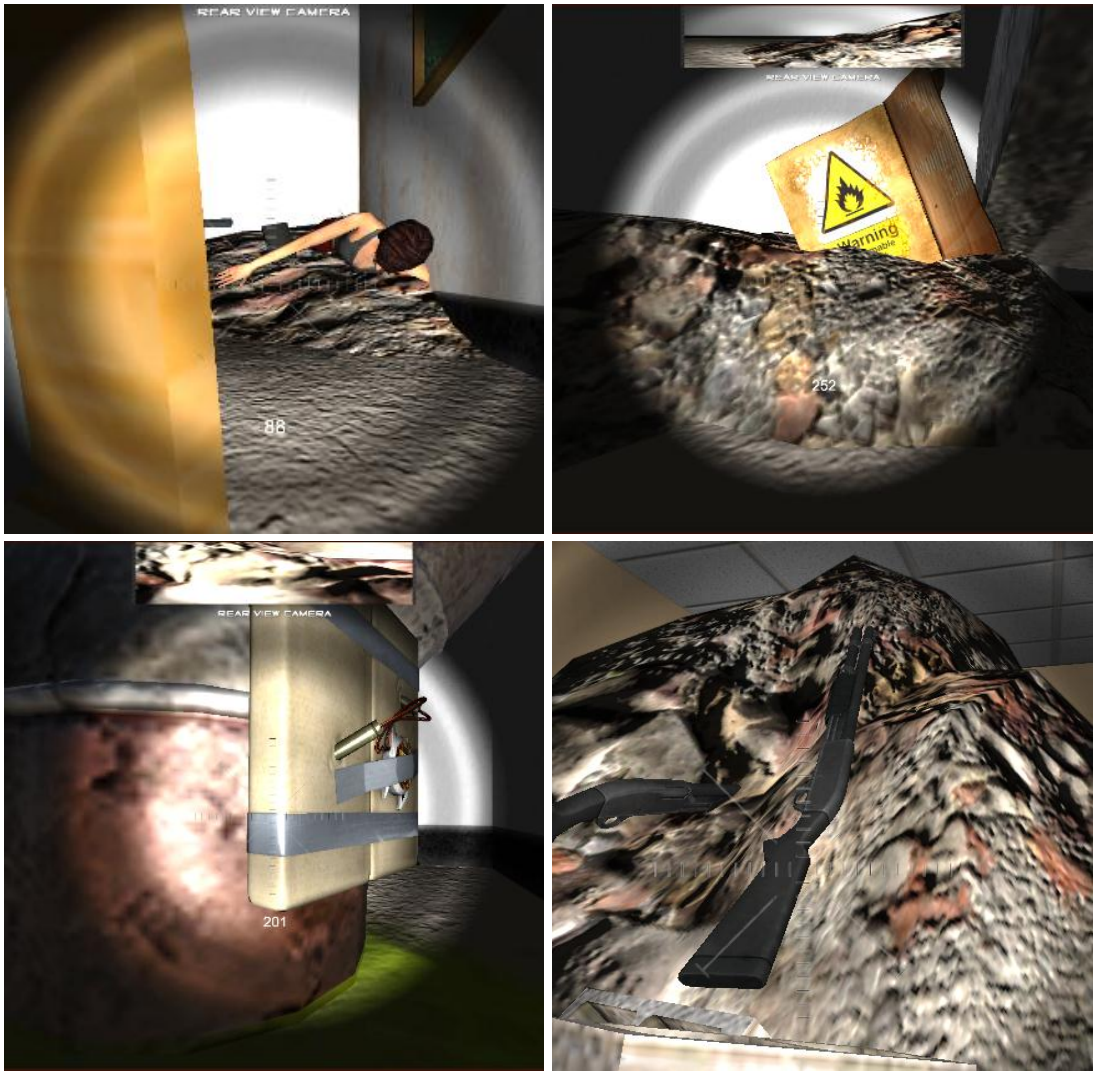


Figure 35 – Examples of targets in the environment. Top left to bottom right: Victim, hazard sign, bomb, weapons

bombs, or weapons. Victims were not injured or disfigured to not unnecessarily unsettle or upset participants. Generally, hazard signs consisted of different symbols such as biohazard, flammable materials, etc., and bombs looked like self-made plastic explosives. Targets were distributed equally in all environments to keep participant runs comparable. Figure 35 depicts the original interface view of participants, the environment was very dark and difficult to search.

Triggers – A collider can be configured as a trigger. If a collider is a trigger other objects can penetrate it and "trigger" another programmed event (script). In the rescue scenarios these triggers were temperature zones or CO_2 zones. For example, if the robot collides with the temperature collider, it can penetrate the object as if it was not there, but a trigger is activated. In the case of the temperature collider the trigger updates the temperature gauge to the set trigger vale (e.g. the temperature rises from 50% to 60%). So each collider, which is configured as a trigger, can activate any script, behaviour or actions needed. Figure 36 depicts such a collider in the editor view (green lines). The collider is not visible to the participants.

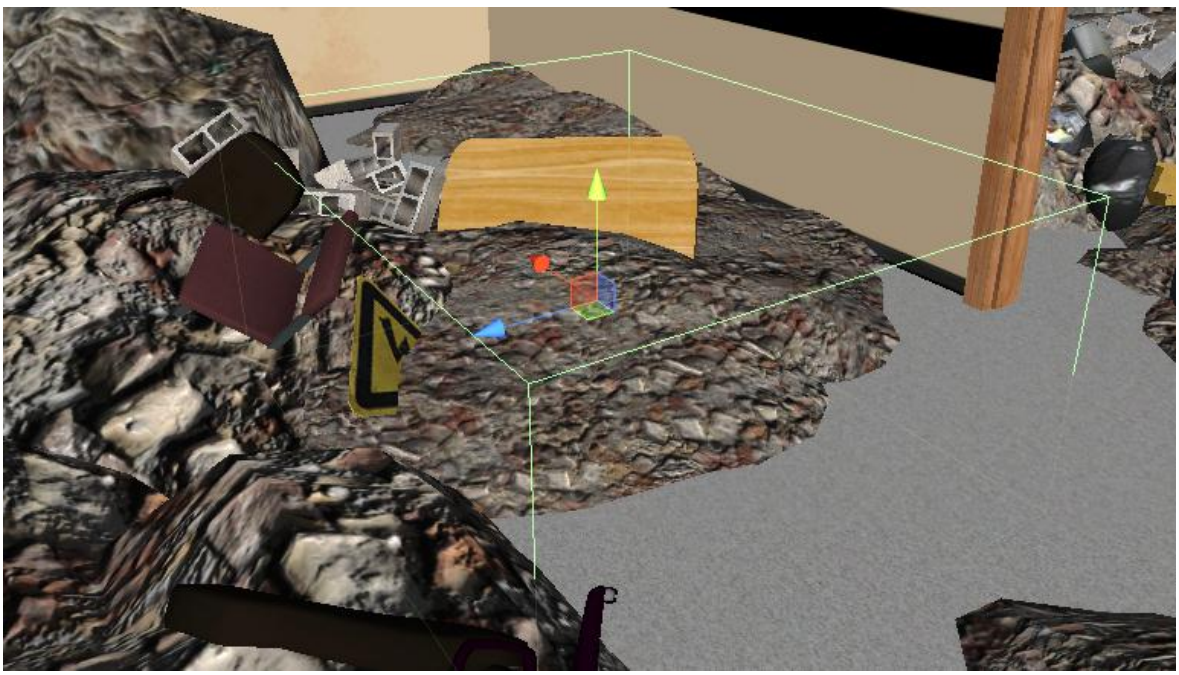


Figure 36 - Collider configured as a trigger. Collider is invisible to the participant.

An example script of a trigger is shown below (see Figure 37).

```
void OnTriggerEnter(Collider other)
{
    if (other.tag == "ROBOT")
    {
        Debug.Log("Emtering tempzone");
        labelTemp.text = "Temp.\n" + Temperature + "%";
        labelCO2.text = "CO2\n" + CO2 + "%";
        scrollBarTemp.value = 1-Mathf.InverseLerp(0, 100, Temperature);
        scrollBarCO2.value = 1- Mathf.InverseLerp(0, 100, CO2);
    }
}

void OnTriggerExit()
{
```

Figure 37 - Script of temperature trigger (collider)

The script shows, when the trigger (`collider`) is entered (`OnTriggerEnter`) and it is entered by the robot (if-statement) then the interface changes the text (`labelTemp.text`) and the scroll bar of the temperature gauge (`scrollBarTemp.value`) according to the provided value of the trigger (`Temperature`).

3.5.2.2.2 Robot

At this point it is essential to understand that the robot's behaviours were pre-programmed and entirely simulated. Therefore, the algorithms are developed for research and not for real world applications. The robot system's faults and successes were pre-programmed. Full control over the robot's behaviour was necessary to eliminate the interference of undesirable robot performance. The autonomous robot navigation could only be overwritten by the participant when using manual control.

Robot navigation

The robot's route was pre-planned by numbered waypoints that the system would work through. Waypoints are objects in the environment with a trigger (see Figure 38).

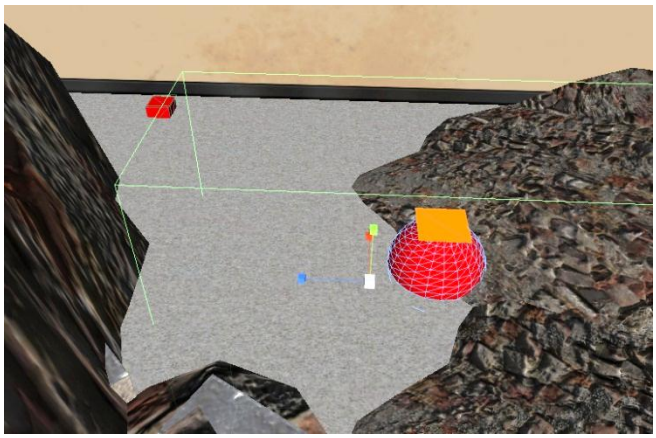


Figure 38 - Waypoint with trigger

If the robot triggered that waypoint, the waypoint was reached and the robot proceeded to the next waypoint on the list. Additionally each waypoint had a square red Look-at-box (see Figure 38, top left corner). This Look-at-box represented a LookAt-function. A LookAt-function made the robot, when reaching the waypoint, turn and focus with the camera on this square red box (not visible to the participant). By this procedure the robot gives the impression to look around in the environment.

The waypoints with colliders (green lines in the environment) and the list of waypoints (right in the inspector window) are shown in Figure 39.

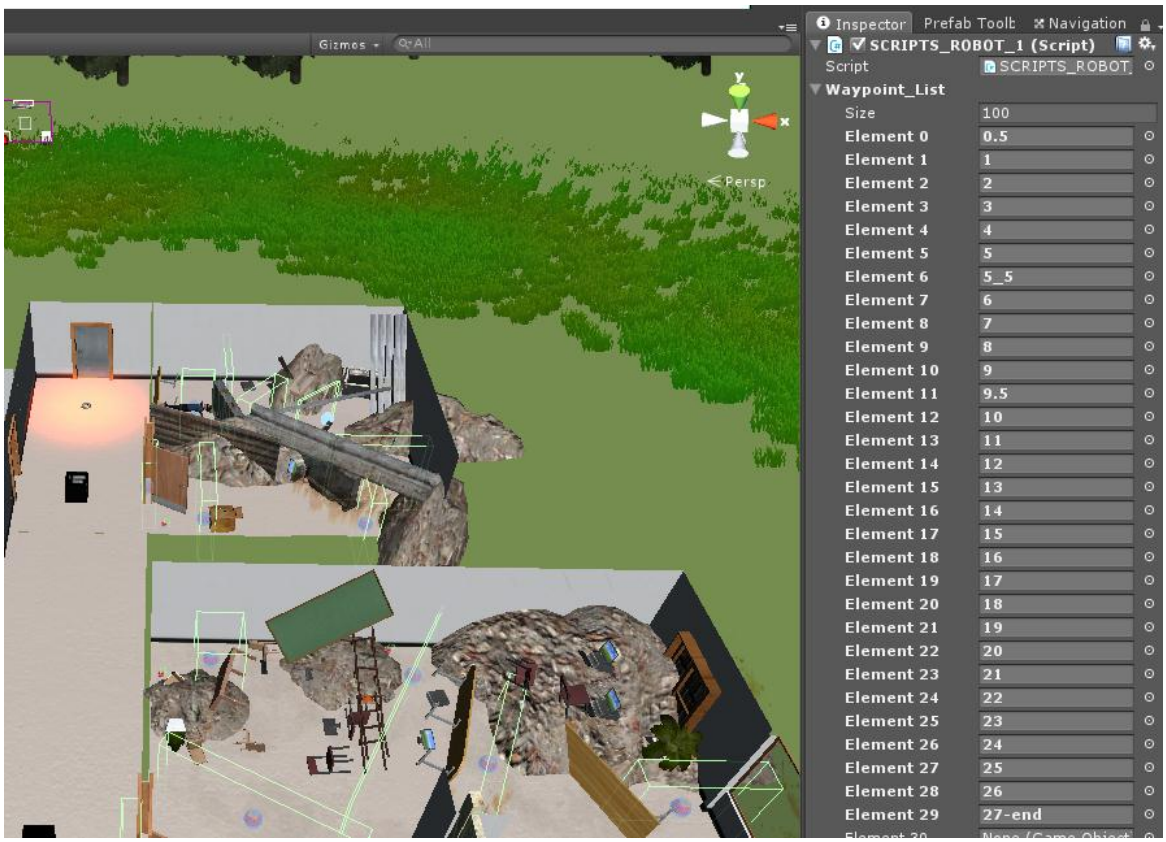


Figure 39 - Waypoints in the environment with waypoint list

The waypoint indicators and the layout of the rooms were designed so that the operator had to follow that pre-planned route. Therefore, when the robot was in manual mode and participants passed a collider of a waypoint, this waypoint got deactivated. If the participant decided to go back to auto mode the robot would drive to the next active waypoint, on the shortest way possible.

The script for a waypoint is depicted in Figure 40. If this function is triggered the current waypoint to reach (`Waypoints[currentTarget]`) is set as the next target. Then the Look-at-box is retrieved to get the information where the robot needs to look next (`lookTarget`). The script will check if the look-at-box is activated. If it is deactivated the robot will just drive to the waypoint but not turn to look at around (`if (lookTarget.activeInHierarchy)`). Additionally, this step is only executed when the robot is in auto mode (`if (AutoMode)`).

```
IEnumerator NextWaypoint()
{

        Debug.Log("NEXT WAYPOINT " + Waypoints[currentTarget]);
        GameObject target = (GameObject)Waypoints[currentTarget];

        CurrentTarget = target;
        GameObject lookTarget = target.transform.GetChild(0).gameObject;

        if (lookTarget.activeInHierarchy)
        {
            CurrentLookTarget = lookTarget;
        }
        else
        {
            CurrentLookTarget = null;
        }

        yield return new WaitForSeconds(1f);
        if (AutoMode)
        {
            agent.SetDestination(target.transform.position);
        }

}
```

Figure 40 - Script for reaching the next waypoint

Collision detection during auto mode was done by baking a navigation mesh onto the environment. This is a mesh (surface) where the robot is able to drive on by nestling around obstacles (see Figure 41).

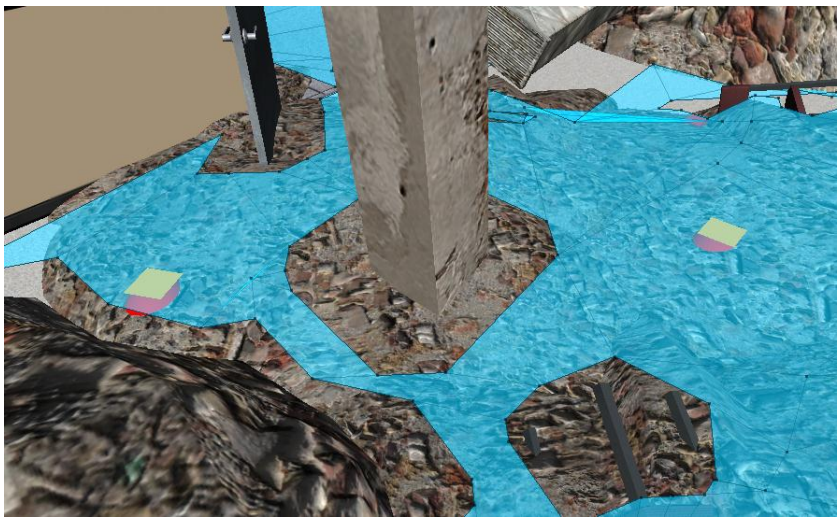


Figure 41 - Navigation mesh visualised in blue, overlaid with the environment

Robot target recognition

The robot was not able to detect or function like a real robot system and had no real target recognition. However, to give the participant the

impression that the robot can identify targets, the Look-at-boxes were used. A Look-at-box can have the characteristic to imitate a target. This procedure will be made clearer with an example: The target will be positioned as planned. Then a waypoint is positioned next to the target and the Look-At-box will be placed directly on the target and assigned the characteristic of being a target. If the robot, approaches the waypoint and starts too look at the Look-At-box a trigger activates the interface to display the message that a target was found. The script of the message is provided in Figure 42. How this message is shown on the screen of the participant is provided in Figure 44.

```
IEnumerator AlertVictim()
{
    if (AutoMode)
    {
        yield return new WaitForSeconds(3f);
    }
    Debug.Log("Displaying Alert Message");
    alertVictim.enabled = true;
    yield return new WaitForSeconds(2f);
    alertVictim.enabled = false;
    LogLabel.Add ("Victim found");
    Target_Value = Target_Value + 1f;
    TargetLabel.text = Target_Value + " ";
}
```

Figure 42 - Script for displaying a message

The coding for such an event works as followed: If the robot is in auto mode it will look at the victim and give the impression to recognise the victim by giving a short delay (`WaitForSeconds`). Then the message for finding a victim is enabled (`alertVictim.enabled`). The message will be displayed for two seconds (`WaitForSeconds`) and then be deactivated again.

3.5.2.2.3 Interface

The interface is the only element that the participant will be seeing from this simulation. For the subsequent studies different interface elements were utilised. In general the Interface is generated from a camera that is coupled with the robot's 3D model. Since it is dark and foggy in the environment the robot uses a torch like light source to illuminate the surroundings. The following elements could be used in the interface:

Figure 43 – Rescue robot interface with all elements visible

- Current task: Starting on the top left of Figure 43, the current task that is in progress is displayed. Therefore at the moment the robot is searching for targets.
- Rear view: The rear view camera gives a slightly squashed image of the environment behind the robot. This centred position on the top of the screen was used because most people are familiar with driving a car, where the rear mirror is similarly located.
- Status: In the top right corner the status of the robot is indicated with a green smiling face that the robot is working properly. If the robot malfunctioned, a sad face coloured in red and the type of error would be displayed.
- Battery: Percentage indicator of how many battery power is left.
- Signal: Percentage indicator of how strong the signal of the robot is at the moment.
- Top view map: This is a map that the robot creates with the help of proximity sensors. The sensors are positioned 360 degree around the

robot. If the proximity sensor hits an object, the edge of the object will be shown in the map. The waypoints are visualised as orange squares as well as the shortest distance between them by an orange line.

- Current reliability level: This indicator shows how confident the robot is about its own performance. Or in other words how reliable it performs. The indicator can show high, middle, and low reliability levels.
- Robot log: Different trigger boxes will write log entries into the box. Here the robot indicates what it is doing. For example it indicates that it enters a new room (e.g. "Enter room 1") or that it looked behind an object and it did not find anything (e.g. "Behind object = area clear").
- Mission info: This box counts how many targets have been found so far, an estimated number of targets in the environment, and how many rooms have to be searched. In addition it shows in seconds, how long the scenario is running for and how much battery time is left.
- System mode: Located in the bottom right the system mode indicates in which mode the robot is at the moment (auto or manual). The mode that is active appears read and the deactivated mode appears grey.
- Environmental indicators: These indicators show on a scroll bar how hot the environment is and how much CO_2 is present. To the participants was explained that these are percentage values. The temperature shows how much heat the robot can endure. Therefore a heat of 100% would damage the robot. The CO_2 shows in percentage the survivability of victims. A CO_2 level of 100% means, that there is too much CO_2 present that no victim could have survived this area.

During a scenario different messages from the robot can appear on the participant's screen.

Figure 44 - Message from the robot displayed on the interface

These were displayed when the robot found a certain type of target: "Victim/Weapon/Hazard found, Position marked". An example of such a message, on the screen of the participant, is illustrated in Figure 44. In addition to the message a sound is played.

The individual configurations of the environments, robot behaviours, and interfaces are explained in each study chapter that used the virtual rescue scenario approach.

3.6 Chapter summary

This chapter gave an overview of the measures of performance and trust. Furthermore, the methods used in this PhD were presented. For each measure and method a justification and detailed explanation was given. Each aim and objective was addressed by a certain study with a certain method in order to answer the overall research question. This chapter also showed how the virtual rescue scenario was developed and what functions and features it provides. All scenarios were designed and programmed in

Unity, a game development engine. The components of the program are explained in detail and examples of scripts are explained. The progress has shown that using Unity without having extensive knowledge in coding is possible. Unity proved to be a very useful and versatile research tool.

4 Study I - Urban Search and Rescue field work

4.1 Chapter overview

This ethnographic study aimed to gather information about USAR technicians, their training, tasks, working environment, currently used equipment, behaviour and culture. Over a period of two weeks eleven delegates of the USAR Level 1 technician course were observed in order to gather requirements and implications for robots in terms of features, behaviours, interface design, and robot implementation in the U.K. Fire and Rescue Service. Furthermore attitudes and traits of the technicians were collected. The background information was used to inform the subsequent studies. The chapter concludes with recommendations of robot usage in the USAR domain as well as a set of search and rescue scenarios.

4.2 Introduction

USAR is the abbreviation for Urban Search and Rescue. USAR teams are specialised incident/emergency response teams for rescue in urban areas. USAR teams have the expertise to localise casualties in collapsed structures, such as houses or tube tunnels, provide first aid and extricate casualties safely. Their equipment aids them in lifting, cutting and removing rubble as well as to shore up/support urban structures which are in danger of collapse. USAR teams consist of specially trained firefighters, who perform USAR additionally to their normal fire service duties. Overall, 20 fully trained USAR teams (each 30 firefighters) are strategically distributed across the UK. Performance standards demand that they have to respond within 45 minutes to emergencies (West Midlands Fire Station, 2013).

After the terrorist attacks on the World Trade Centre in New York of 11 September 2001, the U.K. government developed a new programme to enhance the Fire and Rescue Services' capacity to respond to terrorist and other large-scale incidents. The new programme was called "New

Dimension" and it aims to build resilience against catastrophic, chemical, biological, radiological, nuclear or conventional terrorists' incidents and make emergency response rapid, effective and flexible. For the United Kingdom Fire and Rescue Service (UKFRS) the programme includes revised command and control structures, new vehicles, equipment, and training necessary to respond appropriately to large-scale USAR incidents.

Particularly difficult in terms of USAR incidents are the unpredictable environments and the possible size of the incidents (Casper & Murphy, 2003; Y. Liu & Nejat, 2013). Incidents can have a variety of causes, which can be natural disasters, road accidents or terrorist attacks, involving unstable or collapsed urban structures. Flexibility and expanded skill-sets are necessary to cope with this uncertainty and the high induced stress levels (National Audit Office, 2008).

One of the core disciplines include collecting as much information as possible about the incident site through searches (visual and technical) and mapping of the area. In technical searches common used equipment are ultrasonic sensors and cameras. With advancing technology also robots with autonomous features could be used for USAR missions. The main advantage of USAR robots is that they could be sent into highly dangerous areas, while rescuers can stay in a safe place. For example, reconnaissance robots could explore inaccessible terrain, voids and instable structures to map the environment (e.g. with 3D scanning technology), locate victims and provide information for more accurate rescue plans and whilst rescuers can keep a safe distance. However, these systems are not used often due to cost issues, no standardisation and low trust levels between operators and robots. Despite the potential benefits of making rescuers' work safer, the author is only aware of one known USAR Team in the world who uses a rescue robot: New Jersey Task Force 1, a USA state team (Murphy, 2014, p. 53).

From the literature review trust emerged as an important factor in human-robot interaction. Trust is necessary to use the full potential of rescue robots and presents a challenge for design and implementation (Groom, Takayama, Ochi, & Nass, 2009; Sanders, Oleson, Billings, Chen, & Hancock,

2011). In particular, this study aims to develop the author's own knowledge about the nature of work and workers in USAR and ensure that experimental stimuli and tasks that are developed have ecological validity and are representative of the real world. Furthermore, the study aims to provide a basis for guidance for design of robots that emerges from this thesis and identify a potential user group of robots within the UKFRS because it is still unclear where such robot technology can be implemented.

In the later sections the author will occasionally refer to herself with "I" and "me" in order to emphasise on the autoethnographic nature of this study. Most sections are accompanied by personal and participant's quotes to bring across feelings, stress and complex constructs of rescue work.

The observation study took place at the Fire Service College (FSC) in Moreton-in-Marsh. The Fire Service College provides leadership, management and advanced operational training courses for senior fire officers from the United Kingdom and other foreign fire authorities. The FSC offers different courses regarding USAR, varying from USAR initial tool skills to Technician level 5 timber shoring course. I attended the two-week USAR technician 2 course, which aims to further develop USAR knowledge, skills and understanding of operations across the range of core disciplines including breaching and breaking, lifting and moving, shoring and technical search. This course was chosen due to the technical aspect of rescue and because the technicians are a potential user group of robots.

The report will give an overview of the working environment, tasks and tools rescuers are using and about different aspects that shape their behaviour. It will conclude with implications for USAR robots.

"It was just an amazing experience to see how USAR technicians actually work. Every researcher has an idea of what their target group is doing and how they behave, but actually "doing" the whole training is a very different experience with all the ups and downs, with all the dirt and dust." – Katharina Gabrecht

4.3 Methodology

4.3.1 Participants

The eleven delegates from the USAR Technician 2 course 2014 had an average age of 41 years (SD=5.99). All delegates were from the same Fire Station and knew each other. None of the firefighters had experience with robots in the Search and Rescue (SAR) or Urban Search and Rescue (USAR) environment.

4.3.2 Materials

The researcher needed a journal and a video camera to record the experiences at the USAR course. Furthermore, printed questionnaires were used to capture the attitudes of the rescuers towards robots (see Appendix A). Study information and consent form can be found in Appendix K - - Digital Appendix I (p. 404). The author also needed the complete kit of personal protective equipment, which is explained in Section 4.4.1.1.

4.3.3 Experimental design

In order to gather background knowledge an autoethnographic approach was used. Autoethnography comes from "auto" (self), "ethno" (culture) and "graphy" (writing) (Munro, 2011) and is a qualitative method that combines autobiography and ethnography (Ellis et al., 2011), therefore a combination of personal experience and observation. With the background knowledge gathered, future experiments can be adequately informed and designed to reproduce a real-world like scenario. The researcher wrote a details journal to collect the experiences and note observations.

In order to collect attitudes towards robots and further data of the delegates, in order to inform the aims and objectives of this study, the firefighters were asked to complete a consent form and a general questionnaire. Additionally the general questionnaire comprised a Negative Attitude toward Robot Scale. The "Negative Attitude Toward Robots Scale" (NARS) has been applied in the area of autonomous and telepresence robots (Tsui et al., 2010). In this study the scale aimed to provide a baseline of the general attitude of participants towards robots in the SAR context.

Occasionally, the researcher undertook informal interviews. Quotes are provided throughout this chapter.

4.3.4 Procedure

On the first day of the course the researcher explained the aims and objectives of the study. Then participants completed the consent form and the questionnaires. Over a period of two weeks the researcher accompanied the eleven firefighters and documented the experiences in a journal and with a video camera.

4.4 Results

The results section is divided into collected background knowledge, emerged factors of USAR work and the results from the questionnaires. The collected background knowledge was compiled from the observations/experiences of the researcher. The emerged factors of USAR work present in detail parts of the journal (indicated with *personal note*) the author wrote to collect the study data.

4.4.1 Collected background knowledge

During the course the researcher collected relevant knowledge about USAR work, management, tasks, and equipment. In addition, literature and course materials are cited to complement the background knowledge.

4.4.1.1 **Working environment**

Since USAR teams consist of Firefighters, they risk their lives to help others (Cowman et al., 2004). The environments they are working in are hostile and dangerous. In addition, major incidents and disasters are unpredictable (Y. Liu & Nejat, 2013). For instance, heavy transport incidents, confined space rescues (e.g. mine accidents), collapsed buildings of any kind, floods or terrorist bombings can happen without a warning.

Personal note: *Environments can include fire, water, dust, hazardous materials or explosive devices. The USAR technicians need to be highly flexible to respond to all sorts of events. I was quite overwhelmed by all the different things you have to look for. The search environment is mostly unstructured, cluttered and very complex. Different training scenarios*

showed that rescue approaches are very different, because of the different complexities of the task. For example, searching an intact area of the disaster site [disaster training ground at the Fire Service College], where less rubble and obstacles are present, is much easier compared to a collapsed room, which can only be searched with cameras through little entry points. I was totally helpless when trying to map a room [at the training site] through little holes in the rubble. I needed to concentrate very hard.

In order to be protected by the environment rescuers need to wear personal protective equipment (PPE), these include:

- Helmet
- Hearing protection
- Full eye protection
- Work gloves
- Dust mask (half/full) or breathing operator
- Knee and elbow pads
- Ankle supporting steel load safety boots

During an operation a variety of tasks need to be performed, ranging from search management to lifting and moving rubble. These tasks depend on the nature of the incident, resources and personnel available.

4.4.1.2 **Management, Tasks and Tools**

Personal note: *The main tasks of the rescuers were to localise, give aid and extricate casualties from the incident site. In order to accomplish these tasks a variety of skills were required. It included search management, technical and dog searches, shoring, and lifting, moving, breaking, breaching of rubble or other materials.*

For the purpose of an overview these skills are briefly described in this section.

4.4.1.2.1 Search management

A search is organised by the six stages of rescue, called REPEAT (The Fire Service College, 2014). The acronym stands for:

1. **R**econnaissance and Survey
   Gathering of as much information as possible, which include numbers of persons missing, possible locations of casualties, mapping of the area, existing hazards, structural assessment, cause of collapse and resource management.
2. **E**limination of Utilities
   The environment needs to be safe for rescue operations, therefore utilities such as water, gas, electricity or oil needs to be isolated.
3. **P**rimary Surface Search and Rescue
   This is an initial search to check how safe the area is and for saving all visible casualties that are lightly trapped. Additionally, equipment for longer searches is prepared and rescuers hail and listen for further casualties.
4. **E**xploration of Voids and Spaces
   Dependent on equipment voids under a rubble pile or other difficult reachable areas will be explored. This equipment can be listening devices, cameras or even robots/drones (see 4.4.1.2.2 Technical search). The goal is to locate casualties, possible survivable areas and their entry points.
5. **A**ccess by Selected Debris Removal
   At this stage the exploration goes even further with more detailed structural assessment, shoring and air quality test.
6. **T**ermination by General Debris Removal
   The last stage consists of using heavy lifting equipment for removing debris to recover remaining casualties. Occasionally reassessment of structures and shoring will be still necessary.

Casualties' chance of survival is mainly dependent on time. Statistically 91% survive the first 30 minutes, 81% the first day, 37% the second day and only 7% after 5 days (The Fire Service College, 2014).

Personal note: *The environment and the tasks are changing over and over again. It feels like an iterative process. There are very complex and difficult searches that require special equipment and skill. There are also tasks that are very easy but still require vigilance. Search complexity changes all the time and can be very demanding.*

4.4.1.2.2 Technical search

Technical equipment for victim localisation can be visual search devices (e.g. cameras), vibration detection equipment (e.g. seismic detectors) or scent detectors (e.g. search dogs or oxygen detectors). However, the primary search devices are the rescuer's ears and eyes.

"Always use your ears and eyes. Your eyes can see better than the camera, your ears can hear better than the microphone." said the instructor several times.

Important for visual search devices is that they are organised and systematic, because the rescuer is only able to see a part of the void at a time. Common used equipment in the U.K. are gas monitors, cameras, and life detection systems.

SnakeEye

Figure 45 - SnakeEye monitor (left) and goose neck extension (right)

SnakeEye is a remote visual inspection system and is also used in technical inspections (e.g. turbines or pipes). The camera (see Figure 45) can be mounted on a wand with a swivel head or onto a goose neck to be able to look into more difficult reachable areas. The system can take pictures, short videos and can be used under water.

Search Cam

Figure 46 - Search Cam 3000

The search cam is especially build for search and rescue operations. The camera has a variety of extensions for variable length of the pole and an articulated head to look to both sides. Also video, voice and pictures can be recorded. Figure 46 depicts the basic configuration of the search camera.

Personal note: *It is very useful to take pictures of casualties or items that might look like improvised explosive devices (IEDs) to they can be passed on for further professional inspection.*

DELSAR – Life detector

Figure 47 - DELSAR Life Detector LD3 with sensor channel monitor (left) and a sensor on a rubble pile (right)

The DELSAR is an acoustic and seismic search system. Sensors are deployed over the incident site (see Figure 47) to can pick up seismic (sound that travels through solid materials) and acoustic sounds (sound that travels through air) and the monitor will display these vibration or noise in a visual bar graph for each deployed sensor in order to locate the casualty.

Personal note: *By occasional shouting the casualties were instructed by rescuers to make noises. This could be any kind of noises such as scratching on a surface, banging on concrete, or screaming.*

Gas monitor

Figure 48 - Gas monitor Impact Series from Honeywell

This portable device, as shown in Figure 48, monitors the atmosphere for different gases. It can be used to check the atmosphere of a void or other spaces, which can indicate if there are hazardous levels of certain gases or if the atmosphere is survivable. It is standard to use this device before breaching into a void.

No robots are currently deployed in any USAR teams across England, Wales, Scotland or Northern Ireland.

4.4.1.2.3 Dog searches

Personal note: *Dogs are used for casualty detection and localisation. They are deployed mainly in Stage 4 (Exploration of Voids and Spaces) of rescue operations but they can be used for hasty searches in Stage 1 and 3 as well as repeated searches in Stage 5.* (Stages of rescue see 4.4.1.2.1, p.98).

Figure 49 - Handler with search dog during training

Across England and Wales there are a minimum of 20 USAR canine teams. A USAR canine search team consists of a certified dog and its handler (see Figure 49). The handler knows how to deploy the dog, is aware of its capabilities, and can read the dog's body language (Nuttall, 2008).

Dogs are able to locate, with their superior sense of smell, humans over the air transported scent. This fact includes a variety of constraints. Scent particles are affected by wind, temperature and humidity. Figure 50 illustrates how far and dispersed scent of a live casualty can be.

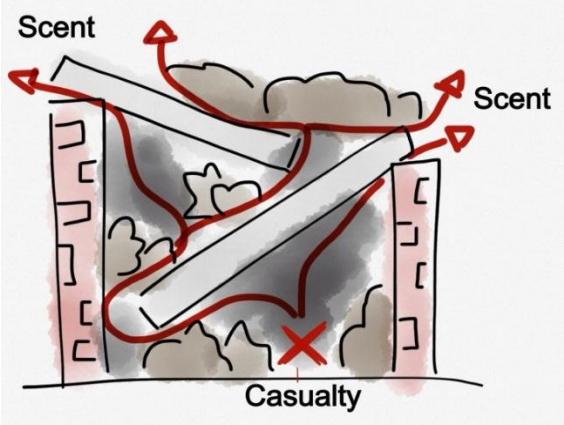


Figure 50 - Casualty under a collapsed structure and scent movements

Personal note: *Dogs are trained to find the strongest location of the recognised scent and to bark at that location. That will not indicate the actual location of the casualty, but a qualified handler is able to give indications for casualty locations. During training the dog handler demonstrated a dog search and instructed the rescuer how to behave if a search dog is present.*

Advantages of search dogs are that they are able to detect deeply buried casualties or even unconscious casualties. Dogs are lighter than humans and have more contact points (four legs) and therefore using dogs can be quicker and safer for searching an incident site (Nuttall, 2008).

4.4.1.2.4 Shoring

Personal note: *When rescuers are operating at the incident site and the structure or structures around them are unsafe, rescuers need to support these structures. This temporary support is called shoring. The aim of*

shoring is to protect the rescuers and casualties from further collapse and provide rescuers secure access to trapped casualties. For these operations USAR technicians need to be able to distinguish different types of loads and constructions for selecting the appropriate shoring technique.

Figure 51 - Two post vertical shores

The materials used for shoring are timber and Paratech. Paratech are heavy duty struts which are specially designed for rescue operations (The Fire Service College, 2014). Paratech consists of a strut which is height adjustable, further extensions and different types of base plates. An example of two post vertical shores is shown in Figure 51. The shore is constructed with Paratech and timber. There are a variety of different types of shores, each suitable for a certain type of structural problem.

4.4.1.2.5 Lifting and moving

Personal note: *To gain access to casualties, huge masses of rubble, floors, walls or other obstructing elements need to be moved carefully. Depending on the weight of the load, different equipment has to be used. For lighter loads pure physical strength, crow bars or rope hauling systems are appropriate.*

Figure 52 - Casualty extrication with crow bars and wedges (left) and Paratech tripod for lifting (right)

Figure 52 depicts the extrication of a casualty from a rubble pile with the aid of leavers and wedges. If loads are heavier the use of tripods (see Figure 52) or bipods is necessary. During the process of lifting the load needs to be secured at all times to avoid backwards movements or crashing. This is done with wedges, staked timber or other support structures.

4.4.1.2.6 Breaking and Breaching

Sometimes it can be required to breach through floors or walls made of different materials when a casualty is located behind them. Tools for removing sections of concrete can be hydraulic breakers, chipping/rotary/demolishing hammers or concrete chain saws (see Figure 53).

Figure 53 - Hydraulic concrete chain saw

Personal note: *Before breaching the rescuers need to drill a hole through the concrete and make an oxygen test as well as confirming the position of the casualty (e.g. with eyes/search camera). The casualty might be lying*

too close to the area of breaching. Therefore, dependent on the position of the casualty different kinds of breaches are used (e.g. dirty or clean breach).

The breach will be mostly a triangle of the size to fit the casualty through (see Figure 54). Stick out rebar has to be removed or bend away. A tarp makes ingress and egress easier and safer.

Figure 54 - Concrete breach

The previous mentioned tasks and tools for USAR teams are an overview, a variety of many other methods, tools and processes are available and needs to be used depending on the nature of the incident, because the working environments of USAR teams are highly unpredictable.

4.4.1.3 Work organisation: Organisational structure

The local authority Fire and Rescue Services are responsible for USAR in England. Training and equipment is mostly provided by the government within the New Dimension programme (National Audit Office, 2008).

When it comes to a major incident or disaster the Gold-Silver-Bronze command structure is used. Gold level is the overall strategic command which is not present at the incident site. The Silver level compromises the tactical implementation and Bronze is the operational level. In general the Gold Commander gives the strategic input, which the Silver Commander puts into steps of actions which are executed by the Bronze Commander (HM Government, 2008).

First, an Incident Commander (Bronze) is selected, which is mostly the most senior officer present. If the incident requires multi-agency the Incident Commander starts operating at Silver level. If the incident is even larger, a Gold level needs to be established. When specialist equipment is required, such as USAR or pumps for flood incidents, the Silver Commander may have

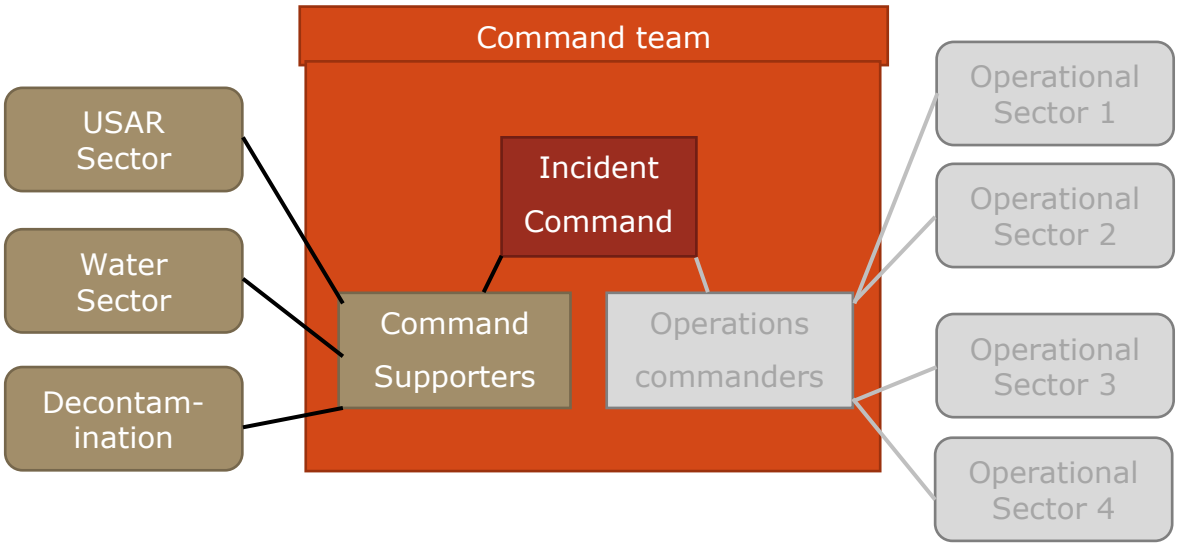


Figure 55 - Simplified incident organisational structure for Fire Service Operations

assistance from a specialist advisor. The specialist advisor is then in the command support and coordinates his/her particular field. Every USAR team will have a team leader who is in communication with the command support. If the co-ordination of more than one USAR team is required, a USAR sector commander is selected, who will coordinate actions with the USAR team leaders. The organisational structure changes with the size of the incident.

A very simplified model of this command structure is depicted by Figure 55. The command team consists of the Incident Commander, the Command Supporters (specialist advisors) and Operations Commanders. Operation commanders are coordinating the regular Fire and Rescue Units, but that area is not in the focus of this work. Each supporter will communicate with their area of expertise and each operation commander will communicate with their assigned operational sectors. If necessary, the strategic input for the Incident Commander will be provided by the Gold level.

Personal note: *During training the command structure was always clear and if in doubt, automatically, the most senior rescuer is in charge. These clear structures left no room for misunderstandings or discussions.*

However, each rescuer in charge was always open to suggestions and acted in the interest of the team.

4.4.2 Emerged factors of USAR work

During the USAR course the researcher wrote a journal. From this journal a variety of factors that are important in search and rescue emerged. These factors are outlined below and underpinned with personal notes from the journal entries.

4.4.2.1 Mental Fitness

Firefighters and therefore USAR personnel are confronted with very stressful, unpredictable, life threatening situations. During operations they have to do tasks under enormous mental and physical stress and make decisions in a very small time frame. They have to stay calm and confident to master all sorts of dangerous events. Further rescuer may encounter cruel scenes of dead or dying people which put huge emotional stresses on them. During the application process for being a firefighter the confidence and resilience of an applicant is tested to investigate if he or she is suitable for the job.

> *"We see a lot of stuff, especially hard it is, - when children are involved." The rescuer (delegate) stares thoughtful into the distance.*

Personal note: *In their everyday working life these people have to make very hard decisions which can depend on life or death. This was clear to me when it came to the final exercise, when a whole rescue mission was planned and performed: Five of the six rescuers and me were deep down in a void of a collapsed building (e.g. Figure 56), they just strapped a casualty on a stretcher, but the small tunnels in the void made it difficult to manoeuvre the casualty towards the entry point. All of a sudden there was an alarm sound which indicated the immediate danger of collapse and therefore indicated the rescuers to leave the void as soon as possible. A short debate began regarding whether*

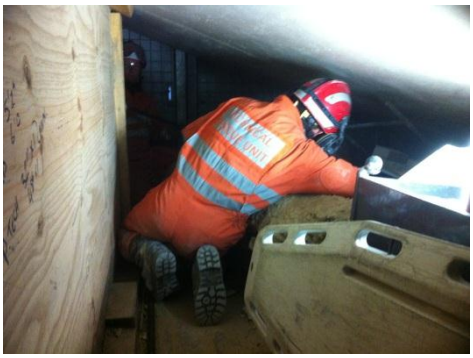


Figure 56 - A rescuer in a void

to take the casualty with us. Obviously it would have taken more time to do so. The group decided with a heavy heart to leave the casualty behind. The rule is: First my (rescuers) safety, than the team (rescue team), than the task (includes rescuing casualty). Even though this decision is debateable it was justified. After the exercise the sixth team member was beside himself because five rescuers were not able to rescue one casualty. The concerned rest of the team were obviously not happy with leaving the casualty but defended their decision. A short and slightly heated debate broke free. This was "just" an exercise but still motivation, eagerness, guild and frustration were present. The whole heart is involved in their work. They have to live with the decisions they have made. This emphasises on the importance of understanding the competing elements that influence rescue work, such as stress, physical strain, organisational and social aspects.

4.4.2.2 **Education level**

As mentioned before USAR technicians are regular firefighters with additional training for USAR missions. There is no formal qualifications required to become a firefighter. However, applicants need to pass a series of written and aptitude tests ("Fireservice Recruitment [Website]," 2014).

The delegates participating in the observed course were asked about their level of education/training. Five of the participants answered "GCSE or equivalent", three answered "A-levels or equivalent" and one answered "Degree or equivalent". One of the participants achieved none of the education levels (see Figure 57).

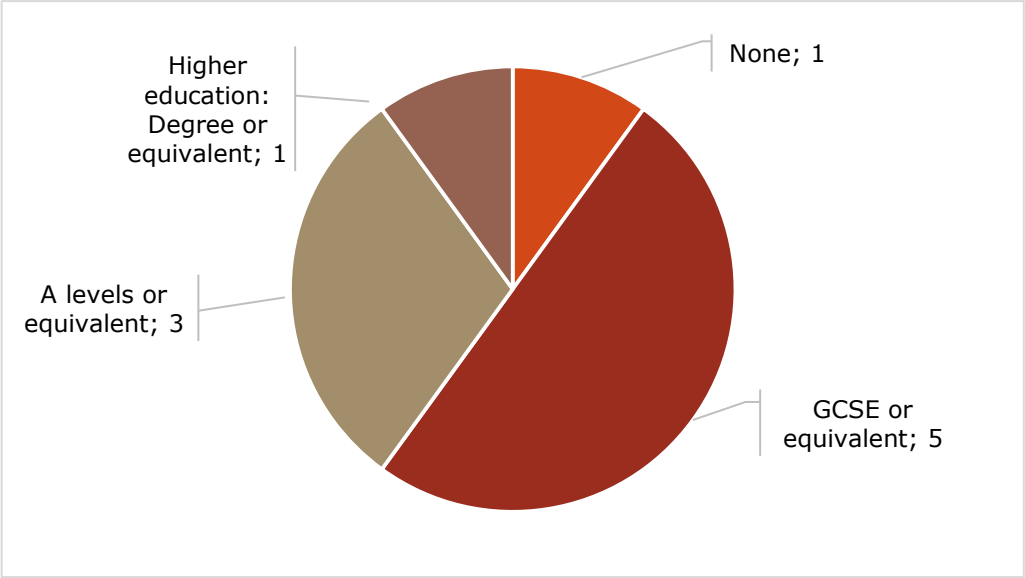


Figure 57 - Education level of USAR technicians

Four of the ten USAR technicians are performing a second job next to their fire and rescue duties. Four had an apprenticeship in another area before joining the rescue service.

Personal Note: *The USAR technicians are down to earth and focussed rescuers. They are highly qualified in the tasks necessary for rescue personnel, although some new technology (e.g. GPS devices) can be challenging to use for them.*

4.4.2.3 Physical Fitness

USAR rescuers are exposed to physically demanding situations, which may include operating heavy equipment (e.g. a hydraulic breaker which weights ca. 34 kg) or moving large pieces of rubble to free casualties and carry this person to a safe place. At the same time a rescuer needs to wear full personal protective equipment, such as a respirator, helmet, glasses, ear protection, safety shoes/boots and gloves. In addition, work might be performed in dark areas, confined spaces or at heights, as well as in atmospheres with low oxygen levels or hot and humid environments. Physical Fitness is an essential part of being a USAR rescuer, it is important to perform duties safely and efficiently and needs to be maintained. For this purpose fitness programs and equipment is provided by each fire brigade.

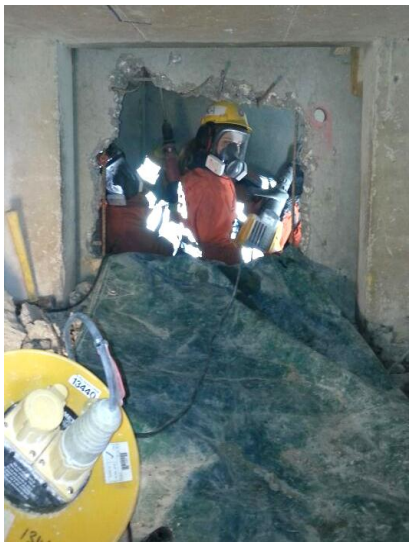


Figure 58 - Tool usage upside down in confined space

Personal note: *I was aware that certain tools need to be robust and needed to be deployed in really rough and difficult accessible environments. However, not just that these tools are very heavy (e.g. a hydraulic breaker which weights ca. 34 kg), they are also deployed in environments which are cramped and difficult to access. For example, it can be required to use tools upside down (e. g. breach concrete in a ceiling). Figure 58 shows two workers using a 10 kg demolition hammers overhead in confined space. In this position rescuers are not able to work for more than 20 minutes without being exhausted (I was exhausted after 5 min!). The officer in charge will enforce a strict working rotation. After a rescuer has been on a tool he/she has to have a break. These breaks are important to counteract exhaustion and potential errors/slips which could lead to injuries.*

Even after a hard day on the training ground, most of the delegates were going to the gym or to the swimming pool. It seems most of them really like to perform sports and fitness as an integral part of their everyday lives. These activities were also made in the group or in smaller sub-groups, which shows the existing team cohesion.

4.4.2.4 Teamwork

One requirement to become a firefighter is "Working with others". Working effectively with other is an important attribute in this field. Teamwork is necessary to cope with the everyday stresses and risks. It is also necessary

to carry out firefighter duties. One firefighter alone is not able to extinguish a fire. The whole team needs to work together, everyone in his/her role has to be able to accomplish firefighting and person rescue successfully. The same is valid for USAR members.

The delegates of the course very much depict the Personal Qualities and Attributes Framework of the UKFRS ("The Personal Qualities and Attributes [Website]," 2014) (excerpt):

- Commitment to Diversity and Integrity
- Openness to Change
- Confidence and Resilience
- Working with others
- Effective Communication
- Commitment to Development
- Problem Solving
- Situational Awareness
- Commitment to Excellence

Personal note: *My first appearance resulted in me looking at very muscular and mostly bald heads, which intimidated me a bit, since they all were male and I was the only female. I was sitting, a bit alien, in the back row of the class room and took notes of the presentations about building structures. In the coffee/tea break some of the delegates were very friendly and asked questions about my work and what I was aiming for. And very soon the first impression of a very cool and distant atmosphere turned into a very warm and friendly one. After asking them to provide some information about their attitude towards SAR robots and giving their consent for being photographed and filmed, we had our first outdoor exercise with a search dog to see how they work and how to behave in front of them. At that point I was not that alien anymore. I also got my PPE which integrated me even more.*

The next day we were doing our first practical training which involved shoring of a house entrance. I was still a bit shy and was more observing than laying hands on. However, I am a very practical person and have a variety of manual skills, so I was tempted to help with shoring. I received

sceptical looks and was uncertain what to do. But, after they realised I am not the stereotype of an untalented office sitter the relationship between me and the group changed rapidly and they accepted me as a full team member. They even provided me with a real USAR overall, so I could blend in very well. I felt very accepted. I think if they see your motivation and team spirit you are easily part of the team.

Especially the team spirit impressed me most. The entire group functions as a unit. Surprisingly there is no structure of power; also there is not a structure of the strongest. Everyone has certain strengths and weaknesses which are communicated and visible in the team. Through that process the team can dynamically use their members in terms of their individual strengths and weaknesses to perform tasks highly effective and efficient. Team members are not afraid to ask if something is in question or not afraid to ask for help. Knowledge is shared willingly, accurate and effective. I felt people were very pure and honest. They have to deal with saving lives, there is asking questions in training no shame at all. Important is the goal and the teams are immensely goal oriented. Of course conflicts are present; however they did not compromise the work itself. Conflicts were solved uncomplicated at an appropriate moment.

> *"We sometimes take the piss out of each other, this is how we are. But it's never serious." said one of the firefighters.*

Nevertheless, the atmosphere is rough and uncensored as the hostile working environment around them. Some communication seems harsh and unfriendly, but it is necessary to have a fast and clear communication in dangerous situations. Still, there was always a strong sense of cohesion within the team. Also after work, during the two weeks course, most of the activities were done together or in smaller parts of the group.

4.4.3 Questionnaires and robot attitude

The eleven delegates from the USAR Technician 2 course 2014 had a mean age of 41 years (SD=5.99). All delegates were from the same Fire Station and knew each other. They were asked how frequently they are using computers. Two answered "1-2 times a week", the rest "everyday". None

of the firefighter had experience with robots in the SAR or USAR environment.

Additionally the general questionnaire compromised a Negative Attitude toward Robot Scale (NARS). Delegates were asked to answer the questions with regards to SAR robots.

4.4.3.1 Negative Attitude Toward Robots Scale

In this study the scale aims to provide a baseline of the general attitude of the participant towards robots in general. The entire questionnaire can be found in Appendix A (p. 362). The scale is divided into three subsets which ask about different aspects:

- Subset 1:
  Negative Attitudes toward Situations and Interactions with Robots
- Subset 2:
  Negative Attitudes toward Social Influence of Robots,
- Subset 3:
  Negative Attitudes toward Emotions in Interaction with Robots

The subsets of the questionnaire were compared to each other to identify if a certain subset is of more concern than the others. This can help to determine which area of attitude delegates are more or less inclined to have negative emotions to. This subset analysis was also used by Bartneck and colleagues (2006), who identified that, for example, female participants had significantly higher positive attitudes towards the social influence of robots than their male counterparts.

The Shapiro-Wilk test showed that not all of the recorded data sets met the assumptions of normality, therefore non-parametric tests were used. For a better comparison of the subsets, which have different item counts, the scores are reported in percentages.

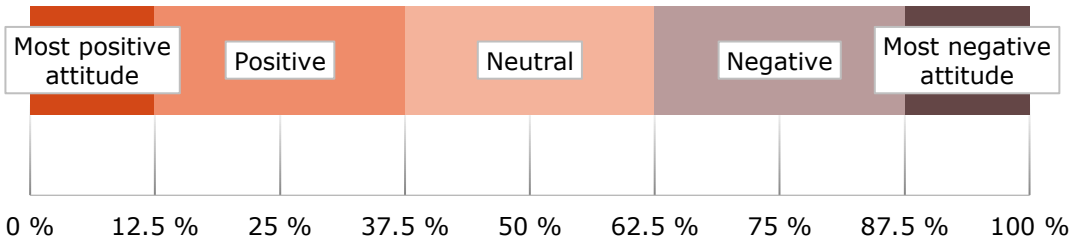


Figure 59 – Percentage indication of NARS questionnaire percentage scores

The following scale will aid to understand the relative scores of the NARS (see Figure 59). The lower the negative attitude score, the more positive participants about robots.

Multiple comparison with Wilcoxon signed-rank tests and Bonferroni correction (see Figure 60) showed that there is a significant difference between Subset 1 (Mdn=38%) and Subset 2 (Mdn=46%) (Z=-2.675 p<.016, r=-.57). However, there was no difference between Subset 1 (Mdn=38%) and Subset 3 (Mdn=42%) (Z=-2.201, p=.028, r=-.47) and no difference between Subset 2 (Mdn=46%) and Subset 3 (Mdn=42%) (Z=-.582, p=.560).

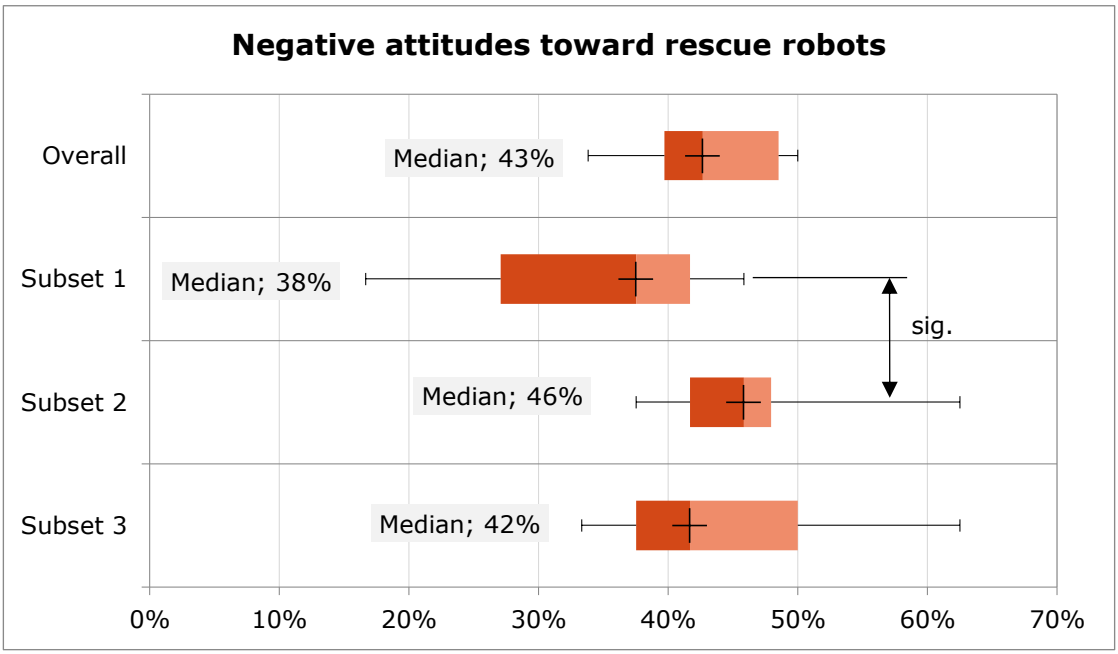


Figure 60 - NARS relative median scores for each subset and the overall score

In general participants were more positive towards situations and interactions with robots (Mdn=38%) compared to social influence of robots (Mdn=46%) and emotions in interactions with robots (Mdn=42%). The social influence of robots was the subset with the most negative responses (Mdn=46%). Nevertheless, in general the attitude towards robots was neutral (Mdn=43%) with a tendency towards a positive attitude.

4.5 Implications for USAR robots

USAR robot is a wider term for several types of robots: bomb-disposal robots, heavy lifting robots, or unmanned reconnaissance drones. This study focussed on unmanned reconnaissance ground and air vehicles. These are robots with the aim of exploring inaccessible areas for providing more accurate information and locating casualties.

The implications mentioned in this section are derived from the personal experience of the author and organised into organisational factors, tasks/functions of the robot, and the robot interface. Organisational factors are implications for when and where robots should be used. Task and functions are capabilities which are required from a robot according to the tasks of USAR technicians. The interface section will illustrate some factors which might be important when rescue technicians are interacting with robots.

Organisational factors: Robot operators in USAR teams

So far the UKFRS have never obtained or used a USAR robot. Therefore no knowledge base or implementation plans exist. Generally, a full USAR team has 33 operational members. Each team member is a trained USAR technician with one or more additional skill sets, which could be for example Advanced Shoring Specialist, Trench Rescue, dog handler or Technical Search Management. This is very important in order to be able to respond flexibly enough to any kind of incident. One of these skills could include operating a USAR robot. It can be one of the special skill sets and not every member needs to be trained in robot handling. Since the USAR team members are normal firefighters and the USAR work is additional to their normal duties it is important that the training required to use the robot is

very low. Establishing a new robot operator position outside the team is not recommended due to an additional source of information for the sector commander and additional management of people directly operating at the incident site, as well as higher personnel costs. For a fast deployment it is proposed that the robot is stored in the modules provided for USAR operations (Module 1 First Strike alongside technical search equipment). That implies that the robot is packable and able to be stored easily and safe.

Bearing the search management REPEAT in mind (cf. 4.4.1.2.1 Search management, p.98), these robots might already be used in the phase of Reconnaissance and Survey (Stage 1), when solely information is gathered. This might not be possible for all robots, since they would need to be intrinsically safe. The main deployment stage would be for Exploration of Voids and Spaces (Stage 4). Robots can already start with searching during the Primary Surface Search and Rescue stage (Stage 3), because they are able to take more risks than humans or dogs. However, the deployment of a robot is very much dependent on its capabilities and the equipped sensors.

Furthermore, deployment time is a very critical factor.

Personal note: *The task was to breach through a concrete wall in confined space. In order to determine the position and type of breach, the space behind the wall needs to be searched for casualties. In order to do that a small spy hole on the top will be drilled into the concrete. Using one of search cams takes a bit of time, because the system needed to boot up. This could take up to 2-3 minutes. In training the rescuers didn't want to wait that long and they got out their phone, pushed it through the small hole and made a picture of the void behind the concrete. This was a much faster way of localising a casualty.*

Therefore it is important that a robot is very fast and easy to deploy.

Tasks/Functions of the robot:

The main tasks of the operator and robot are exploring the extremely cluttered incident site and search for victims. The more equipment/features the robot has the bigger its physical body. It is a challenging task to find trade-offs between functionality and physical dimensions.

From the observations in this study it is clear that one of the most important feature is the camera. However, there are more features that might support the work of rescuers.

Necessary features:

- Camera: With photo and video function
- Microphone: Direction of detected sound

Additional useful features:

- Air quality sensors
- Infra-red sensors
- 3D scanning/automatic mapping of the environment
- Two way Communication with casualty (e.g. Survivor buddy in Murphy et al., 2011)
- Automatic navigation over irregular terrain: Path planning and collision avoidance
- IED + casualty identification

Not all features are always usable, for example if a collapsed building is very hot due to fires an infra-red camera will not detect warm bodies, because the heat of the building will overlay that signal. This suggests to use a customisable robot where sensors can be added, substituted, or removed. This list emerged from features of existing equipment and due to the observation of training scenarios as well as hands-on training. The list is not complete and just provides main headings that should be covered in future developments.

Interface

Since the rescue environment can be wet, hot, cold, or dusty the interface must be easy to read and controls need to be able to be operated with PPE (e.g. gloves, respirator, etc.).

Since the education level of firefighters is mainly GCSE level or equivalent the interface should not provide too much complexity. Because of the cluttered environment and the high visual demands the display needs to be ordered and not too overloaded. Key information should always be visible.

Most of all the picture needs to be clear and as big as possible, because it needs to be considered that rescuers may be deprived from sleep and experience constant stress during deployment.

Providing the state of the robot is necessary to identify problems and current reliability/needed attention level. Furthermore mapping and orientation is vital for gathering accurate information. ***Personal note:*** *One of the training tasks is to map a void through small spy holes. By using the search cam operators lost orientation and had to restart the search action again. Another team even mapped an area as one room, although in reality it consisted of two rooms.*

Other remarks:

The robot size must be appropriate to fit through the breaking and breaching zones and shoring constructs the technicians make. These dimensions usually are as big as the casualty to extricate. However, smaller robots can be useful to feed through drilled spy holes in order to gather information as early as possible, and not waiting for the breach through the concrete.

Since the environment is unpredictable, the robot needs to be safe against water, heat, and dust. Moreover the robot should be able to drive over uneven ground with big and small rubble pieces. An advantage would be if the robot is packable/wearable, because sometimes rescuers have to bring the robot to its deployment area which is already deep in the "hot zone".

The meta-analysis of factors affecting trust by Hancock et al. (2011) suggested is the main influencing factor on trust is the robot's performance. This means that everything that influences the perception of performance is related to the level of trust operators will have towards the robot. Therefore, the software needs to be as robust and stable as possible. This does not mean that the robot needs to perfectly perform to be trustworthy. More importantly, the robot's performance needs to be constant, then the operator will "trust" the robot to make certain mistakes (Freedy & de Visser, 2007).

Furthermore, task complexity did vary greatly during training missions and influenced the choice of equipment, number of technical equipment operators and it had influence on the stress of the operators and me.

4.6 Discussion

The researcher used an autoethnographic approach and was able to produce valuable background knowledge by participating in the USAR technician course at the Fire Service College in Moreton-in-Marsh. Previous literature did not report such insight into tasks, processes, and equipment of USAR units in the U.K. The data and knowledge gathered can inform the design of the subsequent research of this thesis. This study collected and documented the main tasks and attitudes towards robots of USAR personnel and derived implications for rescue robots.

Main tasks of USAR technicians:

- Identify hazards/failing structures/collapse patterns/loads in the urban search & rescue environment.
- Apply the "6 Stages of Rescue".
- Assess the incident ground.
- Mapping and planning of operations.
- Shoring of unstable structures.
- Breaking and breaching methods.
- Lifting and moving methods.
- Effectively select and operate all USAR technical search equipment.

Attitudes and traits of USAR technicians:

- The Overall NARS scores indicated that USAR technicians are neutral towards robots. Rescuers were more positive towards situations and interactions with robots (Mdn=38%) compared to social influence of robots (Mdn=46%) and emotions in interactions with robots (Mdn=50%).

Implications for robots:

Organisational key points:

- Usage of robots right from the first phase of search and rescue as well as in the following stages.
- Robot operator skills part of USAR technician skill set.
- Implement at standard USAR operational level; no extra personnel/support team.
- Training required must be a minimum.
- Robot should be part of USAR Module 1 (First Strike), where also the technical search equipment is stored.
- Fast and easy deployment are key.

Function/features/tasks:

- Necessary features:
    - Camera: With photo and video function
    - Microphone: Direction of detected sound
- Additional useful features:
    - Air quality sensors
    - Infra-red sensors
    - 3D scanning/automatic mapping of the environment
    - Two way Communication with casualty (e.g. Survivor buddy in Murphy et al., 2011)
    - Automatic navigation over irregular terrain: Path planning and collision avoidance
    - EOD + casualty identification

Interface:

- Robot controllable with PPE.
- Take into account sleep deprivation, exhaustion and constant stress.
- Uncluttered, clear and big display.
- Key information always visible.
- Provide robot status.
- Mapping and orientation aids.

Other remarks:

- The size of the robot needs to be small enough to be able to access voids and big enough to not fall between rubble.
- Resistance against water, heat, and dust.
- Packable/wearable for easy deployment at operational area.
- For trust consistent performance is more important than perfect performance.

Limitations and future work

However, due to the nature of the study the researcher was not able to constantly document the experiences. For instance, if she was searching for casualties deep in a rubble pile, there was no chance to write notes or analyse the situation in detail. The data gathered is therefore mostly from journal entries that have been written down hours after the experience. ***Personal note:*** *Sometimes I forgot that I am here as a researcher; I was stressed, sweating, physically exhausted and had only one goal: rescue that casualty!*

Future studies have to introduce U.K. USAR teams to real robots and let them use the systems during training and gather first impressions and feedback. ***Personal note:*** *They were very interested in my work and asked a lot of questions. They thought robots would be "cool" and they seemed very open minded.*

4.7 Conclusion

The researcher could gain very valuable background knowledge and insight into USAR work. This study documented the main tasks of USAR teams and showed that they had a neutral attitude towards robots. Similar to other research (Casper & Murphy, 2003; Murphy et al., 2015), this study emphasised on the fact that rescue robots need to be fast and easy to deploy. A variety of recommendations for necessary and optional hardware and features was provided.

The information collected can inform future robot and interface design. The following list shows aspects that were taken into consideration when designing the subsequent studies of this thesis:

- Incorporate sensors such as sound, air quality and casualty identification as well as integrate robot capabilities of path planning and autonomous driving.
    - The interface of the robot was equipped with temperature and air quality indicators. The robot was able to identify certain targets in the virtual environment and was capable of navigating through the environment. The participant had headphones and could hear 3D sounds of the virtual environment (robot motor, general humming noise, fire).
- Consider the influence of task complexity in subsequent studies.
    - With the aid of the collected experiences relevant task complexity elements were selected and manipulated for study III and study IV (please see Section 6.2).
- Give participants the context of their work and a scenario description to work with.
    - The scenarios for study II, III, and IV had be derived from the experiences the author made in this chapter (for example see Section 6.3.2.4).
- Participants need to receive an understandable task description that clearly defines what they have to do and what their responsibilities are.
    - The participants were responsible to find all targets in each of the experiments. They were also responsible to supervise the robot and take over control if they needed to correct the robot.
- Participants should be able to operate the system with a minimum amount of training.
- An element of stress or pressure should be introduced to simulate a more realistic rescue scenario. Rescue work is very complex and a secondary task could simulate the multi-tasking and stress factor.
    - A secondary task was introduced for study II, III, and IV. In addition, each rescue scenario needed to be completed as soon as possible. In study IV the battery time decreased on the display and participants were able to run out of time.
- The design of the virtual rescue scenarios (virtual environment) will be based on the experiences and documentation of this experiment.

4.8 Chapter summary

This chapter provided information about the tasks, working environments, currently used equipment and behaviour of rescuers in the role of an Urban Search and Rescue technician. In addition, recommendations about the implementation of robots in the U.K. Fire and Rescue Service, as well as recommendations about robot and interface design were provided. The collected information will inform subsequent studies.

5 Study II - The influence of robot reliability indication and feedback

5.1 Chapter overview

This chapter examines the influence of different amounts of robot feedback on trust, workload, performance, and participant's perception of the robot. Two robots, each providing different amounts of feedback, autonomously searching an environment for specific targets. Both indicate their reliability level, but one of the robots indicates why it is in a certain reliability level and what type of target it found. This explanatory feedback was perceived as a clearer type of communication and the robot was perceived as more competent, efficient and less malfunctioning. Furthermore, to collect qualitative data about human-robot interactions participants perform retrospective verbal protocols and answer interview questions after the trials.

5.2 Introduction

Urban Search and Rescue (USAR) is the search for and rescue of victims trapped in urban areas, such as collapsed buildings or other structures. USAR operations include finding victims, giving first aid and removing people from danger (see Chapter 4 for details).

With advancing technology, robots with autonomous features could be used for USAR missions. The main advantage of USAR robots is that they could be sent into highly dangerous areas, while rescuers can stay in a safe place (Virk et al., 2008). For example, reconnaissance robots could explore inaccessible terrain, voids and unstable structures to map the environment (e.g. with 3D scanning technology), locate victims and provide information for more accurate rescue plans (Murphy, 2014). Human-robot interaction is an important element and can foster clear communication and shared understanding between the operator and the human in order to ease the use of robots and enhance human-robot team performance (Green,

Billinghurst, Chen, & Chase, 2008; Jung & Lee, 2013; Murphy & Schreckenghost, 2013).

Desai et al. (2013) examined the impact of a robot's confidence feedback and its effects on trust and control allocation. The robot used was an UGV platform from iRobot with customised sensors for research purposes. Confidence level was indicated by a high, neutral, or low interface button. They found that the overall trust levels in the system were the same whether participants received confidence feedback or not. Similar results were found by Chien and Lewis (2012). However, Desai et al. (2013) found a positive influence of confidence feedback on control allocation (when to use manual or auto mode). Participants in the feedback condition switched away from the autonomous mode more often during low reliability. Interestingly, participants also switched away from the autonomous mode when the reliability dropped from high to neutral. Further, Desai et al. (2013) examined whether semantic or non-semantic feedback was appropriate and came to the conclusion that semantic indicators (smileys) of confidence level evoked more sudden control allocation changes. For a more steady trust level they suggested using non-semantic indicators (plus and minus symbols). In a later study Kaniarasu et al. (2013) showed that people generally over-trust automation when no confidence feedback is given.

Another study looked into backchanneling (feedback) of robots in a search and rescue scenario (Jung & Lee, 2013). A fully autonomous humanoid robot used backchanneling verbally (acknowledging and repeating command/request) and non-verbally (nodding, gaze towards speaker). Backchanneling reduced the perceived stress and cognitive load of participants in highly complex tasks, however, it led robots to be perceived as less competent (Jung & Lee, 2013).

The studies outlined above show that a robot's indication of reliability does not appear to negatively impact trust and has the potential to support the appropriate trust calibration. However, not many researchers have studied this topic in detail and further validation is necessary to develop a broader view on understanding the relationship and interaction between humans

and reconnaissance robots. This study investigates the attitudes and behaviour of people towards an autonomous rescue robot. The goal is to understand what shapes people's thoughts and feelings about robots and how to make robots more comprehensible, intuitive to use and predictable in order to establish appropriate levels of trust. In addition, the influence of different amounts of feedback from a robot on perceived trust, performance, workload, and robot characteristics was examined.

As mentioned previously, studies have shown that indication of reliability (e.g. an estimation of how well the robot performs at any given moment) can affect trust alignment (Kaniarasu et al., 2013) and control allocation strategy positively (Desai, 2012). Furthermore, recent literature investigated that trust is, among other variables, influenced by predictability and transparency of the robot (Hancock, Billings, Schaefer, et al., 2011). Robot transparency is a property of an interface to convey the intent, future plans, performance, and reasoning processes (Chen & Barnes, 2014). For example, the lack of background information leads operators to trust robots less and may lead them to use the autonomy inefficiently (Stubbs, Hinds, & Wettergreen, 2007).

This study examines if trust and attention allocation can be further enhanced (calibrated) when participants are provided with explanatory feedback (higher transparency) about the current reliability level and if this affects their workload and perception of the robot characteristics.

In order to investigate these circumstances further the following hypotheses were tested:

H1) The amount of explanation given by the robot will affect an operator's cognitive workload.
H2) The amount of explanation given by the robot will affect task performance.
H3) The amount of explanation given by the robot will affect an operator's perceived characteristics of the robot.
H4) The amount of explanation given by the robot will affect the trust an operator has in the robot.

H5) The indication of reliability will affect an operator's visual attention allocation.

It is also predicted that the workload and trust will increase with the amount of explanatory feedback and that the amount of feedback given will enhance task performance.

Furthermore the collected qualitative data was analysed to support understanding of trust and the influencing factors on trust.

5.3 Methodology

5.3.1 Participants

Participants were recruited using advertisements, emails and posters. 24 of 25 participants successfully finished the study. One participant experienced technical difficulties and was not able to complete the experiment; their data was excluded from this study. Participants were staff (n = 13) and students (n = 9) from different areas of the university as well as from the general public (n = 2). The participants' age ranged from 21 to 50 years with a mean age of 34 years (SD = 9.6). The sample population consisted of thirteen female and eleven male participants. Twenty-one participants were native English speakers. All participants used computers on an everyday basis. Nineteen participants reported playing computer games, app games or console games.

5.3.2 Materials

A maze and a remote controlled unmanned ground vehicle were used to simulate an Urban Search and Rescue (USAR) mission. The ground vehicle was a LEGO Mindstorms robot with a wireless camera which was controlled by a laptop with the aid of a LabVIEW interface. The participants could hear the robot and give voice comments (e.g. to indicate of a robot error) via a headset with a microphone. Due to technical problems the robot could not be controlled by the participant and was therefore completely autonomous. However, it was explained to participants that the voice comments (e.g. indicating a target that was missed by the robot) would be recorded and incorporated in the robot's reconnaissance data.

The autonomous robot was simulated by showing participants a pre-recorded video from the perspective of the robot. This also ensured that each participant saw the same scenes of the robot investigating a collapsed warehouse environment. Therefore, participants were presented with a video rather than a live camera picture from the robot. Each of the two trial videos (each with a different robot) comprised the same path length, number of turns and timing of targets that emerge. Each video lasted for seven minutes. To maintain the impression of a robot actually working next to the operator, the USAR maze was still intact and presented to the participants before starting the trials. The maze (Figure 61) consisted of walls that simulated rooms, rubble (e.g. piles of stones) and other obstacles (e.g. planks, styrofoam). The targets that were required to be found were human clothes, hazard signs and victims. There were three items of clothing and six to seven hazard signs and two victims hidden in the maze. For each trial the number of low reliability phases was the same, but not the location of these phases. A reliability drop consisted of the robot not looking into all corners and missing a target. All signs, clothes and victims were printed on paper, cut out and placed on obstacles or walls in the maze. Irrelevant objects for distraction were scattered throughout the maze.

An information sheet and consent form (Appendix K - Digital Appendix II, p. 404), a general questionnaire (Appendix B, p. 368), and a post-task questionnaires (Appendix C, p. 371) were completed by the participant. The post-task questionnaire asked for robot communication and perception ratings. Robot communication asked three questions about the amount of feedback provided by the robot and six questions about the robot's communication. On a 5-point scale the following six bipolar adjectives were shown: confusing/clear, inconsistent/consistent, hard to understand/easy to understand, unfriendly/friendly, unnatural/natural, and machinelike/humanlike. The items were borrowed from Bartneck, Kulić, Croft, and Zoghbi (2009). Perception ratings consisted of asking about perceived robot intelligence and competence as used by Jung and Lee (2013).

Participants were also given an instruction sheet for the task. A secondary task consisted of another screen showing a variety of blue boxes, where

participants had to count these boxes and click on the corresponding number on the keyboard. In addition, a camera recorded the participant during interaction with the robot.

5.3.3 Experimental design

5.3.3.1 Trial description

Each participant watched two pre-recorded videos of two different robots (within subject design) driving through a maze (see Figure 61).

Figure 61 - The maze in which the video was recorded

Each robot gave a different type of audio feedback. Participants were not aware that they were watching a video. The events of the two videos occurred at the same time, in a different order, and with slightly different timing. However, the first reliability drop was indicated at the same time, because previous research showed that timing of errors can significantly influence trust (Kaniarasu, Steinfeld, Desai, & Yanco, 2012). A reliability drop consisted of the robot not looking into all corners and missing a target.

Figure 62 – Example targets

Participants were required to find hazard signs, human clothes and victims in the maze (see Figure 62). In order to introduce uncertainty into the trial, the participants did not know which type or part of human clothes they needed to find, which aimed to make the trial more realistic (adding uncertainty).

The participants' main task was to indicate and explain an error of the robot into the microphone whenever the robot made a mistake, for example, "Error, this was not a hazard sign." At the same time they were asked to perform a secondary task, if they felt comfortable doing so. The difference between the type of feedback from the two robots participants had to use is explained in the next paragraph.

5.3.3.2 Robot audio feedback types

The following actions were performed by the robot:

- Navigating through the environment (video)
- Identifying possible targets
- Indicating low reliability phases, where the target identification system could be faulty
- Indicating system problems while moving

Two different audio feedback strategies were used to investigate the influence of robot feedback on trust. Robot 1 gave very basic feedback and was called "*Roy*". Robot 2, called "*Parker*", gave more detailed feedback, as shown in Table 3.

ROY – basic audio feedback	PARKER – detailed audio feedback
"Target identification low."/ "Target identification high."	"Target identification likelihood low, because of low lighting levels. "; "Target identification likelihood low, because of unreachable area."; "Target identification likelihood low, because of heat influencing the sensors." / "Target identification likelihood high."
"System got stuck – Recovery."	"Navigation stopped, because right tyre got stuck in debris. Recovering now."
"Target found."	"Hazardous sign found."; "Human evidence found."; "Victim found."

Table 3 - The differences in feedback given by the two robots.

The presentation order of the robots was counterbalanced to avoid order effects.

5.3.3.3 Secondary task

The secondary task was a non-loading task and participants could allocate attention to this task whenever they wanted to. In the secondary task, a certain number of boxes were shown on the screen and participants were required to click the corresponding number (count) on the keyboard. Participants were given brief feedback to indicate if they were right or wrong before the program switched automatically to the next screen with a different number of boxes.

5.3.3.4 Measures and performance

Performance was measured by correctly identified targets and correctly answered secondary tasks. The overall performance score was based on the following scoring system:

- Indicate a missed victim +30
- Indicate a missed target/false identified target +10
- Missed target/false identified target -5
- Secondary task/one correct answer +1
- Secondary task/one wrong answer -1

The scoring system was developed by keeping in mind the main focus of the task. It is most important not to miss human casualties (+30) and other targets (+10). For each correct secondary task answer they earned one point. Penalties were given for not indicating a missed target (-5) and for each wrong secondary task answer (-1). Participants were made aware of the ranks of the scoring system, but not the detailed scoring values.

Video observation of the participants were used in order to capture the visual attention towards the robot. Questionnaires and a semi-structured interview provided data about attitudes, workload (NASA TLX), perceived robot characteristics, and trust. Furthermore the subsequent retrospective verbal protocol was analysed with a theme based content analysis (Neale & Nichols, 2001).

5.3.3.5 Compensation

For introducing a risk factor and an incentive for better performance, compensation was given based on the overall performance. The maximum amount that the participants could earn was 30 GBP. Every participant had a basic compensation of 10 GBP. The overall best scoring participant received further 20 GBP and the second best, an additional 10 GBP.

5.3.4 Procedure

After giving informed consent, participants were asked to complete a general questionnaire which asked for demographics (age, gender, occupation, etc.), and their general trust attitude. Participants were informed about the task itself and the rescue scenario. Participants had a five minute training session supervising the robot while it was searching in the maze. They were asked to practise until they felt comfortable performing the task. Examples of each type of target were shown to them, so they could familiarise themselves with what they were looking for. In addition, participants received training how to perform a retrospective protocol (RVP). After training, they performed the two trials. Each supervision trial of the robot took seven minutes: the goal was to find and mark all relevant targets. After the trials, participants were presented with a post-task questionnaire which asked them to rate their workload and how they generally felt about the robot and how the robot communicated. Then

the video of the interaction was shown to the participants and they were asked to perform a RVP. After a short pause they had to perform another trial with the other robot, answer the post-task questionnaire, and perform another RVP. The study concluded with a short semi-structured interview, which included questions about the two robots and whether participants preferred the first or the second robot.

5.4 Results

The results section reports the findings from the general questionnaire which asked about general demographics and attitudes towards robots, the measures of performance, subjective workload and robot communication. Next, the visual attention allocation, followed by a correlation analyses are presented and the main body of qualitative data is presented in the retrospective verbal protocol and interview section.

5.4.1 General Questionnaire

The general questionnaire incorporated questions regarding participants' general information (age, gender, etc.), and general trust attitude. The following sections will describe the different question sets of the general questionnaire. The complete questionnaire can be found in Appendix B (p. 368).

The participants' age ranged from 21 to 50 years with a mean age of 34 years (SD = 9.6). The sample population consisted of thirteen female and eleven male participants. Twenty-one participants were native English speakers. All participants used computers on an everyday basis. Nineteen participants reported playing computer games, app games or console games.

5.4.2 Trial performance and robot perception

5.4.2.1 Performance

Performance was measures by the scoring system explained in 5.3.3.4, p.132. The relative performance was calculated by how many per cent the participants achieved of the maximum possible score. The performance data did not meet the assumptions of normality and therefore was analysed with

parametric tests with the aid of bootstrapping. Relative mean performance scores showed that there was no significant difference between the task performance of Parker (53.38%) and Roy (51.11 %), $t(21) = .496$, $p > .05$ (see Figure 63).

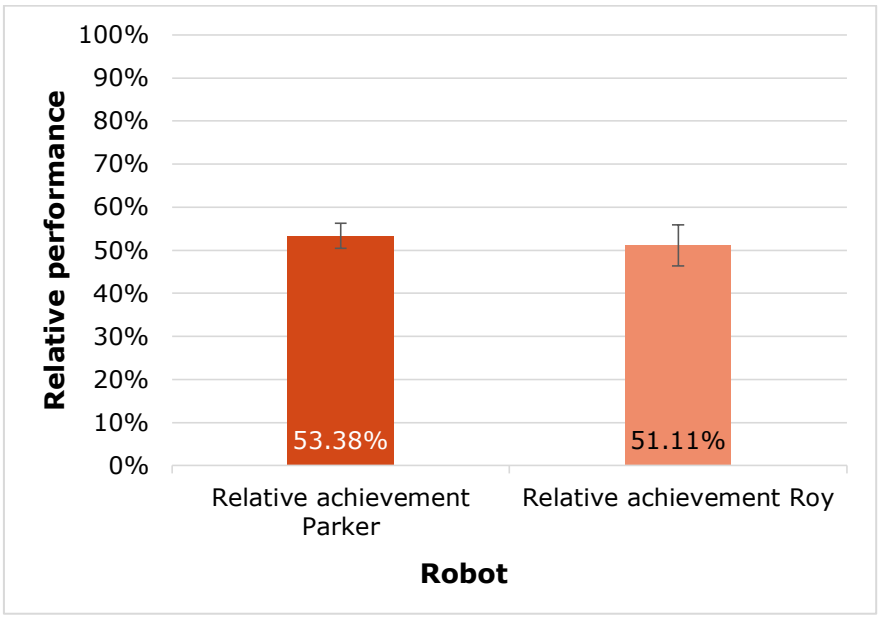


Figure 63 – Relative mean performance bewteen Parker and Roy. Error bars show 95% confidence intervals.

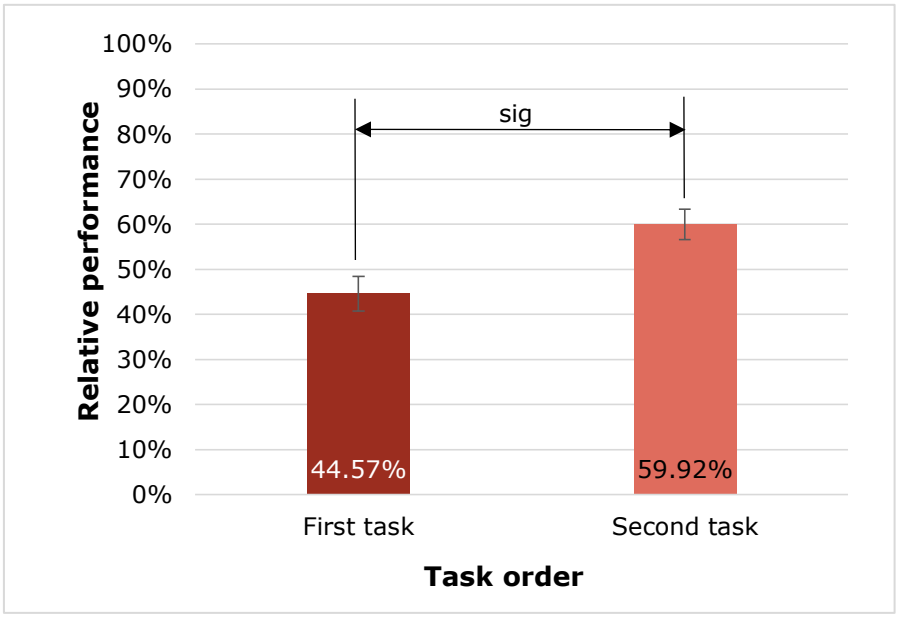


Figure 64 – Relative mean performance between first and second performed task. Error bars show 95% confidence intervals.

However, there was a significant large task learning effect, as the first trial (44.6%) was performed significantly less well than the second trial (59.9%), as shown in Figure 64, $t(21) = -4.18$, $p < .01$, $r = .51$ (paired samples t-test with bootstrap; 1000 samples). The starting task was counterbalanced across the experiment.

Further analysis with an ANOVA showed that there was an interaction effect between robot and task order ($F(1;40)=7.09$, $p < 0.5$, $r = .39$). This means that the effect of the task order on performance was different for Parker and Roy. Post hoc tests (Mann-Whitney U tests) revealed that there was no significant effect between the performances of Parker whether it was used first or second by the participant (see Figure 65). But there was a significant effect on the performance of Roy whether it was used first or second by participants ($U = 13.5$, $p < .01$, $r = 0.66$). It seemed that the performance of Parker remained similar between participants using it first or second in the study. If participants used Roy first their performance was particularly low (37%), but when they used Roy after they have used Parker, their performance was highest (65.18%). This could suggest that participants had a carryover effect from the behaviour of Parker to Roy. Because Parker was explaining the errors and giving additional feedback, with this knowledge in mind participants handled Roy differently and achieved higher performance levels.

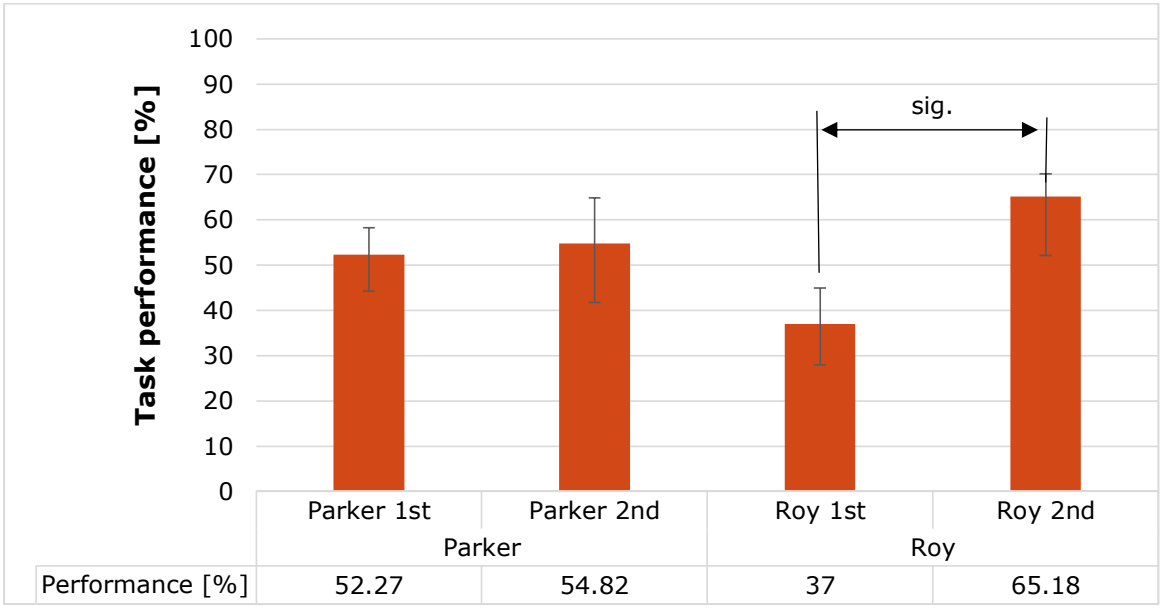


Figure 65 - Robot effect on task order performance with 95% confidence intervals

But it is inconclusive if the amount of information provided by the robot did or did not influence the performance scores.

5.4.2.2 NASA TLX Workload

After each trial participants were asked to complete a NASA-Task Load Index (NASA-TLX) questionnaire (Hart & Staveland, 1988). The NASA TLX is accompanied in Appendix C (p. 371). Wilcoxon signed-rank tests showed no significant differences in any of the items of the NASA TLX Task questionnaire, therefore there were no significant differences in perceived workload across the two different robots (see Table 4), $Z = -0.467$, $p > .05$ (overall workload score). Also, there were no significant changes in the sub scales of the NASA TLX.

Workload ratings across conditions	
Condition	**Mean (SD)**
Parker	55.75 (12.29)
Roy	57.75 (14.38)

Table 4 - Workload ratings across conditions

5.4.2.3 Robot ratings

Robot communication

In addition to the workload measure the participants had to answer twenty further questions regarding the communication and perception of the robot and how they perceived the robot (Appendix C, p. 371).

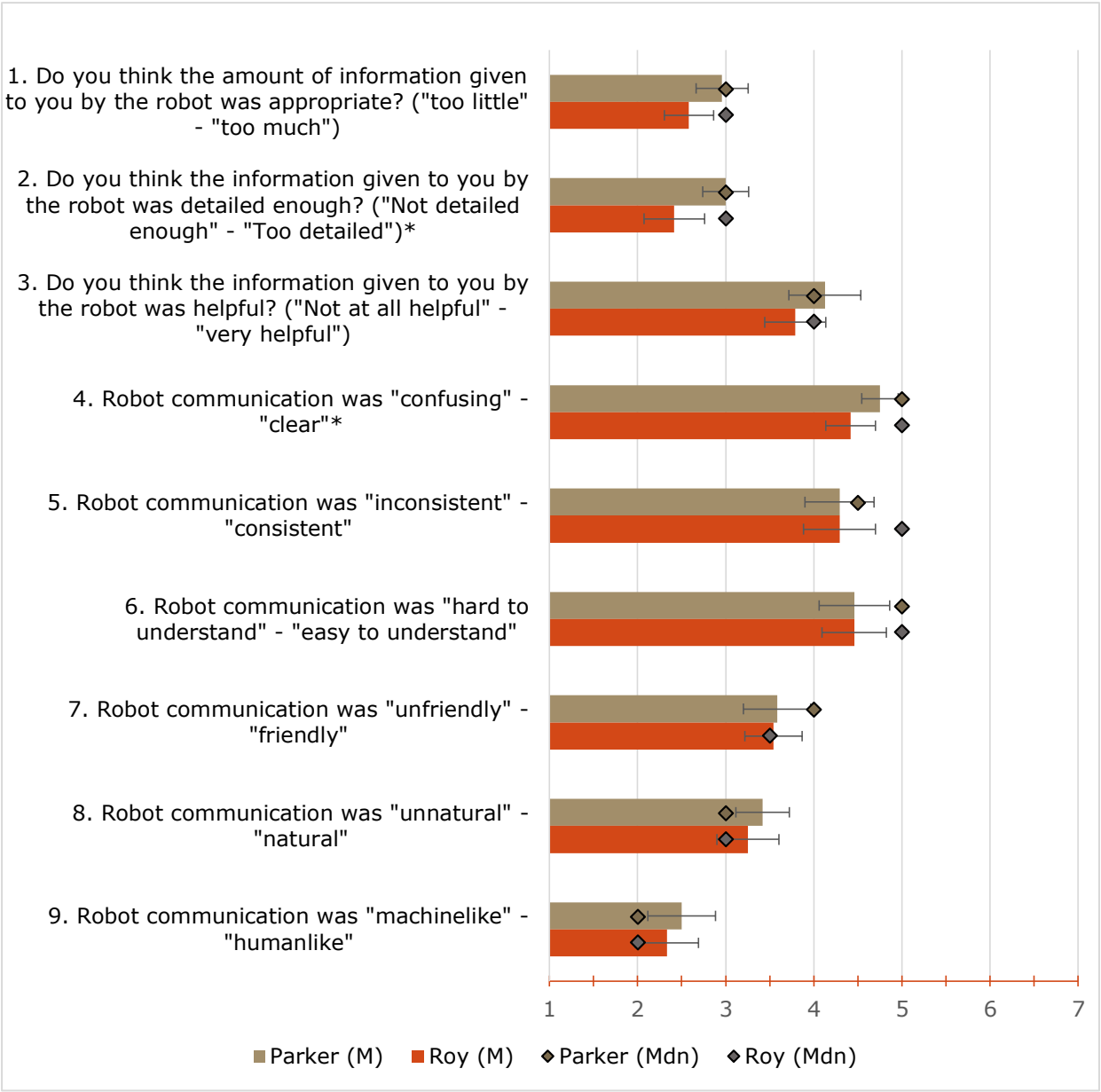


Figure 66 – Robot communication ratings with confidence intervals (*significant difference)

Figure 66 visualises the mean and median ratings for the robot communication ratings. The questions with significant differences are marked with a star (*). Wilcoxon signed-rank tests showed significantly different ratings for question 2 ($Z = -2.18$, $p < .05$, $r = .31$) and question 4 ($Z = -2.00$, $p < .05$, $r = .29$). The difference of question 2 had a medium effect size and question 4 had a small effect size (Cohen, 1988) Participants rated that the information given by Roy was not detailed enough, whereby the information given by Parker was just right (question 2). In terms of the

level of confusion the information given by the robots evoked, Parker was rated clearer than Roy (question 4).

There was a tendency that the amount of information given (see question 1) from Parker was just right and Roy gave "too little" information ($Z = -1.63$, $p > .05$), but the difference was not significant. Further, the information given by Parker was reported as more helpful (question 3) than the information from Roy ($Z = -1.46$, $p > .05$). However, these ratings were not statistically significant, either.

Both robots were rated as generally consistent (question 5), easy to understand (question 6), and friendly (question 7) in their communication. This was anticipated since both robots had the same simulated type of voice. With respect to their articulation Parker was rated slightly more natural ($Z = -1.03$, $p > .05$) and more humanlike ($Z = -1.19$, $p > .05$) than Roy (question 8 & 9). Nevertheless, these were not significant differences.

Robot perception

The ratings for the perception were on a scale from 1 = "strongly disagree" to 7 = "strongly agree" (see Figure 67 and Appendix C, p. 371). With the Wilcoxon signed-rank test significantly differences were found between the answers of question 10 ($Z = -2.07$, $p < .05$, $r = -.3$), question 12 ($Z = -2.50$, $p < .05$, $r = -.36$) and question 15 ($Z = -2.68$, $p < .05$, $r = -.39$). The questions with significant differences are marked with a star (*). All differences showed medium effect sizes. Hence, participants felt that they together with the robot accomplished the task more efficiently (question 10) with Parker then compared to Roy and Parker was more competent (question 12) and malfunctioned less (question 15). Answers to the malfunction question was generally negative, because at the end of both robot trials the robot got stuck on debris. Although, Parker was rated as more intelligent (question 11), more trustworthy (question 13), more dependable (question 18), more reliable (question 19) and participants were more confident in Parker than in Roy, these differences were not statistically significant. Participants also rated, but not significantly different, they would like to operate Parker more than Roy again (question 14). Further the participants were not particularly wary about either of the

robots (question 16) and felt competent operating either of them (question 20).

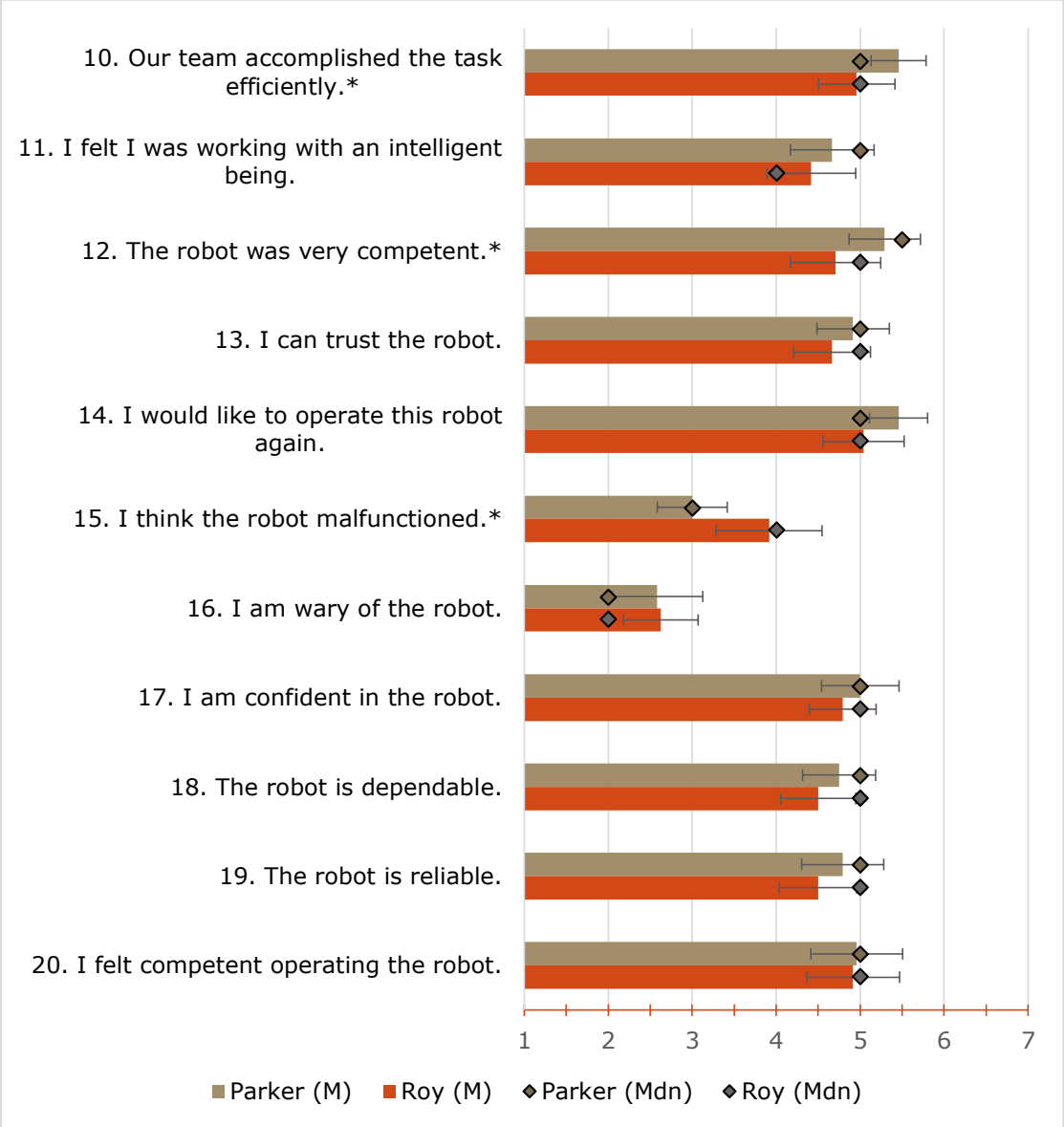


Figure 67 – Robot perception ratings on a scale from 1 = "strongly disagree" to 7 = "strongly agree" with confidence intervals (* significant difference)

Robot task contribution

Also, the question about the extent to what the robot contributed to the task (Figure 68) was higher with Parker than with Roy. Yet, the difference was approaching significance and had a small effect size ($Z = -1.92$, $p = .055$, $r = -.27$). In this particular question Participants could rate in intervals of 10% from 0% to 100%.

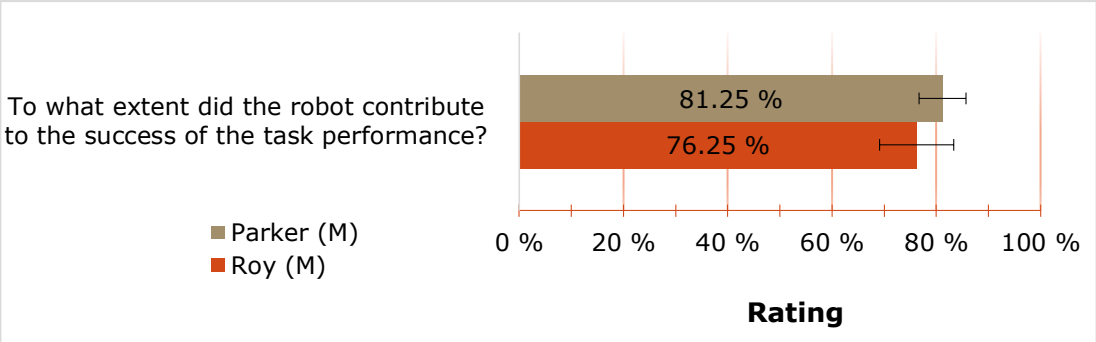


Figure 68 – Percentage rating of robot contribution to task success with confidence intervals

5.4.3 Visual attention allocation

In order to investigate if the robot's indication of reliability can support the visual attention allocation of the participants, the attention towards the robot during low and high reliability phases was examined. Furthermore, it is investigated if the explanatory feedback from Parker can improve the attention allocation. The attention towards the robot was analysed via the recorded video of the face of the participant. A frame-by-frame analyses counted the time (seconds) the participant looked at the robot screen and not glancing towards the secondary task screen. For the purpose of this analysis, a glance is defined as a maximum attention of 500 ms of a second not allocated to the robot and includes the transition times, because the eyes, during transition from one screen to the other, were off the robot.

Both robots had the same amount and duration of high and low reliability phases. An ANOVA revealed that there was a significant effect of robot reliability phase on relative visual attention allocation time, $F(1.\ 21) = 24.86$, $p < .001$, $r = .74$. However, there was no effect of the robot (Roy or Parker), $F(1.\ 21) = 0.96$, $p > .05$. Therefore, Parker (with explanatory feedback) could not improve this attention allocation.

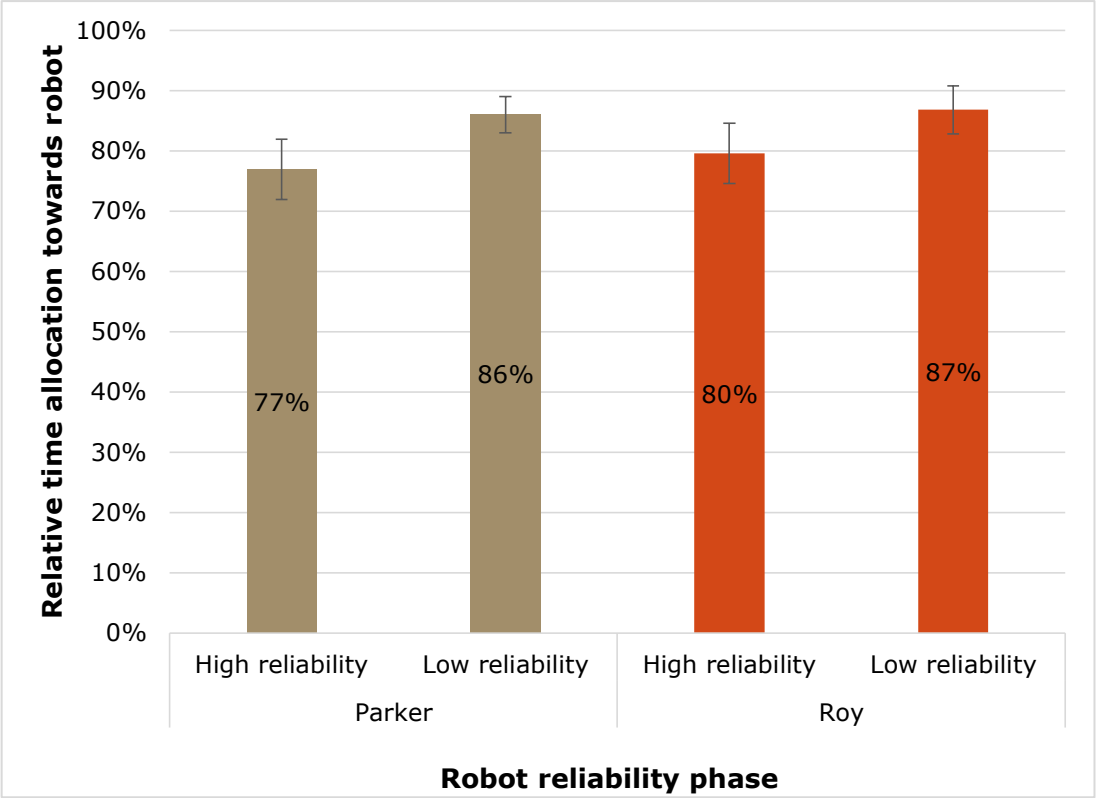


Figure 69 – Attention allocation towards the robot in low and high reliability phase with confidence intervals

Results of a post-hoc paired samples t-test showed that participants generally allocated significantly more time to the robots when in low reliability phases (see Figure 69). In high reliability phases participants supervised Parker on average 77% of the time rather than attending to the secondary task. In low reliability they allocated more time to the robot (M = 86%) compared to the low reliability conditions ($t(21) = -3.624$, $p < .05$, $r = .62$, with bootstrap; 1000 samples). For Roy on average 80% of the participants' time was allocated in high reliability phases. In low reliability phases, they allocated 87% towards Roy ($t(21) = -3.247$, $p < .05$, $r = .58$, with bootstrap; 1000 samples). Both differences were significant and had a large effect size.

5.4.4 Summary of quantitative results

Table 5 provides an overview of the results. Each dependent variable and their results are listed. The robot performance was kept constant.

Dependent variable	Independent variable	Significance	Result details
Trial performance	Explanatory feedback	no significance	Possible learning effect between first and second performed task.
Workload	Explanatory feedback	no significance	
Robot communication ratings	Explanatory feedback	significant	Explanatory feedback was perceived as a clearer type of communication.
Robot perception ratings	Explanatory feedback	significant	A robot with explanatory feedback was perceived as more competent, efficient and less malfunctioning.
Robot task contribution ratings	Explanatory feedback	no significance	
Visual attention allocation	Explanatory feedback / High and low reliability phases	significant	Independent of explanatory feedback, participants allocated significantly more time to the robot in indicated low reliability phases.

Table 5 - Summary of quantitative results

5.4.5 Retrospective verbal protocol analysis

The retrospective verbal protocol (RVP) was used for a qualitative analysis of events rather than for revealing cognitive models or thinking structures. For the latter the length of each trial (about seven minutes) was too long and the disruption of answering the questionnaires after each trial extended the time between doing the trial and the RVP further. The pilot study showed that a concurrent verbal protocol interfered too much with the main task and secondary task. For a practical analysis the RVPs were divided into the following events:

Events:

- Robot succeeded (the robot just identified a target correctly or ignored an incorrect target).

- Robot mistaken (the robot just identified a wrong target or missed a target.
- Reliability indication (the robot indicated a low or high reliability phase).

In-between events (see Appendix D, p. 374):

- Visual attention allocation (participants explained strategies or issues about their attention allocation).
- Robot/Interface characteristics (participants commented on matters concerning the robot or its interface).

Due to the length of the analysis the in-between events were moved to the Appendix (see Appendix D, p. 374).

To analyse each event the theme-based content analysis (TBCA) from Neale & Nichols (2001) was used. Originally this qualitative method was developed for evaluating virtual environments and desktop environments. The interaction in this study via a computer screen can be categorised as desktop environments. TBCA provides information about opinions and behaviours and is able to indicate important issues by meaningful grouping of data.

Under each sub-event the relevant categories/themes and conspicuous issues found in the retrospective verbal protocol are discussed. Additionally only selected citations will be reported, which might best represent the whole picture. An overview of the raw transcript and assigned themes is provided in Appendix K - - Digital Appendix III (p. 404).

It is recommended to refer to the tables (e.g. Table 6) at the beginning of each section in order to capture the quantity and content of comments. In these tables (e.g. Table 6) the **bold** items are explained in more detail (e.g. with quotes) below the table. Only the most often mentioned themes are discussed, due to the limited length of this chapter. Themes in the paragraphs are indicated in *italic*. Numbers provided in squared brackets represent the number of all comments in that theme. Brackets behind a quote indicate the following: participant number, name of robot, timestamp (optional). For example: (P17; Parker; 00:01). Participant is abbreviated

with "P" and the robot is abbreviated with "R". If the context within a quote needs to be explained the relevant information is inserted in squared brackets.

5.4.5.1 Event: Robot succeeded

The following events occurred before and during the time the robot succeeded in finding a target.

5.4.5.1.1 Sub-event: Participant waits for robot [31]

Event:	Robot succeeded	
Theme	**Sub-event: Participant waits for robot [31]**	
Sub-theme	**General [25]**	**Participant declared error [6]**
Raw data theme	• **General waiting [14]** • **P anticipated R's delay [3]** • P anticipated R better performance [2] • P anticipated R to fail [1] • P seeks explanation of delay [1] • P happy to win against R [1] • P is negative about delay [3]	• "I was quicker" [1] • Nearly declared error [4] • Not used to robot [1]

Table 6 - TBCA overview of sub-event: Participant waits for robot

A major issue was that the *participants had to wait for the robot* to identify a target (see Table 6). The theme was mentioned 31 times. Waiting for the robot costs a lot of allocation time and also annoyed participants. Overall, 25 times they mentioned that they *waited for the robot*.

Their *general waiting* comments [14] were:

- "Yes, I was just waiting, squinting again. Hazard sign." (P23; Roy)
- "I waiting for the robot to identify the shoe." (P02; Parker)
- "I am just waiting for it." (P14; Roy)
- "I was spotting it and I was waiting for the robot to identify it." (P17; Parker)

Later on, some already *anticipated the delay* until identifying a target [3]:

- "I was wondering if it gonna miss that. Yeah, I thought, it kind of went away, I thought, maybe it come back around, it usually turns and looks directly at the object, so just waited. And it obviously picked it up." (P15; Parker; 03:42)

- "Because from what happened previously, there seemed to be a slight delay between looking at it and then identifying it, I am not sure how many seconds it was. But it felt like it identified it [Target] within that frames so I didn't say anything." (P02; Roy; 06:00)

Because of the robot's delay to identify a target participants declared or nearly declared a robot error [6]:

- "So here that's the thing, so I saw it before it zoomed in, I saw it like you know as it was moving, the robot, so I thought: Okay I wasn't sure if it was gonna turn back there, so I said it but then it identified it right after I said it, so I was thinking okay: I was just a bit quicker." (P23; Parker)
- "[P declared error] and then it identified it [later]. It came on the picture first and then it panned around and then it came back. So again that's about getting to know the robot […]." (P14; Parker)

These comments implicate that there is a need to visualise the process of identification of the robot. This could incorporate a visual overlay and/or a loading bar as well as the information if the robot tries to get another angle/view upon the target. It also would be useful if operators (if they already identified the object) can abort the identification and declare that it is a target or not a target.

5.4.5.1.2 Sub-event: Robot better [20]

Event	Robot succeeded		
Theme	**Sub-event: Robot was better [20]**		
Sub-theme	**Positive [7]**	**Neutral [8]**	**Negative [5]**
Raw data theme	**(Good/impressed /proud) [7]**	**(P couldn't see/easy to miss/R vision better) [8]**	**(P poor time allocation/I failed/R better) [5]**

Table 7 - TBCA overview of sub-event: Robot better

If the robot performed better than the participant, for example the robot saw the target and the participant didn't, then the feedback was mostly *neutral* [8] or *positive* [7] (see Table 7). For example, *positive* quotes were:

- "Yeah he spotted that one, I was very proud of that one. Because normally I miss all the clothing." (P03; Parker)
- "Generally speaking the robot was one time he finds a shoe, and I was really impressed. Cause I haven't seen that at all." (P05; Roy)
- "[Robot found target] Oh, yeah, right that's true! Oh if it wasn't for the robot, I would have missed that. So that was a good moment, me and the robot." (P23; Roy)

However there were also *negative* comments [5] the participants made about themselves:

- "I think I look away here to do something (secondary task). And the robot sees the shoe. I did not see the shoe. Not until the robot marked it. I was distracted. Very poor time allocation." (P03; Roy)
- "Okay this is the shoe, I failed. And I just disregarded. And then I think he says, Target acquired." (P05; Roy)

If the robot's success is rated *positive* seems to vary among individuals. Participants mainly directed *negative* comments at themselves.

5.4.5.1.3 Sub-event: Found in low reliability [9]

Event	Robot succeeded
Theme	**Sub-event: Found in low reliability [9]**
Sub-theme	**General [9]**
Raw data theme	• **More relaxed [3]** • **Impressed (Still found it!) [3]** • More confident in R [1] • Perfectly happy [1] • more trust = more relaxed [1]

Table 8 - TBCA overview of sub-event: Found in low reliability

When the *robot found a target in a low-reliability phase* (see Table 8) people were *impressed* [3]:

- "And I believe it had two [targets found], oh okay, despite low reliability it is still finding them." (P20; Roy)

They stated to have *more trust and being even more relaxed* [3]:

- "Particularly having known it picked things up in low reliability you relax even more, you got more trust in it." (P17; Roy)

This underlines that a good performance under difficult circumstances is valued by the operator.

5.4.5.1.4 General feelings [23] and general feedback [25]

Event	Robot succeeded			
Theme	**General feelings [23]**	**General feedback [25]**		
Sub-theme	**General [23]**	Positive [19]	Neutral [6]	Negative [2]
Raw data theme	• **More confidence in R/R competent [7]** • R is reliable [3] • Happy [2] • P happy for R [1] • more trusting [2] • **P felt better (relaxed/comfortable) [3]** • Relief [1] • Useful [1] • P assumes R okay [1] • Leave it to it [1] • **P feels useless [1]**	• Good/well done/fine [14] • R works properly [2] • Persistent/thorough [2] • R good in negotiating [1]	• Okay [3] • fairly clear [1] • No need to highlight [1] • Correct [1]	• **Inaccurate [2]**

Table 9 - TBCA overview of sub-event: General feelings and general feedback

In terms of feelings, see Table 9, the robots success of finding a target evoked that participants were *more confident in the robot* [7]:

- "Maybe feel confident that it was working and doing his job." (P13; Roy)
- "This one it spotted. I was thinking at that point: Oh I think it spotted most of them so I was really quite confident with it." (P20; Roy)

And *participants felt better* [3]:

- "The more reliable I found it to be the more I kind of relaxed a bit. Knowing that it was doing a pretty good job." (P08; Roy)
- "[…] I was a bit more relaxed now, that it picked up a few things that I had seen, a bit more confident in it." (P17; Parker)

Yet, one participant claimed to *feel useless*: "I saw this one so that was ok. [Hesitation sound], I think the image was clearer at that point, now it was flashing again so I couldn't see anything. And I felt like: Oh, I am useless." (P25; Roy).

With respect to feedback the robot got *positive* comments (19 *positive* comments opposed to 2 *negative* comments). The only *negative* feedback about the robot happened when participants started with Parker and then used Roy. They saw Roy as *inaccurate* [2]:

- "So it helped that it spot something, but I identified which type of target it is." (P24; Roy)
- "[…] I was waiting for the next one, which was there (hazard sign) and it said: target found, again. And then I thought it is not, it doesn't give you as much information as the other one, but because it said target found but it could be anything." (P18; Roy)

As expected a robot's successful action gained positive comments from the participants and overall contributed to more trust in the robot (directly stated by 2 participants).

5.4.5.1.5 Secondary task [13]

Event	Robot succeeded
Theme	**Secondary task [13]**
Sub-theme	**General [13]**
Raw data theme	• **Carry on/start [9]** • **Do secondary more [2]** • **Do secondary until R finds something [1]** • Focus on secondary task [1]

Table 10 - TBCA overview of sub-event: Secondary task

All of the participants comments indicated that they were *doing the secondary task more* [2], *started it or carried on* [9] with it (see Table 10). Another interesting comment was, that:

- "Yeah [hesitation sound] on this one I saw that it's there [P found target], and I was like, kind of: Oh doing the secondary task, and if the robot doesn't say it I will say it." (P25; Roy) *Do secondary until R finds something*.

So, it happened that participants identified a target before the robot and they used the time until the robot said something to do the secondary task. This is also visible in the videos and supports the implication of being able to abort the robot's identification process and already mark the object in question as target/no target.

5.4.5.1.6 Participant uncertain/unsure [19]

Event	Robot succeeded	
Theme	**P uncertain/unsure [19]**	
Sub-theme	**unsure about robot [4]**	**unsure about target [15]**
Raw data theme	• **Unsure about R feedback [3]** • Unsure about R performance [1]	• **followed R decision [5]** • Closer look for target [2] • **Making sure [2]** • P wants manual control [1] • **Double check/checking sharply [3]** • Confirm with list [1] • **P called error [1]**

Table 11 TBCA overview of sub-event: Participant uncertain/unsure

There were situations where participants were uns*ure about the robot's feedback* and what that actually meant for them [3] (see Table 11). An overview of what the robot can say and what that explicitly means might be useful. When participants were *unsure about a target* they were trying to *make sure* [2] and *double-checked* [3]:

- "I think at the bottle I wanted to make sure, because there was a lot of light, glare on the skull, it was definitely a skull I remember that, and there was a lot of light glare so I just wanted to make sure that it was definitely a skull." (P16; Roy)
- "There has not being an error at this point. Although, you certainly checking very sharply for it." (P10; Parker)

Mostly, if still unsure, they *followed the robot's decision* [5]:

- "There I see a little dot and I don't know what it is, because the light is in the way. [Hesitation sound], but I trust the robot would have picked up on something. So I didn't say anything." (P04; Roy)

- "Here I was a bit curious, cause even I couldn't see, I didn't think I could see that, so I guess the robot was right, he was like low visibility for sure." (P16; Parker)
- "Yeah I couldn't tell what it was. And the robot didn't think it was anything. So I kind of accepted it judgement at that point. […] And I couldn't tell what it was and I was like, fair enough. He looks pretty competent." (P03; Roy)

At one instance a *participant called an error*:

- "[…], it was particularly with bright objects [hesitation sound] this particular robot seemed more susceptible to where I couldn't quite distinguish it [target], so it was enough doubt there to highlight it [called error], I am not sure if it was [a target], it was definitely a bright object which seemed to be the problem." (P10; Parker)

Participants were more likely to trust the robots judgement when they were uncertain about a target. This could lead to over-trust in the robot. It could help to contribute to the human's decision making process by providing a percentage of accuracy (how sure the robot identified/not identified a target).

5.4.5.1.7 Sub-event: 1st target found [23]

Event	Robot succeeded		
Theme	**Sub-event: 1st target found [23]**		
Sub-theme	**Positive feedback [5]**	Improved feelings [9]	**Learning experience [9]**
Raw data theme	• **Good [4]** • **Useful [1]**	• More trust/R more competent [5] • More relaxed [3] • R works properly [1]	• **Voice [6]** • Verify expectations [2] • Targets [1]

Table 12 - TBCA overview of sub-event: Sub-event 1st target found

The first target that the robot acquired in a trial, was quite interesting in terms of the feedback participants gave [23] (see Table 12). There was positive feedback how *useful* [1] and *good* [4] it is to know that the system works:

- "The robot did identify that, yeah that's right. That was useful for me, it's first time that it picked something up, […]." (P10; Roy); *useful*
- "Yeah, I thought that was great, it is working properly. It has done it right." (P18-Parker)
- "When he saw the first one or he identified that it was quite a relief in some way, because I didn't know what was gonna happen, if that make sense. Obviously you expect everything, I had some expectation but I didn't know what voice would be there and the fact that it was quite clear and understandable and you know, [unclear]. It was a good relief." (P01; Parker)

When the first target was found participants learned (*learning effect* [9]) how the *voice* of the robot sounds like [6]:

- "I was still wondering about things. Is possibly now it started, - saw something and it started to talk to me." (P15; Roy)
- "[…] I was confused there, because he said, target identified rather than evidence found, like the other one. I wasn't sure if that was a mistake or whether it was a different robot." (P05; Roy)
- "So here he found it, so here because I didn't expect the audio so I thought it was gonna say: target found. I didn't and I wasn't ready for it, let's say, it said something: hazardous sign or something […]. (P23; Parker)

These comments already implicate that there is a need for continuously visualising the status of the robot and giving a starting message, so that people can familiarise with the voice, the level of loudness, and know that the system works properly. Again, an overview of what the robot can say might be beneficial.

5.4.5.2 **Event: Robot mistaken**

The following themes and sub-themes were recorded when the robot made a mistake. The main themes were general comments [43], sub-event: participants wait for robot [5], and sub-event: two mistakes in succession [14]. In this section only detailed analysis was done for the general comments [43] and the sub-event: two mistakes in succession [14]. That

participants had to wait for the robot was similar to the sub-event: participant waits for robot during the event: robot succeeded.

5.4.5.2.1 General comments [43]

Event	Robot mistaken			
Theme	**General comments [43]**			
Sub-theme	**General feedback [7]**	**General feelings [16]**	**Explanation of mistake [12]**	P uncertain/ unsure [5]
Raw data theme	• **Obvious target [4]** • **that's not right [1]** • not very well [1] • R was fast [1]	• **Less confidence [5]** • **More attention [2]** • **P pleased/happy spotting target [4]** • Not happy [1] • **Need a human [1]** • **Good to have (human) backup [1]** • **Humans winning against robots [1]** • Happy about R error [1]	• **Because in LR phase [2]** • Because unclear picture [2] • **Because of difficult environment [2]** • **R needs to get close [2]** • **R needs to see whole triangle [1]** • **R just identifies shapes [2]** • R in a small confined space [1]	• Unsure about object [3] • P wants manual control [2]

Table 13 - TBCA overview of sub-event: General comments

General feedback [7] and General feelings [16]

Some of the *general feedback* [7] showed disappointment or lack of understanding about the robots performance (see Table 13):

- "[…] there were a couple of times when they couldn't see something, which was right in front of him." (P03; Roy); *that's not right*
- "Yes it missed that, it was funny because it did zoom on it and I thought: ok, it's like clear." (P23; Roy); *Obvious target*
- "And there it is very obviously there is a [unclear] of a man. And it just completely misses them. […]" (P03; Parker) ; *Obvious target*

In terms of *feelings* [16] participants felt less confidence in the robot after a mistake [5]:

- "Yes that was a blank triangle, but it picked it up as a hazardous sign. [hesitation sound]. It wasn't. And just be a bit more suspicious again […]." (P10; Parker)
- "[…], I was probably more wary of it being wrong again, […]." (P18; Roy)
- "It made me think it might be less reliable, […]" (P08; Roy)
- "[…] where the second one, it missed, you lose confidence in it." (P10; Roy)

Participants also paid *more attention after a mistake* [2]:

- "Okay. Bearing in mind that I thought it might had missed something, then, so I was probably still engaged, make sure and then relax a bit." (P17; Parker)
- "Yeah I was trying to stay focussed and I got it." (P25; Parker)

Interestingly four of the participants were *pleased about the error*, because they spotted the mistake

- "So I was quite pleased there, I don't think that she [the robot] said anything in a while […]" (P16; Roy)
- "I was quite like pleased […] this one was quite clear and I could see it." (P25; Roy)

Two participants pointed out, that their role is not redundant:

- "[…] when it's looking at something that is behind something and then can't be absolutely sure, so that's where you need a human to, - investigate that further." P21; Roy; *Need a human*
- "I think in this situation it is good to have a [human] backup." P13; Roy; *Good to have (human) backup*

Participant 16 even pointed out that they won against the robot: "I was quite pleased that I gained something above the robot, you know, humans winning or something." (P16; Parker); *Humans winning against robots*

Participant 21 was very enthusiastic to spot an error: "Error, here you go [happy voice] - Thank you." (P21; Roy); *Happy about R error*

In general an error from the robot evoked negative feelings and led the participant to allocate more attention towards the robot and having less confidence in it.

Explanation of mistake [12]

Some of the participants tried to *explain the mistake* the robot made (see Table 13). It seemed some empathised with the robot. The explanations were *low reliability phases* [2] or a *difficult environment* [2].

- "[…] I guess here towards the end it was failing, but it did fail in the low visibility ones, so I can't really blame it." (P16; Roy); *low reliability phase*
- "[…] so even I said that was an error I was waiting for him to correct himself but he didn't at that occasion. Probably because it was hidden." (P14; Parker); *difficult environment*
- "[…] I only just caught the victim I think, it was quite, I think the victim was quite faint." (P12; Parker); *difficult environment*

Participants also assumed certain robot information processing procedures such as the *robot needs to get close* enough [2], the *robot needs to see the whole triangle* [hazard sign] [1], or the *robot can just identify shapes* [2]:

- "[…] yes it is an error but probably just being programmed to identify the shape as opposed to having the information in it [symbol in the target/triangle]." (P18; Parker); *R just identifies shapes*
- "The second one is (missed), is cause the bottom right corner is not very clear, I think. And he needs to see the whole triangle to spot it." (P03; Roy); *robot needs to see the whole triangle*

This suggests that if more information about the robot's internal mechanisms would be available to participants (e.g. how the robot actually identifies the objects), a deeper understanding between robot and operator could be achieved and participants may be able to more accurately predict the robot's actions.

5.4.5.2.2 Two mistakes in succession [14]

Event	Robot mistaken
Theme	**Sub-event: Two mistakes in succession**
Sub-theme	**General**
Raw data theme	• **General mentioning [8]** • **Paying more attention to R [3]** • **Less confidence/trust in R [2]** • **P questions its own performance [1]**

Table 14 - TBCA overview of sub-event: Sub event two mistakes in succession

An overview of the raw data themes is provided in Table 14. Often, participants mentioned they noticed two mistakes in succession (8 times *general mentioning*). This also led participants to *pay more attention towards the robot* [3]:

- "I was checking a bit more often because it just missed those two, […]" (P17; Parker)

And having *less confidence in the robot* [2]:

- "Yeah, yeah towards the end I was definitely doubting the confidence, cause I had to say like three things in quick succession, […]" (P16; Roy)

In case of participant six, two mistakes in succession were a total loss of trust in the robot:

- "I think, in there sort of missing it, missing one before, as soon as he had seen it, one, had been messed up, that was it for me, I think." (P06; Roy); *Less confidence/trust in R*.

Interestingly one participant started even to *question their own performance*:

- "[…] it to have missed two in quite quick succession and I only just caught the victim I think, it was quite, I think the victim was quite faint. I [hesitation sound], that's engaged me more with it, now I am questioning whether or not I have missed anything in the past and I know that I need to concentrate just a little bit more within the future." (P12; Parker)

Making two mistakes in a quick succession seemed to have a bigger impact on confidence in the robot than more timely spaced mistakes.

5.4.5.3 Event: Reliability indication

This event consist of comments that were made by participants when the robot indicated high or low reliability. The reliability indication event was further categorised into *general comments* [14], comments from the *low reliability phase* [77], and from the *high reliability phase* [25].

5.4.5.3.1 General comments [14]

Event	Reliability indication		
Theme	**General comments [14]**		
Sub-theme	General feelings [2]	**General feedback [10]**	Misunderstanding [2]
Raw data theme	• more trust/confidence [1] • sympathetic towards robot [1]	• **No difference between reliability phases [5]** • **Reliability was how good P sees [2]** • Reliability information was useful [2]	• Where is the focus of low reliability? [2]

Table 15 - TBCA overview of sub-event: General comments

General feedback [10]

In this section the *general feedback* comments are explained in detail (see also Table 15). In the *general feedback* it is noticeable that some participants mentioned, that they couldn't recognise any *differences between the reliability phases* [5], but these comments were exclusively mentioned when supervising Roy. This might be down to the fewer amount of feedback which did not provide a reason about the source of the low reliability (e.g. heat).

Reliability was perceived *as good as the participant was able to see* [2]:

- "Yeah reliability was high, but here with the picture breaking up, I was looking at that a lot more focussed on it [robot]." (P10; Roy)

- "[…] but that was more to do with the area it was looking at, rather than the fact that it said like, identification was high. Yeah it had more to do with, what I see." (P07; Roy)

Also the comment "Visibility good enough for me" (P09; Roy) when the robot indicated low reliability, underlines the assumption that a lot of people project their own ability to perform, onto the reliability of the robot.

5.4.5.3.2 Low reliability phase (LR) [77]

Event	Reliability indication			
Theme	**Low reliability phase [77]**			
Sub-theme	**General feedback [21]**	**General feelings [10]**	**P actions [30]**	**secondary task [16]**
Raw data theme	• **More information = good/more trust/useful/ accurate [5]** • **I don't need to know why [1]** • Low reliability make sense in low light [2] • dark/light = anticipated low/high reliability [2] • Visibility good enough for me [1] • Other [10]	• **Sympathetic towards robot [6]** • No complete trust = reliability change makes no difference [1] • Confident about themselves [1] • P could see = feel confident/reliability good [1] • Was heat the environment or robot? [1]	• **More attention/concentration/ checking towards R [25]** • Closer to screen [4] • R took more time to look [1]	• **Stop [8]** • **Less [5]** • Wide space = playing [1] • Still playing [1] • made target out = secondary a bit more [1]

Table 16 - TBCA overview of sub-event: Low reliability phase

Table 16 shows the themes of the low reliability comments. Most participants were *happy with the more information provided* by Parker [5], one stated they *do not need to know why*.

In terms of *general feelings* [10] participants also em*pathised with the robot* [6]:

- "[…] because it couldn't get close enough to the [point at target], because there was, I think it meant to be rubble, in the way." (P18; Parker)

- "Yeah that was quite a difficult scene! To work out visually what was going on and through that grid." (P06; Roy)
- "Yeah that was really hard because it was all the grating and it was very difficult to see anyway what was happening, [hesitation sound]. I wasn't very helpful for the poor robot, […]" (P19; Roy).

When participants encountered low reliability (LR) phases (*P actions*), 25 stated to *pay more attention to the robot*:

- "[LR] So here again full attention to the thing [robot]." (P16; Roy)

And checking more vigilant

- "Yeah, I felt I just had to be extra vigilant when the robot said, their likelihood of spotting a target was low […]" (P06; Parker)

The *secondary task* statements [16] agree with the fact that people paid more attention to the robot in LR than in HR. During LR eight participants mentioned to *stop the secondary task* and five stated to do *less* secondary tasks. Other people continued the secondary task when they felt confident in the environment:

- "Reliability is low again and I am playing on this [secondary task]. Because I, […], it was in quite of broad space." (P07; Roy) *Wide space = playing*

In general, the robot indication of low reliability phases made participants pay more attention towards the robot and in some cases participants empathised with the robot as to why a mistake happened.

5.4.5.3.3 High reliability phase (HR) [25]

Event	Reliability indication		
Theme	High reliability phase [25]		
Sub-theme	General feedback [3]	**General feelings [8]**	**secondary task [14]**
Raw data theme	• Good [1] • Image clear [1] • Still a lot to see [1]	**• more trust/confidence [4]** **• Relaxed [3]** **• Not more confident in R [1]**	**• Do/do more/carry on [10]** **• Always flicking [1]** **• lot of stuff = secondary task, but checking [1]** **• Picture still flashing = keep looking at R [1]**

Table 17 - TBCA overview of sub-event: High reliability phase

An overview of the themes is shown in Table 17. When the robot indicated high reliability the comments about *general feelings* [8] were mostly that people *got relaxed* [3] and had *more confidence in the robot* [4].

- "And then I did feel, when he said back to high reliability that I can relax a bit more […]" (P14; Roy); *relaxed*
- "I was checking a bit more often because it just missed those two, but the high reliability thing I felt confident with." (P17; Parker); *more trust/confidence*

However, one participant stated that he/she was *not more confident in the robot*:

- "[…] I wasn't that bothered whether is high or low. […] I probably think, when it does say high, I did do a little bit more on here [secondary task]. But it's not that I felt more confident in it or anything." (P07; Roy)

Ten participants commented that in HR phases they did *more of the secondary task*:

- "[…]if it said it was picking stuff up, […], then I was probably more inclined to go for the secondary task at that point." (P10; Roy)
- "So I was clicking away [secondary task]." (P17; Roy).

Nevertheless on some occasions they did not commit to the secondary task more:

- "It said high, but it was still flashing and like lots of things around, so I still kept looking, maybe there was something I could see." (P25; Parker); *Picture still flashing = keep looking at R*
- "Even when it said it was in a high identification area I never once solely looked at the secondary task. I was always flicking my head, my eyes between the two (screens) Even though I knew that it was very reliable, […]." (P08; Parker); *Always flicking*
- "[…] it's a balance, so he says it's a high reliability but there is a lot of things to look at, […] I am going about the same pace at the secondary task and checking [robot]." (P14; Roy); *lot of stuff = secondary task, but checking*

High reliability indication led people to be more relaxed and allocate more time towards the secondary task. This agrees with the time participants allocated towards the robot in low and high reliability phases (see Figure 69, p.142).

5.4.5.4 Retrospective verbal protocol conclusions

Due to the huge amount of data collected, a summary of the results from the retrospective verbal protocol is given in Table 18. Results are shown by a bulleted list and conclusions are marked in **bold**.

Event	Summarised results and Conclusion
Robot succeeded	• One of the biggest issue was that participants had to wait for the robot to identify a target. This cost a lot of allocation time and it also annoyed participants; later on some already anticipated the delay and because of this delay participants also declared errors. **These comments indicate that there is a need to visualise the process of identification of the robot. This could be done using a visual overlay and a loading bar as well as by providing information if the robot tries to get another angle/view upon the target. It would also be useful if the operator (if he or she has already identified it) can abort the identification and declare that it is a target/no target.** • If the robot performed better than the participant (e.g. identified the target first) then the feedback from participants was mostly neutral or positive. • When the robot found a target in a low-reliability phase, people were impressed.

	This underlines that a good performance under difficult circumstances is valued by the operator. • In general, the robot's success resulted in positive feedback. • The robot's success led to participants being more confident in the robot and having more positive feelings towards the robot. **As expected a robot's successful action gained positive comments from the participants and overall contributed to more trust in the robot.** • Participants' comments indicated that they were doing the secondary task more, started it or carried on with it. • Participants identified a target before the robot and they used the time until the robot said something to do the secondary task. • When participants were unsure about a target they were trying to make sure and double-checked (e.g. using manual mode or look more carefully). • If participants were unsure they followed the robot's decision (only one participant declared an error). **An overview of what the robot can say and what that explicitly means might be useful.** **Participants were more likely to trust the robot's judgement when they were uncertain about a target. This could lead to over-trust in the robot. The human's decision making process could be aided by providing a percentage of accuracy (how confident the robot identified/not identified a target).**
1st target found	• The event gave participants the impression the system works. • Hearing the voice for the first time was important for some participants. **These comments indicate that there is a need for continuously visualising the status of the robot and giving a starting message, so that people can familiarise themselves with the robot's voice and the level of loudness. Again, an overview of what the robot can say might be beneficial.**
Robot mistaken	• Some of the comments showed disappointment or lack of understanding about the robot's performance. • In terms of feelings, participants felt less confidence in the robot after a mistake. • Participants also paid more attention towards the robot after a mistake. **In general an error from the robot evoked negative feelings and led the participant to allocate more attention towards the robot and it also led to reduced confidence in the robot.** • Some of the participants tried to explain the mistake by saying the robot was in a low reliability phase or a difficult environment. It seemed they empathised with the robot. • Participants also assumed certain robot information processing procedures. **It can be assumed that if information about robot procedures was available to the participants, (e.g. how the robot actually identifies the objects) a deeper understanding between robot and operator could be achieved.** • Participants often mentioned when they noticed two mistakes in succession. This also led participants to more attention towards the robot.

	Making two mistakes in quick succession seemed to have a bigger impact on confidence in the robot than mistakes occurring over a longer time.
Reliability indication	• Some participants mentioned that they could not recognise any differences between the reliability phases (Roy only). This might be down to the lesser feedback which did not provide a reason about the source of the low reliability (e.g. heat). • Robot reliability was perceived as good as the participant could see in the remote environment. **A lot of people projected their own ability to perform onto the reliability of the robot.** • In low-reliability, participants were happy with the greater amount of information provided by Parker and they empathised with the robots. • If low reliability was indicated, participants paid more attention to the robot. **In general, feedback indicating low reliability made people pay more attention towards the robot.** • When the robot indicated high reliability, the feedback was mostly that people were relaxed and had more confidence in the robot again. • Some participants commented that in high reliability phases they did more of the secondary task. **High reliability indication led people to be more relaxed and allocate more time towards the secondary task.**
Attention allocation	• They mostly switched from the robot to the secondary task when they could see a clear and wide area in front or when the robot was advancing towards an obvious point. • The more familiar and experienced participants felt with the robot, the more they switched to the secondary task. • As soon as some participants identified a target (before the robot) they did the secondary task and at the same time waited for the robot to identify/not identify the target. **Generally, participants switched to the secondary task when the picture was clear and not cluttered. Experience with the task led to more secondary task performance.** • The most stated reason for switching back towards the robot were cluttered and complex environments or when the robot turned towards a new area. • Bad picture quality also led to participants switching back to the robot. • When the robot slowed down, stopped or behaved uncharacteristically, participants allocated their attention to it. **It can be assumed that the robot movement had a huge influence on visual attention allocation, indicating not only what the robot sees, but also how fast or slow it moves. As expected, busy or cluttered environments with potential targets and bad picture quality led participants to watch the robot more.** **The main reasons for not switching were when people felt unfamiliar with the task, when there was bad picture quality, less trust in the robot or they forgot about the secondary task.**
Robot/ Interface character-istics	• The additional feedback from Parker was positive • Further, the robot had a good distance to objects, good speed and an understandable/comprehensible search strategy.

	• Participants asked about physical dimensions of robot • The robot's process of identifying a target was questioned in many different ways by participants **It seems that a more detailed explanation of the robot's victim identification process is necessary. This missing information could be provided during training: a representation of the decision making of the robot and the mechanism of identifying targets (iteration of planes, points, heat pattern, etc.) could be beneficial.** • Three participants commented on having no feedback at the beginning. **It is important to continuously provide the status of the robot and give a starting message, so that people hear the robot's voice and know the robot works properly. In addition, the robot motor sound could positively contribute to the overall understanding of the robot's state. For instance when we are driving a car uphill the motor needs more power and will sound differently, the same could be applied to the robot; if the robot drives over rubble (not obviously visible for the operator) and is therefore uncharacteristic slow, the operator can hear that it might be due to the surface.** • In some cases the robot was too fast for the participants. **Even though the robot might be autonomous it should be possible to slow it down to be adaptable to different skilled operators.** • Ideas from participants: ○ Participants wanted feedback if there was no target, a command to go back/check again and an explanation of what the robot was actually telling the operator. ○ Regarding the camera and movement participants wished that they were provided with a surround view of the area when entering a new section. ○ Other ideas included adjusting the robot's speed to the operator's skills. ○ With respect to the screen/display participants were not sure whether they were still in low or high reliability phases therefore it would be useful to have visible feedback of reliability feedback. ○ Show reliability bars on screen constantly (e.g. light, heat, accessibility and overall reliability) in terms of percentages. ○ Top view diagram of robot indicating faults. **Recapping these ideas the movement of the robot should be adjustable by the operator. If the robot turns towards new areas a surround view is given to the operator. The reliability (low or high) should be visualised on the screen, as well as the light levels, heat levels and accessibility levels. The idea offered was a bar in percent. Furthermore if the robot identifies a target it can also provide a percentage of how sure it is about the identification. If there is no target the robot will indicate that, too. Another feature could be a top view of the robot indicating any faults.**

Table 18 - Retrospective verbal protocol implications and conclusion

The following points were addressed on the subsequent studies (Study III and/or Study IV):

- The status of the robot is visualised constantly (study III: auto/manual mode; study IV: reliability level, status icon, battery percentage, signal percentage).
- Although no starting message was provided, participants had an intensive but short training at the beginning of the experiments. This included adjusting the loudness of the robot, be presented with all possible messages the robot is able to provide, how the robot navigates through the environment, and how the robot identifies targets.
- Providing the robot motor and environmental sounds to provide a richer picture of the remote environment.
- In study IV: When the robot entered a new area it provided a look around the area and indicated if it would get another angle to better identify an object.

5.4.6 Interview analysis

At the end of the two trials a semi-structured interview was conducted with each participant. The answers to each of the four questions were categorised into the possible answer themes (e.g. yes/no/undecided). Each participant could only be in one of these categories. By using the Chi-Square Goodness-of-Fit test the significance of the differences of the number of participants in a certain category were tested. The percentage was calculated from 23 participants who successfully completed the experiment. Each question shows the percentage of people in each of the specific categories. After that, the content of the answers was analysed using the theme based content analysis method (Neale & Nichols, 2001). Content themes (marked in *italic*) and their supporting quotes will illustrate the answers given by the participants. The number of occurrences of each sub-theme is shown with a number in square brackets (e.g. [5]).

5.4.6.1 Question 1

The first question asked if participants noticed the different amount of feedback given by the robots. Both robots had the same type of voice, but gave different amounts of feedback.

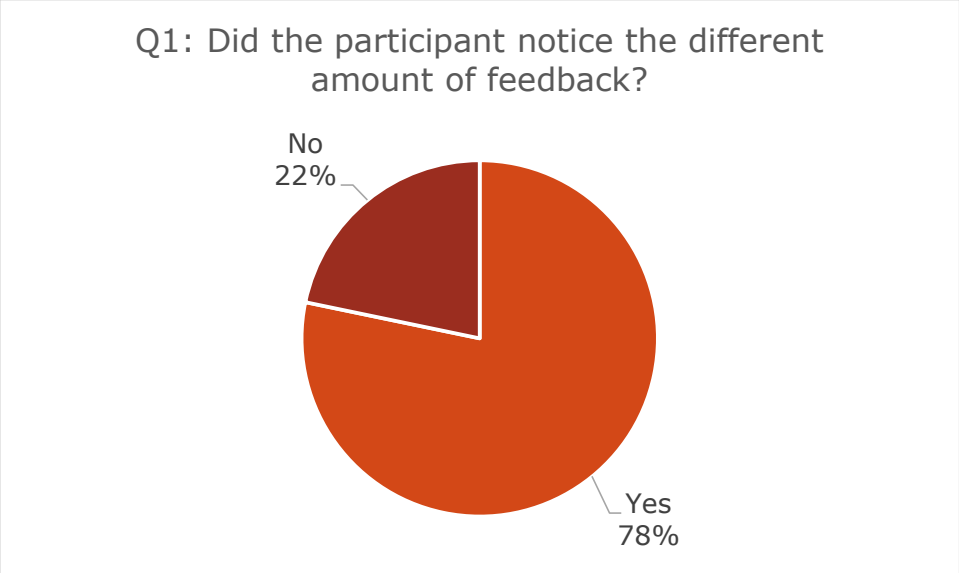


Figure 70 – Did participants realise a difference between the robots?

Figure 70 shows how many of the participants realised that the robots gave different depths of detail in their feedback. Significantly more participants (78%) were aware of the difference than not ($X^2(1, N = 23) = 7.35$, $p < .05$, $w = .57$).

Content themes

Content themes	Count
Feedback difference realised	18
Focussed on other differences	2
No difference realised	1
Not listening	2

Table 19 - TCBA content theme overview of question 1

An overview of the content themes can be seen in Table 19. Some participants did *not realise there was a difference between the robots*, because they were *not listening* what the robot was saying [2]:

- "The voice were the same for me." (P01)
- "I suppose I didn't take any notice of the first voice either, so." (P14)

It seemed that participants who did not realise that there was a difference in feedback were *focussing on other traits of the robot* [2]:

- "[..] both seem to take about the same amount of time to identify […] seem to take similar speed […] one of them was lower to the ground?" (P02)
- "[…] maybe one had wheels and [the other] tracks […] one didn't seem to react so good […] they both highlighted the same things […]" (P21)

5.4.6.2 Question 2

The second question asked participants which robot they preferred. Significantly more participants preferred Parker (65%) compared to Roy (22%) ($X^2(1, N = 20) = 5$, $p < .05$, $w = .5$). 13% of the participants were undecided (see Figure 71).

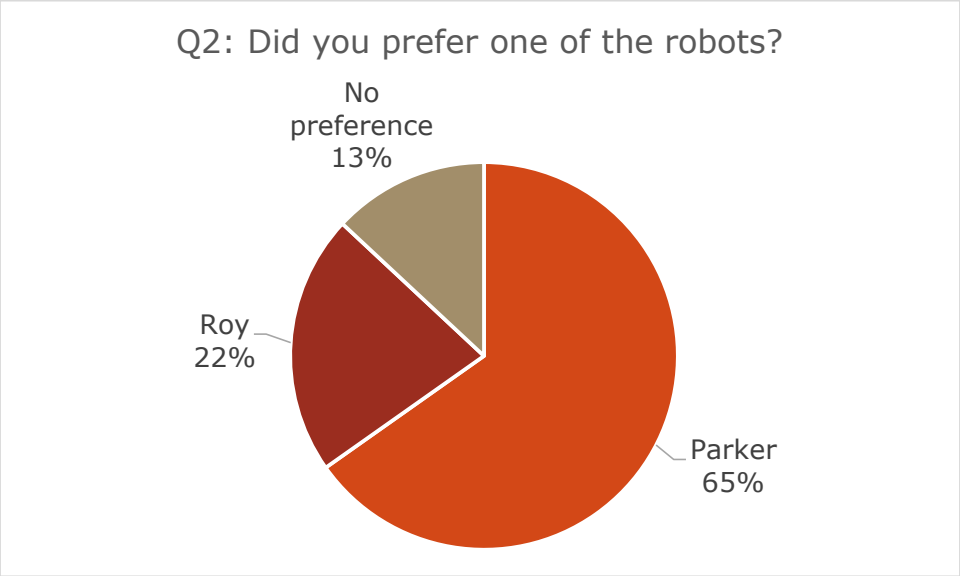


Figure 71 – Which robot is preferred over the other?

All participants with no preference also did not realise there was a difference between the amounts of feedback the robots gave.

Content themes

Table 20 provides an overview of the emerged content themes.

Content themes	Count
Preferred Roy	5
Preferred Parker	15
Preferred Parker, because of additional information	8
No preference	3
Familiarity had impact on preference	2

Table 20 - TCBA content theme overview of question 2

Some of the reasons why people chose Roy were as follows [5]:

- "The first one [Roy] even though I wanted to speak more. But not speak like the second one." (P07)
- "[Roy] was more consistent for longer." (P16)
- "[…] but then if you have only got a limited amount of time to do your rescuing, […] I would go for the second one [Roy], just because it is faster." (P18).

Participants who preferred Parker appeared to favour Parker due to the *additional information* provided [8]:

- "[…] first one [Parker], there was more information about what the fault was than the second one [Roy]." (P02)
- "[Parker], because of the more information it gave. More confidence in that." (P05)
- "Because it gave me more detailed information. But I didn't think it was any more reliable." (P06)
- "It kept the signs clearer what he was finding and giving me a lot more information. […] And If I had another task to do, then I could rely on this [information] a little bit more." (P13)

There also seemed to be a learning effect which might have biased their preference [2]:

- "The second one [Roy], […] because I felt more comfortable, maybe that was more to with me feeling more familiar with the system and what I was doing, then the robot and stuff." (P14)

- "I would say second one [Parker] but that might be only because I already knew what I was doing." (P25).

5.4.6.3 Question 3

Regarding the third question about which asked participants which robot they would trust more (see Figure 72), most of the participants were undecided (48%). 39% would trust Parker more and 13% would trust Roy more. However, these differences were not significant ($X^2(2, N = 23) = 4.52, p > .05$).

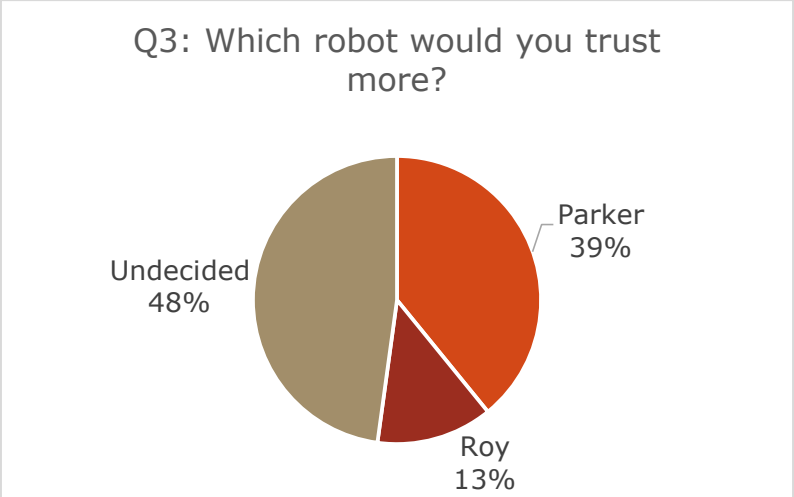


Figure 72 – Which robot would you trust more?

Content themes

Content themes	Count
General performance	4
Reliability	3
Consistency	1
Effectiveness	1

Table 21 - TCBA content theme overview of question 3

Participants associated trust with certain *general performance* related robot traits [4]. Further, participant mentioned that *reliability* [3], *consistency* [1], and *effectiveness* [1] were the main attributes of the robot they associated with trust (see Table 21):

Performance in general [4]:

- "I thought the first one [Parker] performed better." (P01)
- "I think they are both fairly equally trustworthy as to the information they can detect." (P15)

Reliability [3]:

- "[…] I think that the reliability was comparable, […]." (P19)
- "Probably the first one (Parker), because it seemed to do a more thorough job of looking around everything, even though they were both equally reliable." (P18)

Consistency [1]:

- "[…] the second robot [Parker] seemed much worse in the conditions where it was too crowded. Whereas, that was consistent, whereas with the other robot [Roy] it didn't seem to had a consistent pattern of where would it recognise things […]." (P19)

Effectiveness [1]:

- "I don't think I trust any different, because they both missed something […]. I am not sure which one was the most effective. […]. I didn't feel I put more trust in it because it was giving me more information. But it [Parker] helped me to do my part of the job more effectively." (P17)

5.4.6.4 **Question 4**

The final question asked which robot was more intelligent (see Figure 73).

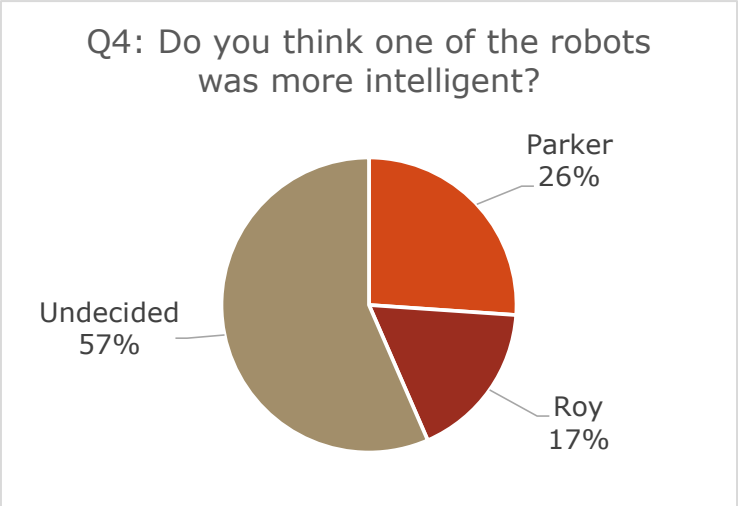


Figure 73 – Which robot is more intelligent?

More preferred Parker (26%) over Roy (17%) and 57% were undecided. However, the differences were not significant ($X^2(2, N = 23) = 5.83$, $p = .054$).

Content themes

Content themes	Count
Performance	10
more engaging robot	1
More information	4

Table 22 - TCBA content theme overview of question 4

The Table 22 shows what participants associated with intelligence. Intelligence appeared to be associated mostly with *performance* [10] but also with the *additional information* provided by the robot [4]:

Performance [10]

- "But from judging what they did recognise [targets], they seemed even." (P05)
- "You got hunches to say the second, just because the shear amount of feedback, but I think […] the first one [Roy] definitely seem to have less errors and therefore would be more intelligent." (P10)
- "Both are just as intelligent. […] They found the targets for the same success rate and they assessed the risks and did that with the same

level of competence, […] I don't think that is intelligence, - that is computation." (P20)

Additional feedback given [4]:

- "Possibly the second one [Parker] because it had the ability to reason, to know why things were going on, I would say that made it more complex." (P08)
- "I have to say the second, only because it gave me more information but that might not be true at all." (P13)
- "[Parker] was more intelligent, […], it was able to give you the information as to what was malfunctioning." (P15)

5.5 Discussion

This study aimed to investigate the influence of the amount of feedback provided by a robot on workload, performance, perceived robot characteristics, visual attention allocation and trust. The two robots used in this study were named Parker and Roy. Both gave reliability feedback as to how reliable they are at the given moment. Parker gave more detailed feedback (explaining the reason for low reliability, stating the cause of faults, and identifying the type of target found) than Roy.

5.5.1 The influence of additional robot feedback

Relating back to the research questions the following results were obtained.

H1) The amount of explanation given by the robot will affect an operator's cognitive workload.

The amount of explanation from the robot did not significantly affect the reported subjective cognitive workload. Desai's (2012) experiments showed that the introduction of reliability feedback itself produced significantly higher workload. In this study Roy and Parker both indicated reliability feedback but Parker gave a more detailed explanation as to why the reliability was low. This additional information did not influence subjective workload ratings significantly. This suggests that reliability feedback produces higher workload ratings (Desai, 2012), but additional explanation does not increase workload significantly further.

H2) The amount of explanation given by the robot will affect task performance.

The amount of information provided by the robot did not significantly influence task performance scores. However, further data analysis revealed a significant learning effect between the first and second task performed. Therefore, the learning effect could have blurred the results.

H3) The amount of explanation given by the robot will affect an operator's perceived characteristics of the robot.

The intelligence ratings showed no significant difference between Parker and Roy. However, the competence rating differed significantly: Parker was rated more competent than Roy.

H4) The amount of explanation given by the robot will affect the trust an operator has in the robot.

Overall, the depth of explanation provided by the robot had no influence on the trust participants had in the robot. Yet significantly more participants would prefer to use Parker instead of Roy and most of the participants were undecided which of the two robots they would trust more. Parker's communication was rated as being significantly clearer than Roy's communication and Parker was seen as more competent and malfunctioning less than Roy. It seems that providing more information and the reason for the state of the robot can make robots appear more competent.

H5) The indication of reliability will affect an operator's attention allocation.

The indication of reliability influenced the visual attention allocation significantly. Participants supervised the robot more when it was in a low reliability phase. This is in agreement with Chien and Lewis (2012) as well as Desai et al. (2013), who found that participants switched control modes when the reliability of the robot dropped and the indication of the robot's reliability improved the efficiency of the human-robot interaction. Participants in this study rated that the human-robot team accomplished

the task significantly more efficiently with Parker, who provided more information.

5.5.2 Combined discussion

A visualisation of the results is given in Figure 74. The additional explanatory feedback had influence only on items of perceived robot perception and communication. A robot with additional feedback was also not perceived as more trustworthy.

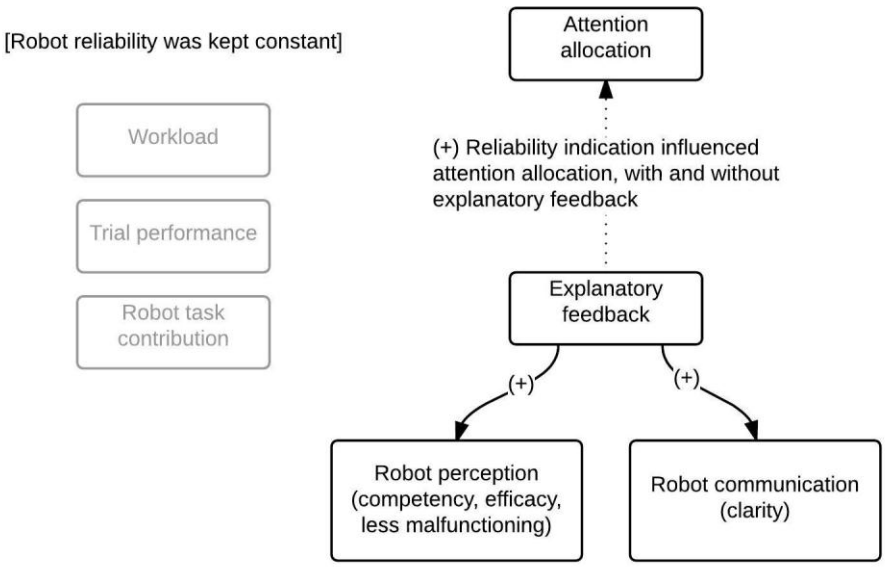


Figure 74 - Qualitative overview of research results of study II; positive influences are indicated with (+), negative influences indicated with (-)

Workload, trial performance and rated robot task contribution was not affected by the additional information given. Therefore other factors have greater influence on these variables. The main influencing factor in human-robot interaction, which is robot reliability, was kept constant. But the indication of reliability by the robot, with and without explanatory feedback, contributed to a better visual attention allocation of the participants.

5.5.3 Qualitative data analysis

The retrospective verbal protocol provided a large amount of data regarding the human-robot interaction. This method had not been previously used in conjunction with autonomous remotely operating robots. It needs to be

considered that this is a purely qualitative analysis which is naturally biased by the interpretation of the researcher. The study revealed the impact of different aspects of robot features and behaviour, such as illustrating the importance of visual cues about robot running processes and the influence of robot speed/movement on human attention. In general, the data suggests that a higher degree of robot transparency is necessary for the operator in order to understand the robot's actions.

Participants favoured Parker (65%) due to the additional information provided. Associated attributes of the robots regarding trust were performance, reliability and consistency. The result is in accordance with other researchers who investigated trust factors in human-robot teams (Hancock, Billings, Schaefer, et al., 2011; Park, Jenkins, & Jiang, 2008). Interestingly, also robot intelligence was associated with performance and with the additional information provided by the robot.

5.5.4 Limitations and future work

It has to be acknowledged that the participant sample was not representative of actual search and rescue workers. Rescue personnel might have different approaches to tasks and a different self-confidence level in performing them, which can have a different influence on trust and workload. Future studies, if possible, should incorporate rescuers in their studies.

Even though the participants had five minutes of training (and more if requested), the learning effect between the first and second task was significant. Therefore the learning effect could have influenced the significance levels of the other dependent variables. In addition, the sample size was relatively small and might not be sufficient for a generalisation of these findings. However, the qualitative analysis provided rich information about participants' strategies and feelings when dealing with a robot.

Trust was measured with a single rated question and was further asked about in the semi-structured interview. This was done due to the time limitations between the end of the trials and the beginning of the RVP. A more detailed questionnaire about trust that is quantifiable will be used in

subsequent studies. By doing so it might be possible to identify sources of distrust and change the robot's interface design and behaviour accordingly.

Furthermore there were not many significant differences which could be an indicator that the independent variable of this experiment was not strong enough to elicit major differences in the perception of the robot. Since both robots had the same performance, which is the most influencing factor on trust (Hancock, Billings, Schaefer, et al., 2011), the additional information (explaining, providing a reason) given by the robot might just have had a small influence on the overall construct of trust. In addition, sometimes participants did not observe an error and the robot had a higher perceived performance than intended by the researcher. This issue will be addressed in the subsequent studies.

A variety of recommendations about robot characteristics and robot behaviour emerged from the quantitative data. Future studies can use these recommendations to develop and test new interfaces and robot designs. Furthermore it would be useful to investigate the factors affecting the perception of performance and how these affect the trust between the human and the robot.

The harsh search and rescue environment must not be neglected. Collecting more information about the working conditions and possible target groups (users) in the U.K. would aid the design process. For example, by developing guidelines for the design of robot interfaces with the aid of focus groups consisting of subject matter experts.

5.6 Conclusion

Reconnaissance robots in search and rescue missions can make rescuers' work safer and allow them to search areas that are too dangerous or inaccessible for humans to investigate. This experiment aimed to investigate the thoughts, attitudes and behaviours of robot operators when interacting with an autonomous reconnaissance robot. The study showed that robot transparency is of importance for the operator to understand the robots' states and actions.

Trust in the robot is mainly influenced by the performance shaping attributes of the robot, which is in accordance with previous research. Both robots had the same performance levels which might have been the reason why trust did not change across experimental conditions.

Both of the tested robots, Roy and Parker, indicated their reliability level during the trials. Parker gave additional feedback as to why the current reliability level was low. The robots' indication of reliability levels positively influenced the visual attention allocation (whether to supervise the robot or attend to the secondary task). In low reliability phases, participants watched the robot more thoroughly and in high reliability phases they relaxed and scored more on the secondary task.

In general it might be useful to provide participants with additional explanations of the robot states. Although the additional feedback from Parker did not increase performance or trust levels, the perceived workload of participants did not increase either. Parker was significantly favoured over Roy and was perceived as malfunctioning less, communicating more clearly and being more competent. However, these findings were obtained in a constant task without varying the task complexity or task difficulty. Different task complexity levels are very likely encountered in the search and rescue domain and might interact with workload and performance measures.

In conclusion, reliability indication can more accurately inform participant's appropriate visual attention allocation and providing additional explanatory feedback can enhance the quality of communication between the operator and the robot without risking higher levels of workload. Nevertheless, the most influencing factor regarding trust is the robot's performance. If performance is kept constant it is likely that trust levels will not change significantly.

5.7 Chapter summary

The chapter tested the influence of different amounts of robot feedback, in addition to reliability feedback, on trust, workload, and performance. This chapter showed that there was no influence on trust, workload, or

performance, which could have been occurred due to a flaw in the study design. However, the study suggests that reliability indication is a valid method to support better control allocation strategies of operators. Furthermore, a variety of qualitative data and their implications for robot design and behaviour was collected.

6 Study III - The influence of robot reliability and task complexity

6.1 Chapter overview

The chapter presents a virtual search and rescue scenario with a semi-autonomous robot system which examines the influence of robot reliability and task complexity on workload, performance, and trust. A post-task questionnaire collects data about trust, subjective workload, robot characteristics, and participant's experiences. This study also informs about possible measurements of performance in semi-autonomous robot systems. In addition, two trust questionnaires and their correlations are compared with each other and recommendations about their application are provided.

6.2 Introduction

Search and rescue tasks are complex by nature, primarily due to their safety and time-critical characteristics within dynamic and unpredictable environments (Wegner & Anderson, 2004). Therefore it is important to investigate how task complexity influences trust, workload, manual mode usage, participant's perceptions, and overall human-robot team performance.

The literature has repeatedly shown that the higher the reliability of the robot, the higher are the levels of trust of operators in the robot (Chen & Terrence, 2009; Chien & Lewis, 2012; de Visser & Parasuraman, 2011; Desai et al., 2012; Robinette et al., 2015). In the previous study (Chapter 5) the reliability of the robots was constant. This chapter examines how failures of autonomy (reliability) influence trust and if this interacts with task complexity. This short literature review starts with discussing robot reliability in terms of effects of errors and their influence on trust. Next, task complexity factors are reviewed.

There are different effects that reliability (errors made by the robot) can have on trust. First, the initial impression matters: Fallon et al. (2005) and

Freedy & DeVisser (2007) found that initial system errors produced lower levels of overall trust than highly reliable first encounters. Desai et al. (2013) found by using a remote controlled robot in a search scenario that early errors of the robot had a more negative impact on trust than later occurring errors. Therefore, not only does the general level of reliability have an influence on trust and control allocation (participants allocation of control to the robot or themselves), so does the timing of the errors (Desai et al., 2012). Additionally, Desai et al. (2013) found that trust after an error recovers slower than trust would develop without an error. Furthermore they suggest that early reliability drops might confuse participants and lead them to poor control allocation. The use of automation is also influenced by a positivity bias of novice users (Desai et al., 2012; Dzindolet et al., 2003). Novice users are more willing to trust automation initially (Robinette et al., 2015).

Second, predictable errors are trustworthy: Freedy & DeVisser (2007) found that trust is not only concerned with the expectation of "correct performance" but also with the expectation of "level of performance". For example, if the user knows that at a certain stage the robot will fail to perform, he or she "trusts" the robot to fail and overall trust remains even if errors occur. Nevertheless, in situations where users could not "trust" the robot to fail in a distinct situation, overall trust decreased over time (Freedy & de Visser, 2007). Therefore it might be the case that the robot's correct performance is not always the most influencing factor on trust, the predictability and consistent behaviour may be more important.

However, there can also be a discrepancy between intended robot reliability and the perception of the operator. In multi-robot control where a human has to supervise and control more than one robot, there is a need for automated aids in order to shift the operators attention between the robots in an effective and efficient manner (de Visser & Parasuraman, 2011). In the Chien and Lewis work (2012) system alarms (robot requests for assistance) directed the operator's attention where needed. They introduced misses and false alarms of the robots. Results showed that there was no difference in trust because it was difficult for operators to discriminate between low and high reliability as well as spotting failures that

did not alarm them. Notably, in high reliability conditions participants focussed their attention on dealing with the robot's help requests rather than with the task of finding victims, which increased the rate of unmarked victims (errors).

However, still some literature reports that there is a discrepancy between the reported trust towards the robot and the actual use of the autonomy features of the robot. Participants sometimes used the robots features although they reported not trusting them (e.g. Robinette et al., 2015).

A recent study from Kaniarasu and Steinfeld (2014) investigated the effects of error attribution or better known as blame. In their study the robot assigned blame either to the user, to itself or to the team. Participants did not rate differently in real-time or post task trust questionnaires. When they ranked the robots, which one they trust most, there was no clear majority. In general blame attribution by the robot lowers the trust in the robot. Furthermore some participants did not trust the self-blame robot, which must be like a co-worker who always points out their negative performance, because this co-worker is likely to be seen negatively (Kaniarasu & Steinfeld, 2014).

Summarising the previous points made, a robot's initial performance is vital for continued trust (Robinette et al., 2015). Poor initial performance has a strong negative impact on a person's trust in the robot. Furthermore, operators' expectations have to be adjusted appropriately for the system to be used.

The other variable that is tested in this study is task complexity. Search and Rescue tasks are complex by nature, primarily due to their safety and time-critical characteristics within dynamic and unpredictable environments (Wegner & Anderson, 2004). Therefore, it is important to investigate how task complexity influences trust, workload, manual mode usage, participant's perceptions, and overall human-robot team performance.

The previous studies (Chapter 4 and Chapter 5) of this thesis emphasised that task complexity is an important factors in USAR. For example, Desai (2012) examined the influence of task difficulty on mode switching

behaviour and trust in human-robot teams. Even though task complexity and task difficulty are similar concepts but they are not the same. Different and contradicting concepts/definitions of task complexity and task difficulty exist in the literature (Bedny, Karwowski, & Bedny, 2012; Braarud & Kirwan, 2011; Campbell, 1988; Hendy, Liao, & Milgram, 1997). According to Liu and Li (2012) task complexity refers to objective characteristics of the task and task difficulty focusses on the perception of the difficulty of the task by the performer. A difficult task does not need to be complex but it is likely that a more complex task is more difficult (Braarud & Kirwan, 2011). Therefore task complexity in the present study is defined as follows:

"Task complexity is the aggregation of any intrinsic task characteristic that influences the performance of a task." (P. Liu & Li, 2012, p. 559)

As mentioned earlier, Desai (2012) investigated the influence of task difficulty on mode switching behaviour and trust. Desai (2012) used a real-world robot system and asked participants to drive around boxes in a corridor. Compared to his base-line experiment he increased the width between boxes to present an easier task. The easier task did not influence trust ratings or workload ratings but as expected the overall performance and the participant's self-performance ratings were better. According to the task complexity definition above Desai (2012) in fact did change task complexity in order to influence the difficulty of the task. However, Desai (2012) only made obstacles smaller so it is easier to navigate the robot. USAR missions require not only navigation but also searching the scene for targets.

USAR missions are immensely complex that there are a variety of task factors contributing to task complexity and eventually to task difficulty. Therefore this experiment examined complexity levels that are relevant and most likely to occur during USAR missions.

Liu and Li (2012) developed a task model (see Figure 75) with five components that influence task complexity. Each task component has complexity contributing factors. To vary task complexity across objective task characteristics in the present study two of five task components of this model varied across experimental conditions. The selection of the factors is

based on the knowledge gathered in study I (Chapter 4) and the literature review (Chapter 2). *Process* and *presentation* were kept constant, because rescuers are highly trained in their relevant working steps (see Chapter 4). Furthermore participants received training with the robot and the interface until they felt comfortable to do so. In addition, *time* constraints were not changed across complexity levels. Higher complex tasks needed to be performed within the same timeframe as less complex tasks.

Each mission is different and goals will change accordingly (Wegner & Anderson, 2004). Therefore, *Goal/Output* was changed in *clarity*, *quantity*, *and redundancy*. Participants had to search for more different targets, some target information was ambiguous or redundant. The *input* was changed in *clarity*, *quantity*, and *diversity*. The environment, which represents the main input of the task, was more or less cluttered and the quantity as well as the diversity of objects cluttering the environment changed. These two task components were chosen because they are most challenging in remote rescue operations: diversity of the hostile environments and the hard to see or unrecognisable targets.

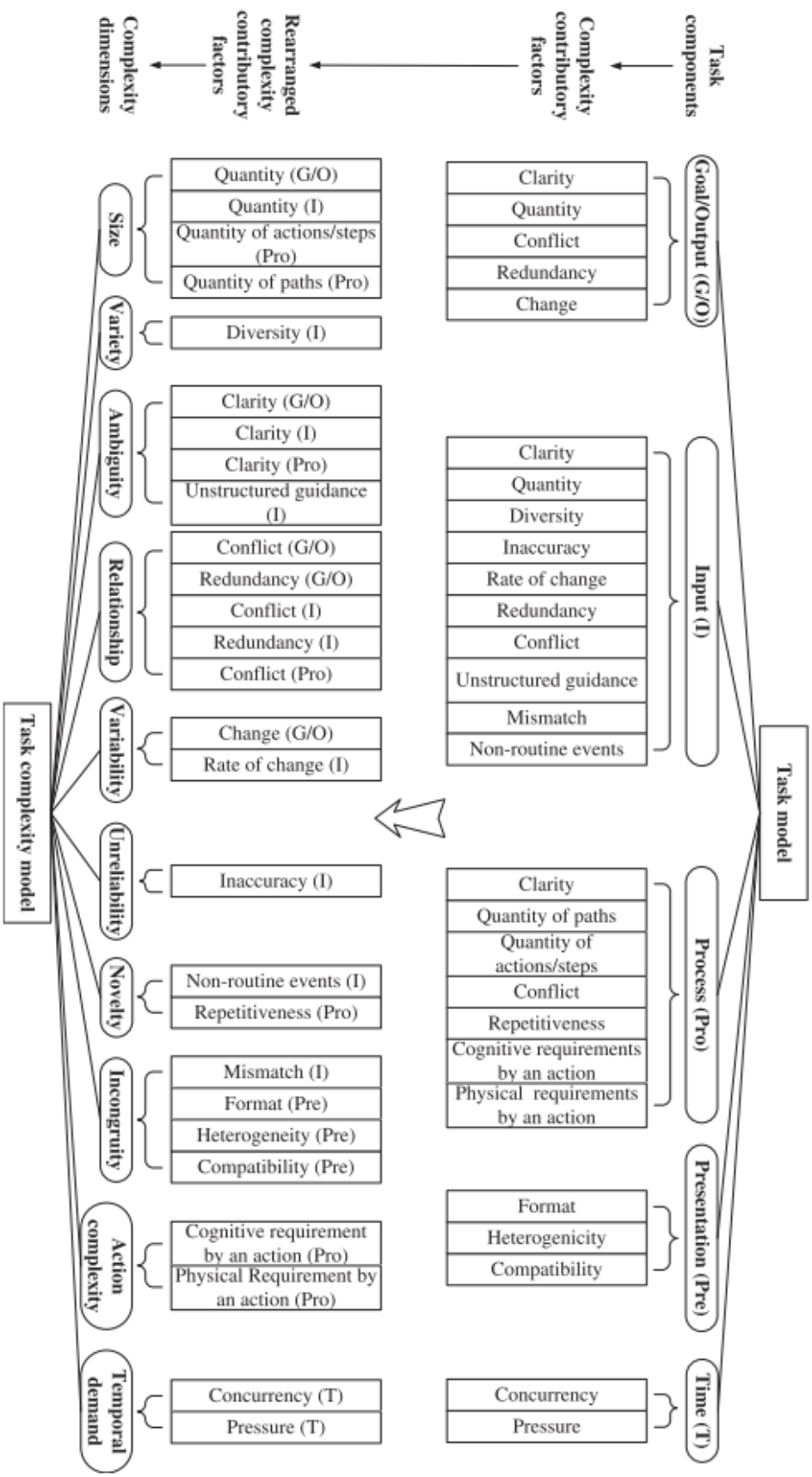


Figure 75 - Task-component-factor-dimension framework (P. Liu & Li, 2012)

In addition to task complexity, the robot's reliability was varied across conditions. This aimed to investigate the interaction effects between task complexity and robot reliability. A robot's performance is claimed to be the most influencing factor on trust (de Visser & Parasuraman, 2011; Oleson et al., 2011; Robinette et al., 2015) and is very likely to vary among new technologies (that have mostly not been tested in the field), such as search and rescue robots, because these robots will encounter numerous unexpected situations that will be unique and extreme. Subjective workload data was collected because it is an important factor in the USAR domain: operators are working under high levels of stress and time pressure and have high workload levels. A reduction in workload can result in higher performance levels. In addition, information about participant's personality traits was collected. According to a meta-analysis from Hancock et al. (2011) personality traits are also likely to influence trust ratings. It is novel to test personality traits together with task complexity levels with respect to remote controlled semi-autonomous rescue robots.

Hypotheses:

Task complexity is changing during an USAR mission. Task complexity influences self-performance ratings and easy tasks foster a better control allocation strategy (whether to use manual or auto mode) (Desai, 2012). However, previous research did not often vary task complexity (Cesa, Farinelli, & Iocchi, 2008; Doroodgar, Ficocelli, Mobedi, & Nejat, 2010; Larochelle & Kruijff, 2012), or only one aspect of complexity (Desai.2012). This study changed several aspects of task complexity that are relevant to USAR missions. If it is possible to understand the influence of task complexity, different sets of robot behaviours in different task complexity levels, might be able to mediate control allocation strategy, performance, workload, trust, and self-confidence. Furthermore, this study will investigate if the robot reliability level will interact with task complexity regarding trust. For example, a bad performing robot in a less complex task might not influence performance or workload, but an unreliable robot in a very complex task might.

This experiment evaluated the influence of task complexity and robot reliability, which establishes the first five hypotheses:

H1) Task complexity will positively influence trust, subjective workload ratings, manual mode usage, and trial times. Task complexity will negatively influence performance measures.
H2) Robot reliability will positively influence trust, performances measures, trial times, and negatively influence subjective workload ratings and manual mode usage.

It is suggested that lower robot reliability levels will elicit lower levels of trust and performance, as well as increase the subjective workload. Also, low robot reliability will increase the manual mode usage and trial times as found in previous literature (Desai, 2012).

Further, it will be investigated if the magnitude of the effect of robot reliability is higher than the one of task complexity. Also, the possible effects of robot reliability can be compared to other studies, which used real robot systems (real-world approaches) to see whether virtual reality approaches produce similar results or differ.

H3) Task complexity will positively influence rated robot performance and negatively influence rated self-performance.

Previous literature showed that higher task complexity reduced the rated self-performance (Desai, 2012).

H4) Robot reliability will influence rated robot performance and rated self-performance.
H5) There will be an interaction between task complexity and robot reliability regarding trust.

Further, participant-rated task difficulty was of interest and formed the following two hypotheses:

H6) Participants will rate more complex tasks as being more difficult.
H7) Participants will rate lower robot reliability as being more difficult.

Additionally, the influence of personality traits was examined:

H8) Personality scores will correlate with trust, performance, subjective workload ratings, and manual mode usage.

The last hypothesis dealt with the differences between the manual mode only and the mixed mode group. This will show whether semi-autonomous features, as used in this study, can enhance human-robot team performance and reduce operator workload.

H9) The manual user group will have higher levels of workload and lower performance measures than participants who used the semi-autonomous robot control (mixed mode).

Furthermore, two trust questionnaires, Muir (1989) and Schaefer (2013), were compared in order to select the appropriate trust questionnaire for a semi-autonomous robot system interaction and qualitative data was collected via interviews after participants performed the trials. Qualitative data was gathered to elaborate on quantitative findings and infer recommendations for future robot interface designs and behaviours.

6.3 Methodology

6.3.1 Participants

39 Participants were recruited from the University of Nottingham. Staff and students were approached via adverts and contacted via phone or e-mail. They were screened to fit the requirements for the study (over 18 years and no vulnerable members of the public). After participants completed the experiment they were compensated with a £10 Amazon Voucher.

To determine the required number of participants in order to have enough statistical power an a priori analysis with the program G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) was performed. The power (1 - β) was set at 0.95 and the α level to 0.05 for a medium effect size of 0.5 (Cohen, 1988). The a priori power analysis indicated that a total sample size of 39 would be sufficient to detect a significant interaction effect between three groups and nine measurements with a correction among repeated measures of 0.5.

All participants were coded with a participant number (PN) and the data was stored under the PN and not under their name. The code links between PN and name were stored separately. The study was approved by Faculty of Engineering Ethics committee.

6.3.2 Experimental design

The study was designed as a 3 x 3 mixed-subject design. There were three participant groups. The between subject conditions were the reliability levels of the robot and the within subject conditions were the complexity levels. Each of the three tasks lasted about 5-7 minutes and the average study time for each participant was approximately 70 minutes. Table 23 shows how the independent variables were grouped and how each condition was named. All conditions were counterbalanced.

Independent Variables		**Between subject factor**		
		Group A (Reliability A)	Group B (Reliability B)	Group C (Reliability C)
Within subject factor	Low task complexity	Condition 1	Condition 4	Condition 7
	Middle task complexity	Condition 2	Condition 5	Condition 8
	High task complexity	Condition 3	Condition 6	Condition 9

Table 23 - Independent variable table

6.3.2.1 Robot reliability (between subject factor)

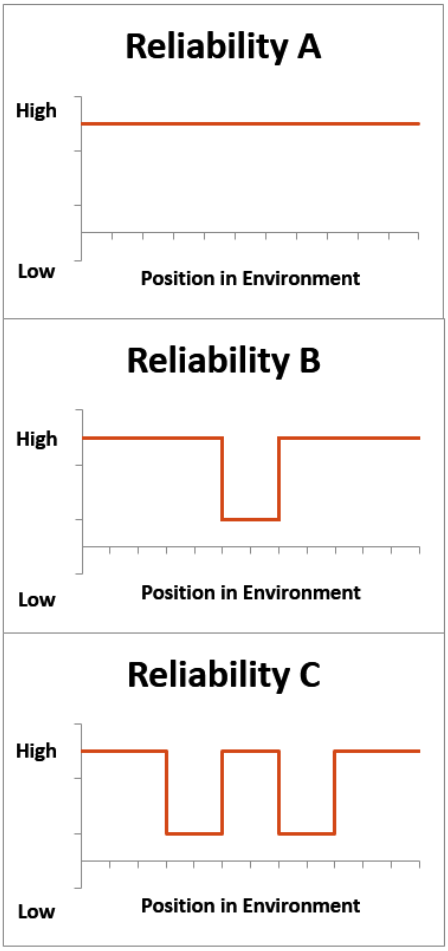


Figure 76 - Overview of reliability profiles

Three different reliability profiles were used to investigate how the performance of the robot influenced the overall performance of the human-robot team and how it influenced the participant's behaviour. The autonomy mode time and location of the robot was the basis for the reliability profiles. Therefore, all reliability profiles were location-linked within the virtual environments. A reliability drop lasted for 45 seconds. The place for the drops was fixed for all three environments, which ensured that they were comparable.

Reliability profile A simulated a robot that did not make any mistakes. Reliability profile B simulated a robot that had one reliability drop in the middle of the course and missed one target. Reliability profile C had two reliability drops during the course and missed one target and wrongly identified another target. Figure 76 shows these profiles. Desai (2012) found that reliability drops immediately at the start or end of a scenario can significantly influence trust, therefore high reliability phases were incorporated at the beginning and at the end of each task. A reliability drop consisted of the robot navigating inaccurately (failing to look at certain corners/areas) and missing a target or identifying a wrong target.

6.3.2.2 Task complexity (within subject factor)

As previously mentioned, two of the five factors (*goal and input*) influencing task complexity (P. Liu & Li, 2012) were modified in order to increase objective task complexity. Table 24 shows what factors were modified and to what extent.

Complexity level	Goal/Output	Input	Secondary task
Low	1. Find casualties	Uncluttered environment	Loading task
Middle	1. Find casualties 2. Find hazard signs	Cluttered environment	
High	1. Find casualties 2. Find hazard signs 3. Find evidence for terrorist attack (weapons, bombs, etc.)	Highly cluttered environment	

Table 24 - Task complexity modification overview

In terms of *Goals and Output*, the participant was required to find fewer types of objects in the low complexity condition compared to the middle complexity condition or high complexity condition. This represented a change in goal *quantity*. *Uncertainty* and *less clarity* were induced for the middle and high task complexity by asking participants to find hazard signs. The participants did not know what the hazard signs looked like, and what colours they might have (e.g. orange, yellow, red or electricity hazard, explosion hazard, bio hazard). Additionally, in the high complexity level, *redundancy* and *less clarity* were introduced by telling people to look for evidence of terrorist attacks, namely bombs and weapons. However, participants did not know how the bombs might look and there were no weapons present in the environments. With respect to the *input* of the task, the higher the complexity level, the more objects (*quantity*) were present in the environment and the more *diverse* were the types of objects.

The following screenshots visualise the different environments that corresponded to the respective task complexity (low, middle, and high). Figure 77 depicts the low task complexity environment which had between 20 to 30 objects per room. Most objects were rubble piles, tables and computers. Figure 78 shows the middle task complexity environment, which had between 40 and 50 objects in each room. Compared to the low complexity environment, this environment had additional types of furniture and concrete elements in the environment. Figure 79 shows the high task complexity environment, which was cluttered with 60 to 70 objects per

room. The environment contained even more diverse objects, such as wooden pallets, books and shelfs. These differences made the environment gradually more cluttered, more difficult to navigate in, and challenging to detect targets in.

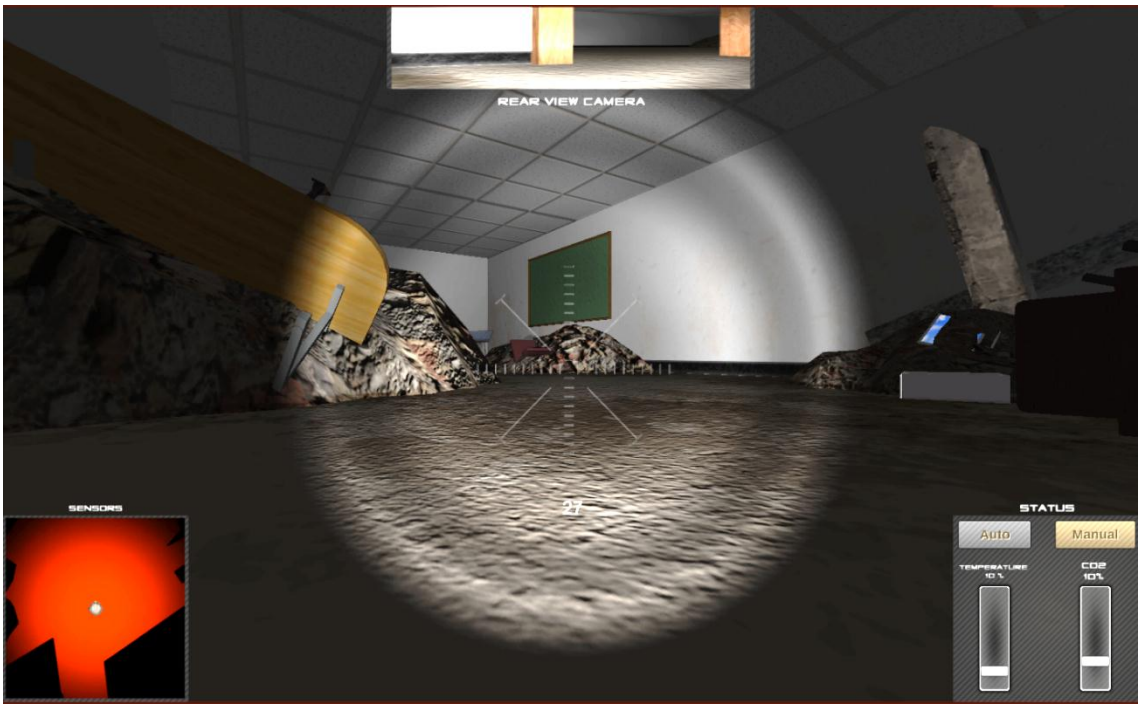


Figure 77 - Screenshot from the low complexity task

Figure 78 - Screenshot from the middle complexity task

Figure 79 - Screenshot from the high complexity task

6.3.2.3 Virtual rescue scenario

The development of the virtual environment und the underlying programming structures are explained in detail in Section 3.5.2.

Virtual Robot

The simulated robot had proximity sensors (360 degree) which were visualised by a top view map. A front and rear camera were provided. Furthermore it had a target identification system which was enabled to find specific targets in the simulated environment. The robot could further show CO_2 and temperature levels. All three trials were performed in a mixed mode, which is explained below.

MIXED MODE

Mixed mode meant that participants were free to choose between using a manual or an autonomy mode. However, the researcher encouraged participants to use autonomy mode.

Manual mode:

- Participant was in charge of the robot's movements and target identification.

- Participant could see the goal navigation points the robot would navigate to in the map.

Autonomy mode:

- Robot was in control of driving.
- Target identification was active.

After 39 participants used the mixed mode an extra 13 participants were recruited to use the manual mode only (without seeing the goal navigation points) to establish a baseline for the experiment.

Virtual environment

Each virtual environment resembled an office complex with one corridor and adjacent rooms. Each room contained rubble, desks, chairs, computers, and other objects you might find in an office with collapsed structures. In distinct places were hidden targets that the participant was required to find. The environment varied across the independent variable of task complexity. The design of the different complex environments was explained in the Task complexity section (6.3.2.2, p.189). The training environment consisted of four rooms of different complexities, so that the participants could familiarise themselves with all complexity types they will encounter.

Figure 80 - Example of victim (left), hazard sign (middle), and bomb (right) in the environment

Targets varied due to complexity level. The low complexity environment incorporated three victims, which were not realistically injured, to be found. An example of a victim in the environment is shown in Figure 80 (left). In the middle complexity of the environment two victims were required to be

found and additionally two hazard signs. Such a hazard sign is depicted in Figure 80 (middle) as well. During high complexity tasks the participant had to find victims, hazards and terrorist indicator, which could be weapons or bombs (self-made plastic explosives). Such a bomb is depicted in Figure 80 (right).

Interface

The robot interface showed front view, rear view, proximity map, control mode, oxygen levels, temperature and a robot damage map. Depending on which mode is active the "AUTO" or "MANUAL" button was highlighted in red. The interface is depicted in Figure 81. In addition, the proximity map had navigation goal points (NGP). These are orange squares that indicate where the will robot drive next. This was used because in the previous study participants liked predictability of the robot and were more relaxed when the robot was advancing towards an obvious point in the environment.

Figure 81 - Interface of the rescue robot

6.3.2.4 Tasks

Participants were presented with an Urban Search and Rescue (USAR) scenario. The primary task for participants was to find all targets as fast as

possible. The following scenario description was read to participants to ease them into the simulation:

"An explosion has occurred in an office complex with an attached warehouse. There are reports of survivors inside. The building is concrete construction and rescue personnel have identified an entry point for the robot. The structure is highly unstable and smaller explosions are occurring at irregular intervals. The safety manager and engineers have determined that the robot is the only safe option for reconnaissance at this time. Your task is to perform a very thorough search of the first three rooms of the office building for the targets. Although the robot can navigate the building safely and has features which can identify targets, only you can decide whether it is a target or not using the cameras on the robot. You will be controlling the robot from a safe location outside the office."

Furthermore, the participant had to attend to a secondary loading task. The secondary task consisted of an extra screen which showed a certain number of blue boxes every 25 seconds. The participants had 5 seconds to input the correct number of boxes on the keyboard, before the boxes disappeared.

Participants interacted with the robot via a X-BOX 360 controller. During each task participants started using the robot in autonomous mode. The participant could decide at any point which mode (manual or autonomy) they would like to use.

6.3.3 Materials

A Laptop (Acer Ultrabook TimelineU i5 with 1.7 GHz, 4GB RAM) with an external 17" screen for the participant was used. The USAR simulation program was created in UNITY, a multi-platform game creation system. The participant could interact with the virtual robot via an X-Box 360 controller. Paper questionnaires, pens and two digital cameras (Sony DCR-SR58E) were also used. The experimental setup is shown in Figure 82. The left screen provided a view through the robots' camera, which is controlled by the X-Box controller. The right screen shows the secondary task. The screen shows a certain number of blue boxes every 25 seconds. The participants had 5 seconds to input the correct number of boxes on the keyboard, before

the boxes disappeared. Figure 83 shows a screenshot how the secondary task looked like.

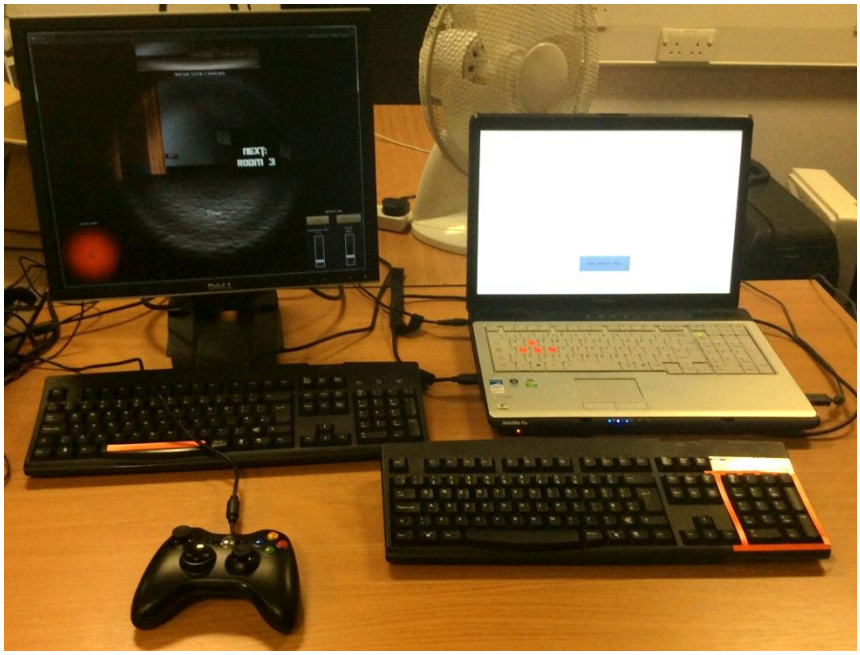


Figure 82 - Experimental setup

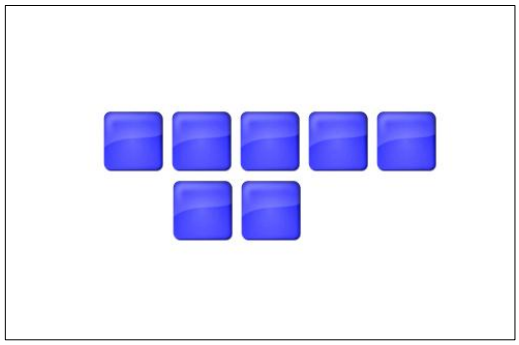


Figure 83 - Secondary task screenshot

After giving informed consent (see Appendix K - - Digital Appendix IV, p. 404, for consent form and study information) participants were asked to answer a general questionnaire. The general questionnaire, provided in Appendix E (p. 387), consisted of two parts. The first part asked about general demographics, use of technology and experience with robots. The second part of the general questionnaire compromised a personality test. As mentioned in Hancock et al. (2011), personality can influence trust ratings. The experiment, reported in this chapter, gathered further empirical data to examine the relationship between personality traits and trust ratings. The items of the personality questionnaire are based on Goldberg's

Big Five factor structure (Goldberg, 1992) and was shortened and modified by 'IPIP' (Gow et al., 2005). The shortened 50-item 'IPIP' personality questionnaire was used because it showed good internal consistency and related strongly with the known dimensions of personality (Gow et al., 2005). The five factors are extraversion, conscientiousness, neuroticism, agreeableness and openness. Each 'Big Five Factor' corresponds to ten questions of the questionnaire. Each factor has a score ranging from low 10 to high 50 points. Participants answered each item on a five point Likert scale with the anchors "Very Inaccurate", "Moderately Inaccurate", "Neither Inaccurate nor Accurate", "Moderately Accurate" and "Very Accurate".

After each task, participants were asked to complete a NASA-Task Load Index (NASA-TLX) questionnaire. The NASA TLX is a multi-dimensional scale used to obtain subjective workload (Hart & Staveland, 1988). Additionally after each task, participants completed two trust questionnaires. The first trust questionnaire was from Schaefer (2013), called the human-robot trust scale (HRTS) which aims to measure trust perceptions specific to human robot interaction. The trust rating delivers a percentage trust score. The second trust questionnaire includes the Muir trust questions (Muir, 1989). Two trust questionnaires were used for the purpose of comparing the results. The aim is to see which of the questionnaires is more sensitive to changes in trust regarding semi-autonomous robot systems. An example of the post-task questionnaire is provided in the Appendix F (p. 391). The post task questionnaire concludes with questions about self-performance, robot performance, mode preferences, difficulty/complexity of task, and self-confidence.

6.3.4 Procedure

Each participant took approximately 70 minutes to complete the experiment. The introduction and pre-questionnaire took approx. 15 minutes, followed by a 10 minute training run with the robot. The three conditions each participant had to complete took, including post-task questionnaires, approx. 45 minutes.

Before starting the experiment each participant was briefed about the study and its aims, who was conducting the study, and the procedure. After giving

informed consent, each participant was requested to complete the general questionnaire (age, gender, occupation etc.) and answer a short version of the Big-Five personality scale. After this, participants received instructions about the task and how to use the robot. In addition, participants were encouraged to use autonomy mode during the trials if they were comfortable doing so.

Each participant had a training session on how to navigate the robot in the virtual environment and how to mark targets. The training session lasted until the participants felt comfortable controlling the robot. The training session started in manual mode, so the participant had to learn how to drive the robot manually. After a certain level of familiarisation, the researcher switched the robot into auto mode and explained how the robot navigates and which search strategy it uses. Following this, participants had two rooms in the training environment to freely try switching between manual and auto mode.

After the training session the researcher read a scenario (Section 6.3.2.4, p. 194) to the participants to ease them into the role of a rescuer. Overall the participant had to complete three conditions; each of the three conditions had the same procedure; just the complexity of the environment changed. Each scenario lasted about 5-7 minutes. The participant and the robot had to find all targets in the environment as quickly as possible.

After each search scenario, participants completed a post-task questionnaire which asked about self-performance, robot performance, post-task workload, post-task trust, mode preferences, difficulty/complexity of task, and self-confidence (see Appendix F, p. 391). Furthermore data about completion-time, mode times, and errors were recorded.

6.3.5 Measures

In this section the different measures and their calculation are explained. During the scenario each participant encountered six targets, each encounter represents an event. Depending on the robot's and participant's behaviour these events were categorised. The categorisation depended on the robot mode they used (auto or manual), whether the robot or the human

found the target, and whether the robot or the human were responsible to find the target. If the robot was in auto mode it was responsible for finding the target but in manual mode the human was responsible. Furthermore when the robot missed a target, the human was responsible for correcting that mistake. Also, it was important whether the human was aware of the robot's mistake or not. An overview of the events and the corresponding category number is provided in Table 25.

Event category	Mode	Robot	Human	Awareness of mistake	Respon-sibility
①	Auto	Found	Acknowledged	Yes	Robot
②	Auto	Missed	Found	Yes	Robot/Human
③	Manual	-	Found	Yes	Human
④	Auto	False target	Acknowledged	Yes	Robot/Human
⑤	Auto	Missed	Missed	Yes	Robot/Human
⑥	Manual	-	Missed	Yes	Human
⑦	Auto	Missed	Missed	No	Robot/Human
⑧	Manual	-	Missed	No	Human
⑨	Auto	False target	Missed	Yes	Robot/Human
⑩	Manual	Manual/false target	Missed	No	-

Table 25 - Overview of event categories

How the performance measures with the aid of these event categories were calculated is explained in the next section.

6.3.5.1 Observed[1] performance measure

The challenge with semi-autonomous remote controlled systems is that when participants supervise a robot in auto mode and the robot misses the target and the participants miss it as well (and are not aware of the robot's mistake), then the perception of the scenario is different to that intended by the researcher. Therefore two performance measures were introduced: observed performance and objective performance. The observed performance illustrates the actual witnessed performance by the participant. The objective performance demonstrates how many targets of the maximum possible targets were found.

Observed robot performance (Equation 1): How many targets the robot found in the trial (%) that the human was aware of. Not included are the number of targets found in manual mode because these were found by the human and not by the robot. N represents the number of times an event occurred.

$$\frac{\text{Event category } [\text{N}(①)]}{\text{Event category } [\text{N}(①)+\text{N}(②)+\text{N}(④)+\text{N}(⑤)+\text{N}(⑨)]}$$

Equation 1 - Observed robot performance

Observed human performance (Equation 2): How many targets the human found and responded to (%).Not included are the targets found in auto mode because these were found by the robot and not by the human.

$$\frac{\text{Event category } [\text{N}(②)+\text{N}(③)+\text{N}(④)]}{\text{Event category } [\text{N}(②)+\text{N}(③)+\text{N}(④)+\text{N}(⑤)+\text{N}(⑥)+\text{N}(⑨)]}$$

Equation 2 - Observed human performance

Observed team performance (Equation 3): This is the combined observed performance of both the robot and the human. Therefore this is the percentage of targets found that the human was aware of, whether the target was found by the robot or by the human.

[1] The observed performance refers to the experienced robot performance of the participant.

$$\frac{\text{Event category [N(①)+N(②)+N(③)+N(④)]}}{\text{Event category [N(①)+N(②)+N(③)+N(④)+N(⑤)+N(⑥)+N(⑨)]}}$$

Equation 3 - Observed team performance

6.3.5.2 Objective performance measure

The objective performance measure is independent of the awareness of the human and represent the percentage of targets found by the robot the human and both of them.

Objective robot performance (Equation 4): How many targets the robot found in the trial (%).

$$\frac{\text{Event category N(①)}}{\text{N (Maximum targets)}}$$

Equation 4 - Objective robot performance

Objective human performance (Equation 5): How many targets the human found in the entire trial (%). Event category (4) is included, because the human needed to acknowledge that the robot marked a false target. They did that by switching to manual mode and delete the last marker. However sometimes participants did not complete this procedure and just told the researcher that this was a mistake by the robot and there was no target; this represents event category (9).

$$\frac{\text{Event category [N(②)+N(③)+N(④)+N(⑨)]}}{\text{N (Maximum targets)}}$$

Equation 5 - Objective human performance

Objective team performance (Equation 6): This is the combined objective performance of both the robot and the human. In other words it represents how many targets were found in the entire trial (%).

$$\frac{\text{Event category [N(①)+N(②)+N(③)+N(④)+N(⑨)]}}{\text{N (Maximum targets)}}$$

Equation 6 - Objective team performance

Example

In order to better understand the concept, the following example in Table 26 presents a possible scenario with six events. Each event is categorised with the previously explained event categories (see Table 25). For the explanation of this example the encounter number is labelled A to F. A represents the first target encountered in the scenario and F the last target encountered.

Encounter no.	A	B	C	D	E	F
Event	Auto: Robot found	Manual: missed target	Manual: found target	Auto: Robot miss/ human found	Auto: Robot found	Auto: Robot miss/ human miss
Event category	①	⑧	③	②	①	⑦
Respon-sibility	R	H	H	R/H	R	R/H
Human aware-ness	Y	N	Y	Y	Y	N

Table 26 - Example scenario with six events

First the observed performances are calculated. For all the observed performance calculations, the encounter B and F were excluded because the human was not aware of these events/mistakes.

- Observed robot performance: The robot's responsibilities were the encounter A, D and E. The robot managed to successfully fulfil encounter A and E and failed in in encounter D, as shown in Table 26. Therefore the robots fulfilled 2/3 of its responsibility because it found 2 of 3 targets, which leads to an observed robot performance value of 66.67%.
- Observed human performance: The responsibilities of the human were the encounter C and D. Hence, the human successfully found encounter C and D. The participant fulfilled 2/2 of their

responsibilities, which calculates to an observed human performance value of 100%.

- Observed team performance: The human was aware of the encounters A, C, D, and E. All of these encounters were found, either by the robot or the human, therefore 4 of 4 targets were found which calculates an observed team performance of 100%.

Objective performance measures take into account all targets present in the scenario without any assigned responsibilities. In this example, there were six targets (encounters) overall.

- Objective robot performance: The objective robot performance was calculated from the robot's successful encounters, which in this example is the encounter A and E. Therefore the objective robot performance is measured against the maximum number of targets and the robot found 2 of 6 targets (33.33%).
- Objective human performance: The objective human performance includes only the targets found by the human, which is the encounter C and D. Thus the human found 2 out of the 6 possible targets, which leads to an objective human performance of 2/6 (33.33%).
- Objective team performance: The objective team performance comprises all successfully found targets divided by the maximum number of targets. In the example from Table 26 these are encounters A, C, D and E. The team therefore found 4 out of the 6 targets, which equals an objective team performance of 66.66%.

In addition to the main task the participants had to attend to a secondary task. The calculation of the secondary task performance is explained in the next section.

6.3.5.3 **Secondary task performance**

The secondary task appeared every 25 seconds. A participant with a short trial time encountered fewer secondary tasks and vice versa. Hence, the maximum number of secondary task encounters was counted as well as the successfully answered encounters. The secondary task performance divides the successful encounters by all encounters and delivers the percentage of successfully answered secondary task encounters (Equation 7).

$$\frac{\text{N (Successfully answered encounters)}}{\text{N (All secondary task encounters)}}$$

Equation 7 - Secondary task performance

6.4 Results

The results are divided into three sections. The first section discusses the pre-tests, which examined the correlation between the different trust questionnaires and the testing of correlations between the performance measures. In the second section the results of the participants who used the mixed mode robot are presented. The third section is about the differences between manual mode groups and the mixed mode group. The result section finishes with the description of the interview data.

6.4.1 Pre-test: Trust questionnaires comparison (Schaefer/Muir)

In order to select the appropriate questionnaire to measure trust in the subsequent experiments, two trust questionnaires were tested. The Muir (1989) questionnaire (see Appendix F, p. 391) has four statements that were answered on a 10-point scale from "Not at all" (1) to "Completely" (10) and was originally developed to measure trust into automation. The Schaefer (2013) trust questionnaire (see Appendix F, p. 391) consists of 40 questions starting with "What percentage of the time did the robot…" and could be answered from 0% to 100% with 10% intervals.

First the correlation between the two questionnaires was tested followed by testing the correlations of each trust questionnaire with the relevant study variables (see Figure 84). A Spearman's correlation test was used throughout to determine the relationship between the Schaefer trust ratings and Muir trust ratings. The trust scores for both questionnaires were not normally distributed and showed significant outliers. Therefore, another reason for selecting the Spearman correlation test was that it is rank based and robust to outliers (Croux & Dehon, 2010). In order to account for multiple correlation testing, the alpha value was adjusted to 0.01 to be

significant. Very weak correlations with a coefficient below 0.3 were excluded and greyed out in the figures.

There was a very strong, positive monotonic correlation between Schaefer and Muir (r_s = .81, [.730, .860], N = 117, p < .01). This suggests that the much shorter questionnaire from Muir, which entails four questions, showed similar data behaviour compared to the much longer 40 item trust questionnaire from Schaefer.

However, the correlation is solely rank based and the questionnaires were not similarly correlated to all other measures of the study variables. Trust correlations in Figure 84 present the differences between the trust questionnaires and the collected data. Both questionnaires were correlated with the programmed reliability level, observed robot performance, objective robot performance, manual mode time, and robot performance rating (see Figure 84) but the magnitude of correlation did differ. Muir's trust questionnaire correlated weakly with rated self confidence in the task. The Muir questionnaire seems more sensitive towards the subjective measures (rated self-performance and rated self-confidence) than the Schaefer trust questionnaire. The Schaefer questionnaire was weakly correlated with programmed reliability and objective robot performance. Both questionnaires were not correlated with the objective team performance (r_{sS} = -.04, N = 117, p > .05; r_{sM} = .04, N = 117, p > .05), which means that the questionnaires, as intended, focused more on the robot's performance rather than the end result of the trial. Both questionnaires showed that the more targets were found by the robot (objective robot performance) the higher was the trust in the robot. Some variables are not present, due to their collinearity with the other measures taken.

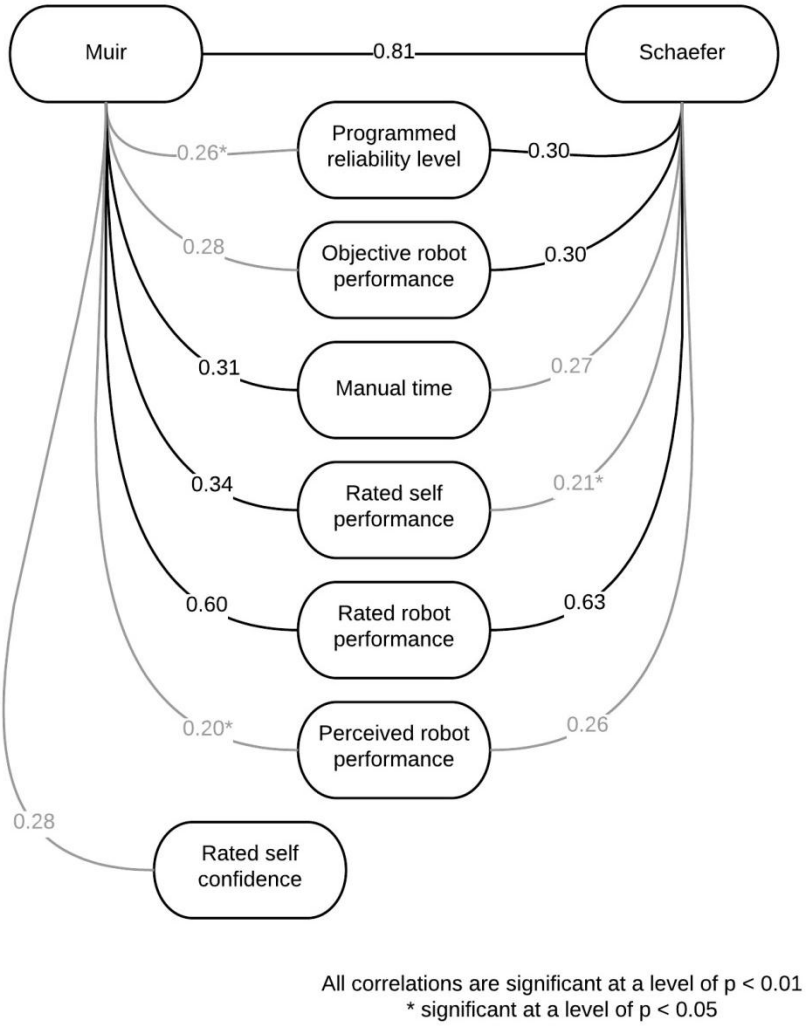


Figure 84 – Comparison of correlations with study variables of the Muir trust score and Schaefer trust score (weak correlations are greyed out)

Later analysis showed that robot reliability and task complexity had a significant main effect on the Schaefer trust questionnaire. A mixed ANOVA (outliers P31 and P33, excluded) with the Muir trust scores showed no significant differences, $F(2, 68) = .748$, $p=.748$. Although the Muir trust scores correlated with the Schaefer trust scores there was no significant effect of task complexity or robot reliability on the Muir trust scores. The results suggests that the Muir trust questionnaire is not as sensitive as the Schaefer trust questionnaire.

In order to test the independent variables and their effect on trust in more detail, the Schaefer trust questionnaire was selected for further use.

6.4.2 Mixed mode results

The influence of robot reliability and task complexity on trust, workload and performance as well as subjective ratings was tested. This section will present the results of the participants who used the mixed mode robot system.

In all post-hoc tests no Bonferroni correction was used. Since this small scale study has already low levels of observed power, the use of Bonferroni (Cabin & Mitchell, 2000) or sequential Bonferroni corrections (Holm, 1979) can further substantially reduce the statistical power and increase a Type II error (Nakagawa, 2004; Perneger, 1998). All variables were carefully selected to avoid performing more tests than necessary.

Although the programmed reliability was fixed in each condition and not a dependent variable, the fact that participants were free to choose between manual and auto mode gave a different robot reliability profile for each run for each participant. A detailed overview of the programmed values of the robot's reliability can be found in Appendix G (p. 396). This problem was also addressed with the observed performance measure (see 6.3.5.1 Observed performance).

6.4.2.1 Trust

The influence of robot reliability and task complexity was examined. In order to not violate the assumptions of ANOVA the extreme outliers, participants 19 and 26 were excluded from this analysis. A mixed ANOVA with post hoc tests was performed across task complexity and robot reliability.

The influence of robot reliability on trust

A significant main effect of robot reliability was found, $F(2, 34) = 3.66$, $p < .05$, $r = .32$. Post-hoc independent sample t-tests demonstrated that the significant difference occurred between the Schafer trust scores of high reliability ($M = 86.99$, $SD = 8.46$) and low reliability ($M = 75.91$, $SD = 12.84$) ($t(73) = 4.37$, $p < .001$, $r = .46$) and between middle reliability and low reliability ($t(73) = 2.31$, $p < .001$, $r = .26$) (see Figure 85). Between high and middle reliability the p value approached significance ($t(70) = 1.99$, $p = .05$, $r = .23$).

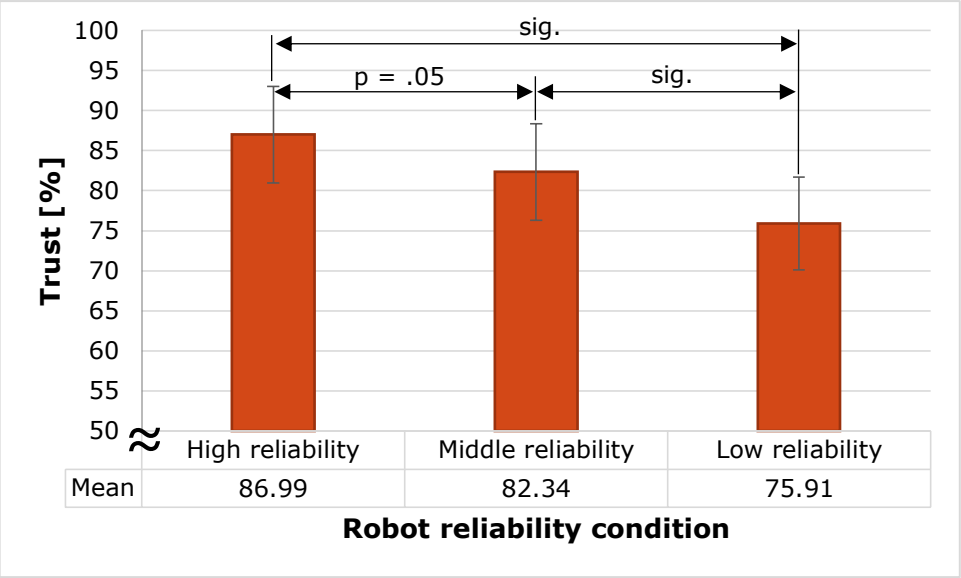


Figure 85 - Schaefer trust scores across robot reliability with 95% confidence intervals

The influence of task complexity on trust

A mixed ANOVA showed that there was a significant main effect of task complexity on trust, $F(2, 68) = 3.50$, $p < .05$, $r = .22$. Contrasts revealed that there was a significant quadratic trend, $F(1, 34) = 6.04$, $p < .05$, $r = .39$. This indicates an interaction with another variable and will be discussed later in this chapter (Section 6.5.2, p. 242).

Post hoc paired samples t-tests confirmed that there was a significant difference between low task complexity ($M = 82.81$, $SD = 10.06$) and middle task complexity ($M = 79.69$, $SD = 13.18$) ($t(36) = 2.93$, $p < .05$, $r = .44$). However, the differences in trust ratings between low complexity ($M = 82.81$) and high complexity ($M = 82.26$, $SD = 12.21$) ($t(36) = 0.46$, $p > .05$) as well as middle complexity ($M = 79.69$) and high complexity ($M = 82.26$) were not statistically significant ($t(36) = -1.73$, $p>.05$).

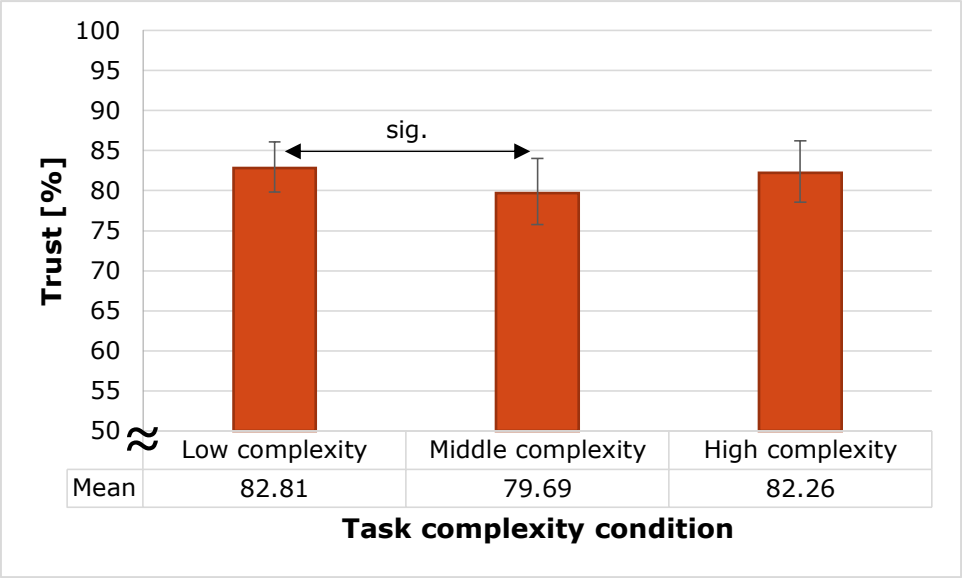


Figure 86 - Schaefer trust scores across task complexity with 95% confidence intervals

As shown in Figure 86 the middle task complexity led people to make a lower rating of trust in the robot compared to the low complexity condition. However, during the high complexity tasks participants rated the trust nearly as high as in the low complexity task.

The interaction of task complexity and robot reliability on the Schaefer trust questionnaire

A mixed ANOVA revealed there was no interaction effect between robot reliability and task complexity on trust, $F(4, 68) = 1.76$, $p > .05$). Therefore, another variable might have influenced the quadratic data trend of trust across task complexity levels.

6.4.2.2 Workload

Workload was measured after each trial with the NASA TLX. The effect of the independent variables is investigated in this section.

The influence of robot reliability on workload

The samples were not equally distributed, therefore a pairwise comparison with the Mann-Whitney test was used. A significant difference between the middle and low reliability levels ($U= 512$, $p < .05$, $r = -.28$) was found. The other comparisons were not significant (high to middle) $U= 624$, $p > .05$; (high to low) $U= 599$, $p > .05$.

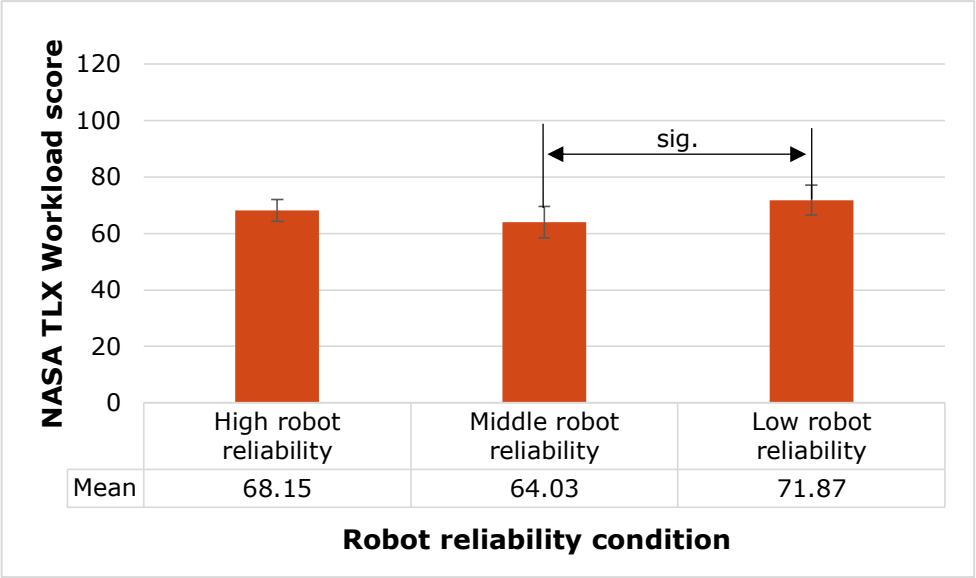


Figure 87 – Workload (NASA TLX) across robot reliability; with 95% confidence intervals

Participants experienced more subjective workload in the low reliability conditions, with a median of 71.87 (SD = 16.37) compared to the middle reliability conditions (M = 64.03, SD = 17.37). However, data showed only a small effect size of r = -.28. Figure 87 shows a bar chart of the workload ratings. A subscale analysis revealed that especially the physical demand and the frustration increased with lower robot reliability.

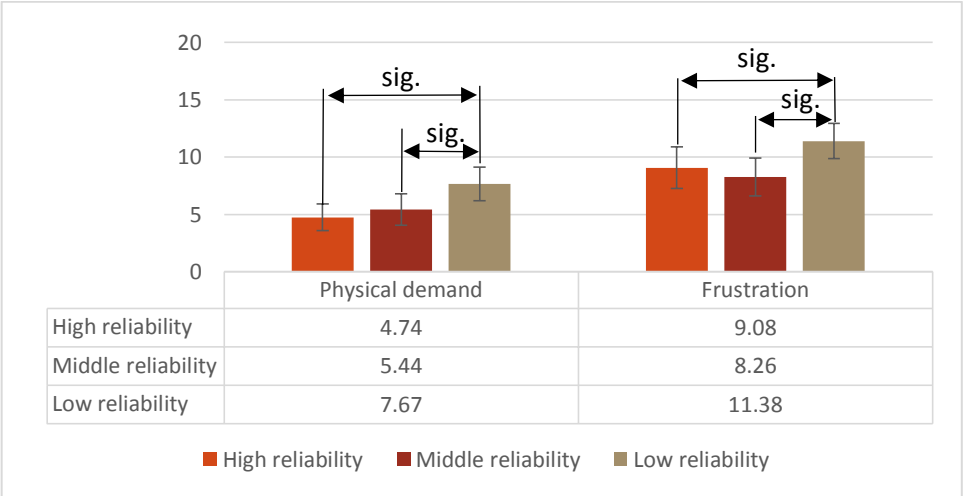


Figure 88 - Significantly different workload (NASA TLX) subscales with 95% confidence intervals

As shown in Figure 88, there was a significant increase in physical demand ($U = 517.5$, $p < .05$, $r = .28$) and frustration ($U = 491$, $p < .01$, $r = .31$) between the low reliability condition and the middle reliability condition. In

addition, there was a significant increase in physical demand (U = 442, p < .01, r = .36) and frustration (U = 564.5, p = .05, r = .22) between the low reliability condition and the high reliability condition. In general, the lower the reliability of the robot the more physical demand and frustration the participant experienced. The effect sizes were between small and medium.

The influence of task complexity on workload

The Friedman test revealed that there were no significant differences among task complexity levels regarding subjective workload ratings (χ^2 (2) = .842, p > .05). For values see Table 27. A subscale analysis showed no significant differences between the conditions.

Workload ratings across task complexity	
Condition	**Mean (SD)**
High task complexity	86.05 (17.04)
Middle task complexity	67.00 (15.32)
Low task complexity	69.00 (14.74)

Table 27 - Workload ratings across task complexity

6.4.2.3 Performance measures

6.4.2.3.1 Objective (team) performance

The objective team performance was the percentage of the targets found by both the robot and the participant. The influence of robot reliability and task complexity on the objective team performance is illustrated in the following paragraphs.

The influence of robot reliability on objective team performance

A non-parametric test was used because the data was not normally distributed. A pairwise comparison with Mann-Whitney tests (see Figure 89) showed significant differences between the high reliability level (M = 97%, SD = 9%) and the low reliability level (M = 88%, SD = 12%), U = 471.5, p = .001, r = -.39. This difference had a medium effect size. The differences between the high (M = 0.97) and middle reliability levels (M = 91%, SD = 12%) were also significant, but with a small effect size (U = 579.5, p < .05, r = -.27). There was no significant difference between middle (M = 91%)

and low reliability levels (M = 88%) (U = 650.5, p > .05). The data shows a declining trend: the lower the reliability of the robot the lower was the overall objective team performance. Teams scored significantly higher in the high reliability condition compared to the middle or low reliability condition.

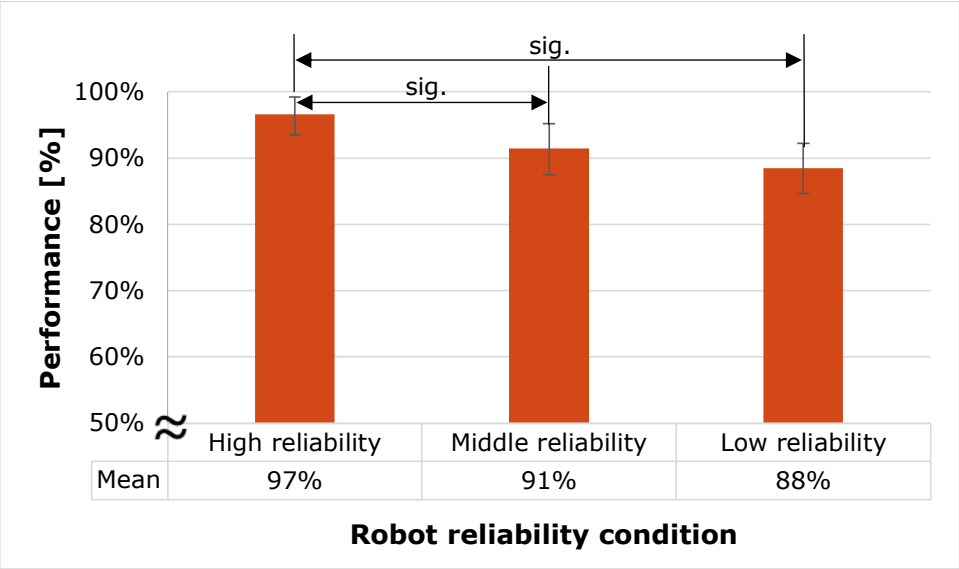


Figure 89 - Objective team performance across reliability levels with 95% confidence intervals (bootstrapped)

The influence of task complexity on objective team performance

The objective team performance changed significantly across task complexity. Wilcoxon signed-ranks tests revealed significant differences between middle complexity and high complexity (Z = -3.05, p < .05, r = -0.35), as well as low and high complexity (Z = -2.28, p < .05, r = -.26). The larger effect occurred between middle and high complexity (medium effect size).There was no significant difference between low and middle complexity (Z = -1.02, p > .05).

Therefore the human-robot teams had a significantly lower performance in the high complexity task (M = 88%, SD = 13%) compared to the middle (M = 96%, SD = 9%) or low complexity task (M = 93%, SD = 11%) (see Figure 90).

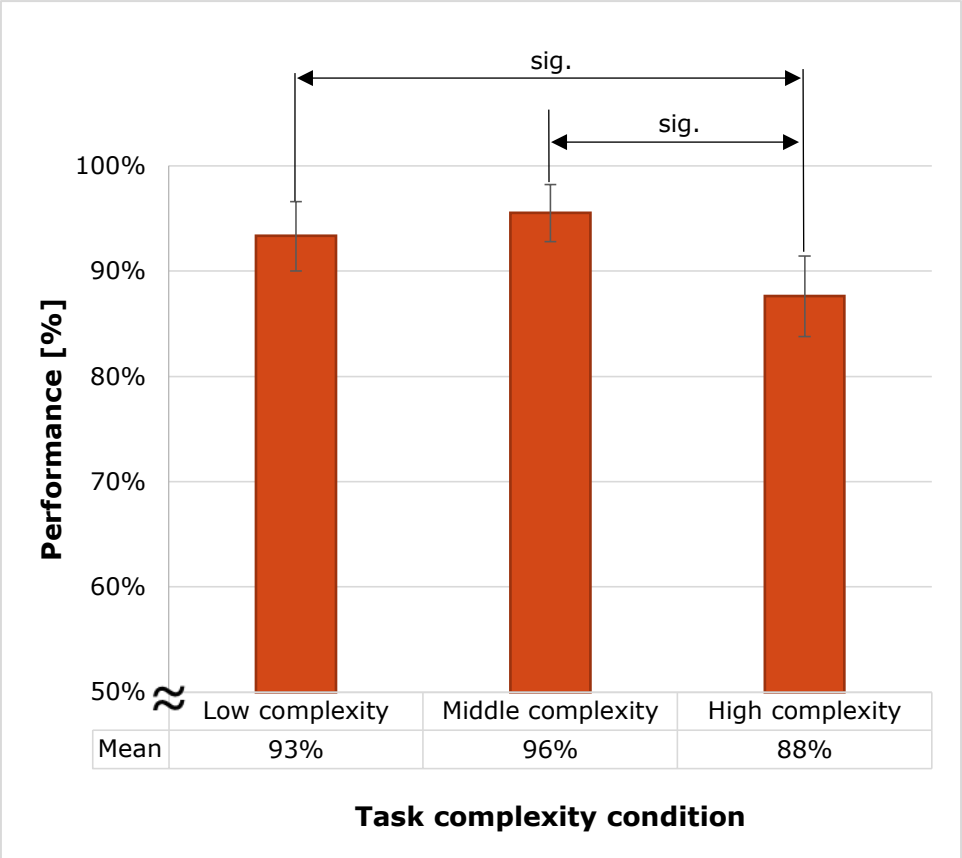


Figure 90 - Objective team performance across complexity levels with 95% confidence intervals (bootstrapped)

6.4.2.3.2 Observed robot performance

The observed robot performance was the performance of the robot as perceived by the participant, for example, the number of targets participants thought the robot found.

The influence of robot reliability on observed robot performance

A Kruskal-Wallis test showed that there was a highly significant difference in observed performance between the different robot reliability levels, $\chi^2(2) = 82.336$, $p < .001$, with a mean observed performance of 99% (SD = 4%) for high reliability, 81% (SD = 10%) for middle reliability and 61% (SD = 19%) for low reliability. A Mann-Whitney test revealed that all differences were highly significant: high to middle ($U = 179$, $p < .001$, $r = -.75$), middle to low ($U = 205$, $p < .001$, $r = -.65$) and high to low ($U = 27$, $p < .001$, $r = -.89$). All effects had a large effect size.

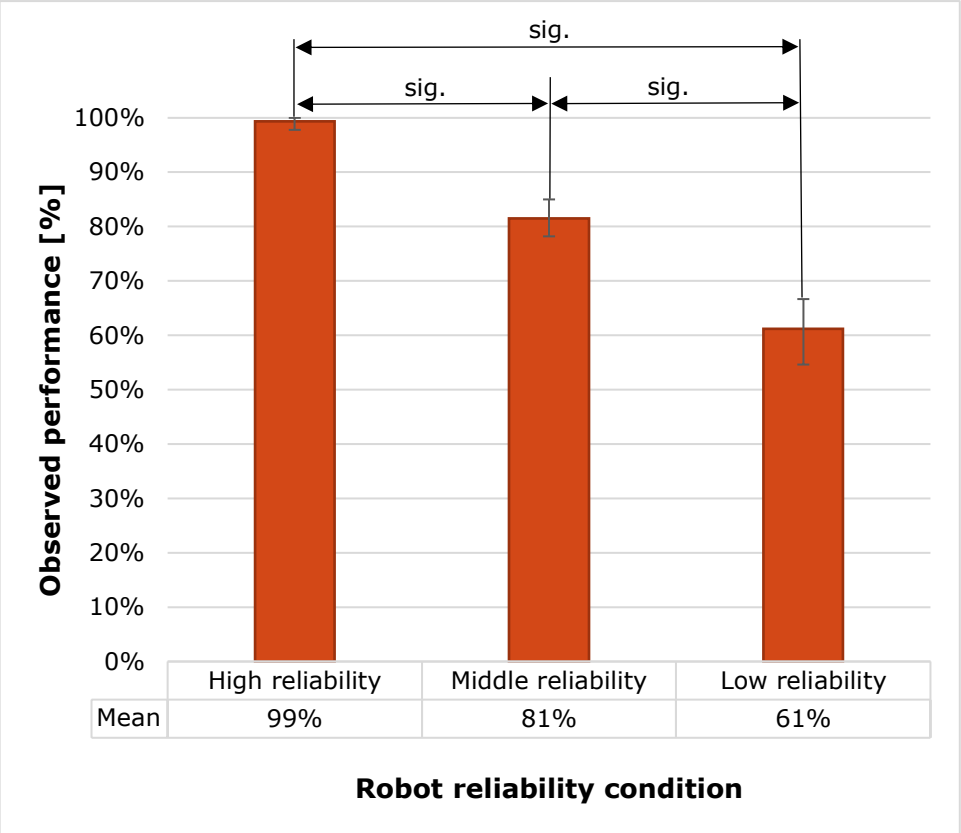


Figure 91 - Observed robot performance across reliability with 95% confidence intervals (bootstrapped)

Although there was the possibility that the participant could not observe the robot's performance due to using manual mode or missing a target too, Figure 91 shows that the observed robot performance declined significantly across the reliability levels: the lower the robot reliability, the lower was the observed robot performance.

The influence of task complexity on observed robot performance

A Friedman test indicated a significant difference across task complexity levels ($\chi^2(2) = 17.732$, $p < .001$). A multiple comparison with a Wilcoxon signed-rank test showed that there was a significant difference between high and low task complexity ($Z = -3.086$, $p < .01$, $r = -.35$) and middle and high task complexity ($Z = -2.351$, $p < .05$, $r = -.27$). There was no significant difference between low and middle complexity ($Z = -1.132$, $p > .05$). Figure 92 shows that the lowest observed performance was present during low task complexity with a mean of 77% (SD = 26%). The middle task complexity had an observed performance mean of 79% (SD = 22%),

which is lower than the high task complexity condition with a mean of 86% (SD = 15%).

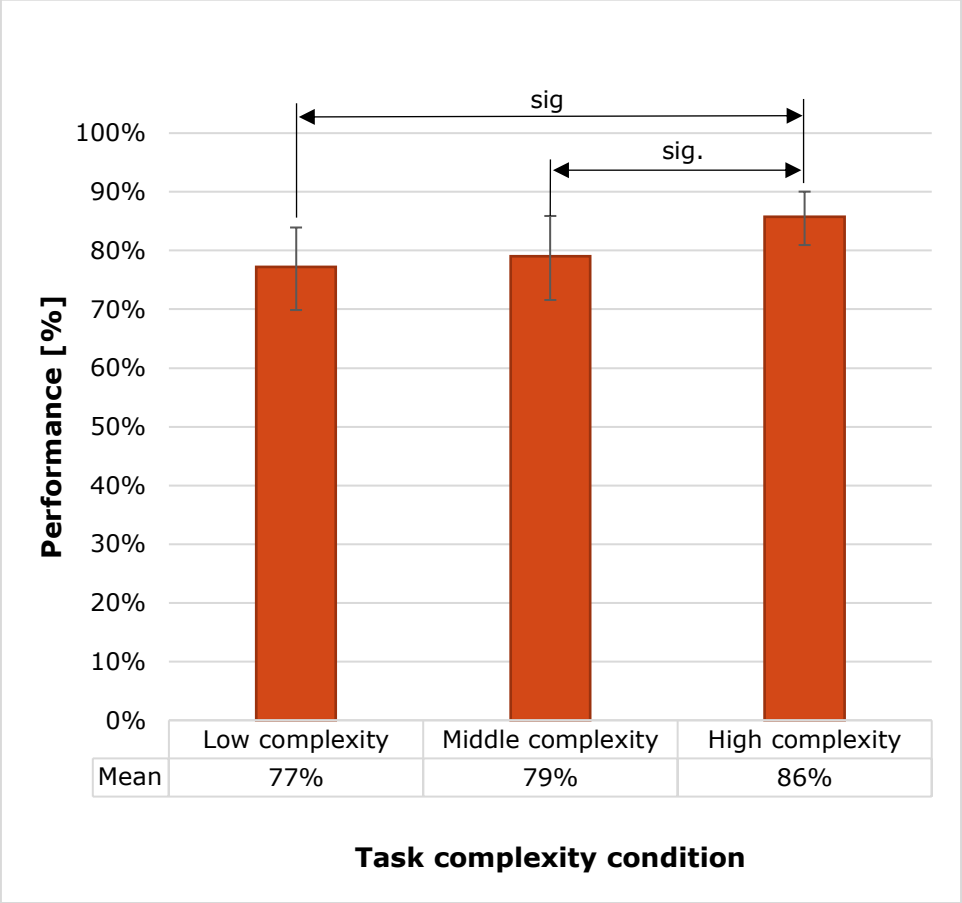


Figure 92 - Observed robot performance across task complexity with 95% confidence intervals (bootstrapped)

6.4.2.3.3 Secondary task performance

Secondary task performance data did not meet the assumptions of ANOVA, therefore non-parametric tests were used for the analysis. There was no significant difference found across reliability levels (χ^2 (2) = 3.755, p > .05). A Friedman test also showed no significant differences of secondary task performance across task complexity levels, χ^2 (2) = 2.258, p > .05.

6.4.2.4 **Manual mode usage**

The amount of time participants spent in manual mode was relatively low because participants were encouraged to use auto mode. A percentage value of time spent in manual mode was calculated and compared across robot reliability and task complexity.

The influence of robot reliability on manual mode time

There was a significant difference between high (M = 12%, SD = 13%) and middle robot reliability (M = 19%, SD = 11%), U = 445, p < .05, r = -.36. The second significant difference was found between the low (M = 25%, SD = 21%) and high reliability (M = 12%), U = 470, p = .01, r = -.33. The Mann-Whitney test did not reveal significant differences between the middle and low complexity tasks, U = 673, p > .05.

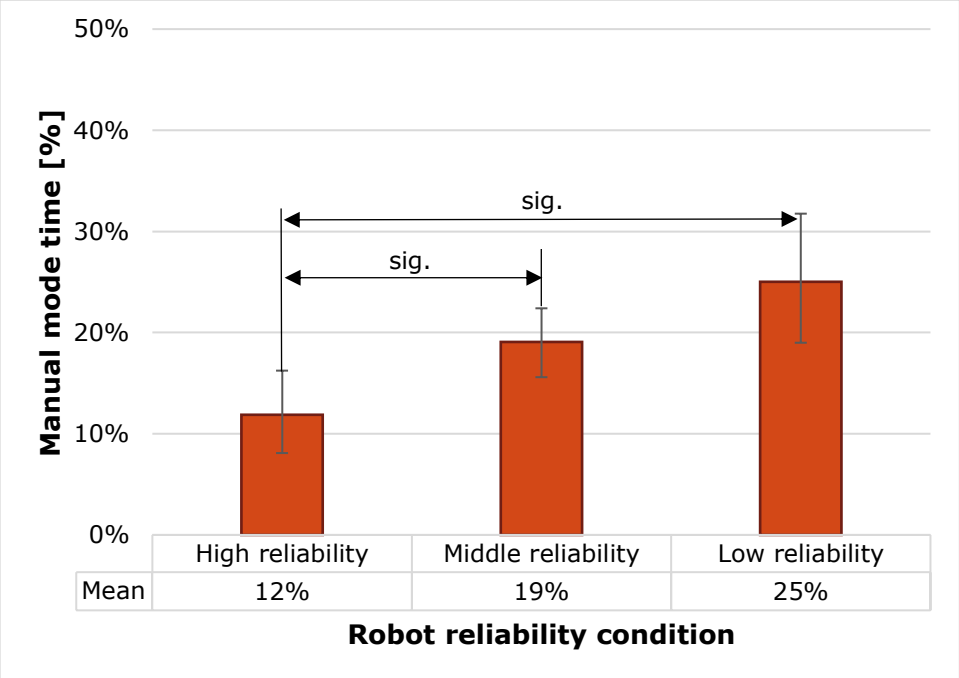


Figure 93 - Manual mode times across robot reliability with 95% confidence intervals (bootstrapped)

Participants used significantly more manual mode when the robot was less reliable. This was expected because participants needed to take over manual control when the robot made a mistake. Figure 93 shows this trend.

The influence of task complexity on Manual mode time

Across complexity levels a Friedman test showed no significantly different manual mode times, χ^2 (2) = 0.712, p > .05.

6.4.2.5 Trial times

Trial times are the percentage of time that the participant and the robot (the team) needed to complete the scenario compared to the time the robot would have taken to complete the scenario in exclusively auto mode

The influence of robot reliability on trial times

A Kruskal-Wallis test showed that the differences between trial times across reliability were found to be significant, χ^2 (2) = 20.235, p < .001. Post hoc tests (Wilcoxon signed-rank tests) determined that differences were significant between high robot reliability and middle reliability (Z = -4.227, p < .001, r = -.48), and high reliability and low reliability (Z = -3.555, p < .001, r = -.40). During high robot reliability participants needed 12% (SD = 13%) more time than they would have required in constant auto mode (see Figure 94). The trial times were higher for middle (M = 26%, SD = 18%) and low robot reliability (M = 28%, SD = 23%). This was expected since participants needed to intervene in lower reliability conditions with manual mode and possibly drive back and correct the robot or check areas the robot neglected.

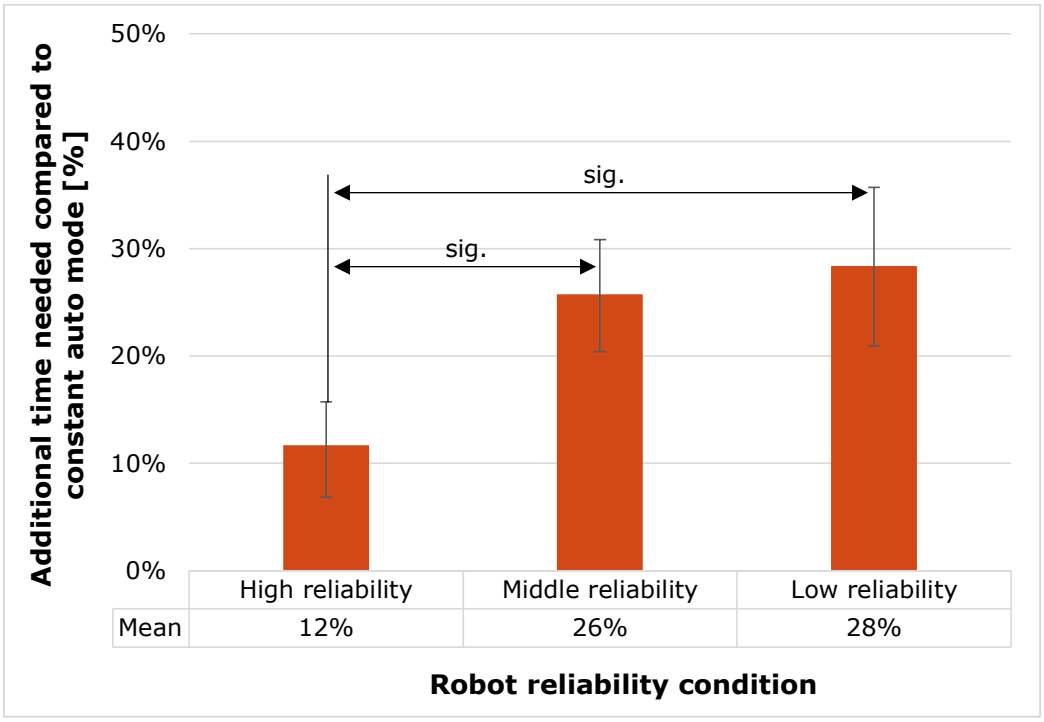


Figure 94 - Trial times across reliability with 95% confidence intervals (bootstrapped)

The influence of task complexity on trial times

There were no significant differences of trial times across the task complexity levels, χ^2 (2) = .384, p > .05. Participants took the same amount of time to complete the task regardless of the complexity of the task.

6.4.2.6 **Subjective ratings**

6.4.2.6.1 Rated task difficulty

Participants were asked to rate "How difficult did you perceive the task?" on a scale from 1 (extremely difficult) to 6 (not at all difficult). According to Liu and Li (2012) task difficulty focusses on the perception of the difficulty of the task by the performer. This will show if the manipulated task complexity level (objective task characteristics) were perceived as difficult and if low robot reliability is perceived as difficult as well.

The influence of robot reliability on rated task difficulty

A Kruskal-Wallis test showed that there were no significant differences in rated task difficulty across the robot reliability levels, χ^2 (2) = 2.74, p > .05.

The influence of task complexity on rated task difficulty

A Friedman test demonstrated that there was a significant difference between the rated task difficulty for the conditions, $\chi^2(2) = 9.23$, p = .01. A pairwise comparisons by using Wilcoxon signed-rank tests revealed that the significant effect was between the middle and high task complexity condition (Z = -2.287, p < 0.05, r = -.26) but only showed a small effect size. The difference between the low and high task complexity conditions approached significance with a p-value of 0.051 (see Figure 95 and Table 28).

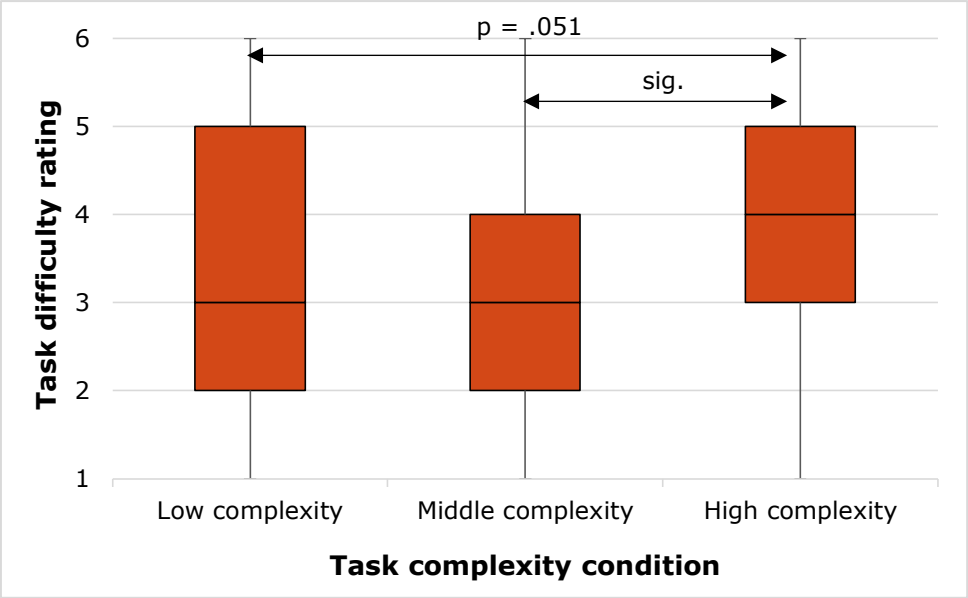


Figure 95 - Rated task difficulty box plots across complexity, whiskers (min/max)

Participants perceived the high complexity task (Mdn = 4) as more difficult than the middle (Mdn = 3) or low complexity task (Mdn = 3).

Rated task difficulty	
Condition	**Median (IQR)**
Low complexity	3 (3)
Middle complexity	3 (2)
High complexity	4 (2)

Table 28 - Rated task difficulty across complexity

6.4.2.6.2 Rated robot performance

The rated robot performance is the performance rating that participants gave the robot after the each trial. They were asked to rate the performance on a scale from 1 (poor) to 6 (excellent).

The influence of robot reliability on rated robot performance

There was a significant difference in rated robot performance between the different robot reliability levels, $\chi^2(2) = 16.46$, $p < .001$. Mann-Whitney tests found significant differences between high to low robot reliability conditions ($U = 381$, $p < .001$, $r = -.45$) and middle and low reliability conditions ($U = 528.5$, $p < .05$, $r = -.28$). Between high and low reliability the effect size was medium but between middle and low the effect was small. The rated robot performance between the high and middle robot reliability conditions was not significant.

The median rating of robot performance was 5 (IQR = 2) in the high reliability condition, 5 (IQR = 1) for middle reliability, and 4 (IQR = 1.5) for low reliability (see Figure 96 and Table 29).

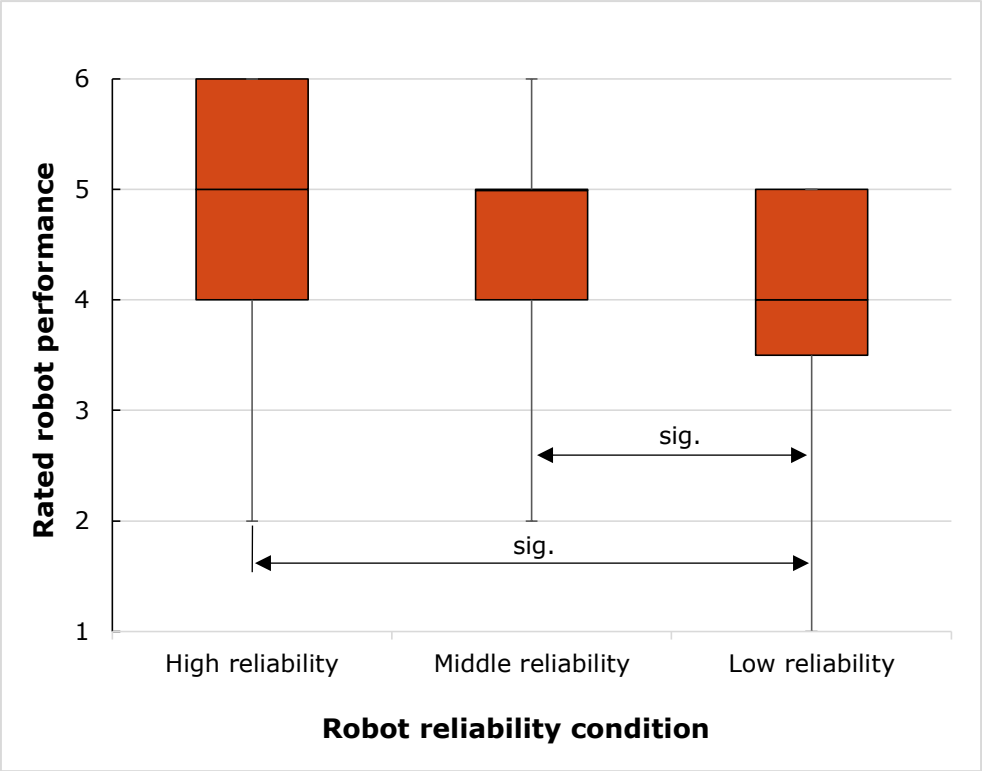


Figure 96 - Rated robot performance box plots across reliability; whiskers (min/max)

Rated robot performance	
Condition	**Median (IQR)**
High reliability	5 (2)
Middle reliability	5 (1)
Low reliability	4 (1.5)

Table 29 - Rated robot performance across reliability

As expected, the higher the reliability of the robot, the higher the participants rated the robot performance. Interestingly participants did not rate the high and middle reliability conditions differently, although the objective and observed performance for these conditions were significantly different.

The influence of task complexity on rated robot performance

According to a Friedman test there was no influence of task complexity on rated robot performance, $\chi^2(2) = 5.029$, $p > .05$.

6.4.2.6.3 Rated self-performance

After each trial the participants rated their self-performance on a scale from 1 (poor) to 6 (excellent).

The influence of robot reliability on rated self-performance

There were no significant differences in rated self-performance between the different robot reliability levels, $\chi^2(2) = 0.084$, $p > .05$.

The influence of task complexity on rated self-performance

A Friedman test showed that there was a significant difference in the rated self-performance across different task complexity conditions, $\chi^2(2) = 7.207$, $p < .05$. A pairwise comparisons using the Wilcoxon signed-rank test demonstrated that the significant effect was between the low task complexity condition (Mdn = 5, IQR = 1) and the high task complexity condition (Mdn = 4, IQR = 1.5), Z = -2.447, $p < .05$, r = -.28. However, the effect was small. All other pairings were not significant.

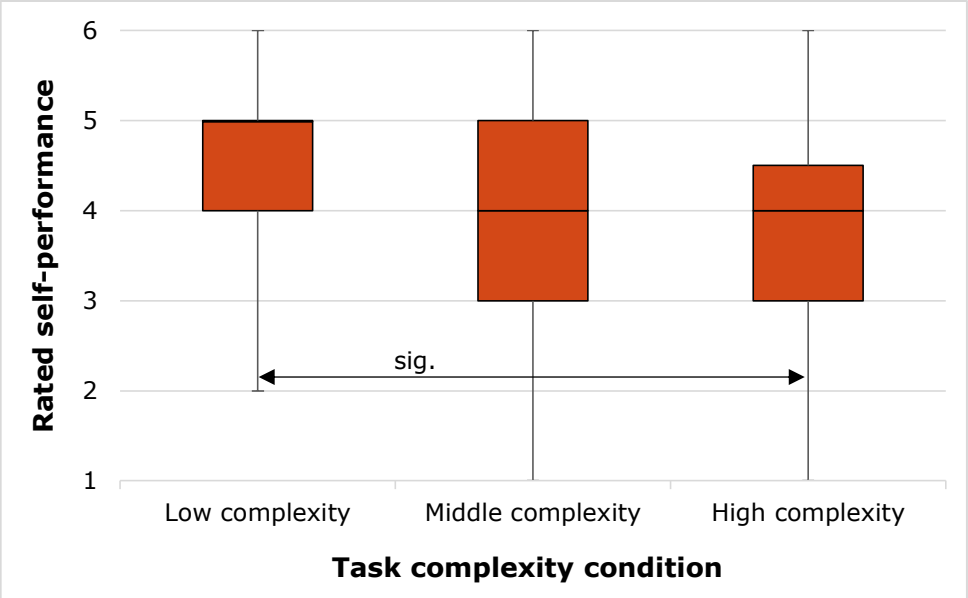


Figure 97 - Rated self-performance box plots across complexity; whiskers (min/max)

Rated self-performance	
Condition	**Median (IQR)**
Low complexity	5 (1)
Middle complexity	4 (2)
High complexity	4 (1.5)

Table 30 - Rated self-performance across complexity

Participants rated their performance significantly lower when the task was of higher complexity compared to a low complexity task (see Figure 97 and Table 30).

6.4.2.7 Personality score correlations

None of the personality scores of the Big-Five questionnaire (Gow et al., 2005) were correlated to the main variables. However, gaming experience was positively correlated with the objective human performance (r_s = .32, $p < .01$) and negatively correlated with the rated task difficulty (r_s = -.30, $p < .01$). Gaming experience was weakly correlated with the rated self-confidence in the task (r_s = .28, $p < .01$). The higher the participant's gaming experience, the better was their human performance, the less difficult they rated the task and the higher was their self-confidence in the task. Although it seems gaming experience can enhance the performance of human-robot teams, it needs to be considered that this is a virtual desktop study and the familiarity of virtual environments and gaming elements, such as the Xbox gamepad, could explain this correlation.

6.4.3 Summary of mixed mode results

Table 31 provides an overview of the results obtained during the mixed mode trials of this study. Each dependent variable and their significant results are listed.

Dependent variable	Independent variable	Significance	Result details (effect size)
Trust	complexity	significant*	LC > MC (r = .44) Trust in the low task complexity condition was higher than in the middle task complexity condition.
	reliability	significant	LR < HR (r = .46); LR < MR (r = .26) Trust in the low reliability condition was lower than in the high and middle reliability conditions.
Workload	complexity	no significance	
	reliability	significant	LR > MR (r = -.28) Participants experienced more subjective workload in the low reliability condition compared to the middle reliability conditions.
Objective performance	complexity	significant	LC > HC (r = -.26); MC > HC (r = -.35) Performance levels were higher in the low and middle task complexity conditions compared to the high task complexity condition.
	reliability	significant	LR < HR (r = -.39); MR < HR (r = -.27) Performance in the high reliability condition was higher than in the low and middle reliability conditions.
Observed robot performance	complexity	significant	LC < HC (r = -.35); MC < HC (r = -.27) Participants perceived the observed robot performance in the high task complexity condition as higher compared to the low and middle task complexity conditions.
	reliability	significant	LR < MR (r = -.65); MR < HR (r = -.75); LR < HR (r = -.89) The observed performance was between all conditions significantly different.
Rated task difficulty	complexity	significant	MC < HC (r = -.26) Participants rated the middle task complexity as less difficult as the high task complexity.
	reliability	no significance	
Manual mode usage	complexity	no significance	
	reliability	significant	LR > HR (r = -.33); MR > HR (r = -.48)

			Participants used less manual time in the high robot reliability condition compared to the low and middle robot reliability conditions.
Trial times	complexity	no significance	
	reliability	significant	LR > HR (r = -.40); MR > HR (r = -.48) Participants needed less time to complete the trial in high robot reliability conditions compared to the low and middle robot reliability conditions.
Rated robot performance	complexity	no significance	
	reliability	significant	LR < HR (r = -.45); LR < MR (r = -.28) Participants rated the robot performance in the low robot reliability condition as lower compared to the high and middle robot reliability conditions.
Rated self-performance	complexity	significant	LC > HC (r = -.28) Participants rated their self-performance higher during the low complex tasks compared to the high complex tasks.
	reliability	no significance	
* Result might have been influenced by the observed robot performance.			

Table 31 - Summary of mixed mode results

6.4.4 Comparison between manual and mixed mode results

The manual operating group (13 participants) used the same virtual environments and conditions from the middle reliability conditions as the mixed mode group from the previous section (Condition 4, 5, and 6, see Table 23, p. 188). Participants used the same experimental setup, process and interface. The only difference was that they were not able to use the robot in auto mode, instead they were only able to drive and mark targets manually. This section compared the mixed mode participants from the three reliability conditions from the main study with the 13 participants from the manual condition. This is particularly important because the following analysis shows at what point a semi-autonomous robot can contribute to a

higher performance or at what point it even contributes to a decrease in performance.

6.4.4.1 Workload

Mann-Whitney tests showed that there was no significant difference in Workload between the conditions (manual to high reliability, U = 691.5, p > .05; manual to middle reliability, U = 585, p = .05; manual to low reliability, U = 726, p > .05). Therefore participants did not experience a significant difference in workload between manual mode or any of the mixed modes (low; U = 725.5, p > .05, middle; U = 583.5, p > .05, high reliability robot; U = 691.5, p > .05).

6.4.4.2 Objective team performance

A between subjects t-test (equal variances not assumed and 1000 samples bootstrapped (Field, 2013)) showed a significant difference between the not normally distributed datasets of the manual group and the high robot reliability group, t(53.14) = -2.97, p < .05, r = .38.

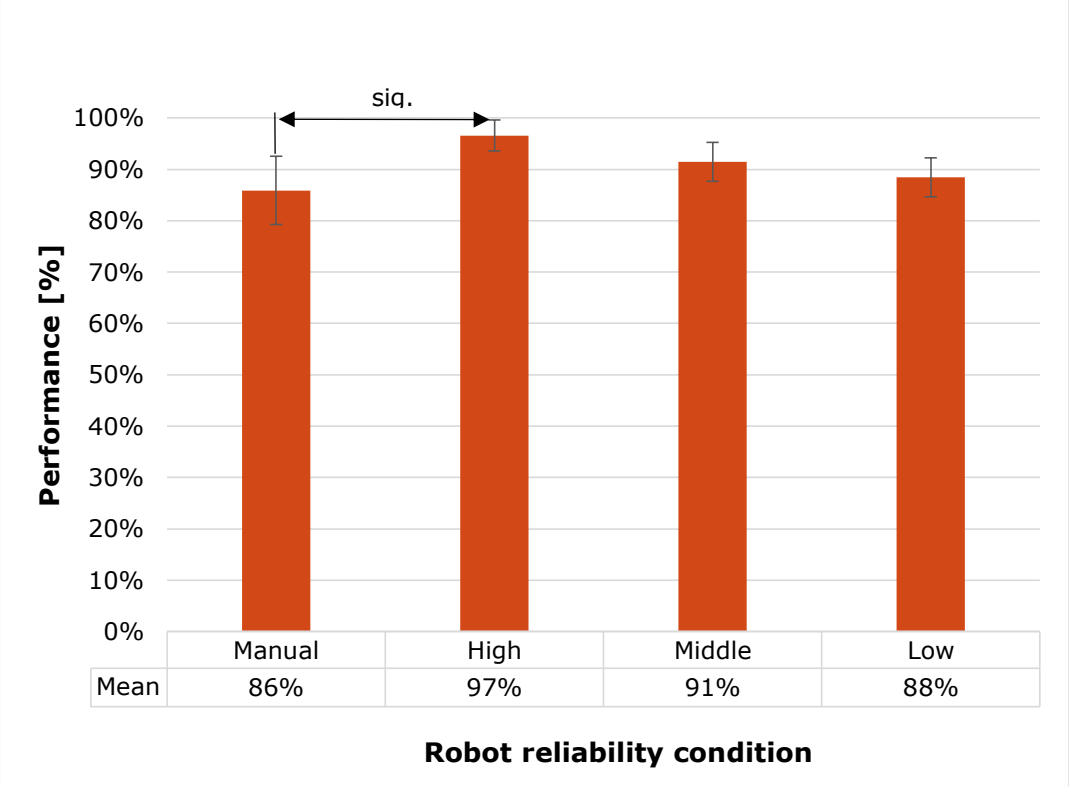


Figure 98 – Comparison of Objective team performance between manual a with 95% confidence intervals (bootstrapped)

As shown in Figure 98 the manual mode produced a mean of 86% (SD = 20%) and the high robot reliability with a mean performance of 97% (SD = 9%). Participants who used the 100% percent reliable robot achieved a significant higher (medium effect size) objective performance than the participants who had to drive manual mode only.

In addition the performance was tested by using independent samples t-tests across the task complexity levels. There were no differences in objective team performance between low (93%), middle (96%), or high task complexities (88%) compared to the manual performance (86%).

6.4.4.3 **Secondary task performance**

An independent samples t-test (equal variances not assumed and 1000 samples bootstrapped) showed a significant difference between the manual group and the low reliability group ($t(55.78) = -2.84$, $p < .05$, $r = .36$) and the manual group and the middle reliability group ($t(49.8) = -3.06$, $p < .01$, $r = .40$). The manual group had a mean secondary task performance of 55% (SD = 33%), which was much lower than the mean performance of 72% (SD = 13%) in the middle reliability and 72% (SD = 16%) in the low robot reliability group (see Figure 99). These findings suggest that in middle and low robot reliability the participants had the capacity to correctly answer more secondary task questions compared to the manual group.

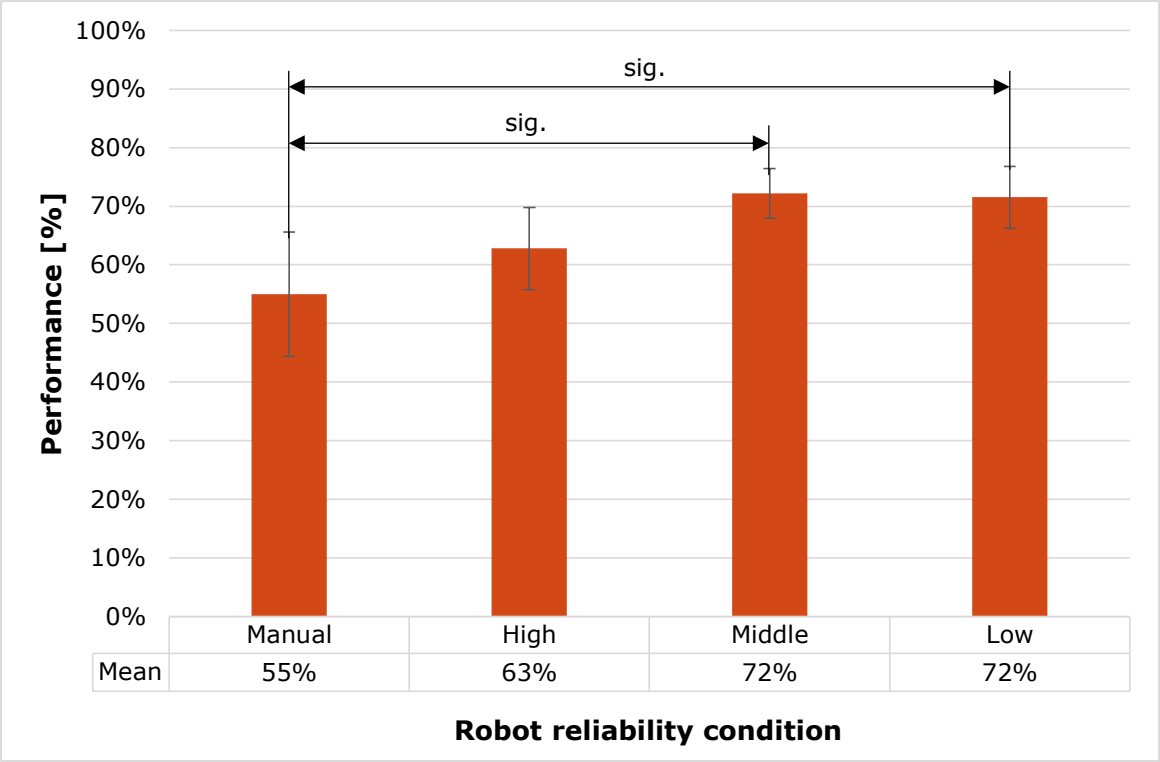


Figure 99 - Comparison of secondary task performance between manual and mixed mode group with 95% confidence intervals (bootstrapped)

This was expected, because driving a robot manually and answering counting questions on another keyboard is quite challenging. However, it seemed that high robot reliability did not allow participants to answer significantly more secondary task questions compared to the manual group. Which suggests that a high reliability robot might take up nearly as much attention as if participants steered the robot manually.

6.4.4.4 Trial times

Several between subjects t-tests (equal variances not assumed and 1000 samples bootstrapped) revealed that there was a significant difference between the manual robot group and the high robot reliability group, $t(55.12) = -4.428$, $p = 0.001$, $r = .51$) with a large effect size. The manual group needed 34% (SD = 35%) more time to complete the trial. Compared to that the group who used the high reliability robot only needed on average 12% (SD = 13%) longer. Therefore participants steering the robot manually needed significantly longer to complete the task than participants with 100% reliable robot. The significance is depicted in Figure 100.

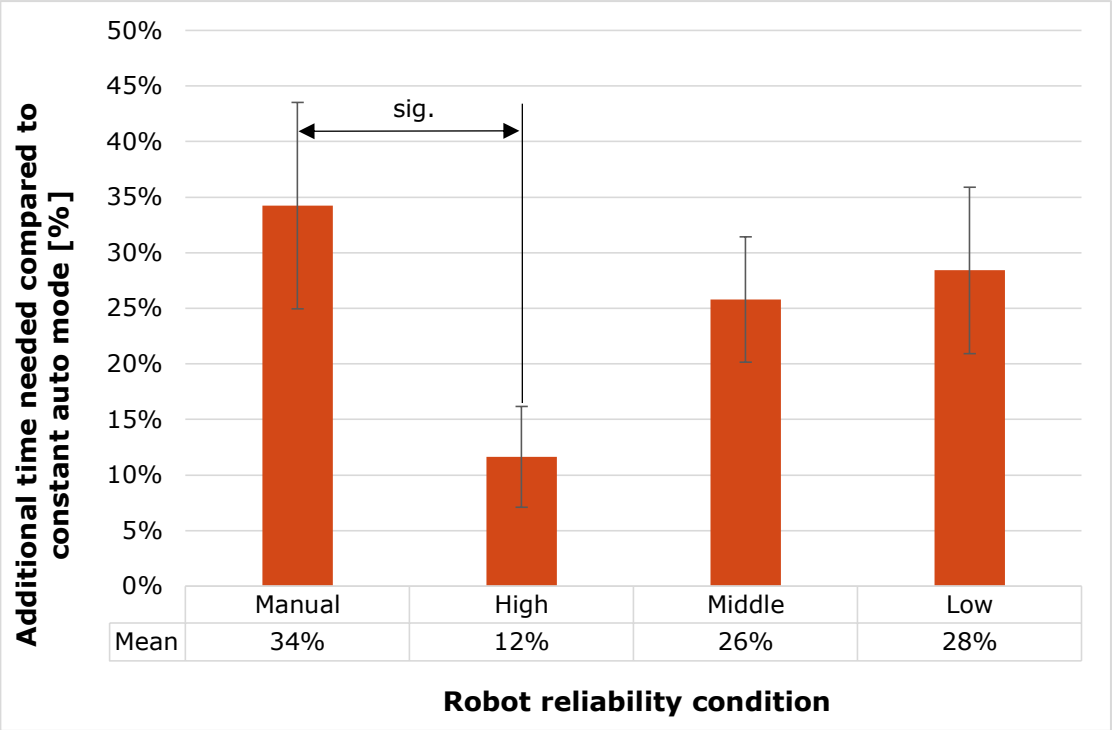


Figure 100 - Comparison of trial times between manual and mixed mode group with 95% confidence intervals (bootstrapped)

6.4.4.5 Rated task difficulty

Several Mann-Whitney tests tested the high middle and low robot complexities against the manual condition. Data showed that there was a significant difference between the manual participant group and the low robot reliability group, $U = 559$, $p < .05$, $r = .23$). Participants did experience a significant higher task difficulty when interacting with a low reliability robot (Mdn = 4) compared to driving the robot entirely manual with a median of 3 (see Figure 101 and Table 32). But this difference had a small effect size.

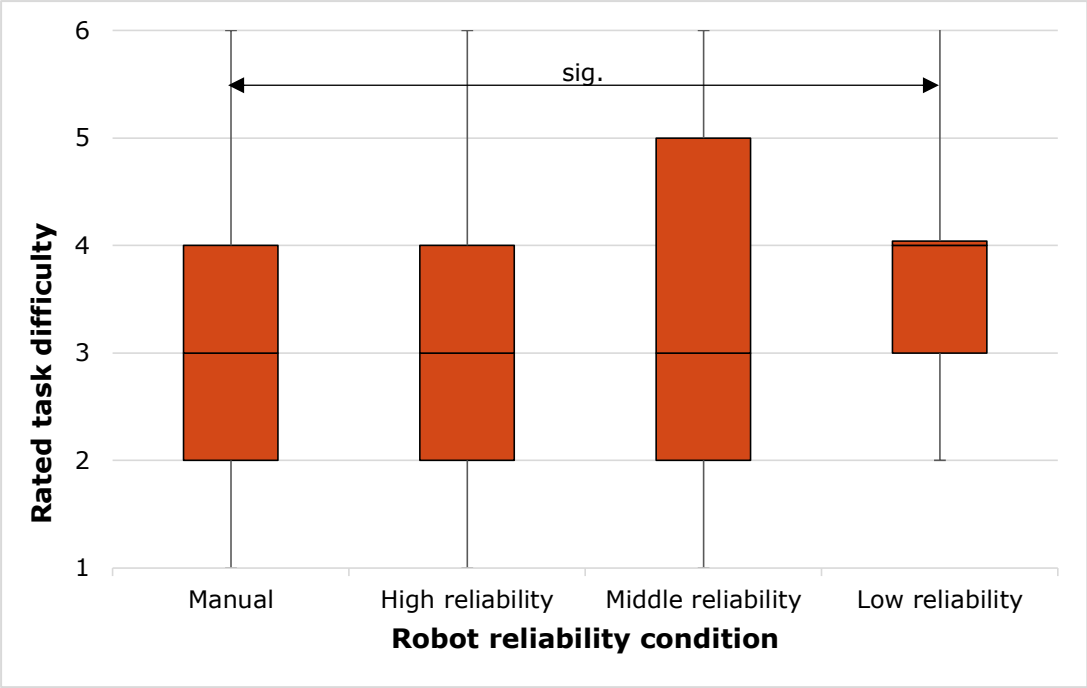


Figure 101 - Rated task difficulty box plots across reliability levels and manual mode, whiskers (min/max)

Rated task difficulty	
Condition	**Median (IQR)**
Manual	3 (2)
High	3 (2)
Middle	3 (3)
Low	4 (1)

Table 32 - Rated task difficulty table across reliability and manual mode

6.4.5 Interviews

6.4.5.1 Mixed mode: Mode switching behaviour

After a participant performed all three conditions they were questioned about their experience in a semi-structured interview. With the aid of a theme based content analysis (TBCA) (Neale & Nichols, 2001), the themes of the answers are visualised from Figure 102 to Figure 106. The full transcript can be found in Appendix K - - Digital Appendix V (p. 404). The numbers in squared brackets in the figures and in the text indicate how often this theme was mentioned by participants. If a quote is provided the participant number and the related robot reliability condition is mentioned

in brackets (LR = low reliability/MR = middle reliability/HR = high reliability). The quotes are representative examples of the individual theme mentioned. A theme is indicated in the text in *italic*.

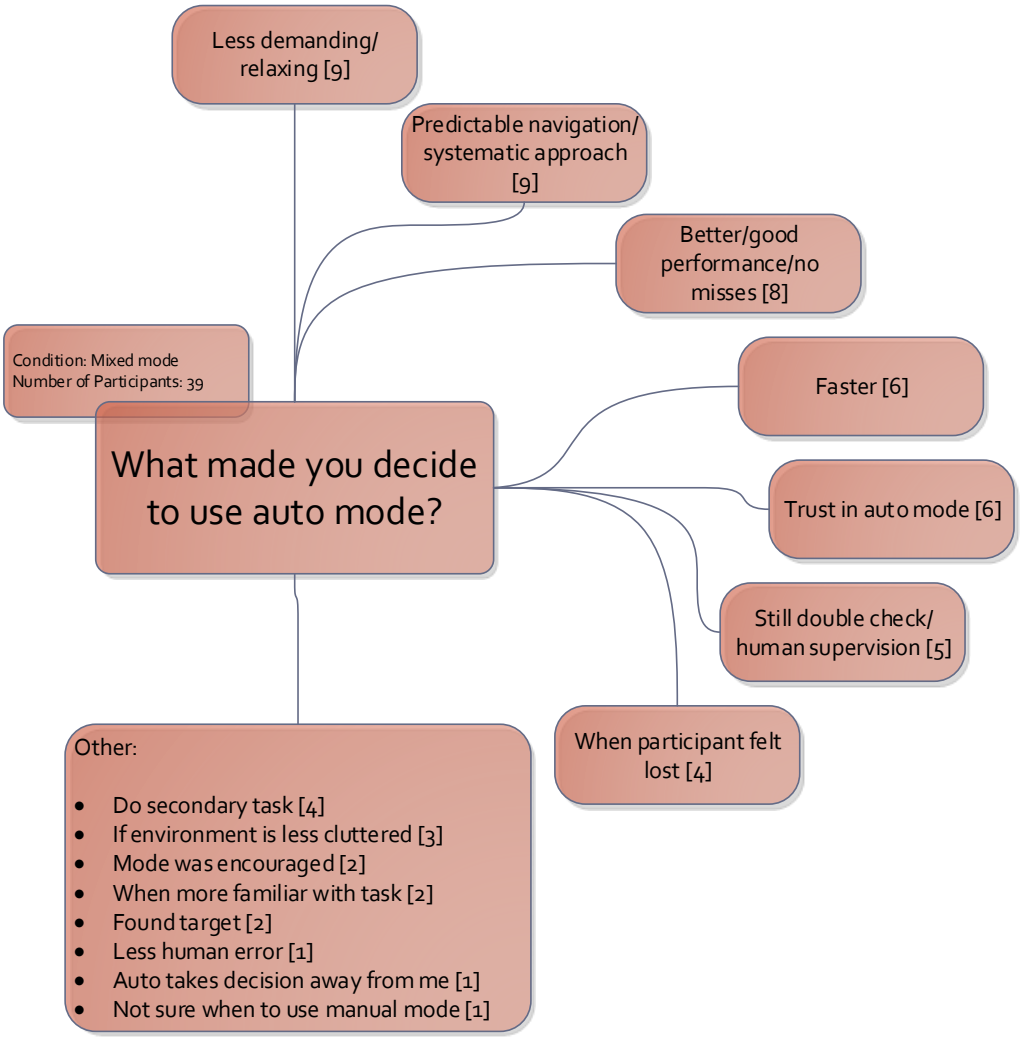


Figure 102 - TBCA of the question as to why participants used auto mode, with item count in brackets

Most participants said that using auto mode made them feel *more relaxed* and the task was *less demanding* [9]. For example:

- "I went for auto mode because it required me to do less, […] I didn't have to waste my time controlling through the whole scenario. But I could, out of the corner of my eye, make sure it was picking up on things." (P5, MR)

Participants were able to predict the movements and how the robot navigated (*predictable navigation*) [9], which made them stay in auto mode. One participant commented:

- "It moves at a predictable pace, and it does things more or less in the same pattern, every time. So you kind of know what it's doing, so yeah, if it misses something you can stop it and do what you need to do with it." (P10, HR)

Eight participants mentioned that the *robot performance in auto mode was better* [8]. As expected these themes emerged mostly from participants in the high reliability condition [5], than in middle [2] or low reliability conditions [1]. Others commented that it would be *faster*, than them having to steer the robot [6]:

- "I saw that the robot was navigating through the rooms, as I would expect it to. When it was finding a target, it was detecting it correctly, so I didn't feel the need to tinker with the robot's correct functioning. And I thought I couldn't do better than what it was doing on its own, so I would be slower because I would have to use a button to align it and then press it; I might even make a mistake and pressing the wrong button." (P16, HR)

However, across all reliabilities participants mentioned that there was still a need to *double check* what the robot was doing [5]. Four participants used auto mode when they *got lost* and let the robot navigate.

With respect to the interview data it can be inferred that automatic robot features can reduce subjective workload (*less demanding/relaxed*). *Predictability* was an important factor for using auto mode. Although, some participants thought that auto mode was *faster* and had *better performance*, they still believed that the robot needed *human supervision*.

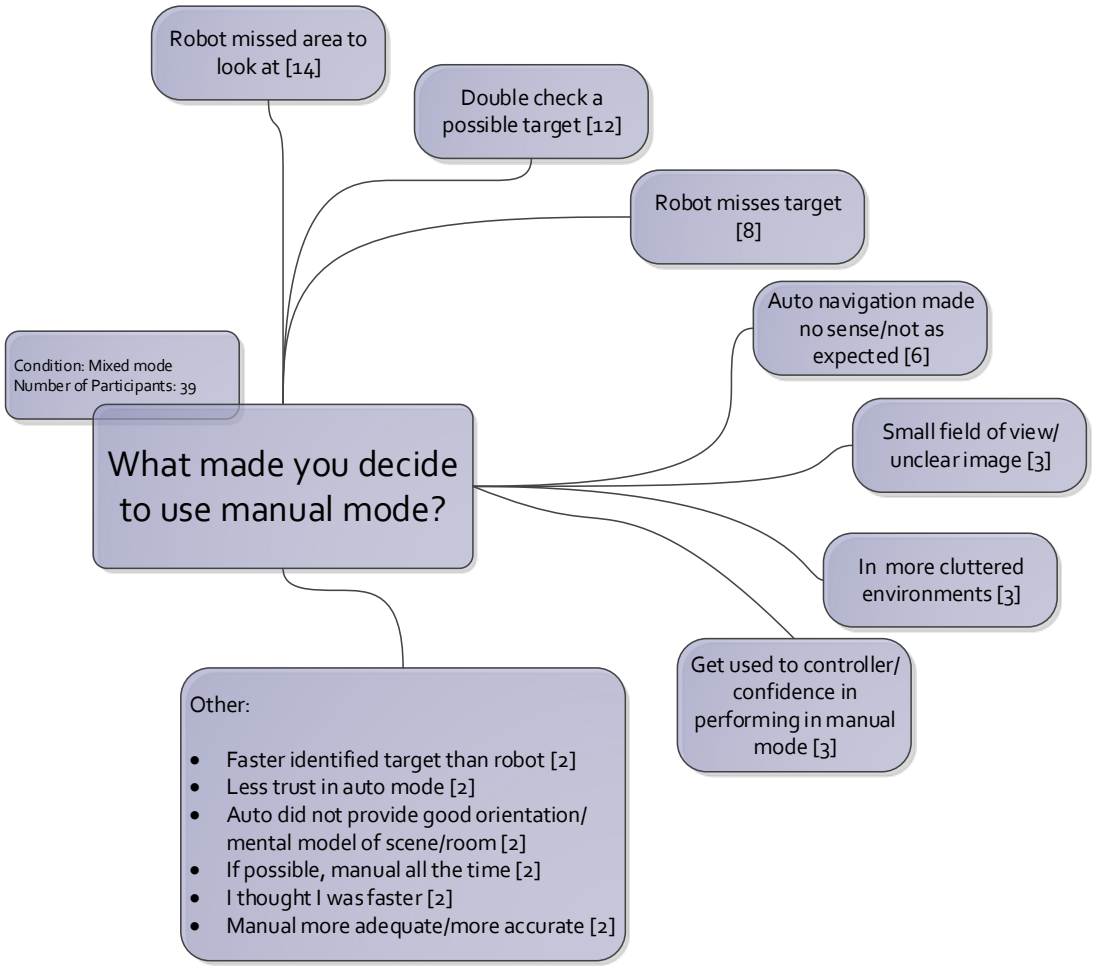


Figure 103 - TBCA of the question as to why participants used manual mode, with item count in brackets

Figure 103 depicts the reasons why participants went into manual mode. Obviously the participants took over in three main situations: when the *robot missed an area* [14], a participant thought that there was a possible target and wanted to *double check* [12], and when the *robot missed a target* [8]. The following quote underlines this:

- "[I used] manual when I felt that either, well, in two situations really, one when I didn't think it got into the corners properly enough, and the other one when I noticed something of the corner of my eye, that it might have missed, […]. When there wasn't a marker coming up any time soon, so it probably haven't picked up on it, so I just went into manual, marked it, go back into auto, keep going." (P8, MR)

For some participants [6], the *automatic navigation was not predictable* and the robot did not do what they expected; if that was the case participants

drove mostly in manual mode. Most of the participants mentioning that auto mode was unpredictable were assigned to the low reliability condition [4], rather than to the middle [1] or high condition [1]. The reason for that could be the fact that in the low reliability condition the robot only looked into corners or to the side infrequently because it was more often in low reliability zones. Likewise, only participants in the low reliability condition [3] mentioned a *small field of view* or an unclear image. Some participants [3] only used manual mode after they *got used to the controller* and more *grew confident in the task*. This means that a certain self-confidence is necessary to take over robot control.

The data suggests that if *the robot missed areas* [14], did not inspect objects properly or the participant had to *double check* [12], the *robot missed a target* completely [8], or lacked predictability (*auto navigation made no sense*) [6], participants used more manual mode.

A follow-up question asked participants if they thought auto mode was useful. 92% of the 39 participants thought auto mode was useful only 8% said no.

6.4.5.2 Mixed mode: The use of robot navigation goal points

This question explored what people thought about the map feature that showed the navigation goal points (NGP), which the robot used to navigate through the environment. The robot drove from point to point, at each point it had a look around. Figure 104 shows the map from the interface with the NGPs in orange (squares).

Figure 104 - Top view map from the robot interface with navigation goal points (orange squares)

Participants were asked whether they used the NGPs, if they were useful and for what reasons they were used (see Figure 105).

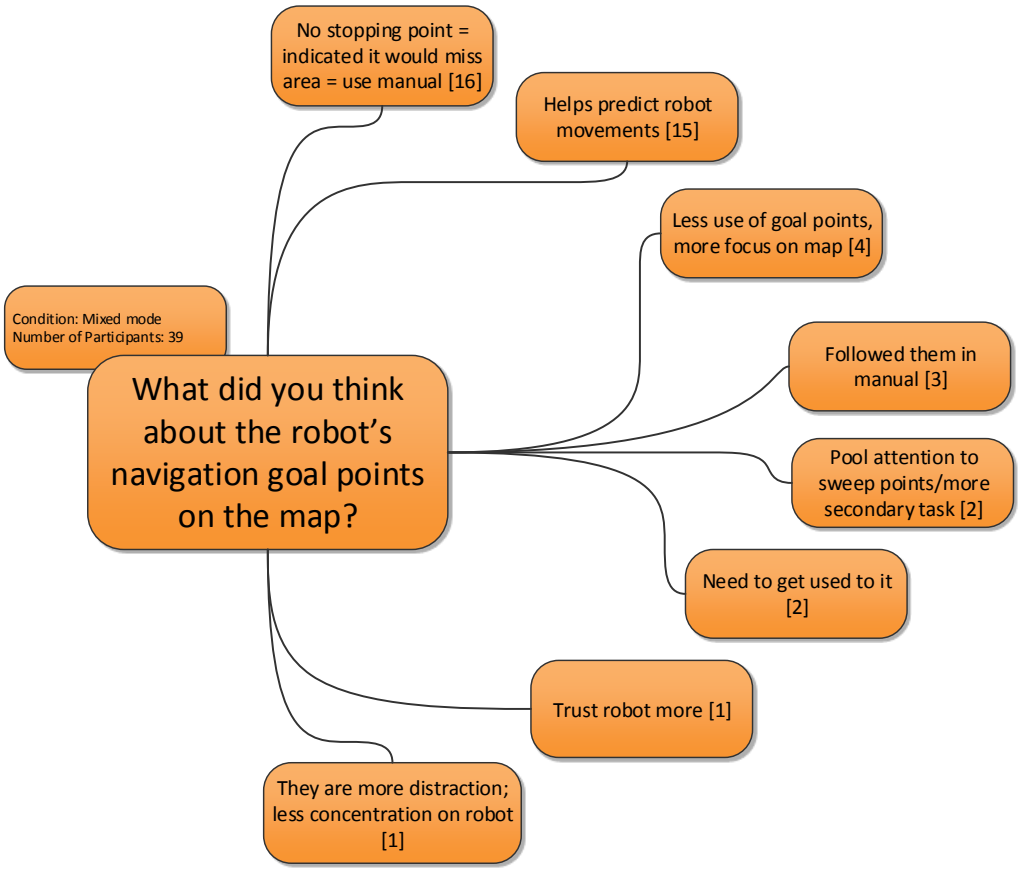


Figure 105 - TBCA of the question as to how participants used NGPs, with item count in brackets

Most participants [16] used the points to see whether they had to use manual mode, because for a longer period of time there was no NGP where

the robot would look around (N*o stopping point = indicated it would miss area = use manual*). For example, participant 5 explained:

- "It was good to know where the robot would stop and have a look around because if I looked on the map and there was an area where was no stopping point, I would switch back to manual and have a look around myself, just to make sure." (P5, MR)

Additionally, the NGPs provided the ability to *predict the robot's movements* [15]. Other participants were more *focussed on the map* [4] rather than the NGPs.

Two participants paid attention only when the robot was at a NGP in order to score more highly on the secondary task (*pool attention to sweep points/more secondary task*). One participant had the opinion that the points were more a *distraction* than helpful.

Overall, this map feature, seemed to be very useful for participants to predict the robot's behaviour and support participant's work, as well as foster better mode switching behaviour.

6.4.5.3 **Manual mode group: Trust ratings**

Since the manually operated robot had no decision making capabilities or any other automatic features (no target identification or automatic navigation), the participants in manual mode were asked what influenced them to give certain trust ratings. The indicator 'M' after participant numbers indicates 'manual' group participant.

Figure 106 shows that most [7] of the 13 participants mentioned that the trust they rated was more about *their own performance* rather than in the robot, for example one participant responded with:

- "[…] I just think, sort of like, actually controlling it, it is sort of is you, […]. You look through his eyes sort of thing. [...], it is how much you trust yourself a little bit." (P13M)
- "Based on whether I found a target." (P7M)

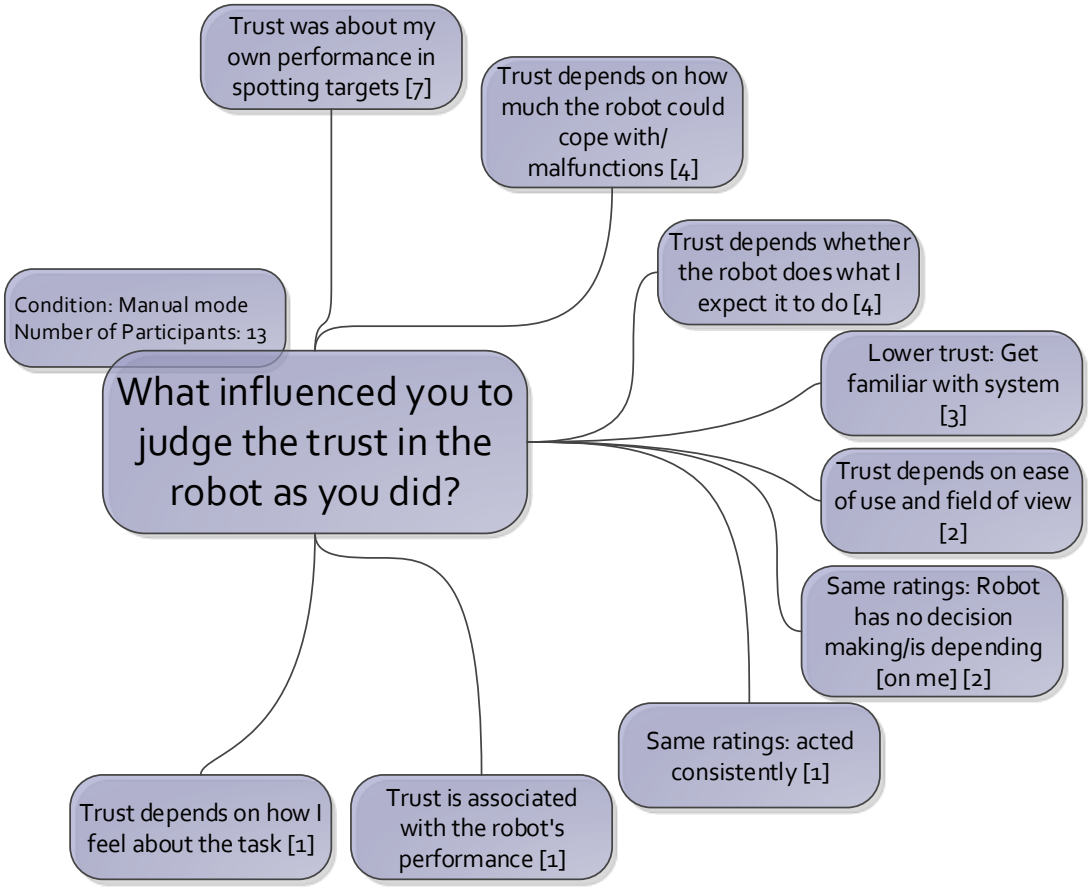


Figure 106 - TBCA of the question why participants gave different trust ratings, with item count in brackets

Another influencing factor of trust was how much the *robot could cope with the environment and not malfunction* [4]:

- "I think I just trusted it a bit more and when it flagged around things [driving around obstacles] and see how much it could cope with [the environment], how much it went up and down." (P3M)
- "I guess depending on the obstacles on the way, I would say the tighter spaces, where it couldn't get through or it got stuck, I trusted [the robot] less." (P12M)

If the *robot acted as the participants expected* it to do [4], their trust was higher. For example one participant said:

- "Because it acted consistently, so it didn't change my expectation of what I can expect from it." (P9M)

Three participants mentioned that their trust was lower because they were not yet familiar with the system (*lower trust: get familiar with system*).

Summarising the points made above, trust ratings in the manual group depended mainly on the subjective self-performance of the participant, since they were in charge of finding all targets and rated the trust according to how many targets they found. Further, the mechanical capabilities/reliabilities and the predictability were relevant in the rating of trust. It seems that the trust ratings towards robots with less autonomous behaviour reflected the trust of the participants' own ability more, compared to autonomous robots with decision making capabilities.

6.4.5.4 **Manual mode: Robot support**

The following list represents what type of support, features and sensors of the robot participants would have wished to have:

- *Target identification [6]*
- *Grabber for rubble or bomb disposal [5]*
- *Move camera independently from driving direction [4]*
- *Show possible moving grid [3]*
- *Torch [3]*
- *Gas detection [2]*
- *Heat map [2]*
- *Save image/screenshot [1]*
- *Second supervisor [1]*
- *Show covered areas [1]*
- *Flying [1]*
- *Fire extinguisher [1]*
- *Wider field of vision [1]*

Most participants wished for support in the target identification:

- "You know the camera, I think it can target people's faces automatically, and target it automatically, and give me some alarm saying this might be a target. This can be easier I think, because I need to do multi-task, then if it can help me to do something, than

can make the whole process more efficient, maybe I won't miss these targets." (P10M)

Furthermore they would have liked to have some sort of grabber in order to remove rubble or diffuse a bomb.

Four of the 13 participants in manual mode wished to be able to move the camera independently from the robot's driving direction. For example:

- "I would have liked to look up and down […]" (P1M)
- "So you could move and look into a different way." (P3M)
- "Vision could move independently from the direction of the robot, that way you can sweep the area faster because you can just like go straight but take a good look 360 [degree], all around the room." (P9M)

In order to navigate better in manual mode, participants suggested that the robot could show where it is able to drive or which areas were already covered. Other features mentioned were better robot light (torch), gas detection and a view of a heat map to locate targets easier.

6.5 Discussion

This study investigated the effects of task complexity and robot reliability on trust, workload, operator's perception of the robot, and team performance in the context of semi-autonomous Urban Search and Rescue robots. Qualitative interview data was collected to describe thoughts and behaviours of robot operators. Furthermore, in terms of their benefits for the rescue mission, remote semi-autonomous robot control was compared to exclusively remote manual robot control.

The discussion is divided into six parts, the first part is looking into the influence of robot reliability and the second part is concerned with the influence of task complexity. This is followed by a discussion about personality scores and rated task difficulty. The fourth part discusses the comparison with the manual operating user group. After that, the findings from the interviews are reviewed. The discussion will conclude with study limitations and future work required to advance this topic.

6.5.1 The influence of robot reliability on independent variables

Three different reliability profiles were used to investigate how the performance of the robot would influence the independent variables of this experiment. A reliability drop consisted of the robot navigating inaccurately (failing to look at certain corners/areas) and missing a target or identifying a wrong target. In high reliability the robot did not miss any of the targets. During middle robot reliability the machine had one reliability drop and missed a single target. The low reliability level consisted of two reliability drops and subsequently the robot made two errors.

The first two hypothesis stated that the robot reliability will influence trust, performances measures, subjective workload ratings, manual mode usage, and trial times. Additionally, the third hypothesis asked their influence on rated robot performance and rated self-performance.

There was a clear trend between high and low robot reliability levels. The less reliable the robot was the lower were the trust ratings. Although participants might have been unaware of some robot mistakes, which can lead to insignificant results among trust (Chien & Lewis, 2012), observed robot performance measures (observed performance refers to the experienced robot performance of the participant) showed a significant difference between all reliability levels. The positive influence of reliability on trust was expected, since robot performance is the main influencing factor on trust (e. g. Desai et al., 2012; Hancock, Billings, Schaefer, et al., 2011). Yet, this proves that the use of simulated rescue robot scenarios developed in UNITY can produce similar results compared to real-robot systems (e.g. Desai et al., 2012). This was also demonstrated by Robinette et al. (2015) who used a UNITY simulation to investigate the effects of robot performance on automation usage.

Regarding performance, the lower the reliability of the robot the lower was the objective team performance. Objective team performance was increased by six percent between middle and high robot reliability but the increase in performance between low and middle reliability was not significant. These results show that a good, or very good, working robot can

enhance the performance of a USAR mission but differences between lower reliability levels (between 62% and 78% robot reliability) did not show a significant increase in performance. These results are similar to de Visser and Parasuraman (2011) who found that imperfect automation had a performance benefit if the reliability was over 70%. In addition they found that it can have some benefits even if reliability is as low as 30%. However, the study presented in this chapter did not show how badly performing robot systems might influence the mission performance.

Between the middle reliability and low reliability condition workload increased significantly. Already 1999 Endsley and Kaber found that higher levels of automation produce lower workload ratings but this study showed that failures in automation (requires manual correction) can cause this benefit to vanish. This might have been influenced by the higher levels of workload due to more manual operation. It is known that using manual mode (teleoperation) produces higher levels of workload (see also Chen & Terrence, 2009). Therefore, a very good robot that made no mistakes was rated as similarly demanding as a robot that made one mistake. This finding is different to Desai (2012), where workload was significantly lower for the 100% reliable robot compared to situations where the robot made one reliability drop. That low and middle reliability workload was not significantly different could also be due to the characteristics of a vigilance task, where too high levels of sustained attention (supervising the robot) induce hypostress and result in high levels of subjective workload (cf. Bainbridge, 1983; Warm, Parasuraman, & Matthews, 2008). This could have been the case for the situation where the robot made no mistakes and participants did not need to interfere/take over manual control. Data was unlikely influenced by the secondary task, as suspected by de Visser and Parasuraman (2011) who claimed that during performing the main task, if workload dropped, participants might have used this extra capacity to perform the secondary task or use different strategies to allocate attention. Because the secondary task performance in the manual condition was not significantly different to the secondary task performance in the high robot reliability condition. Which supports the notion of high levels of sustained attention during high robot reliability conditions.

As mentioned previously, during low reliability participants used manual mode the most. This was a significant increase compared to the high robot reliability condition. However, this was expected because the robot in low reliability made more mistakes and therefore the participant needed to take over control more often to correct the mistakes. It is unclear if this continuous increase in manual mode usage (from high to low robot reliability) can be attributed to the decrease of trust levels. A more detailed analysis of mode switching behaviour and strategies would be necessary to investigate this matter further. A similar pattern was observed for the trial times, the more manual mode was used, the more time participants needed to complete the task. These findings are in agreement with Desai (2012).

After each trial participants had to rate how well the robot performed. The measure was a validation variable in order to check whether participants perceived a low or high robot performance induced by changing reliability levels. The rated robot performance showed a significant decrease from high to low robot reliability. This was expected and validates that the participants actually experienced the different reliability levels. However, there was no significant differences in the participants rated self-performance. This is different to the findings of Desai (2012) who found a significant decrease in rated self-performance between 100% reliability and a robot who had one reliability drop. But Desai also had significantly different rated levels of self-performance due to the length of interaction. Perhaps the time of interaction in this study was too little to show any effects of robot reliability on rated self-performance.

In conclusion, data showed that robot reliability influenced trust, workload, performance, manual mode usage, and trial times.

6.5.2 The influence of task complexity on independent variables

Task complexity was influenced by changing the required quantity and types of targets as well as the quantity and difficulty of the search environment. Low task complexity only required to find casualties in a fairly uncluttered environment. Middle task complexity consisted of finding casualties and hazard signs in a medium cluttered environment. The high task complexity

level introduced a third type of target to find: evidence for terrorist attacks. This target introduced uncertainty because it could be a weapon or a self-made explosive device, whose appearance is unknown. In addition the high task complexity environment was the most cluttered and obstacle rich environment in this study.

It was hypothesised that task complexity will influence trust, performances measures, subjective workload ratings, manual mode usage, and trial times. And indeed, participants rated the robot on lower complex tasks as more trustworthy than middle complex tasks. But participants also rated low and high complexity tasks as similarly trustworthy. This shape of the data was not expected and might have been influenced by other variables. An explanation for these unexpected trust ratings could be that the observed robot performance was significantly higher in high complexity conditions compared to low and middle complexity and showed a constant increase from low to high complexity. Therefore the highest perceived performance of the robot was during high task complexity. This could explain the high trust ratings in the high task complexity condition. Consequently, it is important to collect data of the observed robot performance (what the participant actually witnessed) and the objective team performance (how many % of all targets were found). While the observed robot performance increased with increasing task complexity, the actual objective team performance showed a different picture and decreased from low to high complexity. The fact that participants failed to see a robot's mistakes because they missed them too, led them believe that the robot was more reliable and they perceived the robot's performance as higher. During the experiments of Rovira et al. (2007) and Chien and Lewis (2012) it was suspected that because of the absence of alarms system failures were not detected and hence participants could not easily discriminate between low and high system reliability.

Nevertheless, the human-robot team performance (objective team performance) was significantly declining between low and high complexity and middle and high complexity. Therefore, teams performed worse when the task was highly complex. This is in agreement with recent literature. When Desai (2012) changed task complexity it showed that the more

complex tasks were, the worse was the performance with the semi-autonomous robot system. There was no significant performance difference between low and middle complexity. It could be that the difference between low and middle task complexity was not big enough to produce significant differences. Because the same pattern is shown in rated task difficulty where participants did not perceive a difference in task difficulty between low and middle complex tasks.

There were no significant differences across task complexity in terms of workload, secondary task performance, manual time or trial times. This means that participants, no matter how complex the task was, had similar levels of workload, secondary task performance, and they also needed the same time to complete the task, and used the similar amount of manual mode usage.

In addition the second hypothesis stated that task complexity will influence rated self-performance and rated robot performance. Rated self-performance varied significantly between low and high task complexity. The more complex a task was, the lower they rated their own performance. This is in accordance with Desai's (2012) experiment where he changed task complexity. But the rated robot performance showed no significant differences across the complexity levels, which is also in accordance with the studies from Desai (2012). This data indicates that participants did not think the robots performance declined when a task got more complex, which might be the effect of the perceived performance, which increased with complexity level. Participants attributed the decrease in performance to themselves by rating their own performance worse during high task complexity.

In hypothesis H5 it was hypothesised that task complexity and robot reliability will interact regarding trust. However data showed that there was no significant interaction effect.

6.5.3 Personality scores and rated task difficulty

It was further hypothesised that personality scores will correlate with trust, performance, subjective workload ratings, and manual mode usage. This was not the case. None of the Big-Five personality factors correlated with any of the data. Still, a significant correlation was found with respect to the gaming experience of participants. If participants were experienced in gaming they tended to have better performances and they rated the tasks less difficult as well as claimed to have a higher self-confidence in the task. It needs to be noted that this was a virtual desktop study and the search environment was very similar to a gaming environment (e.g. computer screen, Xbox controller, etc.). The similarity could have caused this significant correlation. For a real rescue scenario the skills of gamers might only partly support their interaction with the robot.

The seventh and eighth hypothesis declared that participants will rate more complex tasks more difficult and rate lower robot reliability more difficult as well. That participants will rate more complex tasks as more difficult could not be proven in this study. Although Liu and Li (2012) said that a complex task does not need to be difficult (but it is likely that it is), participants perceived the high complexity task as more difficult than the middle or low complexity task. But there was no significant difference between low and middle task complexity. Since objective team performance, observed robot performance and rated self-performance were not significant between low and middle complexity conditions it is comprehensible that rated task difficulty did not vary either. This fact is strengthening the suggestion that the independent variable was not strong enough to elicit significant differences between low and middle complexity. In terms of the hypothesis that participants will rate lower robot reliability as more difficult, participants did not indicate that the task was more difficult when the robot was unreliable. Therefore an additional task component, in this case steering the robot manually, made the task maybe more complex but not necessarily more difficult.

6.5.4 Combined discussion

The results showed that, next to robot reliability, task complexity is an important influencing factor on the number of targets found. Figure 107 qualitatively visualises the results of this experiment.

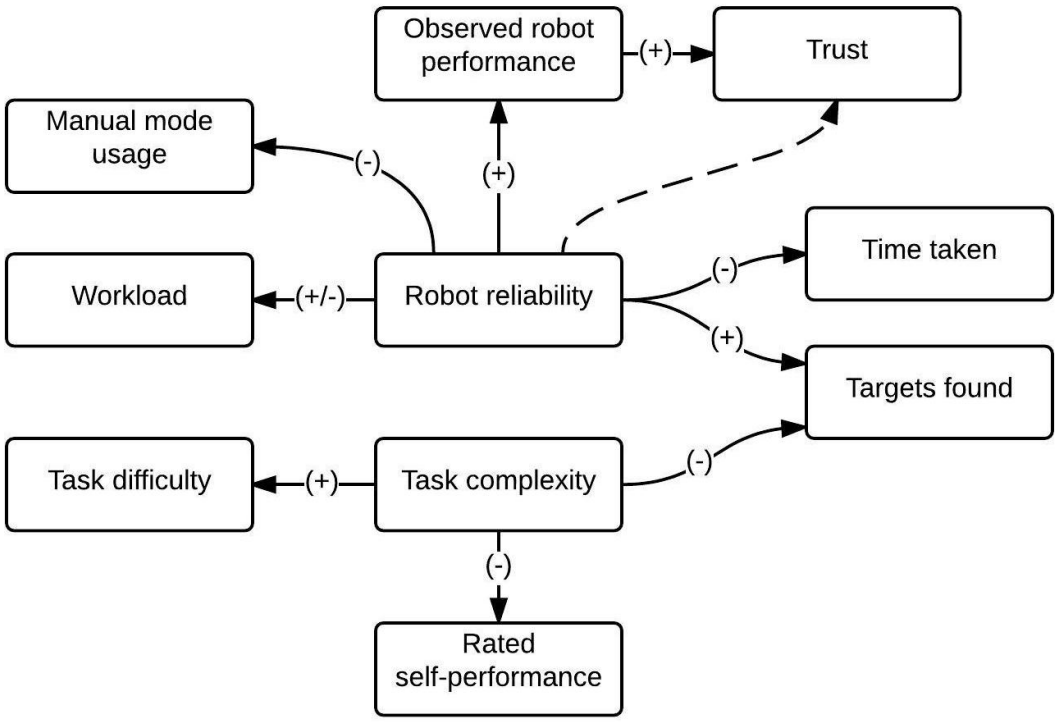


Figure 107 - Qualitative overview of research results of study III; positive influences are indicated with (+), negative influences indicated with (-)

Robot reliability positively influenced the number of targets found and the observed robot performance. It also influenced trust, however, there is evidence that if an operator does not notice a robot mistake they will rate their trust in the robot differently than intended by the researcher. Therefore, the observed robot performance, which is dependent on the robot reliability is the determining factor on trust. The higher the robot's reliability the less time was needed to complete the scenario and the less participants used manual mode. However, this is very much dependent on the level of performance. Low or very low robot reliability can have an adverse effect on time taken, workload, and targets found. Also, the relationship between robot reliability and workload was inconclusive. Participants experienced more subjective workload in the low reliability

condition compared to the middle reliability condition. There was no significant difference compared to the high reliability condition.

Task complexity negatively influenced the number of targets found and the rated self-performance of the participants. The higher the task complexity, the higher was the perceived task difficulty. In general, task complexity and robot reliability seem to be independent concepts regarding trust, targets found, and time taken. However, it needs to be noticed that in real rescue scenarios the task complexity of the environment is likely to influence the reliability of the robot.

6.5.5 Comparison between manual and mixed mode groups

The last hypothesis compared 13 participants, who steered the robot during the entire trial manually, with the mixed-mode groups (low, middle, high reliability). It was expected that the workload for a manually controlled robot is higher than supervising an autonomous robot. However, no significant difference was found. This is contradicting to the findings from Chen and Terrence (2009) who found that remote controlling robots induced higher workload levels than automated systems. In addition Endsley and Kaber (1999) found that, in terms of automated systems, that a higher level of automation induces substantially lower subjective workload. A reason for this differences in findings could be the different types of task. Secondly, that if participants had low levels of workload they engaged more into the secondary task performance and may have balanced out their subjective cognitive workload levels.

The hypothesis also stated that performance measures will be lower when participants exclusively used manual mode. This could only be proven partially. There was no significant difference in objective performance between the manual group and both, the middle reliability robot and low reliability robot group. Nevertheless, there was a significant increase in performance when participants used the high reliability robot. Therefore only the use of a 100% reliable robot resulted in significantly increased performance. Also Desai (2012) showed that 100% reliable robots in semi-autonomous systems have significantly higher levels of performance than

robots with automation faults, unfortunately he did not compare these results to manual operation. Not many HRI experiments compare automated robot systems additionally to exclusively manual performance. De Visser and Parasuraman (2011) compared a reconnaissance mission with two robots (UAV and UGV) during manual, static automation and adaptive automation conditions. Nonetheless they were not able to show any differences between these conditions in terms of performance (detection performance). The study results and previous literature can question if automation really can enhance human-robot team performance in terms of semi-autonomous rescue robots, because 100% reliable systems are not possible to develop in the near future.

A similar picture gave the distribution of trial times. The 100% reliable robot condition was significantly faster in completing the trials compared to the manual group. There was no significant difference in completion time between the manual and the middle and low reliability robot condition. Therefore only a perfect robot could help to achieve the task faster than in manual operation.

The secondary task performance showed that participants had significantly more capacity to answer the secondary task in the low and middle robot reliability conditions compared to the manual group. This was expected since they needed to drive the robot manually which demanded their hands constantly on the controller rather than on the keyboard to answer the secondary task.

Although workload showed no differences between manual and mixed mode groups, participants did experience a significant higher task difficulty when interacting with a low reliability robot compared to driving the robot entirely manually. Therefore one can suspect that unreliable robots might be able to reverse the beneficial effects of automation. As the other results showed, merely a 100% reliable robot did provide a benefit in terms of performance measures.

6.5.6 Qualitative data analysis regarding auto and manual mode usage, robot features, and trust

After completing the experiment participants took part in a semi-structured interview. People were asked why they used a certain robot mode (auto or manual), used certain robot features, and rated the trust as they did.

At the beginning of the interview participants in the mixed mode group were asked why they used auto mode. Their main reasons were that auto mode was less demanding and relaxed them. Furthermore the robot was predictable and followed a systematic approach. Another reason for using auto mode was that participants felt that the robot did a good job and performed well.

Participants were also asked if they used the navigation goal points (NGPs) of the robot. Most participants used the points to see whether there was a stopping point where the robot would turn any time soon, or if they had to take over manual control in order to not miss looking in all the corners. Participants liked that the NGPs made the robot's movements more predictable and they even followed them in the manual mode. Future robot systems should be able to visualise such an aid in order to help the operator to predict the robot actions. Predictability emerged time and again to be a very important factor for trust and collaboration. Kruijff et al. (2014, p. 12) claimed: *"And before we can even talk of common ground, of collaboration, one of the most fundamental lessons we have learnt recently is that this all stands and falls with that robot's autonomous behavior being transparent."*

The 13 manual mode-only participants were asked what influenced them to give certain trust ratings. Most participants said that they rated their own performance rather than the robot's performance. Further, participants said that trust in the robot depended on how much it was able to cope with the environment they were steering it in. Also, participants claimed that their trust in the robot depended on whether the robot did what they expect it to do. This shows that trust has no fixed parameters. Trust in manual operated machines focussed on technical reliability and trust in robots with automation features focusses on the decision making capabilities. This is important for the decision which trust questionnaire is more appropriate to

use and whether trust is measured in the technology or in the artificial intelligence of the robot.

The main features or sensors participants wished for in order to support the search task were a target identification system and a grabber to remove rubble or diffuse a bomb. For some participants it was very important to be able to move the camera independently from the driving direction to have a better field of view and situation awareness. Other ideas were to show where the robot might be able to drive (e.g. possible moving grid) and show areas that have already been covered.

When the manual mode-only group was asked about their feelings towards an auto mode of the robot, nearly half of the participants said that human supervision was still required. Additionally, it would very much depend on the capabilities of the robot. However, participants thought that automation would require less workload but manual mode would still be safer. It seems that participants liked to be in control and did not want to give it away easily depending on their own confidence in performing the task. Especially professional rescue personnel are highly qualified and taking away control from them might be much more difficult and maybe not wanted at all.

6.5.7 Limitations and future work

Programmed reliability level was diluted by the fact that some participants did not perceive certain mistakes of the robot and therefore their impression of the robot was different than desired by the independent variable, which changed the participants' ratings accordingly. That is why this study developed a measure (observed robot performance) that will take this into account. The effect was especially visible between the task complexity conditions where with the increase of task complexity the observed robot performance increased, too. This could have also led to the fact that trust ratings, objective performance, rated task difficulty had a u-shaped data trend and influenced the correlations and violated linear based statistical models. This might be the reason why some correlations were solely weak or moderate and why the independent variable interaction was not significant. It is advisable to test task complexity and minimise the

difference between intended robot performances and observed robot performances to eliminate other influences on the dependent variables.

Although an a priori power analysis showed that 39 participants are sufficient, more participants might have shown more significant result, since a clear reoccurring data trend was visible. Especially the number of participants in the comparison between the manual and mixed mode with each group consisting of 13 participants (26 in total) was insufficient to obtain significant results. This data was not included in the a priori power analysis.

Furthermore, the participants tested were university students, staff and from the general public. The target group, which are firefighters, have a different self-confidence in the task at hand. They are highly trained in these type of tasks and might have different attitudes and interaction behaviours. Because self-confidence is an important factor regarding trust when working together with a robot that is capable of autonomous behaviour (Chen & Terrence, 2009; see Lee & Moray, 1994).

It needs to be considered that the main performance measure in this study was the number of targets found. The rated robot performance was moderately correlated with the objective robot performance, which suggests that not only the total numbers of targets found influenced the rated robot performance. It may be that other variables such as the robots movements or the robot search strategy influenced rated robot performance as well. Also, trust was weakly correlated to the objective robot performance. Although performance is the main influencing factor of trust, there are other components of a robot's performance that are important.

Future studies might determine which performance shaping factors are important when interacting with a semi-autonomous rescue robot system. Further, in future experiments with different designs the measure of the observed robot performance might be useful to distinguish between objective and observed performance. Finally, in this experiment the robot performance was generally quite high, it would be of interest to test robot performances that are very low, or even a robot with no success at all, in

order to see to what extend that is influencing the human-robot team performance and the behaviour of the participant.

6.6 Conclusion

Urban Search and Rescue missions are highly demanding and dangerous for rescuers. They have to deal with unpredictable and very complex tasks and environments under constant time pressure. Their work is greatly performance oriented, because peoples' lives depend on it. Reconnaissance robots can help with rescue tasks and can give support in finding victims in areas which are too dangerous or inaccessible for rescue personnel.

This study aimed to determine the effects of task complexity and robot reliability on trust, workload, operator perception, and performance. The study also investigated the manual mode usage and gathered qualitative interview data to shed light on thoughts, preferences, and behaviours of robot operators. Furthermore the utilisation of autonomous robot features were compared to exclusively manual operated (remote controlled) conditions in terms of their benefits for the rescue mission. In addition, a semi-autonomous performance measure was developed.

As expected and reported in previous literature, the higher the robot reliability was, the higher was the trust in the robot and the higher was the human-robot team performance. This demonstrated that virtual rescue scenarios are a valid method to examine human-robot teams. A very unreliable robot induced higher levels of subjective workload compared to a more reliable robot. The robot performance was the most influencing factor on trust. Especially the strong correlation of the subjectively rated robot performance with the trust scores underlines this result.

With respect to task complexity it can be concluded that task complexity did influence trust ratings and performance but it did not influence the subjective workload ratings. Highly complex tasks resulted in a drop in the performance of the human-robot teams compared to other task complexities. However, these results were diluted by the fact the observed robot performance was significantly higher in high task complexity conditions.

Most interesting was the comparison between the operator group who used the robot exclusively in manual mode and the operator group who utilised the autonomous robot features (semi-autonomous/mixed mode). Only the 100% reliable robot was able to yield significant higher performance levels and less task completion time compared to the manual group. These results illustrated that unreliable robot systems did not show benefits for the overall task performance. It needs to be taken into account that 100% reliable robot systems are unlikely to exist and therefore the use of low reliable robot systems is fruitless.

It does not mean that every autonomous feature will be futile but it is of utmost importance to keep in mind the performance outcomes. The most viable variable in rescue missions is the mission performance. Many studies neglect to prove that the semi-autonomous/autonomous robot systems under examination provide real benefits compared to pure tele operation. Only if we can achieve an improvement in performance we can consider autonomous features as meaningful and future-oriented for Urban Search and Rescue.

6.7 Chapter summary

This chapter examined the influence of robot reliability and task complexity on trust, workload, and different performance measures. Results showed that trust was mainly influenced by the robot reliability. Task complexity did influence trust ratings. However, the effect is likely to be diluted by the observed performance because the observed robot performance (how reliable the robot seems) changed across task complexity which most likely influenced the trust ratings. Furthermore, unreliable robots induce higher levels of workload and decreased performances. An interesting comparison of mixed mode operators (auto and manual operation) and manual operators (manual operation only) showed that only the robot with 100% reliability could contribute towards significantly higher performances and reduced task times compared to the manual operating group.

7 Study IV - The influence of robot transparency and task complexity

7.1 Chapter overview

This chapter's study uses the same virtual environment and robot as the previous study. It examines the influence of robot transparency and task complexity on workload, performance and trust. The quantitative data of the previous study showed that robot transparency is of importance for the operator to understand the robots' states and actions. Transparency levels in this chapter consist of two different interfaces with different levels of feedback and scenario information. In addition, the study aims to see if the results from the previous study can be validated and have the same effects on trust, workload, and performance. Interview data examines and quantifies which elements of the interface were actually used and why, in order to understand the benefit of presenting more or less information for higher transparency levels.

7.2 Introduction

Despite the fact that robotic agents are becoming increasingly autonomous and sophisticated, the human still holds an essential role in autonomous systems (Lyons & Havig, 2014). But it can be challenging to operate them because new autonomous systems demand new forms of interactions and most processes of the system are invisible to the operator. Therefore it is crucial to take a human-centred approach and develop robotic team mates that are understandable and supporting. The lack of background information of the robot's behaviour/functioning leads operators to trust robots less and may lead them to use the autonomous features inefficiently (Stubbs et al., 2007). It is important that operators gain insight and understanding of the robotic agents' actions in order to calibrate their trust into the system. The previous study (Chapter 6) showed that participants positively mentioned features that supported them to predict the robot's actions. Predictability is

an integral part of a system's transparency (Colombi, Lenfestey, Cring, & Colombi, 2009; Ososky, Sanders, Jentsch, Hancock, & Chen, 2014).

Recent literature focussed on transparency because it might also influence trust calibration and could enhance situation awareness resulting in an improved task performance (Ososky et al., 2014). One can simply put that transparency is the provision of more information to the user (Sanders et al., 2014). Although, this definition is not sufficient because the term transparency includes that the system is supplementing expected outputs, reveals how a system works and what it is doing in a format that is understandable and intuitive (Ososky et al., 2014; Preece, Rogers, & Sharp, 2002). For example, proving more information by visualising the running source code of the robot is rather confusing than transparent. It may be possible to increase transparency by providing more information, even though a too high level of feedback can lead to overload and confusion (Finomore et al., 2012). According to Chen et al. (2014, p. 2) transparency is "the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process.". With this definition it is clear that transparency needs to be adjusted to a comprehensible and performance increasing level. Also, the adjustment has to be fitted to the goals and tasks of the field of application (Lyons & Havig, 2014).

Ososky et al. (2014) emphasised on the importance of mental models because varying mental models can influence the interpretation and understanding of the robot. In the rescue domain this is especially challenging, since the time for the operator to build a mental model (e.g. training time) is very limited. More transparency can support the formation of a mental model and foster better human-robot interaction (Ososky et al., 2014). Noteworthy in this context is that trust is not based on what the robot can or cannot do, it is based on what the human perceives it to be capable of (Ososky et al., 2014). Which emphasises not to underestimate the importance of observed robot performance measures and perceived robot performance measures. This paragraph showed how important the creation of an appropriate mental model with the aid of transparency is.

Similarly to Ososky et al. (2014), Lyons and Havig (2014) were looking to foster transparency with regard to shared awareness and intent by introducing straightforward implementation approaches. For a shared intent they suggest that the robot has to inform the human as to why it functions non nominal at the given moment (Why did the robot something different, with what goal?). The robot can also show intent by social cues, for example with a directed gaze (also see Kwok et al., 2012). Robots may convey good intent by communicating with benevolence, which means that they should promote the belief that the robot will act in the best interest of the operator.

As the data from the second (chapter 5) and third study (chapter 6) of this thesis suggested, as well as other previous work (Kruijff et al., 2014; Seppelt & Lee, 2007; Stanton, Young, & Walker, 2007), operators are reluctant to use robotic tools with autonomy because of issues regarding the understanding of the robot's states and actions. Hence, in terms of shared awareness Lyons and Havig (2014) stated that they recommend the robot should give detailed information about the actions and tasks or task steps it is currently doing. In that way the operator knows why the robot is doing something and what it will do next.

Further, according to Lyons and Havig (2014), the system should communicate limitations, constraints, and why it is failing (e.g. share the reason for the failure). In a study in human-robot teaming in the USAR domain lack of transparency of the robot behaviour was suggested as being the reason of less autonomy usage by the operator (Kruijff et al., 2014). Their findings demanded, exactly as Lyons and Havig (2014) did, for transparency of the robot's state, current and future tasks, behaviour, explanation of failing or succeeding, existing knowledge and capabilities. Additionally, the second study of this PhD (Chapter 5) found that a reliability indication positively influenced visual attention allocation towards the robot. Other literature showed that reliability or confidence indication of the robot enhanced control allocation strategy (Desai, 2012; L. Wang, Jamieson, & Hollands, 2009). Another intriguing finding was made by Dzindolet et al. (2003): If the capabilities of an automated system were communicated to the operator it positively mediated trust recovery after an error.

Consequently, communicating the robot's confidence and capabilities proved to be useful and added to transparency.

Furthermore, Lyons and Havig (2014) advice that some basic information to cultivate transparency should be visible. For example the robot's health status and environmental changes that are or can influence the system's performance (e.g. sensors, gauges). In order to clarify the team status the robot could communicate for which task it is responsible and for what tasks the human is accountable. This is, for instance, important when the robot can have different modes/levels of autonomy or sliding autonomy.

The qualitative analysis of the previous study, reported in Chapter 6, showed that the participants' situation awareness was often lacking and sometimes kept participants from using manual mode, even if this would have been the appropriate choice. Chen et al. (2014) developed a situation awareness based agent transparency model (SAT). Many of the previous suggestions from literature regarding transparency can be incorporated in this model. The model was used in this experiment to differentiate between the transparency levels. The model consists of three information levels (see also Figure 108, p. 262):

- Level 1 information conveys the current state, goal and process of the robot. For example the path of the robot is displayed and a green/red light provides the current robot status.
- Level 2 consists of the information as to why level 1 information is like it is. For instance, the robot gives information about resources available (battery life) or shows the constraints of the environment (why it cannot drive a certain path).
- Level 3 information gives insight about the projected future status of the robot. This information could be a visualisation, in percentages of how sure the robot is about the target it identified or a general indication of current reliability status.

Selkowitz et al. (2015) used the SAT model in order to determine the influence of transparency on trust, situation awareness and workload. They showed that more information transparency allowed operators to calibrate their trust better without experiencing more workload, but it did not support

their situation awareness. However, Helldin (2014) tested different aspects of transparency, such as the visualisation of automation parameters, uncertainty, reliability and ability, and found that workload increased with parameter detail and decreased with providing automation uncertainty. Furthermore different representation styles (text or bar chart) influenced trust ratings and performance as well.

Finally it is critical to understand that transparency is a resource that supports trust calibration by matching the user's expectations to the actual robot behaviour (Ososky et al., 2014). Transparency is therefore the key to predictability.

Hypotheses:

This experiment will investigate how different levels of task complexity in combination with two different transparency levels influence trust, workload and performance by using virtual rescue scenarios. Previous studies of this PhD have shown that task complexity is an important factor to consider in search and rescue missions (see Chapter 4 and Chapter 6). The previous study (Chapter 6) encountered difficulties to test task complexity and its influence on the dependent variables. Therefore this study minimised the difference between intended robot performances and observed robot performances in order to measure the influence of task complexity on trust, workload, and performance.

H1) Robot transparency influences trust, performance and workload ratings.
H2) Task complexity influences trust, performance and workload ratings.

Furthermore, it is important to see whether task complexity interacts with robot transparency, because workload can increase with transparency detail (Helldin, 2014) and further visual demand in searching a complex scene might lead to decreased performance levels

H3) Task complexity interacts with robot transparency to influence trust.
H4) Task complexity and robot transparency influence subjective ratings of participants towards the robot.

Furthermore, the events that occurred during the trials were classified and their distribution analysed in order to look for differences between the experimental conditions. A post-task interview presents qualitative data about the participant's interface preferences and the individual interface items used.

7.3 Methodology

7.3.1 Participants

Thirty participants were recruited via e-mail and posters. The study was open to staff, students and the general public. They were screened to fit the requirements for the study (over 18 years and no vulnerable members of the public). Overall the study took 35 - 45 minutes to complete. Participants were reimbursed with a 5 GBP Amazon voucher. The average age of the 12 female and 18 male participants was 29.27 years (SD = 6.7). All participants were assigned a participant number and the data was stored under this ID and not under their name. The study was approved by Faculty of Engineering Ethics committee.

7.3.2 Experimental design

The study was organised in a 2x2 mixed subject design (see Table 33). Robot transparency was the within factor with two levels (low and high). The independent between factor was task complexity with two levels (low and high). The dependent variables were trust, workload, and performance measures, as well as participants' feelings about themselves and the robot.

Conditions		**within**	
		Low transparency (LT)	High transparency (HT)
between	Low complexity (LC)	LT-LC	HT-LC
	High complexity (HC)	LT-HC	HT-HC

Table 33 - 2x2 mixed subject design with the variables task complexity and robot transparency

The differences between low and high transparency/complexity are explained in the Virtual rescue scenario. LT-LC, HT-LC, LT-HC, and HT-HC are called trials.

All environments had exactly the same floor plan and possible driving routes. This aimed for a better comparison among trials. To avoid that participants would get familiar with the route during their two trials (changing transparency levels), they had to drive in one trial from start to end and in the other from end to start. In addition, the position and type of objects/clutter/rubble in the environment changed for each trial.

7.3.2.1 Interface transparency levels

Standard elements of the display are rear view and main view of the robot. The robot also showed if it was in auto or manual mode as well as the readings of CO_2 and temperature sensors. These elements were in every condition present.

The two different interface transparencies were designed with the guidance of the situation awareness based agent transparency model (SAT) (Chen et al., 2014). The SAT model has three information levels, as shown in Figure 108. Each interface will be explained in detail in the next two sections.

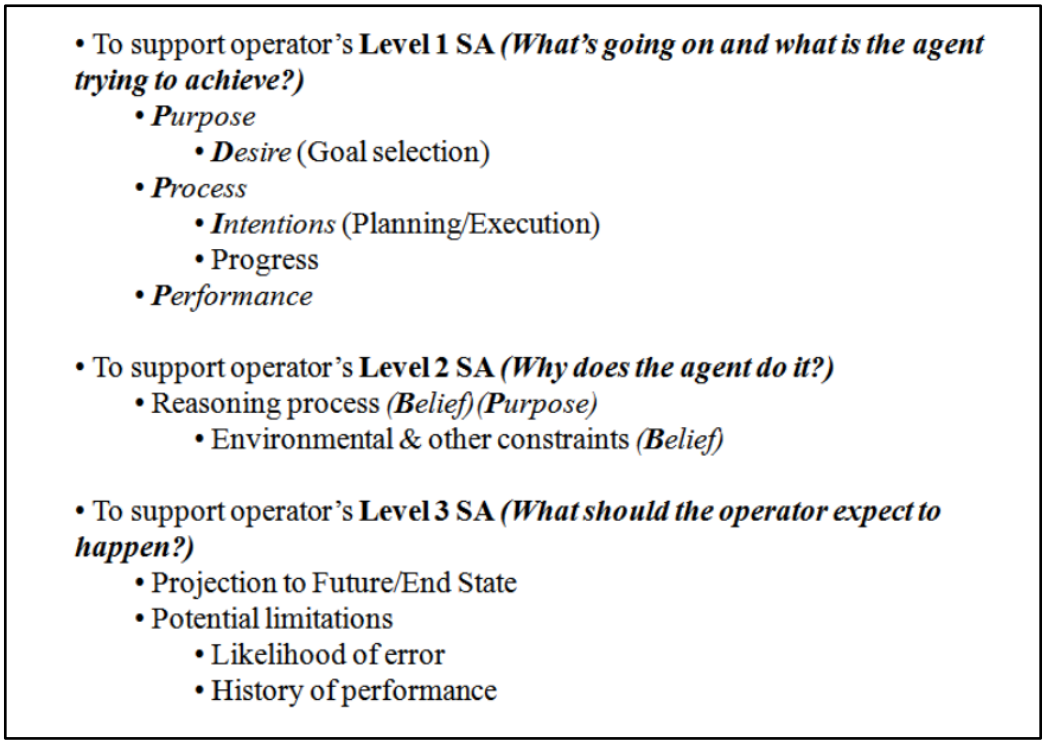


Figure 108 - Situation awareness-based Agent transparency model (Chen et al., 2014)

Low transparency interface

The low transparency interface incorporated Level 1 (L1) situation awareness (SA) of the SAT model. The low transparency interface is shown in Figure 109. If an element of the display is explained, the element number is written in brackets.

- L1 Purpose/Process: The rescue robot shows the operator what its current goal is by displaying the goal it is trying to achieve (1). Furthermore, the orange squares, called navigation goal points (NGPs), on the map (4) show the intention of the robot by visualising where it wants to navigate next.
- L1 Performance: The current status (2) of the robot is depicted with a smiley and a word. The green word "working" and a smiling face indicates the system is working properly. A red sad face and the word "system failure" indicates a faulty system. Furthermore the battery status and the signal strength is displayed (3).

Figure 109 - Low transparency interface with highlighted display elements

High transparency interface

The high transparency interface incorporated SA Level 1+2+3 of the SAT model. The high transparency interface is shown in Figure 110. The SA level 1 is depicted in the section above. In addition to the low transparency interface components (SA Level 1) the high transparency interface incorporates the following elements:

- L1 Process: The process is additionally supported within an enhanced map view (5). The map not only shows the NGPs, but also the shortest line between the points, so called 'path lines'. This aims to show a proper visualised route through the environment. The robot intentions can be seen in the mission log (7) (e.g. as to why the robot is turning towards a certain direction). It additionally uses gaze (turning camera towards an object) to give away its intent (e.g. Coradeschi et al., 2006).

- L2 Reasoning process: The robot log depicts the reasoning as to why the robot did certain things (e.g. the robot turns towards a door and displays: "Door found, save position, continue search"). It also provides additional information regarding the mission parameters in the mission info box (8). For instance it shows an estimated number of targets in the room, number of rooms to search, how long the scenario is running, and how much seconds of battery power is left.
- L3 Projection to future: The robot will show in the mission log (7) in which room it is (e.g. Entering room 1) and the mission info box (8) shows the estimated number of rooms. The participant can therefore estimate how long the scenario might last.
- L3 Potential limitations: The robot is able to visualise its current reliability level (6). It can display how confident the system is in its work. With this function it can provide the information about potential limitations in its performance. It also gives an indication of performance history by depicting in the Mission Info box (8) how many targets have been found so far.

Figure 110 - High transparency interface with highlighted display elements

7.3.2.2 Task complexity levels

Each trial had its own environment. In all environments six targets were present. The position of targets was predetermined by the scenario timeline. Every participant, when in auto mode, encountered targets and experienced robot errors at the same time in all trials. This was pre-programmed because timing of errors had an impact on trust (Desai et al., 2013). In the following paragraphs the different task complexity levels are explained.

Low task complexity environment

Low task complexity utilised 30-40 objects per room. A bird's eye view during development in Figure 111 depicts an example view of a low task complexity environment.

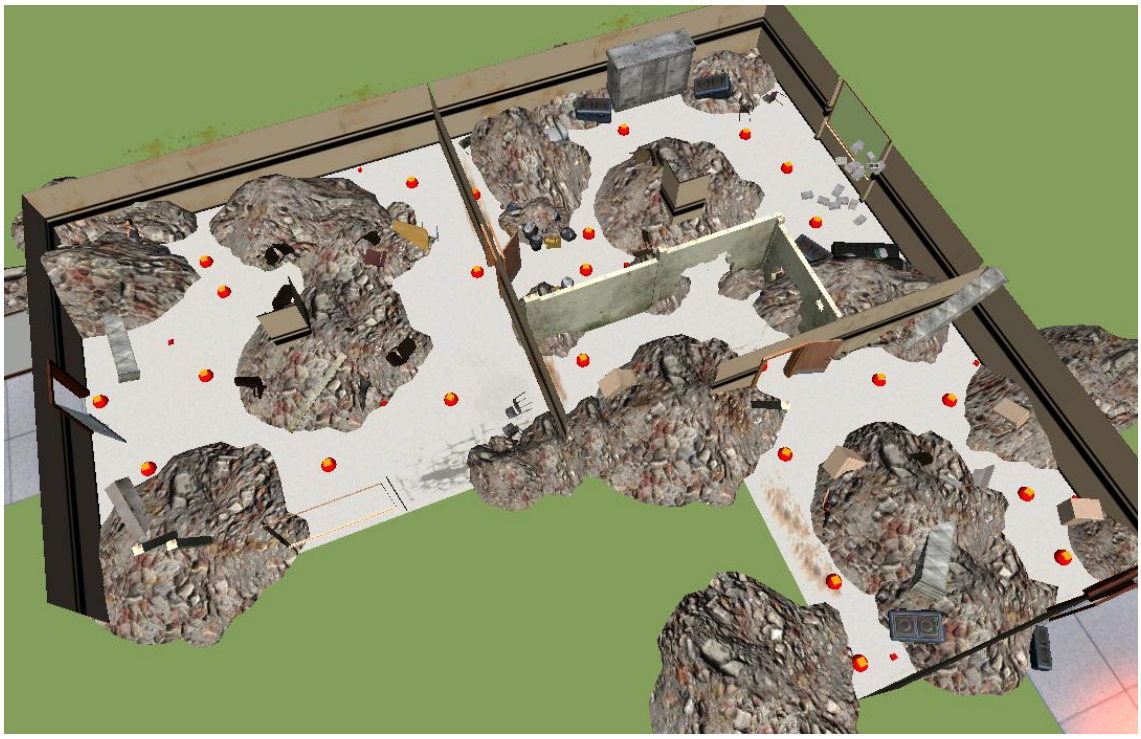


Figure 111 - Low task complexity (editor view of LT-LC) with waypoints visualised.

Targets in the low task complexity conditions were victims only. An example of such a victim is visualised in Figure 112.

Figure 112 - Example of victim in a low complexity environment

High task complexity environment

In a high task complexity condition were between 50 and 60 objects placed in each room. A bird's eye view of a high task complexity condition is shown in Figure 113.

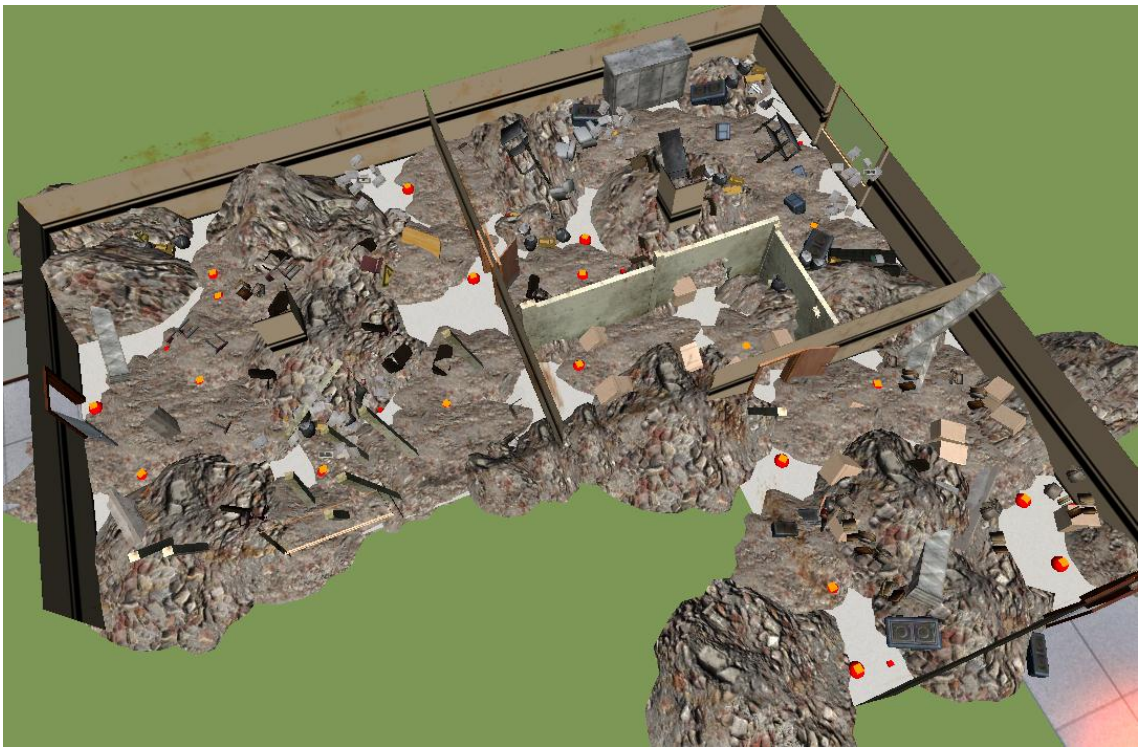


Figure 113 - High task complexity (editor view of LT-HC) with waypoints visualised.

During high task complexity participants had to find three different kinds of targets: terrorist indicators (weapons/bombs), hazard signs, and victims. In Figure 114 examples of these three types are illustrated.

Figure 114 - High task complexity targets (left to right: weapon, hazard sign, victim)

7.3.2.3 Virtual Rescue scenario

The Virtual rescue scenario used the same hardware and software as the previous chapter (Study III, Chapter 6). The development of the software is explained in Section 3.5.2.

Virtual robot

The simulated robot had proximity sensors (360 degree) which were visualised by a top view map. A front and rear camera were provided and visualised on the display. The robot's target identification system was enabled to find specific targets in the environment. The robot further had CO_2 and temperature sensors. All four conditions were performed in a mixed mode, which is explained below.

MIXED MODE

Mixed mode means that participants were free to choose between using manual and autonomy mode.

Manual mode:

- Participant is in charge of robot's movements and target identification.
- Participant can see the goal navigation points the robot would navigate to in the map.

Autonomy mode:

- Robot is in control of driving.
- Target identification is active.

7.3.3 Materials

A Laptop with an additional 17" screen was used. The program was created in UNITY, a multi-platform game creation system. The participant could interact with the virtual robot via a X-Box 360 controller. In addition, paper questionnaires and two cameras with tripods were used.

7.3.4 Procedure

Participants were required to give informed consent before starting the study (see Appendix K - - Digital Appendix VI, p. 404). They began by completing a general questionnaire (see Appendix H, p. 398) which asked for demographics and their propensity to trust robots. Next, participants had a five minute training session where they learned how to use the robot

manually, in auto mode and in mixed mode. Additionally, they received a full explanation of the two interfaces. After the training, participants performed two trials. Participants were told after each trial how many targets they have missed and with that in mind they had to answer the post-task questionnaire. The post-task questionnaire (see Appendix I, p. 400) consisted of the Schaefer trust questionnaire (short), a NASA TLX to measure subjective workload, and questions about their performance.

7.4 Results

7.4.1 Trust

First, propensity to trust regarding robots was measured with the trust propensity scale used by Merritt (Merritt, 2011). The six statements could be answered on a 5-point Likert-scale ranging from "strongly disagree" to "strongly agree". The participant sample had a median of 3.33 (IQR = 0.29) which indicates a very neutral trust propensity towards robots. The trust propensity between the groups (low complexity and high complexity) was not significantly different.

Robot trust after the trials was measured with the short version of the Schaefer (2013) trust questionnaire (14 items) as shown in Appendix I (p. 400). The data of participant 23 violated the assumptions of analysis of variance by being a significant outlier. Therefore this participant's data was excluded from the dataset. Data was tested with a 2x2 mixed analysis of variance test.

The influence of task complexity on trust

Data showed that task complexity had no significant main effect on trust, $F(1, 27) = 2490.65$, $p > .05$ (see Table 34).

Schaefer (short) trust scores across task complexity	
Condition	**Mean (SD) [%]**
High task complexity	77.90 (9.60)
Low task complexity	75.38 (12.12)

Table 34 - Schaefer (short) trust scores across task complexity

The influence of robot transparency on trust

Robot transparency had a highly significant main effect on trust ($F(1, 27) = 15.94$, $p < .001$, $r = .61$). Low transparency with a mean of 74.46 (SD = 9.6) was rated lower in trust than the high transparency robot with a mean of 80.71 (SD = 8.8). Hence, during high robot transparency participants rated the trust in the robot higher than compared to the low transparency condition. This is visualised in Figure 115.

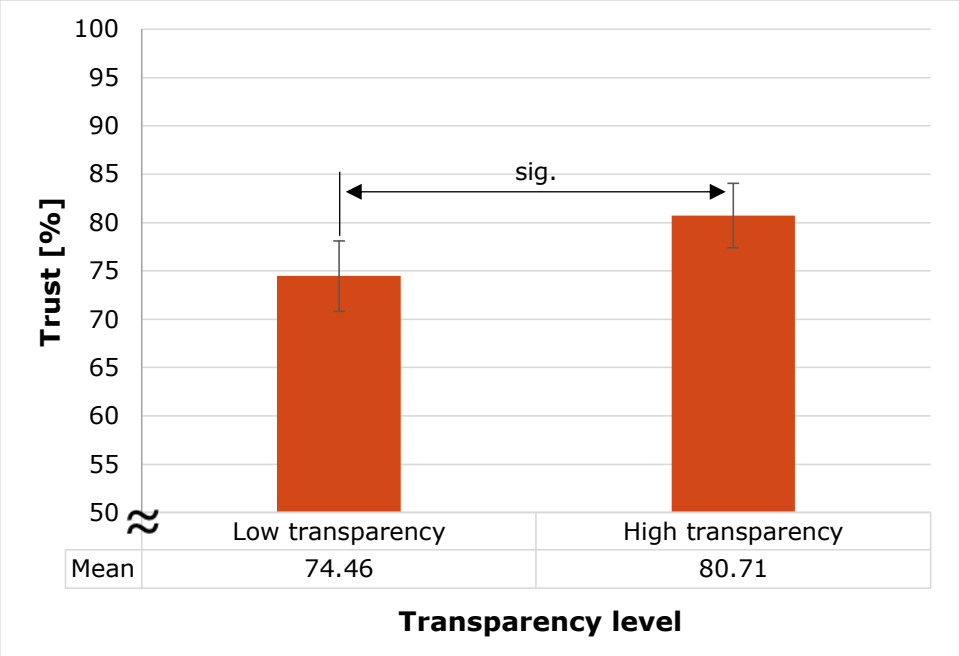


Figure 115 - Schaefer (short) trust scores across the two transparency conditions with 95% confidence intervals

The interaction between task complexity and robot transparency

There was no interaction effect of complexity and transparency regarding the trust measure, $F(1, 27) = .007$, $p > .05$.

7.4.2 Objective performance

The performance was measured in terms of how many percent of the six possible targets were found.

The influence of task complexity on performance

A Mann-Whitney test (see Figure 116) revealed that there was a significant difference of performance among the complexity levels with a medium effect

size (U= 271.5, p < .01, r = -.36). In the lower complexity tasks, participants had a significantly higher level of performance (M = 91%, SD = 13%) compared to the high complexity task (M =81%, SD = 15%).

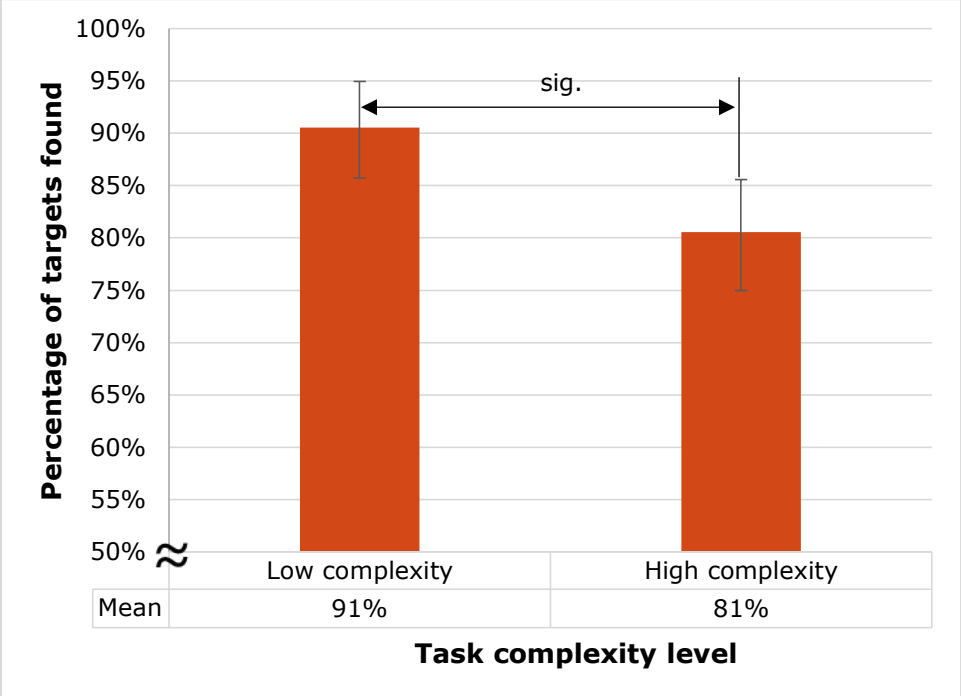


Figure 116 - Performance across the two levels of complexity with 95% confidence intervals (bootstrapped)

The influence of robot transparency on performance

The data (see Table 35) showed that robot transparency had no significant effect on performance. The data was tested with a Wilcoxon signed-rank test (Z = -1.8, p > .05).

Objective performance scores across robot transparency	
Condition	**Mean (SD) [%]**
High robot transparency	82.78 (16.66)
Low robot transparency	88.33 (11.70)

Table 35 – Objective performance scores across robot transparency

7.4.3 Observed robot performance

The observed robot performance is the performance of the robot the participant actually witnessed, for more information about this measure please see the previous chapter (Section 6.3.5) In three of the 60 trials

conducted (30 participants, each 2 conditions) the robot was not able to demonstrate any performance because all targets were found in manual mode or were missed entirely by the participant. These cases were excluded listwise for statistical tests. Furthermore, data was not normally distributed, therefore a Mann-Whitney test was conducted.

No significant difference between the observed robot performances across task complexity levels have been found ($U = 302$, $p > .05$). A Wilcoxon signed-rank test showed there were no differences across transparency levels ($Z = -1.24$, $p > .05$), therefore the observed robot performance did not vary significantly among conditions (see Table 36).

Observed performance across conditions	
Condition	**Mean (SD) [%]**
High task complexity	63.33 (22.37)
Low task complexity	61.67 (23.31)
High robot transparency	60.30 (22.80)
Low robot transparency	64.82 (22.63)

Table 36 - Observed performance across conditions

7.4.4 Workload

Workload was measured with a raw NASA TLX. A 2x2 mixed analysis of variance showed that there was no significant differences in workload across complexity ($F(1, 28) = .276$, $p > .05$) or transparency ($F(1, 28) = .01$, $p > .05$). In addition, no interaction effect could be found, $F(1, 28) = .389$, $p > .05$. Therefore it can be assumed that participants did not experience significantly different levels of workload between any of the conditions. For values see Table 37.

Workload ratings across conditions	
Condition	**Mean (SD)**
High task complexity	54.27 (17.60)
Low task complexity	51.03 (17.80)
High robot transparency	52.77 (17.70)
Low robot transparency	52.53 (17.84)

Table 37 - Workload ratings across conditions

Subscale analysis showed no significant differences across transparency levels but across complexity levels. The performance rating in the NASA TLX differed significantly between the low and high complexity condition ($U = 307$, $p < .05$, $r = .27$). This means that participants perceived themselves to be more successful accomplishing the task in the high complexity condition compared to the low complexity condition (see Figure 117). This pattern was not discovered in the previous study where complexity was tested.

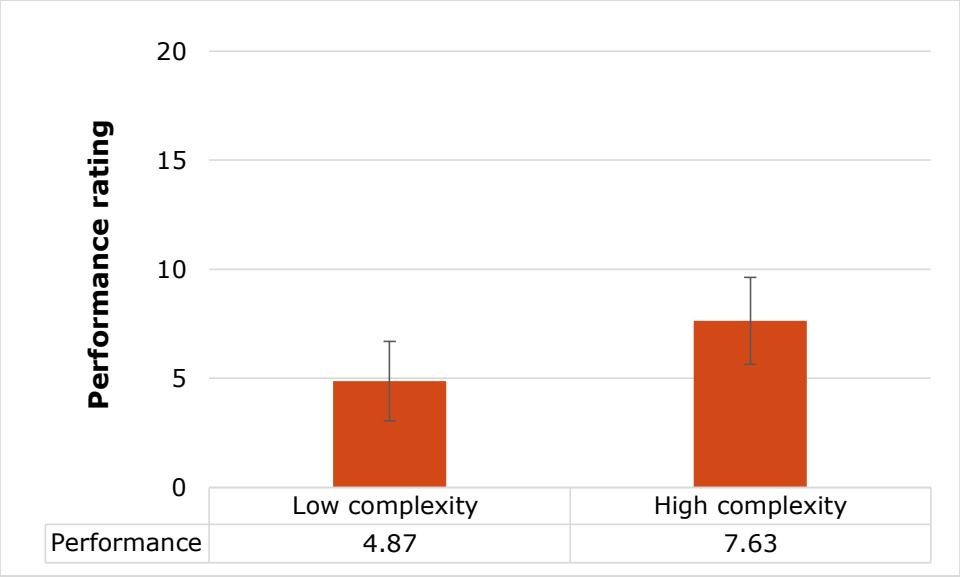


Figure 117 - Significant workload (NASA TLX) subscale analysis of performance with 95% confidence intervals

7.4.5 Subjective ratings

7.4.5.1 Rated task difficulty and complexity

Participants were asked to rate "How difficult did you perceive the task?" on a scale from 1 (not at all difficult) to 6 (very difficult) and "How complex do

you rate the task?" with the given explanation: "Complexity means the simultaneous occurrence of several task components that influence your performance." on a scale from 1 (not at all complex) to 6 (very complex).

Mann-Whitney tests indicated that there were no significant differences between rated task difficulty ($U = 389$, $p > .05$) or rated task complexity ($U = 445$, $p > .05$) across task complexity levels. Using a Wilcoxon signed-rank test revealed that there was no significant difference between the robot transparency levels regarding the participant's ratings of task difficulty ($Z = -1.37$, $p > .05$) or complexity ($Z = -1.38$, $p > .05$). For all values please refer to Table 38 and Table 39.

Rated task complexity across conditions	
Condition	**Median (IQR)**
High task complexity	4 (1)
Low task complexity	4 (1.75)
High robot transparency	4 (1)
Low robot transparency	4 (1.75)

Table 38 - Rated task complexity across conditions

Rated task difficulty across conditions	
Condition	**Median (IQR)**
High task complexity	4 (1)
Low task complexity	3.5 (2)
High robot transparency	3 (2)
Low robot transparency	4 (1)

Table 39 - Rated task difficulty across conditions

7.4.5.2 Rated self-performance

After each trial the participants rated their self-performance on a scale from 1 (poor) to 7 (excellent).

The influence of task complexity on rated self-performance

A Mann-Whitney test revealed that there was a significant difference of rated self-performance among the complexity levels (U= 318, $p < .05$, $r = -.26$). The lower the complexity of the task, the higher participants rated their self-performance, but the size of the effect was small (see Figure 118 and Table 40).

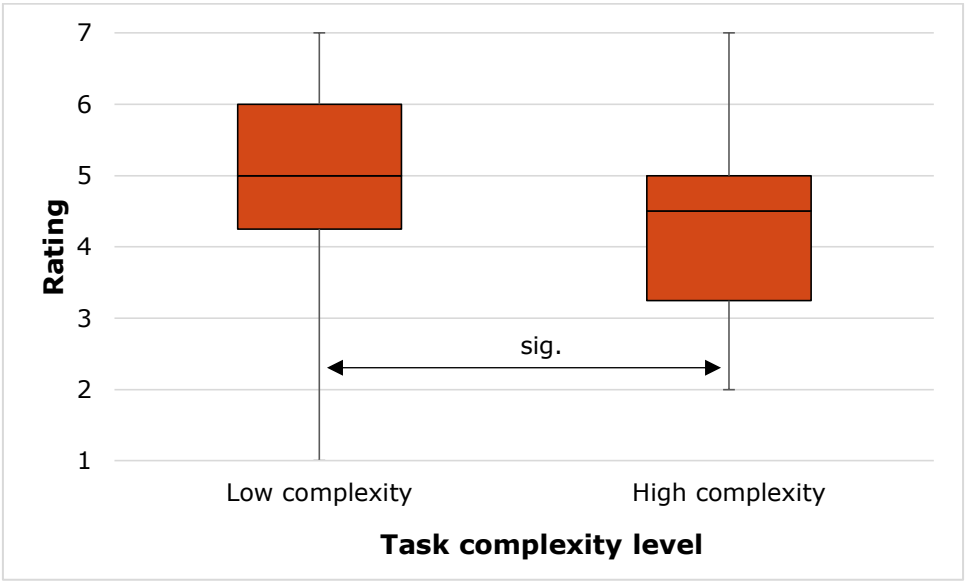


Figure 118 - Rated self-performance box plots across task complexity; whiskers show minimum and maximum values

Rated self-performance	
Condition	**Median (IQR)**
Low task complexity	5 (1.75)
High task complexity	4.5 (1.75)

Table 40 – Rated self-performance across complexity

The influence of robot transparency on rated self-performance

The transparency of the robot did not lead participants to rate their self-performance significantly different (see Table 41). This was tested with a Wilcoxon signed-rank test, $Z = -1.7$, $p > .05$.

Rated self- performance across robot transparency	
Condition	**Median (IQR)**
High robot transparency	5 (1.75)
Low robot transparency	5 (1.75)

Table 41 - Rated self- performance across robot transparency

7.4.5.3 Rated robot performance

The rated robot performance is the performance rating that participants gave the robot after the trial. They were asked to rate the performance on a scale from 1 (poor) to 7 (excellent).

Robot performance was programmed to be the same across all conditions. There was no statistically significant difference between task complexity ($U = 411.5$, $p > .05$) or robot transparency conditions ($Z = -.155$, $p > .05$) regarding the rated robot performance (see Table 42). Therefore participants felt that across conditions the robot's performance did not significantly vary.

Rated robot performance across conditions	
Condition	**Median (IQR)**
High task complexity	4.5 (1)
Low task complexity	5 (1)
High robot transparency	5 (1)
Low robot transparency	4.5 (1)

Table 42 - Rated robot performance across conditions

7.4.6 Event analysis

The event analysis examined the distribution of events for each independent variable. Each target found/missed represents an event. There were 30 participants, each with six targets, therefore there were 180 events for each condition. But in order to understand the circumstances these events occurred, the manual mode usage and the trial time needed analysis. T-tests showed that there was no significant difference in the use of manual mode across complexity ($t(58) = 1.45$, $p > .05$) or transparency ($t(29) = -.528$, $p > .05$). On average participants used 37% of the time manual mode. Trial time was measured by how long participants needed to complete a trial. Trial times showed no significant differences among complexity ($U = 374$, $p > .05$) and transparency levels ($Z = -.062$, $p > .05$). Therefore participants showed no significant differences in time taken whether they had a high or low transparency robot or operated in a high or low complexity task. First task complexity is examined followed by robot transparency.

Event analysis for task complexity

Since there was a significant decrease in the objective team performance across complexity levels, an event analysis was performed to show the cause of the mistakes made.

A chi-square test revealed that the two group distributions are significantly different, $X^2(4, 180) = 11.79$, $p < .05$). Post-hoc tests by using standardised residuals/Pearson residuals (Sharpe, 2015) revealed that there was a significant difference between the values of "Robot found, human acknowledged" ($p < 0.05$), "Manual mode, human found" ($p < 0.001$), and

"Robot missed, human missed" ($p < 0.001$). The significance levels were adjusted with the Holm-Bonferroni method (Cabin & Mitchell, 2000).

An overview of the distribution of the 180 events per complexity level (see Figure 119) shows that participants allowed the robot to find significantly more targets in the high complexity condition (42%) compared to the low complexity condition (31%). The instances when the robot missed the target and the human found it (error detection), were quite similar (28% in low complexity and 26% in high complexity). When the task was less complex participants found significantly more targets in manual mode (31%) compared to the more complex task (13%). During low task complexity participants only chose to use the robot to find 31% of the targets (out of 66%). This suggests that participants relied less (under-reliance) on the system during low task complexity.

Interestingly, during high complexity tasks significantly more often participants missed a target when the robot did. Therefore they were less able to detect robot errors during high task complexity (17%) compared to the low task complexity (5%). These could be an indication of over-reliance during high task complexity.

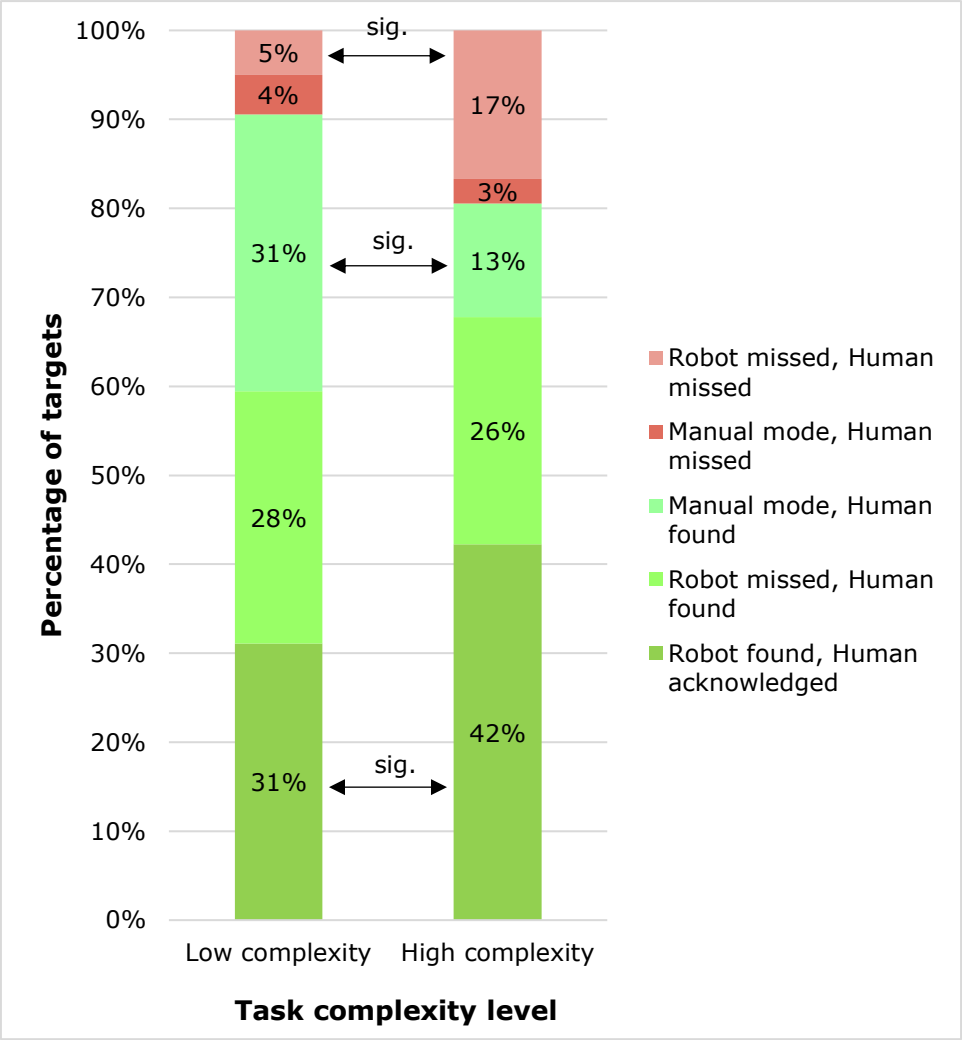


Figure 119 - Event distribution between the two task complexity levels

Event analysis for robot transparency

Although participants' data showed no difference in objective performance between the two transparency conditions, data might reveal differences in the event distribution.

A chi-square test was used to identify any differences between the two group distributions. Results showed that there was a significant difference, $X^2(4, 180) = 69.65$, $p < .001$). Post-hoc tests by using standardised residuals/Pearson residuals (Sharpe, 2015) revealed that there was a significant difference between the values of "Robot missed, human found" ($p < 0.005$). The significance levels were adjusted with the Holm-Bonferroni method (Cabin & Mitchell, 2000). The distribution of events can be seen in Figure 120.

Significantly more times the human detected a robot error in the low transparency condition (32%) compared to the high transparency condition (22%). Therefore more mistakes in error detection were made in the high transparency condition (12% "Robot missed, human missed"; not significant) compared to the low transparency condition (9%). It was expected that the high transparency interface would support the human to easier detect low robot reliability and, therefore, detect more robot errors in the high transparency condition. This was not the case. The finding perhaps suggests that participants' attention was not appropriately directed by the interface to the currently important information (e.g. robot reliability indication). Another reason could be that the high transparency interface displayed too much information at the same time.

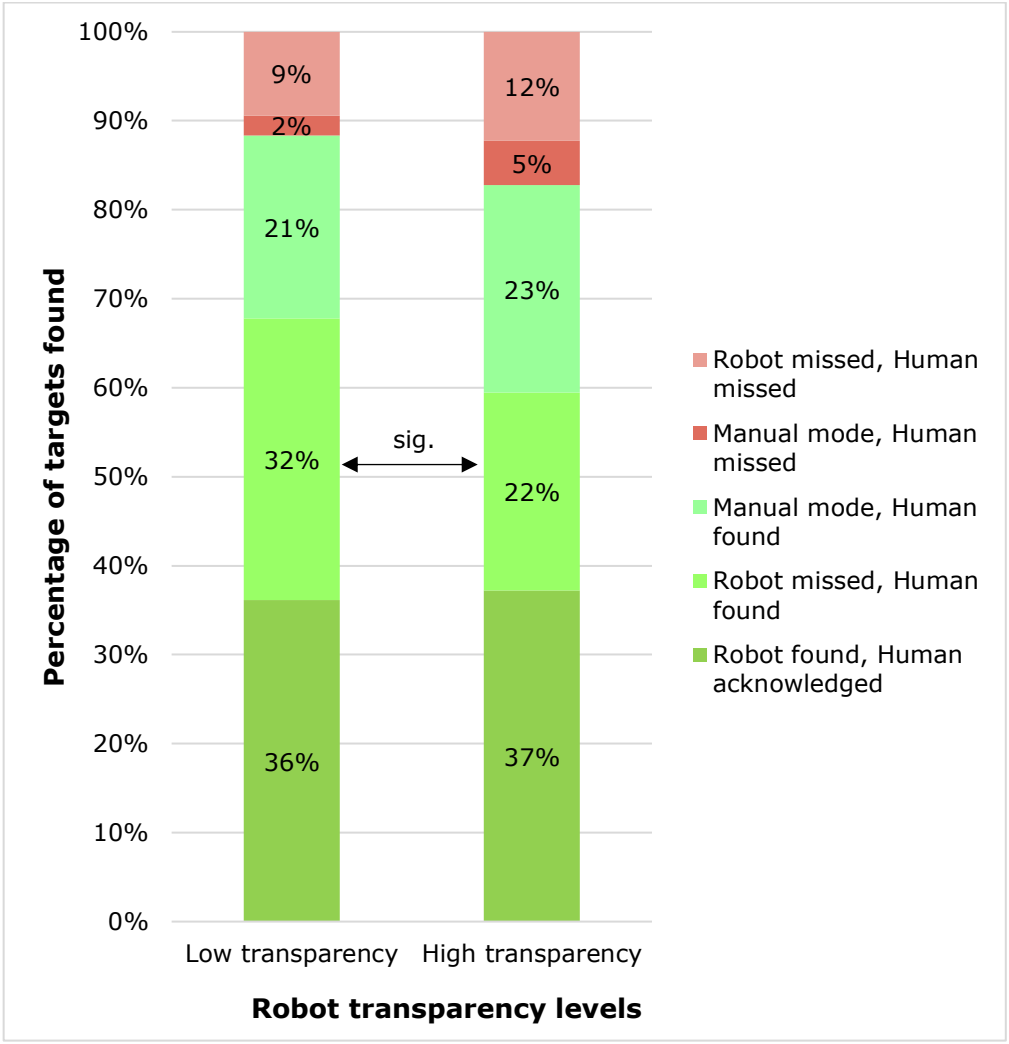


Figure 120 - Event distribution between the two robot transparency levels

7.4.7 Summary of quantitative results

The table below (Table 43) gives an overview of the quantitative results of this experiment.

Dependent variable	Independent variable	Significance	Result details (effect size)
Trust	complexity	no significance	
	transparency	significant	LT < HT (r = .61) Trust in the high transparency condition was higher than in the low transparency condition. However, single items on the trust questionnaire were manipulated.
Workload	complexity	no significance	
	transparency	no significance	
Objective performance	complexity	significant	LC > HC (r = -.36) Low complexity tasks yield a higher performance than high complexity tasks.
	transparency	no significance	
Observed robot performance	complexity	no significance	Robot reliability was constant.
	transparency	no significance	Robot reliability was constant.
Rated task difficulty	complexity	no significance	
	transparency	no significance	
Rated self-performance	complexity	significant	LC > HC (r = -.26) Participants rated their performance higher in low complexity tasks compared to high complexity tasks.
	transparency	no significance	
Rated robot performance	complexity	no significance	
	transparency	no significance	
Event analysis	complexity	significant	- Participants allowed the robot to find more targets in the high complexity conditions. - Participants found more targets manually in the low complexity tasks. - In the high task complexity conditions participants missed

			more robot errors than in the low task complexity conditions.
	transparency	significant	Participants detected more robot errors in the low transparency conditions compared to the high transparency conditions.

Table 43 - Summary of quantitative results

7.4.8 Post-task interview

After each trial participants were asked in a semi-structured interview about the elements of the interface they have used and questions to determine their level of situation awareness. After participants had performed their two trials they were asked which interface they preferred. In order to analyse the interview data the theme based content analysis (TBCA) from Neale & Nichols (2001) was used.

The number in large brackets shows the number of participants mentioned the particular theme (e.g. [8]). Each statement is followed by a selection of supporting quotes. At the end of each quote the participant number is indicated in brackets, for example (P05). The full transcript and the emerging themes are provided in the Appendix K - - Digital Appendix VII (p. 404).

7.4.8.1 Preferred interface

In the semi-structured interview, after completing the two trials, participants were asked which of the two interfaces they would prefer to use. 30% [9] of the participants preferred the low transparency (LT) interface and 70% [21] preferred the high transparency interface (HT), as shown in Figure 121. The difference was tested with a Chi Square goodness of fit test. The test revealed that there is a significant difference with a medium effect size between the values ($X^2(1, N = 30) = 4.8$, $p < .05$, $w = .4$). Significantly more participants preferred the high transparency interface.

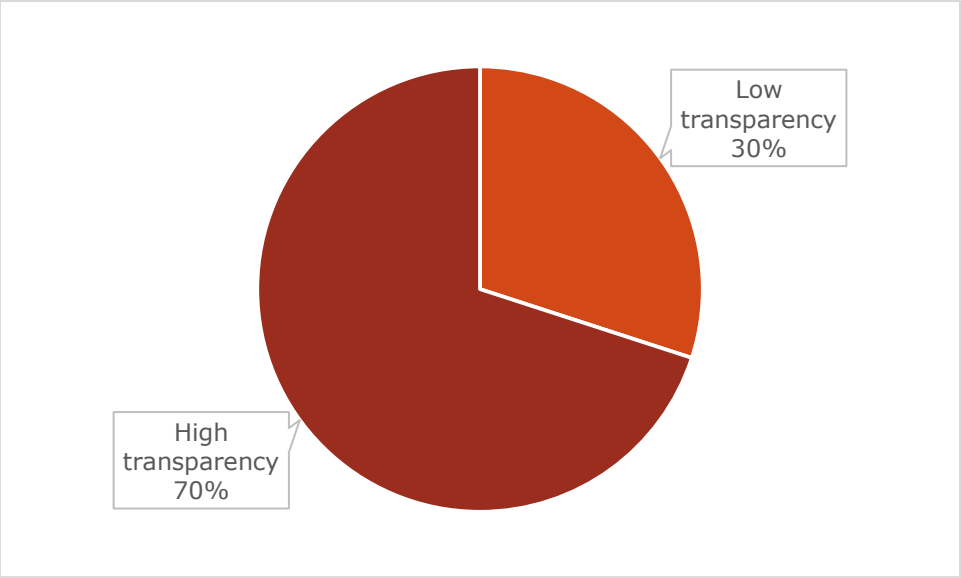


Figure 121 - Pie chart of the interface preference of the participants

Furthermore they were asked as to why they preferred a certain interface. An overview of the themes is provided in Table 44. The comments from participants in each theme are divided into the people who preferred the low transparency (LT) interface and the high transparency (HT) interface. Numbers and themes in **bold** are explained in more detail below the table. Content themes in the detailed analysis are written in *italic*. Direct quotes will indicate the participant number after the statement (e.g. P08).

Content themes	**Count**		
	People who preferred LT interface	People who preferred HT interface	All
HT - too much information	**6**	3	9
LT - easier and less to look at	**3**	1	4
HT - more information (positive)	0	**10**	10
HT - map feature (positive)	1	**6**	7
HT - mission info (positive)	0	**8**	8
HT - reliability indicator (positive)	0	**6**	6
HT - information/robot log (positive)	0	**5**	5
HT - information/robot log (negative)	0	**1**	1
LT - missing elements	**3**	1	4
HT - worry to miss something	1	0	1

Table 44 - TCBA content theme overview of interface preference comments

Six participants mentioned that the reason for preferring the low transparency interface was that the *high transparency interface provided too much information* [6]:

- "I think I may have preferred the first one [LT] with less on it, so you could pay more attention to your surroundings rather than constantly checking. It's useful to have that information, but you don't always need that information." (P13)
- "I didn't check all the information, I just couldn't check everything at the same time." (P19)
- "There was quite a lot of concentration involved, you can't look at everything, all at once [...]" (P26)

Also, three participants mentioned that the *low transparency interface was easier and there was less to look at* [3]:

- "Because it's [LT] easier, the information is more essential." (P01)
- "The first [LT] I think, it had much less to look at. [...] It [LT] had less to concentrate on." (P05)

People who preferred the *low transparency interface mentioned that they still missed some of the elements* [3]:

- Reliability indicator: "I think in terms of saving people's lives the first one [LT] lack the information about reliability." (P01)
- Mission info and mission log: "But obviously if it had the log that would have been good. [...] Probably what rooms and how many victims." (P05)
- High transparency map: "But with the second way-pointing-thing. [the additional lines between the way points on the map]" (P10)

Participants who preferred the *high transparency interface mostly mentioned that they preferred having more information available* [10]:

- "Other than being a gamer I like having as much information, I felt I could make better judgement with more information base." (P04)
- "There was a lot more information in terms of what we are looking for, summaries of what to see and the reliability." (P11)

- "The second one [HT] because it gave me more information. [...] And yea, just having more information, kind of made me feel more confident." (P14)
- "The first one, because even if you have more information it might seem a little bit crowded, is always useful to know." (P16)
- "I liked the first one because it had more information on it." (P24)

Furthermore participants mentioned certain elements in the high transparency interface that they found useful to have:

- *The mission info box* [8]: e.g. "The estimated targets and the targets acquired really useful." (P28).
- *The high transparency map* [6]: e.g. "Also I liked the lines telling me where the robot was going. Because then you know, you can form a map in your head like how the robot is gonna be moving and where you should direct it if you want… at least you know a path. And you can just go somewhere and go back to that path." (P16).
- *Reliability indication* [6]: e.g. "[…] the reliability was important because then you knew how much effort you have to actually put yourself. If it's high, maybe you don't have to be aware of everything." (P16).
- *The robot log* [5]: e.g. "[During LT] When he went behind objects, when you didn't know what he [the robot] was doing, you didn't know when he'd actually done it or if he was just stopping or if he was just glitching or something - when you had the feedback [HT] you know oh he's just determining that that area is clear and now he's gonna move on again. So it's a lot nicer, having that, that sort of feedback. And it felt a lot easier, once I knew, when you know what he's doing, it's a lot easier to kind of let him do it, I guess." (P07). However there was still a participant who did *not like the robot log*: "I found useless, the report of the robot like what it was doing." (P28).

7.4.8.2 **Interface element analysis**

Participants were asked after each trial which elements of the interface they have actively used. This overview does not objectively show which elements of the interface were actually used. The participants did mention the

interface elements they could remember actively using or remembering during the trial. A visualisation of the interface elements can be seen in Section 7.3.2.1 (Figure 109, p.263 and Figure 110, p.265). In the next paragraphs the descriptive analysis for low and high transparency interfaces is discussed.

Low transparency interface

The elements that were mentioned most were battery [26], temperature [25] and map [24]. The signal indicator was mentioned 11 times. There was no great difference between high and low task complexity, as shown in Figure 122.

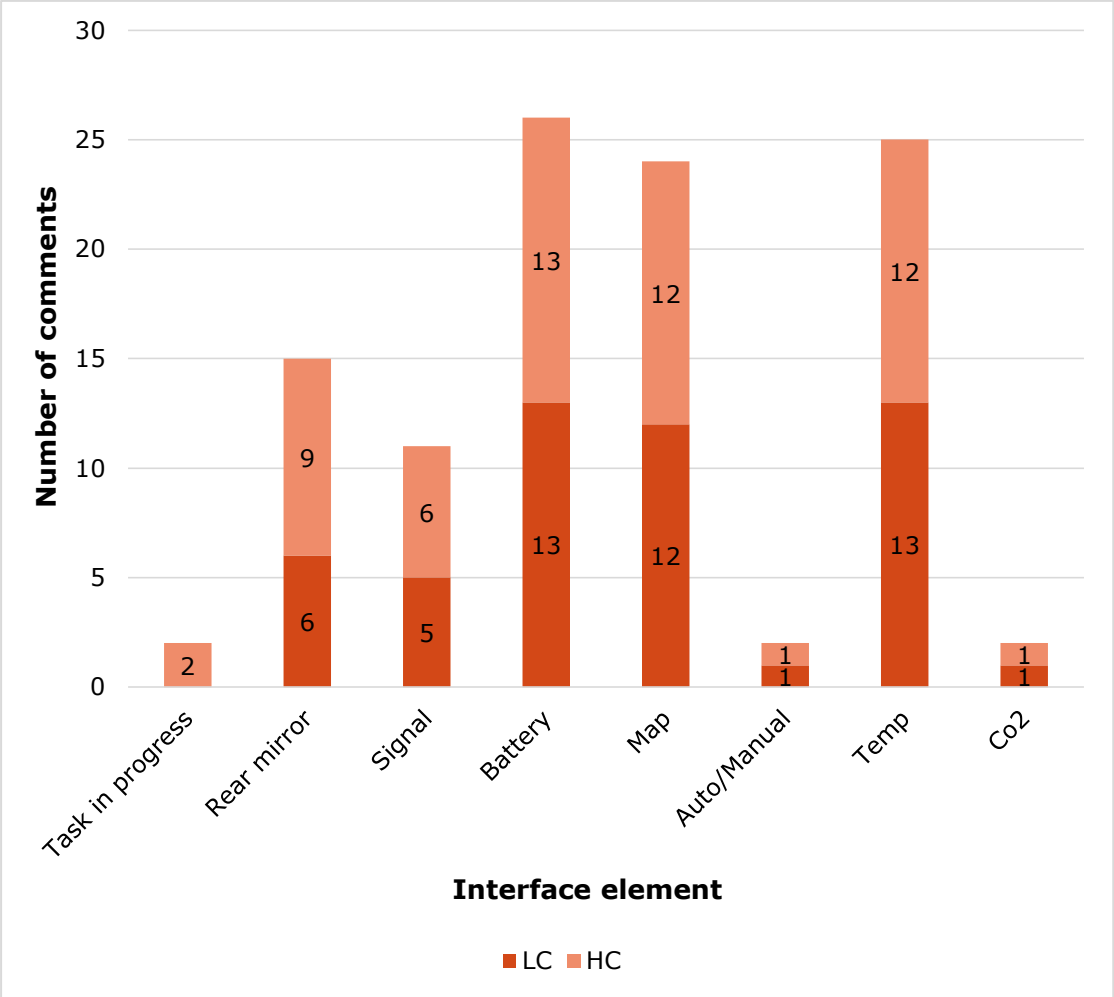


Figure 122 - Number of comments of low transparency interface elements

However, more participants reported to use the rear mirror during high task complexity [9] compared to low task complexity [6].

Below are examples of participants' comments:

- The battery indicator was used to judge the time left before the robot ran out: "I looked at the battery but earlier on, I just wanted to get a feel for how quickly it was dropping." (P02), "I used the battery but just to make sure I wasn't running out" (P11), "I started looking at it probably 30 seconds into it and I saw that it was dropping so I tried to check back with that." (P18).
- The temperature gauge was utilised when people saw a fire in the environment: "When I heard the fire, I checked the temperature." (P09), "When I heard the fire I started looking at the temperature […] and once I stopped hearing it, I saw it dropped pretty quick, so I stopped looking at it." (P18), "Kept my eye on the temperature when there was that fire […]." (P27).
- The map gave participants orientation: "I used the map to see whereabouts I was and the waypoints to get an idea of where I had and hadn't been." (P03), "And also trying to use the map to see kind of the layout of the room to see if the robot missed parts." (P10), "The thing I was using the most actually." (P28). But there was also a participant who thought that the map was confusing: "And the map, I did use but I found more confusing so I sort of had to keep looking which direction I was actually going." (P07).

Participants were briefed upfront that they should make sure that they do not run out of battery and that they take care not to overheat the robot (e.g. driving too near to a fire). This might have led participants to use these elements more than the others. The map was the only orientation aid next to the main view and participants used it to navigate through the environment or to check where they were. The signal was only used when participants experienced lag of the robot (e.g. "Looked at the signal cause it actually got jerky at one point, it lagged." P24).

High transparency interface

As depicted in Figure 123, participants mentioned most the active utilisation of the map [24], the battery [20], the reliability indication [20], and the

temperature [20]. Also the participants said that they actively used the robot log [16] and mission info [13].

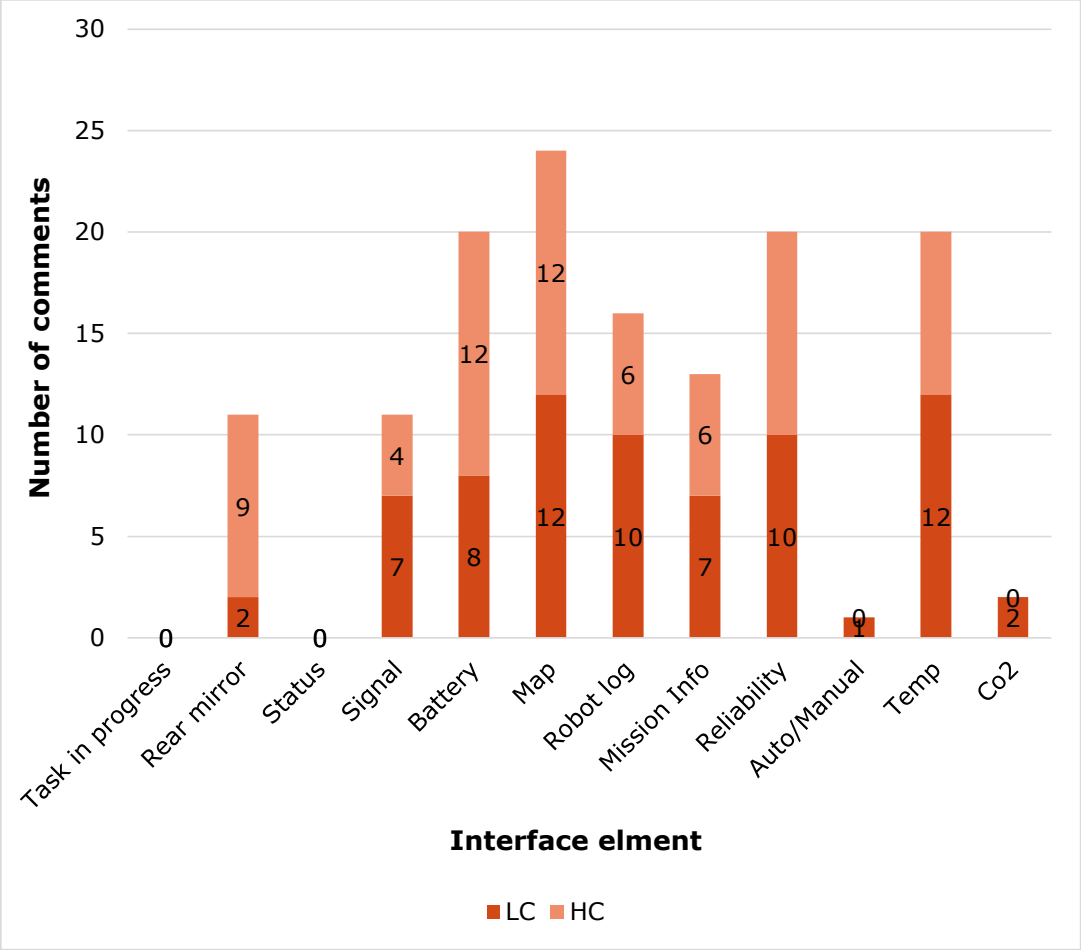


Figure 123 - Number of comments of high transparency interface elements

Examples why participants used some of the interface elements are outlined below.

- Map made the robot more predictable and helped participants when driving manually: " […] and the route again [map], like where I will see where the robot [is] going and why the robot could have a view of this area based on this route, things like that. I kind of predicting where the robot is going to scan this area." (P01), "The lines [lines between the waypoints], I found them very useful, in the plan location [top view map], because you know in which part of your path are. While with the points, especially in the last test [LT], found very difficult when I was turning [manually] to reallocate again. So the lines are very useful for that." (P08), "I used the map to make sure

when there was lines connecting the vantage point [way points] and I used the map to make sure we were searching the whole room and going from one room to another." (P17).

- Battery was used but infrequently: "Just to be aware, I need to finish before the time runs out." (P16), "Probably didn't look at the battery that much cause I knew how much I had used the last time so unless I was running in circles I thought it would be ok so I didn't really look at that until the end." (P27).
- Temperature was used when participants heard or saw a fire: "The temperature, I made sure we didn't get too close to the fire." (P11), "I still used the temperature […] any time I heard fire." (P18), "I looked at the temperature when I went near to the fire. But I had it on auto and I knew it wouldn't drive itself into the fire." (P24).
- The current reliability indication of low, middle and high gave participants an idea of the robots performance and helped them to make decision whether to use auto or manual or search the environment more thorough: "I used the reliability thing, so when it got to medium I was keeping more of an eye out and when it was low and I couldn't really see anything I took control myself." (P02), "I really used the reliability a lot! When it was low I would go to manual and only switch back to automatic when it went high again." (P12), "Yes because if the reliability was low I would attempt to use the manual one if it was high I knew I could rely on the automatic one." (P28). However, it was not entirely clear to all the participants why the reliability level changed: "Oh I did look at that [reliability indication], towards the beginning but I didn't really know why sometimes it was low and sometimes it was high, but it was." (P06), "I used that a little bit. I saw that it went low but then I didn't know why […] so I kind of took note of that and tried to look harder and see if it was missing something." (P19). Showing the explanation of the low reliability in the robot log might help to clarify the robot current reliability level.
- Participants thought the robot log was useful and gave them reassurance: "I checked the log quite a lot, because sometimes like when he saw a door and he said 'I'm saving the position of the door',

I remember where the door is but I'm going on. Then when he turned around and said 'check behind objects and area clear' […]" (P09), "The bit where it told me (indicates mid-bottom) the room I had done, the log, I relaxed a bit when it said everything had been cleared. I checked if it said that while one high reliability. If it'd been low I would have gone back." (P11), "I was looking at the robot-log quite a lot, cause that was quite useful. It was useful cause it would give you kind of ok there has been a couple of people in here, just found the door, here it's blocked, can't go that way […] that was quite useful." (P13), "Just to make sure my passage is not blocked or what's going on. Just checking and making sure everything is going correctly." (P17). There were also a comment that the log as being not useful [1] or understandable [1]: "Because I felt the readouts weren't as useful." (P04). "The log, but I struggled with it at first because I thought it would start at the most recent thing would be on the top - so I was trying to figure that out for a part of the time. So I used it a little bit but then decided that it does not help me because it's like of a past tense thing." (P18). Two participants totally forgot to look at it [2] (e.g. "I don't think I looked at the log at all." P10).

- The mission info was mainly used to look at how many targets they have found so far and how much targets were estimated to be in the scenario: "Particularly when I got to the third room, I could see that I was in the estimated range of targets, so I wasn't worrying that I might have missed loads. Cause I knew it was probably going to be between 5 and 8 targets and I had 6." (P01), "I used the number of victims that had been tagged and the expected number." (P03), "To keep an eye on how many targets I found." (P12), "I looked at how many targets were found." (P15), "I looked at targets [targets found] definitely. I looked at the potential targets." (P24), "I did look at the how many targets found" (P27).

Again, participants were instructed beforehand to take care that they do not run out of battery and not overheat the robot. This might have led participants to use these elements more than the others. The map helped participants to predict the robots behaviour in greater detail and also helped

as a guide when using manual mode. In the low transparency interface condition more participants mentioned to use the rear mirror when they were in the high task complexity condition. The temperature gauge was mainly used on demand, for example if participants heard or saw a fire. The current robot reliability level of low, middle and high supported participants to make decision whether to use auto or manual mode to search the environment more thoroughly. Participants were reassured and better informed by reading the robot log. The mission info box was, for the most part, used to see the number of targets found and the estimated number of targets in the scenario.

7.4.8.3 **Situation awareness**

Participant were asked after each trial through how many rooms they think they drove, how many targets they think they found and how much battery they think was left at the end of the run. This aimed to capture an indication of how aware participants were of the situation regarding the Level 2 of the SAT model.

The average of the percentage deviations of the answers was calculated. For example 31% of the battery was left and the participant answered 25%, this makes a percentage deviation of 6%. The same calculation was done for the number of targets and rooms. The three percentage deviation were averaged across the transparency levels. Pearson residuals (Sharpe, 2015) showed that there was a significant difference between the percentage deviation of the low transparency interface and the high transparency interface ($p < 0.05$).

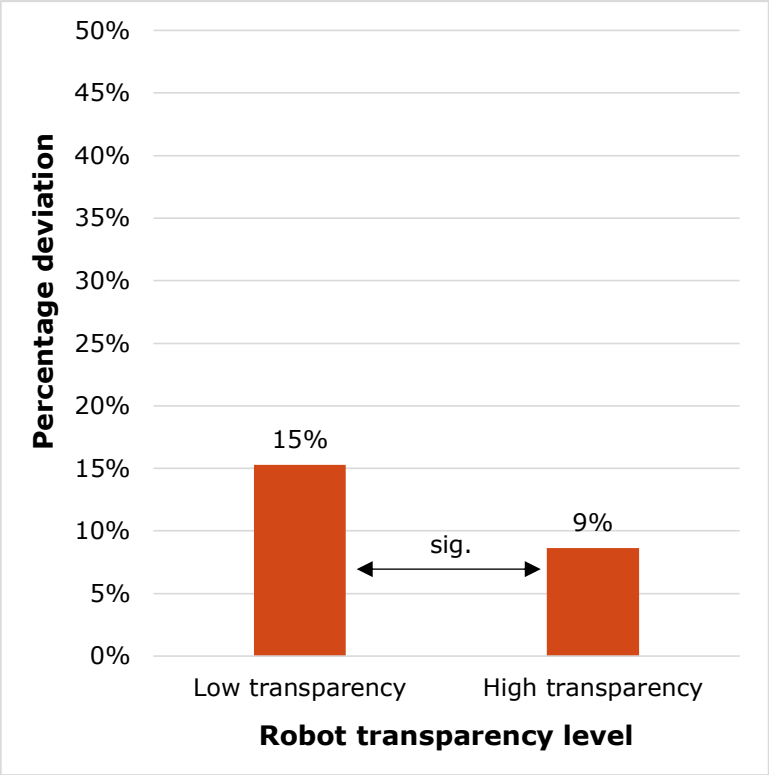


Figure 124 - Situation awareness: percentage deviation across robot transparency levels

As shown in Figure 124, significantly higher deviations from the correct answer were recorded for participants with the low transparency interface (M = 15%), compared to the high transparency interface (M = 9%). This means that participants who used the high transparency interface had a greater level of situation awareness than participants using the low transparency interface. However, room numbers were displayed in the high transparency interface, this could have biased the measurement.

There was no significant difference in situation awareness between high and low task complexity ($p > 0.05$).

7.5 Discussion

The aim of the study was to find out whether robot transparency and task complexity can influence trust, subjective ratings and workload in order to produce higher levels of human-robot team performance.

The first hypothesis stated that robot transparency would influence trust, performance and subjective workload. The interface transparency of the robot, which gave participants different levels of information about the environment, robot status, and robot intent, did not lead the human-robot teams to a higher level of performance. These findings are in agreement with Ososky et al. (2014). Further, there was no change in perceived workload between the conditions.

Trust changed across transparency levels; the lower the transparency, the lower the trust score. This is in accordance with Selkowitz et al. (2015) who found that trust was increasing between SAT level 1 and SAT level 1+2 (more information displayed). They also found that there was no significant change in workload during their conditions. That the workload did not vary can mean that the interface was not too overwhelming for the participants. But interviews suggested that participants might just neglected certain items on the interface.

However, the only great variations in answering the trust questionnaire were among the items "Provide feedback", "Provide appropriate information", and "Communicate with people" (see Appendix J, p. 403). The research question led to the fact that single items on the trust scale were manipulated, which changed the outcome significantly, and therefore these findings needs to be treated with caution.

The event analysis showed that significantly more often participants detected a robot error in the low transparency condition compared to the high transparency condition. This was not expected but suggests that the interface did not appropriately support the participants. It seemed that participants had over-trust in high complexity tasks and under-trust in low complexity tasks. Another reason might be that the high transparency interface was overcrowded and participants spent more recourses on

dealing with the robot, in this case the interface, rather than searching the environment for targets. This was also suggested by Chien and Lewis (2012) who reported a raised rate of unmarked victims when the robot was in high reliability.

The second hypothesis declared that task complexity influences trust, performance and workload. In agreement with the previous study, the objective performance (number of victims found) was significantly different between task complexity levels, as was the rated self-performance. In the low complexity tasks the objective performance and the rated self-performance were higher than in the high complexity task. Since the performance of the robot was the same in all the conditions, the decrease in team performance was attributed to the participant. An event analysis showed that the decrease in performance was mainly due to the fact that the participants failed to see the robot mistakes and therefore missed the targets. This might be due to an over-reliance on the system and miscalibrated trust levels as suggested by Parasuraman and Riley (1997) and others (Atoyan, Duquet, & Robert, 2006; de Vries, Midden, & Bouwhuis, 2003; Lee & See, 2004). Data also suggests that participants under relied on the robot in low complexity conditions.

Whatsoever, task complexity had no influence on trust or workload ratings. This is in agreement with Desai's (2012) findings, where changing the complexity of the robot environment showed no significant influence on trust or workload. On the contrary Adams, Bruyn, Houde, & Angelopoulos (2003) theorised that trust will decrease with higher task complexity, but the theory could not be confirmed by this experiment. In addition, there was no interaction between task complexity and robot transparency on trust.

Further it was investigated if task complexity and robot transparency influence subjective ratings. The rated robot performance was not significantly different across the experimental conditions, this was expected since all trials used robots that were programmed to have the same level of performance. In addition, task complexity and task difficulty were not rated

significantly different across any of the conditions. This was not expected since the task complexity was one of the independent variables.

As expected, the situation awareness questions showed that participants had a better awareness of the battery status, the number of rooms they had searched and the number of victims they had found in the high transparency conditions. Furthermore, participants preferred the high transparency interface. However, participants also mentioned that the interface displayed too much information at the same time. Participants actively used most the battery indicator, temperature gauge and the map in the low transparency interface. In the high transparency interface they mentioned most the map, the battery indicator, the reliability indication and the temperature gauge. A list of six interface design recommendations was compiled and is listed below.

7.5.1 Recommendations for future interface designs

- An essential element in the interface is the map. The map should provide a visible path with navigation points and the direction they came from and they will go to. Best would be to visualise where the robot had already been (e.g. grey out the already driven path lines).
- A robot confidence level or reliability indication can help to adjust the expectations of the operator and help making the decision to use auto or manual mode. Explaining in a short and clear manner the reason for the reliability level might give some participants a better understanding of the robots' situation.
- Battery time needs to be in an observable format (e.g. progress bar), rather than in numbers (e.g. time in seconds) because participants used the progress bar more than the indication of time remaining.
- The target count (how many targets have been found) in the mission info box was very valuable for participants to keep track of the progress.
- Many participants argued that the high transparency interface had too much to look at. Providing information on demand might be trade-off. For example temperature can just be shown when there is

actually too high temperature present, rather than having to check it at all times. Another information than can be optional is the reason for a reliability drop or the robot log.

- There is a need for a customisable interface. Operators should be able to choose the information they want to see or need to complete the task.

7.5.2 Combined discussion

A visualisation of the results of this experiment are shown in Figure 125. Robot transparency positively influenced trust, but the participant's detection of robot errors declined in high robot transparency. Similarly, high task complexity led participants detect less robot errors. The higher task complexity was, the lower was the rated self-performance and the objective team performance (targets found). The reliance on the robot also changed with task complexity: the lower the task complexity, the more participants found targets in manual mode. Vice versa, the higher the complexity of the task, the more participants relied on the robot to find the targets. Furthermore, higher robot transparency fostered the situation awareness of the participants.

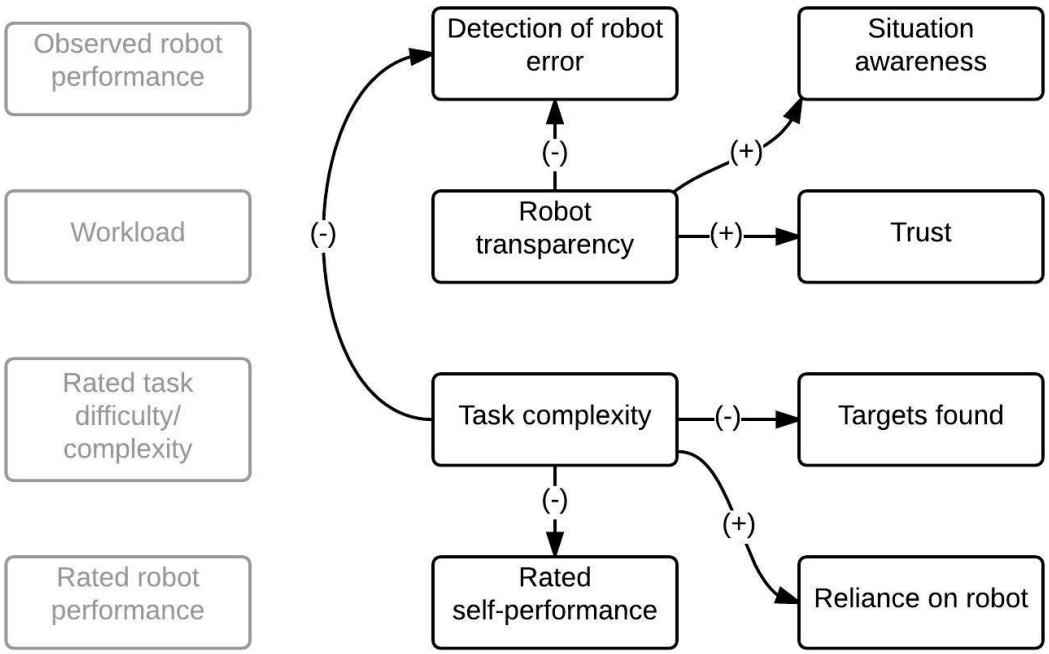


Figure 125 - Qualitative overview of research results of study IV; positive influences are indicated with (+), negative influences indicated with (-)

Since the robot reliability was in all conditions programmed to be constant it is not surprising that the differences for observed robot performance and rated robot performance are not significant.

7.5.3 Limitations and future work

The interface in this study was using the visual channel excessively, because using too many audio cues can be challenging for rescuers in a loud rescue environment with wearing ear protection. The reason why some participants simply neglected elements of the interface might have been caused by an overload of the visual channel. Other interfaces should try to use more channels but keep in mind the environmental circumstances of rescue personnel.

Trust varied significantly between the experimental conditions. This was expected because the only great variations in answering the trust questionnaire were among the items "Provide feedback", "Provide appropriate information", and "Communicate with people" (see Appendix J, p. 403). The research question led to the fact that single items on the trust scale were manipulated, which changed the outcome significantly, and therefore this finding needs to be treated with caution. In order to measure trust more reliably trust games or hypothetical questions, such as "If there was an emergency right now, would you use the robot?" could support the trust measurement and make it more accurate.

As Lyons and Havig (2014) stated, task specific information needed for transparency varies greatly among domains, therefore is it challenging to compare SAT levels across different domains and robotic systems. This can explain why some transparency results differ from other literature.

In regard to the subjective ratings participants did not rate any of the trials as significantly different with respect to task difficulty and task complexity. In the previous study participants rated the high complexity task as significantly more difficult than the low complexity task. A possible explanation for the difference might be the influence of the different robot performances from the previous study and the influence of the robot's transparency in this study.

Furthermore there was an increase in situation awareness when participants used the high transparency interface. This disagrees with Selkowitz et al. (2015) who claimed that increased agent transparency did not support operator situation awareness. The reason for this could be that Selkowitz et al. (2015) encountered in some situation awareness questions ceiling effects or that the measure in this study was biased, because participants were asked to look for the items that were asked in situation awareness questionnaire.

This and the previous study showed that trust in the robot was not influenced by the task complexity. Nevertheless, participants did not rate any difference between the conditions regarding task complexity or task difficulty. This suggests that the robot and its interface might have a larger influence on the rated task complexity than the environment itself. Interestingly rated task difficulty did vary in the previous study. This was not expected since, in both studies, the low and high task complexity levels had the same architecture and characteristics. Further investigation is necessary to determine which variables influence the perceived difficulty of a task.

Furthermore, using verbal protocol technique during a real task scenario with a robot and an operator might shed light on the information that is needed at certain times in a rescue mission.

7.6 Conclusion

Transparency is a mean to support trust calibration and predictability of an autonomous robotic agent. However, producing the right transparency at the right time is a challenging issue. Although this experiment examined two levels of interface transparency of a virtual remote controlled robot, there was no influence of transparency on subjective workload or performance. But transparency did influence trust ratings. Participants had more trust in the robot when the robot provided more transparency. However, only one questionnaire of robot trust was used after each of the conditions. Different measures of trust would give a better picture of trust levels between the conditions. Furthermore, more investigation is needed in terms of the rescue robot interfaces and the required amount and type

of information. There is also the possibility to adopt the method of displaying information on demand.

Task complexity only influenced the objective performance (percentage of victims found): the lower the complexity of the task, the higher the team performance. Since robot performance was programmed to be constant participants were accountable for the decline in performance: they failed to detect robot misses and therefore missed targets entirely. Transparency did not mediate the decline in the performance.

Data also suggest that low task complexity fosters under-reliance and high task complexity over-reliance on the robot.

7.7 Chapter summary

This chapter presented a study that analysed the influence of task complexity and robot transparency on trust, workload and performance. The study demonstrated that trust was influenced by the robot's transparency. The high transparency interface was rated as more trustworthy compared to the low transparency interface. As discovered in the previous study, increased task complexity did influence the human-robot team performance negatively. Workload did not change across any of the conditions. Participants preferred the high transparency interface. However, participants also mentioned that the interface displayed too much information at the same time. A list of six interface design recommendations was compiled. As expected, the situation awareness questions showed that participants had a better awareness of the battery status, the number of rooms they had searched and the number of victims they had found in the high transparency conditions.

8 General Discussion

8.1 Chapter overview

The summary of research findings and the review of aims from the research conducted are discussed in this chapter. Recommendations are made for robot implementation and design. Furthermore, this chapter addresses several discussion points about trust and collaboration in human-robot teams and it highlights the limitations of the research.

8.2 Introduction

The studies in this thesis investigated several aspects of human-robot collaboration and trust within the context of Urban Search and Rescue missions. The idea is to develop robot systems with autonomous features that can support rescuers during their missions by making the rescue work safer and enhancing human-robot team performance. One aspect that seems to play a major role within human-robot interaction literature is trust. The literature review showed that appropriate levels of trust in human-robot teams is the key factor for determining automation usage (Lee & See, 2004), minimising misuse of the system (Parasuraman & Riley, 1997), improving safety and productivity in teams (Hoff & Bashir, 2014), and accepting robot-generated information (Freedy & de Visser, 2007).

So far the literature has concentrated mostly on trust in automation (M. S. Cohen et al., 1998; R. R. Hoffman, Johnson, Bradshaw, & Underbrink, 2013; Muir & Moray, 1996; Muir, 1994; Ross et al., 2007; L. Wang et al., 2009). Recent advances in robotics has led to further research in human-robot teams (Gao et al., 2012; Groom & Nass, 2007; Harriott, Buford, Zhang, & Adams, 2012; G. Hoffman & Breazeal, 2007). With respect to semi-autonomous robot systems, literature looked at autonomy levels (Chien & Lewis, 2012; Larochelle, Kruijff, & Van Diggelen, 2013b), feedback types (Desai et al., 2013; Jung & Lee, 2013; Kaniarasu et al., 2013), varying robot reliability (Chen & Terrence, 2009; Chien & Lewis, 2012; Desai et al., 2012), and system transparency (T. B. Chen et al., 2014; Helldin, 2014; Lyons &

Havig, 2014; Ososky et al., 2014; Sanders et al., 2014; Selkowitz et al., 2015). This thesis took up recent research and further investigated factors influencing human-robot teams: robot reliability, robot feedback, robot transparency, and task complexity.

Furthermore, this thesis contributed knowledge in terms of gathering information on the work carried out by Urban Search and Rescue personnel in the U.K. and the equipment which they currently use along with detailed qualitative analyses of participants interacting with autonomous and semi-autonomous robot systems. Previous research did not investigate tasks, processes, and behaviours of British Urban Search and Rescue teams.

A new approach for measuring performance in semi-autonomous robot systems was developed. During the course of the studies new influencing variables relating to semi-autonomous performance measures emerged: it was possible that participants did not perceive certain robot mistakes, and their observed robot performance differentiated from the intended performance programmed by the researcher. In order to acknowledge the actual witnessed performance of a robot by the participant a new measure of 'observed robot performance' was introduced. Section 6.3.5 (p. 198) details the calculation and application of this measure.

The findings from Chapters 4 to 7 show that autonomous robot features have the potential to support operators and help them to make better choices of function allocation (e.g. letting the robot drive or operate it manually). However, the reliability of the robot strongly influenced whether the system was beneficial to the overall human-robot team performance. This suggests that there are circumstances when a robot's aid is not appropriate and has the potential to decrease performance levels. In addition, the supervision of the robot can take up more attention than the search task itself (Chien & Lewis, 2012), which should not be intended, since the main objective is to find as many casualties as possible rather than accurately supervising the robot's work.

8.3 Discussion of research findings

Figure 126 summarises the key research findings of the individual studies and the corresponding objectives that were discussed at the beginning of this thesis (Section 1.5, p. 7). The arrows in the figure visualise how the information flow and studies influenced each other. Between the first two studies and the last two studies was a longer development phase of the virtual reality game environment as well as the programming of the robot behaviour.

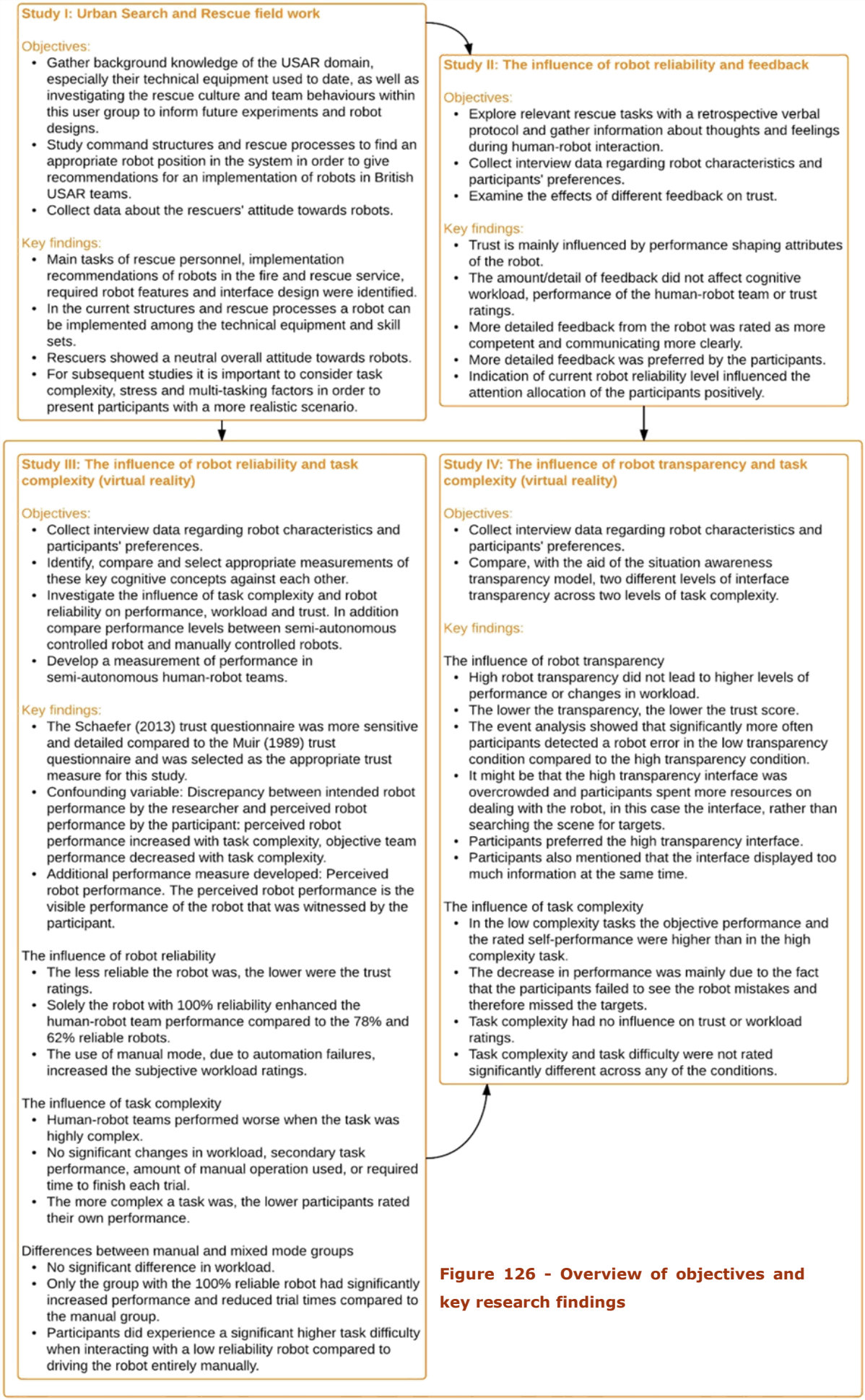


Figure 126 - Overview of objectives and key research findings

The next section discusses each aim and objective in detail regarding the results obtained during this doctoral research.

A visualisation of the findings is shown in Figure 127. The black arrows indicate the findings from the studies conducted in this PhD (see Sections 5.5.2, 6.5.4, and 7.5.2). The orange arrows show that the connections have been verified by other literature. Dotted lines indicate discussion points that are explained in the next paragraphs. Green arrows visualise findings from the literature.

Some concepts incorporate different factors and were combined under an umbrella term. *Appropriate control allocation* is dependent on the *detection of robot errors*, the amount of *manual mode usage*, the participant's *reliance* and the *allocation of attention* towards the robot. There are more factors that influence *appropriate control allocation*, but they were not part of this research. *Mission performance* in this work was mostly determined by the *time taken* to complete the task and how many *targets* were *found*.

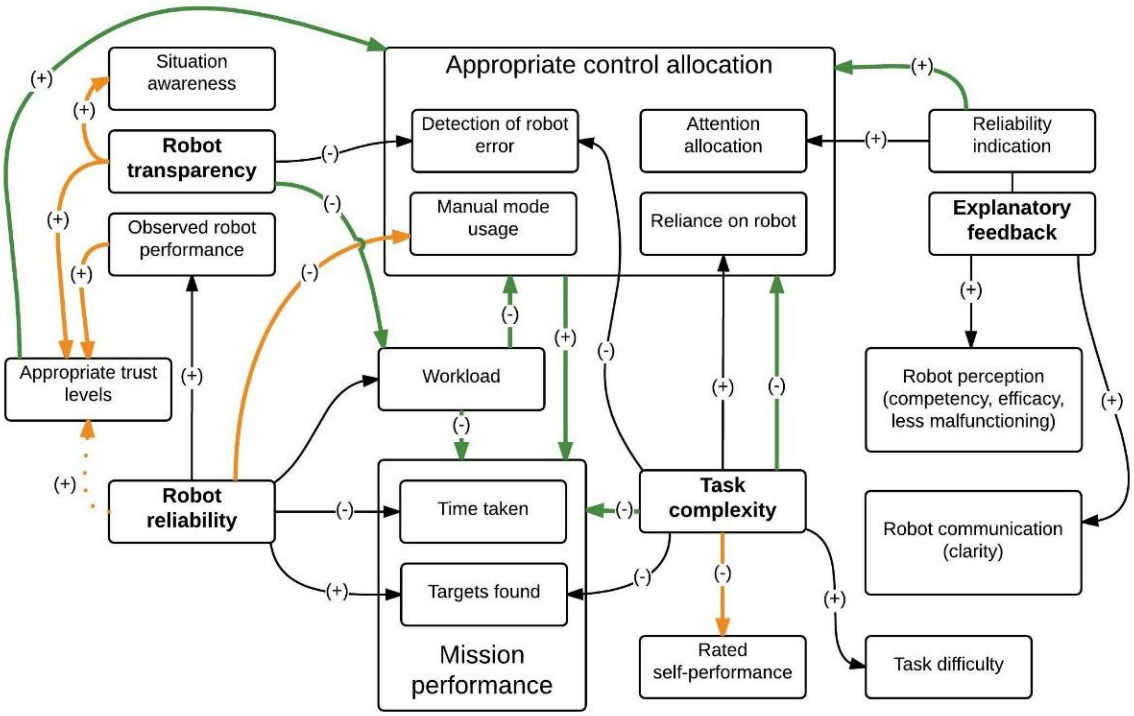


Figure 127 - Summary of research findings; black arrows indicate findings from the studies of this PhD; orange arrows indicate verification of research findings by other literature; green arrows indicate other findings from literature; positive influences are indicated with (+), negative influences are indicated with (-).

Research findings (black and orange arrows)

Some of the research findings of this PhD were confirmed by other literature. The influence of *task complexity* (task difficulty) on *rated self-performance* was also revealed by the HARRT model from Desai (2012). He found that the more complex a task is, the lower participants rate their own performance. He also found that *manual mode usage* produced high levels of *workload*.

Larochelle et al. (2013a) found, that if the expectations of participants about the robot performance (*robot reliability*) were not met, they changed to manual control (*manual mode usage*). In the literature robot reliability is named as a direct influencing factor on trust (Chen & Terrence, 2009; Chien & Lewis, 2012). However, it needs to be considered that the *robot reliability* is the objective measure of the robot's performance, but the actual influence of trust is from the *observed robot performance* (what performance the participant actually noticed). This was also stated by Ososky et al. (2014) who mentioned that trust is not based on what the robot can or cannot do, it is based on what the human perceives it to be capable of.

As found in this PhD research, *robot transparency* influenced *trust*. Likewise other authors noted that the lack of background information for a shared understanding, leads operators to trust robots less or have miscalibrated trust levels (Ososky et al., 2014; Stubbs et al., 2007). Chen et al. (2014) found that system *transparency* provides a better *situation awareness* for participants, which is in agreement with this work.

Further literature findings (green arrows)

Authors also found other connections among the variables shown in Figure 127. All connections that emerged from the literature are indicated with green arrows in order to show the big picture. Desai (2012) showed that *task complexity* and confidence feedback (*reliability indication*) influenced the entire concept of control allocation strategy (*appropriate control allocation*) and of the overall *mission performance* and not only the number of targets found. Desai (2012) also claimed that the higher the *workload*, the less appropriate is the control allocation strategy and the lower is the

mission performance level. Other literature also mentioned that the *appropriateness of control allocation* in a human-robot team determines the *mission performance* (Desai, 2012; Groom & Nass, 2007).

A variety of authors proposed that *appropriate levels of trust* lead to *appropriate control allocation* (Beer et al., 2014; Lee, 2008; Muir & Moray, 1996; Yagoda, 2011). Finomore et al. (2012) suggested that too much information provided by an inappropriate level of *transparency* can lead to increased levels of *workload*.

The following section will provide a review of the aims that were outlined at the beginning of this thesis.

8.4 Review of aims

The overall aim of this PhD was to understand how robot behaviour and interface design can be applied to utilise the benefits of robot autonomy and inform future human-robot collaborative systems. The individual aims were originally described in the introduction (Section 1.5, p. 7). The achievements of each of the aims are discussed below.

- Aim I: Develop a background understanding of the USAR domain and their work as well as describing the real world application of USAR in order to provide recommendations for the implementations of robots in British USAR teams.

An autoethnographic study was completed over two weeks by attending the USAR technician course at the Fire Service College in the UK. The method of an autoethnographic study had not yet been reported in the literature in this application domain.

The data collected produced background knowledge of the domain and gave insights into the work of rescuers. Rescue workers are professional firefighters who are trained in dealing with stressful situations which include high physical and psychological demands. The main tasks of USAR technicians are to extricate casualties out of danger with the help of technical equipment, dogs and by shoring, lifting, moving, or breaking and breaching through obstacles (e.g. rubble, concrete, etc.). The technical

equipment which they currently use consists of cameras (SnakeEye and SearchCam), acoustic and seismic life detectors, and gas monitors.

The recommendation of this research is to implement the robot within an existing USAR technician specialist's team as an additional rescue tool. Robot handling can be part of the special skill sets of USAR technicians and not every member needs to be trained in it. Establishing a new robot operator position outside the team is not recommended due to an additional source of information for the sector commander and additional management of people directly operating at the incident site, as well as higher personnel costs. To be able to use all the benefits of the technology the entire command chain has to be briefed about the robot and its capabilities in order to work efficiently and effectively on site. Communicating the robot's capabilities can also foster better mental models (Fischer, 2014). Knowing what the robot is capable of was also desired by participants (see Study II, Chapter 5). In another study operators interacting with a multi-robot system wanted reference materials of what the robot is capable of (Rule & Forlizzi, 2012).

For a fast deployable robot, it is proposed that the robot is stored in the modules provided for USAR operations (Module 1 First Strike alongside technical search equipment). That infers that the robot is packable and able to be stored easily and safe (see also Booysen & Mathew, 2014).

It is debatable whether the robot should have autonomous capabilities. To ease the acceptance and foster the use of a robot within the fire service, a step wise approach from manually operated systems towards robot collaboration (robots with automation) is advisable. The robot can be potentially used during all stages of rescue (provided it is intrinsically safe, terrain fit and lightweight). That means that right from the start, the robot can be deployed in dangerous zones in order to perform first valuable reconnaissance missions. However, the main deployment stage would be the exploration of voids and spaces, where humans are not able to enter.

Adoption of the system might be very much dependent on the type and performance of the system. The technicians must be reassured that the robot is just another tool and not something that replaces them; it is a tool

that is there to make their work safer and potentially more effective. The attitude of the USAR technicians of the first study (Chapter 4) towards robots was neutral. Rescuers were more positive towards situations and interactions with robots in general, compared to the social influence of robots and emotions in interactions with robots.

The experiences of the researcher and relevant literature informed all subsequent studies in terms of the design of the scenario (robot sensors, contextual information, minimum amount of training required) and the task performed by the participants (varying task complexity, stress inducing elements, realistic search operation).

- Aim II: Improve understanding of underpinning cognitive concepts, thoughts and behaviours of participants while interacting with different autonomous and semi-autonomous robots, in order to inform future robot behaviour and interface design as well as the subsequent studies of this PhD.

The second study (Chapter 5) revealed a variety of recommendations about robot features and behaviour, such as illustrating the importance of visual cues about robot running processes and the influence of robot speed/movement on human attention. In general, qualitative data presented in this thesis (Sections 5.4.5 and 6.5.6) suggests that a higher degree of robot transparency is necessary for the operator to understand the robot's actions and anticipate its next steps. This is in agreement with recent literature (Boyce et al., 2015; Kruijff et al., 2014; Larochelle et al., 2013a). For details of recommendations that emerged from the retrospective verbal protocol please see the list of recommendations provided in a later chapter (see 8.6 Recommendations for robot behaviour and interface design). Retrospective verbal protocols had not previously been used in conjunction with autonomously operating robots.

Furthermore, the interaction with an autonomous robot revealed that participants associated trust with the robot's performance, reliability and consistency. The result is in accordance with other researchers who investigated trust factors in human-robot teams (Hancock et al., 2011; Park, Jenkins, & Jiang, 2008). In another study (Chapter 6) participants

operated the robot entirely manually and they associated trust with their own performance rather than the robot's performance. Trust in the robot depended on how much the robot was able to cope with the environment they were steering it in. This shows that trust does not have fixed parameters. Trust in manually operated machines focussed on technical reliability, and trust in robots with autonomous features seems to focus on their decision making capabilities and associated performance (cf. Chapter 5, Section 5.4.6.3). This is important for the decision as to which trust questionnaire is more appropriate to use and whether trust in the technology or in the artificial intelligence is measured.

When interacting with entirely autonomous robots (Chapter 5) which provided different amounts of feedback (e.g. reporting the reason for the current robot state), participants favoured the robot with more feedback and perceived the robot as more competent, malfunctioning less and as communicating more clearly. When participants interacted with a semi-autonomous robot (participants were free to choose between using manual or auto mode) their main reasons for using auto mode were that auto mode was less demanding and relaxed them (Chapter 6). In addition, they reported that they used auto mode because they felt that the robot did a good job, performed well and that the robot was predictable because the robot followed a systematic approach.

In terms of semi-autonomous robot interaction, participants preferred a more transparent interface (Chapter 7) that provided more information about the robot's current state, explained the robot's current state, and predicted future states. Although participants did not experience increased workload when interacting with the higher transparency interface, they were concerned that the interface displayed too much information simultaneously. Participants neglected some of the interface elements to keep up with the information displayed. This suggests that some information should be displayed on demand.

- Aim III: Investigate how robot and environmental characteristics, influence user cognition, behaviour and performance.

The literature review (Chapter 2) revealed that trust plays a major role in human-robot interaction. When people interact with automated or semi-autonomous systems their subjective trust in these systems can predict and influence the allocation of functions within the human-automation system (Muir & Moray, 1996). An appropriate level of trust is key to the usage of automated systems (Lee & See, 2004), whereby an inappropriate level of trust can lead to misuse or disuse of the system. For example, errors can occur due to over-trusting or under-trusting the system and eventually the potential benefits of the automated system can be lost (Parasuraman & Riley, 1997). This appropriate level of trust is also key to improving safety and productivity (Hoff & Bashir, 2014).

Different concepts and measurements of trust were under examination. Most trust literature can be found in the interpersonal trust research and trust in automation domains. Some of the concepts show similarities to the new area of human-robot trust. Major factors across several trust concepts are human traits (e.g. self-confidence), previous experience, and dispositional trust. Another characteristic that emerged was that trust transforms over time and is therefore dynamic.

It is important not to neglect the technology acceptance model. The main factors associated with accepting technology are the perceived ease of use and the perceived usefulness. To date, only little evidence is available that rescue robots are beneficial in real world emergency scenarios (Matsuno et al., 2014; Steinbauer et al., 2014). It seems that in order to create an intention to use robots, the robot primarily needs to demonstrate usefulness in the field and be perceived as easy to use.

Different measures of trust are discussed in the literature. Most questionnaires aim at certain types of technology (e.g. automation or physical systems) or types of robots (e.g. co-located or remote). After identifying two of the questionnaires appropriate for semi-autonomous remote controlled robot systems, they were compared against each other. This revealed that the Schaefer (2013) trust questionnaire was more

sensitive towards changes in trust than the Muir (1989) questionnaire. Thus, the Schafer (2013) questionnaire was selected to be used in subsequent studies.

This research examined robot and environmental characteristics such as robot feedback, robot reliability, robot transparency, and task complexity in order to understand their influence on trust, workload and performance. The findings of the studies conducted to address Aim III are discussed below.

A new measurement for semi-autonomous robot systems:

There was a discrepancy between the intended robot reliability which the researcher programmed and the actual robot performance that the participants witnessed. This could be due to participants failing to see a robot's mistakes and believing that the robot was reliable. Therefore, they perceived the robot as more reliable than intended. To overcome this shortfall a new measure was proposed: observed robot performance. This observed robot performance measures how the participant perceived the robot's performance by counting the number of witnessed robot's positive and negative actions. Later this was compared to the intended robot performance for a more accurate data analysis.

Robot feedback:

More detailed robot feedback (providing a reason for the current robot state) from an autonomous robot (Chapter 5) did not influence trust, workload or performance (objective team performance). However, there was also a large learning effect present, which could have been the reason for not finding any significant differences.

Task complexity:

In terms of semi-autonomous robot systems, task complexity (number and type of targets and amount of obstacles in the environment) influenced trust only in the third study (Chapter 6) and not in the fourth study (Chapter 7). In the third study participants rated the robot in less complexity tasks as more trustworthy than in middle complexity tasks. This could have been influenced by the observed robot performance (the performance of the robot

that the participant witnessed) because participants observed the robot's performance in high task complexity as being better than the low task complexity robot performance. The different perceived performances could have influenced the trust ratings, since performance is one of the main influencing factors on trust (Hancock, Billings, Schaefer, et al., 2011).

Therefore, observed robot performance should be recorded (e.g. video recording or observation) during experiments because participants failed to see a robot's mistakes because they missed them too, they believed that the robot was more reliable and perceived the robot's performance as higher than intended by the researcher. The same issue occurred during the experiments of Rovira et al. (2007) and Chien and Lewis (2012). They suspected that because of the absence of alarms, system failures were not detected and hence participants could not easily discriminate between low and high robot performance.

Both, the third and fourth study showed that performance declined when the task was more complex. A more detailed analysis of the fourth study showed that the decrease in performance was mainly due to the fact that the participants failed to see the robot's mistakes and therefore missed the targets. This might be due to an over-reliance on the system or miscalibrated trust levels as suggested by Parasuraman and Riley (1997) and others (de Vries et al., 2003; Lee & See, 2004).

At the same time, data suggested that during low task complexity, participants under-relied on the robot. In addition, task complexity had no influence on subjective workload in both studies (study III (Chapter 6) and study IV (Chapter 7)).

Robot reliability:

In the third study robot reliability was varied across three conditions and data showed that the less reliable the robot, the lower the trust ratings. This influence of reliability on trust was expected, since robot performance is the main influencing factor on trust (Desai et al., 2012; e. g. Hancock, Billings, Schaefer, et al., 2011). With regard to reliability and objective team performance, the objective team performance did not increase continuously

between low and high robot reliability conditions. Compared with entirely manual operation, data showed that a good, or very good, working robot can enhance the performance of a USAR mission but differences between lower reliability levels (between 62% and 78% robot reliability) did not show a significant increase in performance. These results are similar to de Visser and Parasuraman (2011) who found that imperfect automation had a performance benefit if the reliability was over 70%. In study III (Chapter 6) only the high reliability robot was able to significantly increase the performance and decrease the time required for the scenario, compared to the manually operated robot. That suggests, unless the deployed robot has very high success rates or reliability, the utilisation of automated robots features in USAR might not be beneficial for the overall mission performance.

The subjective workload with respect to varied robot reliability levels only increased between the middle reliability and low reliability condition. A lack of significant difference between workload between high and middle/low reliability workload conditions could also be due to the characteristics of a vigilance task, where very high levels of sustained attention (supervising the robot in high robot reliability and not interfering with the system) induces hypostress and results in high levels of subjective workload (cf. Bainbridge, 1983; Warm et al., 2008).

Robot transparency:

The fourth study (Study IV, Chapter 7) examined the influence of interface transparency. The interface transparency of the robot, did not lead the human-robot teams to higher levels of performance. These findings are in agreement with Ososky et al. (2014).

However, trust changed across transparency levels: the lower the transparency, the lower the trust score. This is in accordance with Selkowitz et al. (2015) who found that trust increased between levels of transparency (more information displayed). They also found that there was no significant change in workload during their conditions. Similarly, the fourth study of this thesis found that there was no change in perceived workload between the transparency levels.

Interviews showed that participants neglected certain items on the interface. Further, the event analysis showed that participants detected a robot error significantly more often in the low transparency condition compared to the high transparency condition. This was not expected but suggests that the interface did not appropriately guide participant's attention. Another reason might be that the high transparency interface was overcrowded and participants spent more resources on dealing with the robot, in this case the interface, rather than searching the environment for targets. This was also suggested by Chien and Lewis (2012) who reported a raised rate of unmarked victims when the robot was in high reliability mode. This stresses the importance of appropriate system transparency and interface design.

The next section describes the recommendations that emerged from this thesis.

8.5 Review of novel contributions

The proposed novel contributions from the beginning of this thesis can be found in Section 1.6. This section will review the contributions in detail. The PhD addressed a variety of gaps in the literature and aimed to add the following novel contributions to the body of research knowledge:

Only limited data about USAR processes, equipment and command structures are openly available to the public (HM Government, 2008; "The Personal Qualities and Attributes [Website]," 2014), detailed information about rescue work is still missing. Study I (Chapter 4) addressed this deficit and provided an original contribution to knowledge by giving insight into the organisational structures, rescue processes, and currently used equipment of the USAR rescue personnel in the U.K. and linking the findings to robot requirements.

Retrospective verbal protocols had not previously been used in conjunction with autonomously operating robots. Study II (Chapter 5) used this verbalised thought method, collected extensive qualitative insight, and a list with interface and robot behaviour recommendations was compiled which provides another original contribution to knowledge. This thesis could also

show that giving explanatory information in addition to a robot's confidence feedback did not further increase workload.

Many authors concentrated on autonomous machines and robots (Hoff & Bashir, 2014; Merritt, 2011). However, a complex task such as Urban Search and Rescue still needs the operator in the loop and requires operators to take over certain aspects of the search tasks (e.g. identifying casualties) (Virk et al., 2008). In this case semi-autonomous robots are required. This thesis proposed a new semi-autonomous robot team measure (observed robot performance) that can help identify the source of influences on trust and control allocation (Section 6.3.5.1, p. 200).

Search and rescue teams encounter unpredictable environments (Y. Liu & Nejat, 2013) that can be highly complex. Investigation of task complexity relevant to USAR missions is of importance to design useable robot systems (Desai et al., 2013) but received limited attention in previous research. Study III (Chapter 6) and study IV (Chapter 7) investigated the effects of task complexity in virtual USAR missions. This thesis showed that USAR relevant task complexity is of importance and influenced performance and control allocation. In general, task complexity had no influence on workload, but less complex tasks produced higher performance levels and rated self-performance levels (Chapter 6 and Chapter 7). Interestingly, according to study IV (Chapter 7) task complexity influenced the control allocation strategy of participants significantly. Participants allowed the robot to find more targets in the high complexity conditions and missed more robot errors than in the low task complexity conditions. Participants found more targets in manual mode during low complexity tasks. Although trust in the robot was affected by task complexity it can be assumed that participants in study III were influenced by the performance of the robot they observed rather than the actual programmed robot performance (which was the same across task complexity levels). Therefore, this thesis demonstrated that robot performance has a stronger influence on trust and performance than task complexity.

In addition, Study III (Chapter 6) showed that autonomous robot features only benefit the human-robot team performance when the robot's reliability

exceeds approximately 60% - 70%. Therefore, autonomous features do not provide necessarily elevation from workload and enhance mission performance. This finding is of importance for future research and an advice to compare semi-autonomous human-robot team performances more often to manual achievable performances to better show the actual benefits autonomy can or cannot provide in human-robot teams. However, robots still provide the ability to access areas that are too dangerous or unreachable for humans.

Furthermore, transparency is an emerging concept that aims to enhance human-robot team performance and was worth further investigation (Boyce et al., 2015; Lyons, 2013). Study IV (Chapter 7) investigated the effects of robot transparency in semi-autonomous human-robot teams in a virtual USAR scenario. Robot transparency did influence trust ratings but it did not influence the human-robot team performance. It was discovered that participants might use too much attention on supervising the robot (and the provided information) rather than searching the environment for targets. An indication for this could be that participants detected more robot errors in the low transparency conditions compared to the high transparency conditions.

In human-robot interaction several trust questionnaires exist (Jian et al., 2000; Muir, 1989; Schaefer, 2013; Yagoda & Gillan, 2012) but so far literature did not compare these questionnaires with each other regarding their usage in remote controlled semi-autonomous robot systems. This PhD investigated the difference between the Muir (1989) trust scale and the Schaefer (2013) trust questionnaire and found that the Schafer questionnaire is more sensitive to changes in trust but the Muir questionnaire is faster to administer.

This PhD contributed a comprehensive list of recommendations for robot features, behaviours and interface design. Moreover, data suggests that the virtual reality approach of study III and IV (Chapter 6 and Chapter 7) produces similar results to real robot systems (Desai, 2012; Kaniarasu & Steinfeld, 2014; Larochelle et al., 2013a) and seems to be a valid method to investigate semi-autonomous remote controlled robot systems (also see

Chien & Lewis, 2012; Gao et al., 2013; Horsch et al., 2013; Robinette et al., 2015).

8.6 Recommendations for robot behaviour and interface design

The recommendations presented in this section emerged during the development and execution of the experiments. It needs to be considered that in three of the four studies participants were not experts and that some recommendations need to be validated (e.g. focus groups, interviews, etc.) with data gathered from rescuers that have experiences working in the field. The recommendations are divided into organisational, robot physical, robot functional, and robot interface recommendations.

Organisational key points

Robots are replaceable, humans and animals are not. The key advantage of USAR robots is that they can be deployed in dangerous areas that humans and animals cannot access. Robots can be used right from the first phase of search and rescue as well as in the following stages (for the stages of rescue see: Section 4.4.1.2.1 Search management, p. 98). It is advisable to implement a robot within the skill set of a USAR technician. These rescuers already have experience with technology and may be likely to accept a robotic system. Therefore, a robot could be implemented at the standard USAR operational level and can be part of the USAR Module 1 (First Strike), where the technical search equipment is also stored.

Although there is a need to understand a robot's behaviour and functions (or even internal states), training should be short and easy to understand. For a robot to be successfully implemented and accepted within the USAR community, it needs to be fast and easy to deploy and demonstrate usefulness. In order to do that, there needs to be a clear understanding of what the robot can and cannot do. This knowledge should be known along the entire command chain to avoid misunderstandings. Another important point is the distribution of information gathered by the robot. Data should be readily available for the command centre (e.g. pictures/videos of possible explosive devices) without the need to bring a memory device from the

emergency scene to the relevant person in command. The robot needs to become an integrated part of the rescue apparatus.

Physical and functional robot requirements

Perhaps most critical is the easy deployment (packable/wearable) and the size of the robot: small enough to be able to access voids and big enough to not fall between rubble. To have a robot that is universally deployable it needs to be intrinsically safe and resistant against water, heat, and dust.

The obvious feature necessary to do reconnaissance for USAR is a camera. The camera picture should be big and undistorted. There needs to be a trade-off between data volume and video feed quality. Furthermore, sound can be useful for communicating with trapped casualties (e.g. Survivor buddy in Murphy et al. 2011) or to gather more information of the environment (e.g. sounds that indicate trapped casualties). For some participants (non-experts) it was very important to be able to move the camera independently from the driving direction to have a better field of view and situation awareness.

Sound can also be helpful to navigate the robot. Motors will sound different when they are exposed to higher friction. The following additional features are useful to rescue workers: air quality sensors, infrared sensors, 3D scanning/automatic mapping of the environment, small grabber.

The robot's operator control unit needs to be controllable by rescuers who are wearing personal protective equipment. Software features that can support rescuers most can be automatic navigation over irregular terrain including path planning and collision avoidance as well as the identification of explosive devices and casualties.

It may not be possible for a robot to have all of the features described above, but all of the features are necessary at some point. Having a robot that is only made for one task (e.g. capturing video data) might not be as useful as a robot system that is able to be configurable to the needs of the mission at hand (e.g. attach air quality sensors or infrared camera). A modular robot system can be a huge advantage in search and rescue due

to the unpredictable environments that rescuers and their equipment may have to face.

Interface and robot behaviour recommendations

When designing an interface for search and rescue robots it needs to be taken into account that operators may be sleep deprived, exhausted and under constant stress. Due to the nature of their physical and psychological state, the display should be uncluttered, clear and big. The list provided below describes possible interface elements that derived from this thesis:

- Forward facing camera and backward facing camera view.
- Continuously visualising the status of the robot.
- Battery time, air quality, and temperature needs to be in an observable format (e.g. relative indication with a progress bar), rather than in numbers (e.g. absolute information such as time in seconds) because participants (non-experts) used the progress bar more than the indication of time remaining or numbers.
- An essential element in the interface is the map. The map should provide a visible path with navigation points and the direction they came from and where they will go to, which supports the predictability of the robot. It would be best to visualise where the robot had already been (e.g. grey out the already driven path lines, overlay visited areas, etc.). The robot could also provide information about where it might be able to drive to (e.g. possible moving grid).
- The target count (how many targets have been found) in the mission information box was very valuable for participants to keep track of the progress.
- Top view pictogram of the robot can make it easier to indicate the position/part of a technical fault.
- A robot's confidence level or reliability indication can help to adjust the expectations of the operator and help make an informed decision whether to use auto or manual mode. Explaining in a short and clear manner the reason for the reliability level might give some participants a better understanding of the robot's situation.

Giving a starting message, so that operators can familiarise themselves with the robot's voice and the level of loudness can be useful. There is also a need for a customisable interface. Operators should be able to choose the information they want/need to see to complete a certain task. However, it needs to be established which elements need to be mandatory. High transparency interfaces can have too many elements to look at. Providing information on demand might be a trade-off. For example, temperature can just be shown when the temperature is too high, rather than having to check it at all times.

Another feature of the robot that would support rescuers is a system that is able to identify explosive devices and casualties. During this thesis the robot was able to have such a feature and the following recommendations were made by participants (Chapter 5):

- Visualising the process of target identification of the robot can help the operator to understand what the robot is doing. This could incorporate a visual overlay and a loading bar as well as the information if the robot tries to get another angle/view upon the target.
- It would also be useful if operators (if they already identified the object) can abort the identification and declare that it is a target/no target to save time.
- It seems that a more detailed explanation of the robot's victim identification process is necessary. This missing information could be provided during training: a representation of the decision making of the robot and the mechanism of identifying targets (iteration of planes, points, heat pattern, etc.) could be beneficial.

Autonomous robot behaviour also can influence an operator's choices and actions. Robot movements have an influence on attention allocation, not only what the robot sees, but also how fast or slow it moves. Below is a list of recommendations about robot navigation and movement:

- Even though the robot might be autonomous it should be possible to slow down or speed up its autonomous driving speed to be adaptable to different skilled operators and different environments.

- If the robot turns towards new areas, a surround view of this new area would provide the operators with a higher level of situation awareness.
- Navigation goal points which indicate where the robot will drive next was liked by operators. They were able to predict the robot's movements and they even followed the navigation goal points in manual mode.

Participants were more likely to trust the robot's judgement when they were uncertain about a target. This could lead to over-trust in the robot. It could help to contribute to the human's decision making process by providing a percentage of accuracy (how sure the robot was that it identified/did not identify a target) or a general reliability indication. It can allow the operator to make an informed decision about further actions. In general, feedback indicating low reliability made participants pay more attention to the robot and a high reliability indications led participants to be more relaxed and allocate more time to other tasks.

In addition, predictability and transparency emerged time and again to be a very important factor for trust and human-robot collaboration. Before developers and researchers can think of collaborative teams consisting of humans and robots, operators must be able to understand and predict the robot's autonomy (Kruijff et al., 2014, p. 12).

8.7 Personal reflection on trust, human-robot collaboration, and the use of autonomous features

The review of literature highlighted that the meaning of trust as a concept is very unstandardised, flexible, and difficult to define (Mcknight & Chervany, 1996). Furthermore, trust is not easy to measure. Is trust perhaps a synonym for usage or reliance? I rely on my team, because I trust them. Is the word trust misused? We may not intentionally misuse the word trust, but sometimes factors might not be interpreted properly. I believe that trust is a very personal concept: some will ride a rollercoaster, others won't because they do not trust it. We need to focus on what we

really want to know: we want to know if the technology helps to improve our job performance and make work safer and whether the technology is used appropriately to take advantage of all possible benefits. So far evidence is sparse that trust influences performance or workload at a significant level.

Although trust is not yet easy to capture it is not impossible with further research. For example Chapter 7 showed that there could be difficulties in determining if trust or something else was measured. However, is it worth putting effort into determining factors of trust, which change from technology to technology, from situation to situation, and from person to person? Trust is a good start (and an important concept) to find factors that influence our reliance and usage of a supporting robotic technology (Lee & See, 2004). However, research success may be limited by only looking at trust and its associated factors. Especially by not knowing if these trust factors belong to everyone's individual trust concept (Merritt & Ilgen, 2008). There are additional factors which should be considered, such as acceptance, expectations, perceived usefulness, usability, experience, etc. (Davis, 1986; Komatsu & Yamada, 2011; Larochelle et al., 2013a). We need to examine the bigger picture if we want to influence the variables that matter, such as performance and workload. So we have to consider trust and other technology relevant factors, to find a way to enhance human-robot collaboration. This is just a gentle warning that the word trust might be over-used and sometimes misused as a general construct that aims to predict the vast facets of human-robot interaction.

Another question that needs reflection is, if semi-automation in Urban Search and Rescue is useful? Today, remote controlled robots are not robust, versatile and useful enough to be an integral part of a search and rescue team. As research has shown (Chapter 7; Visser and Parasuraman, 2011) only robots that show a certain level of performance can actually contribute to a higher overall performance. On the one hand, currently there are no such semi-autonomous systems that can be used in real-life situations that show high levels of performance (Mioch et al., 2012). On the other hand, there is still a huge advantage of using a robot instead of a

person or an animal in inaccessible and/or dangerous zones. At the current level of technology it is not useful to introduce high levels of automation.

There are other situations where automation can be useful. Less supervision may be required for multiple robots or even swarms that gather information autonomously. For example, a drone can be sent to fly over the incident site and capture aerial view pictures autonomously. On the contrary, semi-autonomous ground robots still need a huge amount of supervision. We need to consider which tasks the human needs to do that the robot cannot. Chien and Lewis (2012) suggested that the overall performance can decline if the robot takes too much of the operator's attention away from the main task (e.g. searching for casualties). This means that robot autonomy is not appropriate in all situations. The supervision of the robot can take up more attention than the search task itself, which should not be intended, since the main objective is to find as many casualties as possible, rather than accurately supervising the robot's work.

The benefit of having autonomous robot features, such as reduced risk of errors and decreased risk for the rescuer, has to outweigh possible performance deficits. If that is not the case, the robot should be operated manually. This also ensures that the operator can gain a higher level of situation awareness. Furthermore, professional rescue personnel are highly qualified and it seems that taking away control from them (e.g. the control over the robot) might be difficult and maybe not desired at all (Virk et al., 2008).

I think, in terms of implementation of robots in the search and rescue services, robotic agents can only be successfully implemented in stages. Rescuers see their equipment as tools and not as team mates (see Chapter 4). Transitional concepts need to be researched. One possibility could be a stepwise introduction to the technology and automation. First, introducing an entirely manually remote controlled robot. Next, introducing features that only support but do not take any decision or control away from the operator (e.g. warnings about missed corners and/or targets). One recommendation is to gradually introduce automation with increased operator experience. But most important of all, the robot needs to prove its

usefulness in the field, if this does not happen, it is futile to start any further attempts of implementing any kinds of robots into Urban Search and Rescue teams.

8.8 Limitations of research

There were several limitations to this research. In three of the four studies the participants were not experts (rescuers) but students and staff from the university. With respect to trust and control allocation (manual and auto), rescuers might react differently when it comes to take over control because their experience may influence their self-confidence in the task, which is known to influence automation usage (Lee & Moray, 1994). Rescuers are highly trained in these types of tasks and might have different attitudes and interaction behaviours. It is of great importance to always include rescuers in the development of rescue robots. Also, the experiments just simulated a rescue scenario. Real-world scenarios are much more stressful and emotionally loaded. This can influence workload levels which can have an effect on control allocation and performance.

Although a power analysis was conducted, the number of participants was in some cases insufficient to produce significant results, especially when comparing the auto and manual group (13 participants in each group).

Another limitation was the amount of training participants received. In Study II the learning effect was a clear confounding variable and might have influenced the insignificant outcome of the study. In addition, depending on the participants' experience with computer games their training effect varied greatly. Therefore, gaming experience of rescuers needs to be recorded to better interpret the result of studies.

The theme based content analyses undertaken throughout this thesis was subject to the interpretation of the researcher. Analysis required inside knowledge of the studies and having watched the participant's recording and experienced the actual situation in which participants answered or talked through the verbal protocol (e.g. being the interviewer and taking notes for each participant). Original quotes are provided in the Appendix, such that the reader or other researchers can make their own analysis.

The intended reliability level of the robot was diluted by the fact that some participants did not perceive certain mistakes of the robot and therefore their impression of the robot was different, which changed the participants' ratings accordingly. To overcome this shortfall an additional performance measure was developed: observed robot performance. This is the performance that the participant actually perceived during the trial.

With respect to further limitations, the measurements of trust and performance need to be mentioned. Performance was measured throughout as the number of victims found. However, it may be that other variables such as the robot's movements or the robot search strategy also influenced the observed and rated robot performance. Trust was measured in Study II with a single question, in study III and IV the Schaefer (2013) trust questionnaire was used. This needs to be considered how trust and performance was measured when comparing experimental results.

Further, trust varied significantly between the experimental conditions in study IV, where the amount of feedback and in general the transparency of the interface was varied. The only great variations in answering the trust questionnaire were among the items "Provide feedback", "Provide appropriate information", and "Communicate with people" (see Appendix J, p. 403). Changing transparency levels led to the manipulation of single items on the trust scale, which changed the outcome significantly. Therefore, in Chapter 7 the trust result needs to be treated with caution. If a different trust questionnaire had been used the results might have been greatly different. In order to measure trust more reliably, trust games or hypothetical questions, such as, "If there was an emergency right now, would you use the robot?" could support the trust measurement and increase its accuracy.

The robot interface in study IV (Chapter 7) relied mainly on the visual channel, because using too many audio cues can be challenging for rescuers in a loud rescue environment. The reason why some participants simply neglected elements of the interface might have been caused by this overload of the visual channel. Other interfaces should try to use more channels but keep in mind the environmental circumstances rescuers have

to work in. Still, there is room to enhance the experience of operators to support them in their dangerous missions and make their work safer.

8.9 Chapter summary

The chapter discussed the key findings of this research and reviewed the aims of this thesis. A list of recommendations for robot implementation, design and behaviour was provided. Discussion was also made on the issue of trust and collaboration in human-robot teams and whether it is useful to further pursue the investigation of trust. The main limitations of this thesis, such as insufficient number of participants and confounding variables, were stated.

9 Conclusion and future work

9.1 Chapter overview

The chapter provides the main conclusions of this thesis in a short concluding statement. Furthermore, possible future work in the area of trust and human-robot collaboration research is outlined.

9.2 Concluding statement

The biggest advantage of using a robot in Urban Search and Rescue (USAR) is that rescuers can remain in a safe place while the robot can run reconnaissance missions in inaccessible or dangerous places. However steering a robot through a rescue environment and looking through "the eyes of robot" into environments that are far from ordered and easy to recognise, in addition, to the emotional and physical stresses associated with rescue work, make this task very demanding. The first study produced a valuable insight into the work of Urban Search and Rescue technicians in the U.K. which can be used by researchers considering how to support their work.

The overall aim of this PhD was to understand how robot behaviour and interface design can be applied to utilise the benefits of robot autonomy and inform future human-robot collaborative systems. This was investigated with the use of a search and rescue scenario, where a robot and a human work together to find targets at an emergency site.

USAR is a performance-oriented task which uses remote controlled robots to collect information about the surroundings. In order to test different robot behaviours and interfaces the influence of robot feedback, robot reliability, robot transparency, and task complexity on trust, workload, and performance was examined, the factors of importance analysed in this work derived from the literature review. By far the most influencing factor on trust, workload, and performance was robot reliability. A semi-autonomous unreliable robot would not support the work of rescuers, it might even make it worse. A robot needs to have a minimum of approximately 60% to 70%

correct performance to be useful and performance enhancing. Primarily, technology needs to prove itself to be useful. Rescuers see the robot as a tool and not as a team mate.

Rescue environments and task demands are unpredictable and vary greatly. This thesis showed that more complex rescue tasks did not influence the trust in the robot or the subjective workload of operators. However, a more complex task decreased the performance of the human-robot team. The decrease in performance was due to the operators failing to see the robot's mistakes. Detailed analysis suggests that low complexity tasks foster under-reliance and high complexity task over-reliance on the robot. However, there is a possibility of a paradox: the more autonomy and supervision demand a robot needs, the less attention is left for the operator to do their task (e.g. search for victims in the environment). A trade-off between these two demands needs to be found to optimise the synergy between human and robot.

This work also investigated robot transparency. There was no effect of transparency on performance or workload. Transparency was preferred by operators and trust increased with transparency. However, the validity of this measure is in question due to the selected method of measuring trust. Although participants stated that the additional information provided by the interface was useful, some participants also said that the robot system displayed too much information and distracted them. This suggests that a balance between providing information and hiding information is needed. Where this trade-off lies needs to be determined from domain to domain and from robot to robot. Interfaces should provide operators with the potential to customise the interface elements to a certain extent. This may be especially useful if the robot is modular and can be equipped with different sensors that are important for the task in question. In addition, a variety of robots can only be used in certain situations (e.g. only for testing if a person is dead or alive). Rescuers are not willing to invest in a robot that cannot be used on a daily-basis (Steinbauer et al., 2014). This barrier might be decreased with a modular, robust, and quickly deployable robot system.

Robot technology can be useful, but robots which cannot live up to operators' expectations, have no clear defined capabilities, have features that require too much supervision, and have low performance levels, will not be accepted or used in the Fire and Rescue Service and will not be a collaborative partner. In conclusion, the interface and robot behaviour can be designed to benefit human-robot teams. Robot confidence feedback can steer the operator's attention where needed; robot transparency can support better situation awareness, and a consistent robot behaviour can contribute to both better predictability and high levels of trust.

9.3 Future work

The findings presented in this thesis suggest to provide the operator with higher levels of robot transparency. However, too much information distracts and confuses the operator. Which key information needs to be visible at what time is a future research area that can be explored in focus groups or experiments with rescuers. The aim is to design interfaces which improve the operator's understanding of the robot's status and actions, but at the same time not overwhelm the operator with unnecessary information. Furthermore, transparency did not lead to higher levels of performance (Chapter 7). If transparency can provide benefits for understanding robot failures or recover faster from errors needs to be investigated in future research.

Further, the experimental work in this thesis showed that trust in the robot was not influenced by task complexity. Nevertheless, participants did not rate the different task complexity levels as more or less difficult/complex. This suggests that the robot and its interface might have a larger influence on the rated task complexity than the environment itself. Further investigation is necessary to determine which variables influence the perceived difficulty of a remote robot control task.

Also new questions about the measurement of performance and trust emerged. Future studies might be able to determine which performance shaping factors are important when interacting with a semi-autonomous rescue robot system. If it is possible to identify these factors, robot performance can be measured in more detail and would show a better

correlation with the operator's rated robot performance. Furthermore, the responsible factors that cause a decreased perceived performance can be identified and positively modified.

In general, future experiments should use the proposed observed robot performance measure (Section 6.3.5, p. 198) to distinguish between objective and observed performances for a better understanding of human ratings, behaviours, and human-robot team performance.

In the experiments the robot performance was generally quite high; it would be of interest to test robot performances that are very low, or even a robot with no success at all, in order to see to what extent the human-robot team performance and the behaviour of the participant is influenced.

For the implementation of robots in the search and rescue service, transitional concepts need to be researched. It is recommended that technology is implemented in small steps because the robot needs to be perceived as useful and not overwhelming or even defective.

9.4 Chapter summary

Research into trust, workload and performance in human-robot teams in the Urban Search and Rescue context has been conducted. This was realised by using autoethnographic and virtual reality approaches. The research findings showed that technology needs to prove itself to be useful and have a certain level of performance to be accepted, used and provide performance enhancement. A variety of recommendations and guidance for future research were given.

10 References

110 First Look [Image]. (2015). iRobot. Retrieved from http://www.irobotweb.com/~/media/Images/iRobot/Robots/Defense/FirstLook/Product Page/specs.jpg?h=450&la=en&w=1200

Adams, B., Bruyn, L., Houde, S., & Angelopoulos, P. (2003). *Trust in automated systems: literature review*. *Defence Research and Development*. (No. CR-2003-096). Canada Toronto.

Angerer, S., Strassmair, C., Rootenbacher, M., & Robertson, N. M. (2012). Give me a hand - The potential of mobile assistive robots in automotive logistics and assembly applications. In *Technologies for Practical Robot Applications, IEEE International* (pp. 111–116).

Atoyan, H., Duquet, J.-R., & Robert, J.-M. (2006). Trust in new decision aid systems. In *Proceedings of the 18th international conference on Association Francophone d'Interaction Homme-Machine* (pp. 115–122). New York, USA: ACM Press. http://doi.org/10.1145/1132736.1132751

Bainbridge, L. (1983). Ironies of Automation *. *Automatica*, *19*(6), 775–779.

Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71–81. http://doi.org/10.1007/s12369-008-0001-3

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2006). The influence of people's culture and prior experiences with Aibo on their attitude towards robots. *Artificial Intelligence & Society*, *21*(1-2), 217–230. http://doi.org/10.1007/s00146-006-0052-7

Bedny, G. Z., Karwowski, W., & Bedny, I. S. (2012). Complexity

evaluation of computer based tasks. *International Journal of Human-Computer Interaction*, *28*(4), 236–257.

Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction*, *3*(2), 74–99. http://doi.org/10.5898/JHRI.3.2.Beer

Bhattacharya, R., Devinney, T. M., & Pillutla, M. M. (1998). A Formal Model of Trust Based on Outcomes. *Academy of Management Review*, *23*(3), 459–472. http://doi.org/10.5465/AMR.1998.926621

Bicho, E., Louro, L., & Erlhagen, W. (2010). Integrating verbal and nonverbal communication in a dynamic neural field architecture for human-robot interaction. *Frontiers in Neurorobotics*, *4*, 1–13. http://doi.org/10.3389/fnbot.2010.00005

Billings, D. R., Schaefer, K. E., Chen, J. Y. C., Kocsis, V., Barrera, M., Cook, J., … Hancock, P. A. (2012). *Human-Animal Trust as an Analog for Human-Robot Trust : A Review of Current Evidence*. (No. ARL-TR-5949). Aberdeen Proving Ground, Maryland.

Booysen, T., & Mathew, T. J. (2014). The Case for a General Purpose, First Response Rescue Robot. In *Proceedings of the 2014 PRASA, RobMech and AfLaT International Joint Symposium*.

Boyce, M. W., Chen, J. Y. C., Selkowitz, A. R., & Lakhmani, S. G. (2015). Effects of Agent Transparency on Operator Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts* (pp. 179–180).

Braarud, P. Ø., & Kirwan, B. (2011). Task complexity: what challenges the crew and how do they cope. In A. B. Skjerve & A. Bye (Eds.), *Simulator-Based Human Factors Studies Across 25 Years: the History of the Halden Man-Machine Laboratory* (pp. 233–251). London: Springer.

Bradshaw, J. M., Beautement, P., Breedy, M. R., Bunch, L., Drakunov, S. V, Feltovich, P. J., … Raj, A. K. (2004). Making Agents Acceptable To

People. In N. Zhong & J. Liu (Eds.), *Intelligent Technologies for Information Analysis: Advances in Agents, Data Mining, and Statistical Learning* (pp. 361–406). Berlin, Heidelberg: Axel-Springer.

Bradshaw, J. M., Dignum, V., Jonker, C. M., & Sierhuis, M. (2012). Human-agent-robot teamwork. *Intelligent Systems, IEEE*, *27*(2), 8–13. http://doi.org/10.1145/2157689.2157843

Bratman, M. (1992). Shared cooperative activity. *The Philosophical Review*, *101*(2), 327–341.

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 708–713). Alberta, Canada. http://doi.org/10.1109/IROS.2005.1545011

Brogårdh, T. (2009). Robot Control Overview: An Industrial Perspective. *Modeling, Identification and Control: A Norwegian Research Bulletin*, *30*(3), 167–180. http://doi.org/10.4173/mic.2009.3.7

Burghart, C., & Steinfeld, A. (2008). Human-Robot Interaction Metrics and Future Directions. *Metrics for Human-Robot Interaction 2008*, (March). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.2966&rep=rep1&type=pdf#page=7

Cabin, R. J., & Mitchell, R. J. (2000). To Bonferroni or Not to Bonferroni: When and How Are the Questions. *Ecological Society of America*, *81*(3), 246–248.

CAESAR robot [Image]. (2015). MR2G Search and Rescue Division. Retrieved from http://mecheng.ukzn.ac.za/Libraries/MR2G/Contrsctible_Arms_Elevating_Search_And_Rescue_CAESAR_robot.sflb.ashx

Campbell, D. J. (1988). Task Complexity: A Review and Analysis. *Academy of Management Review*, *13*(1), 40–52. http://doi.org/10.2307/258353

Casper, J., & Murphy, R. R. (2003). Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics*, *33*(3), 367–85. http://doi.org/10.1109/TSMCB.2003.811794

Castaldo, S., Premazzi, K., & Zerbini, F. (2010). The Meaning(s) of Trust. A Content Analysis on the Diverse Conceptualizations of Trust in Scholarly Research on Business Relationships. *Journal of Business Ethics*, *96*(4), 657–668. http://doi.org/10.1007/s10551-010-0491-4

Cesa, S. L., Farinelli, A., & Iocchi, L. (2008). Semi-autonomous coordinated exploration in rescue scenarios. *RoboCup 2007: Robot Soccer World Cup XI*, 286–293. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-68847-1_27

Chen, J. Y. C., & Barnes, M. J. (2014). Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13–29. http://doi.org/10.1109/THMS.2013.2293535

Chen, J. Y. C., Haas, E. C., & Barnes, M. J. (2007). Human Performance Issues and User Interface Design for Teleoperated Robots. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, *37*(6), 1231–1245. http://doi.org/10.1109/TSMCC.2007.905819

Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. J. (2014). *Situation Awareness – Based Agent Transparency*. (No. ARL-TR-6905). Aberdeen Proving Ground MD: US Army Research Laboratory.

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, *52*(8), 907–920. http://doi.org/10.1177/154193120805201922

Chen, K., & Chan, A. H. S. (2011). A review of technology acceptance by

older adults. *Gerontechnology*, *10*(1), 1–12. http://doi.org/10.4017/gt.2011.10.01.006.00

Chen, T. B., Campbell, D., Gonzalez, F., & Coppin, G. (2014). The Effect of Autonomy Transparency in Human-Robot Interactions : A Preliminary Study on Operator Cognitive Workload and Situation Awareness in Multiple Heterogeneous UAV Management. In *Proceedings of the Australasian Conference on Robotics and Automation* (pp. 2–4).

Chien, S.-Y., & Lewis, M. (2012). Effects of Unreliable Automation in Scheduling Operator Attention for Multi-Robot Control. In *2012 IEEE International Conference on Systems, Man, and Cybernetic* (pp. 321–326). Seoul, Korea.

Chien, S.-Y., Lewis, M., Mehrotra, S., Brooks, N., & Sycara, K. (2012). Scheduling operator attention for Multi-Robot Control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 473–479). Vilamoura, Algarve, Portugal: IEEE. http://doi.org/10.1109/IROS.2012.6386019

Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.

Coeckelbergh, M. (2010). Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations. *International Journal of Social Robotics*, *3*(2), 197–204. http://doi.org/10.1007/s12369-010-0075-6

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates. Retrieved from http://books.google.com/books?hl=en&lr=&id=2v9zDAsLvA0C&oi=fnd&pg=PR3&dq=Statistical+power+analysis+for+the+behavioral+sciences&ots=x5bAbZ9yDU&sig=a4Z-I9n7SeVlVyEoLGexAxmIh-E

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. In *Command and Control Research and Technology Symposium* (pp. 1–37). Washington DC:

CCRP. http://doi.org/10.1.1.90.2591

Cohen, P. R., & Levesque, H. J. (1991). Teamwork. *Nous*, *25*(4), 487–512.

Colombi, J., Lenfestey, A., Cring, E., & Colombi, J. (2009). Architecting Human Operator Trust in Automation for Multiple Unmanned Aerial System (UAS) Control. In *Proceedings of the 2009 International Conference on Software Engineering Research & Practice, SERP 2009*. Las Vegas, Nevada, USA.

Coradeschi, S., Ishiguro, H., Asada, M., Shapiro, S., Thielscher, M., Breazeal, C., … Ishida, H. (2006). Human-inspired robots. *IEEE Intelligent Systems*, *21*(4), 74–85.

Cornwall, W. (2014, July). Causes of Deadly Washington Mudslide Revealed in Scientific Report. *National Geographic*. Retrieved from http://news.nationalgeographic.com/news/2014/07/140722-oso-washington-mudslide-science-logging/

Cowman, S. E., Ferrari, J. R., & Liao-Troth, M. (2004). Mediating effects of social support on firefighters' sense of community and perceptions of care. *Journal of Community Psychology*, *32*(2), 121–126. http://doi.org/10.1002/jcop.10089

Craighead, J., Burke, J., & Murphy, R. R. (2008). Using the Unity Game Engine to Develop SARGE : A Case Study. *Proceedings of the 2008 Simulation Workshop at the International Conference on Intelligent Robots and Systems (IROS 2008)*. Retrieved from http://www.robot.uji.es/research/events/iros08/contributions/craighead.pdf

Crandall, J., & Goodrich, M. (2005). Validating human-robot interaction schemes in multitasking environments. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *35*(4), 438–449.

CRASAR [Website]. (2015). Retrieved February 9, 2015, from http://crasar.org/

Crossman, J., Marinier, R., & Olson, E. B. (2012). A hands-off, multi-robot display for communicating situation awareness to operators. *2012 International Conference on Collaboration Technologies and Systems (CTS)*, 109–116. http://doi.org/10.1109/CTS.2012.6261036

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, *19*(4), 497–515. http://doi.org/10.1007/s10260-010-0142-z

Cuevas, H. M., Fiore, S. M., Caldwell, B. S., & Strater, L. (2007). Augmenting team cognition in human-automation teams performing in complex operational environments. *Aviation, Space, and Environmental Medicine*, *78*(5), 63–70.

CUTLASS EOD robot [Image]. (2012). Northrop Grumman. Retrieved from http://cdn.topsecretwriters.com/wp-content/uploads/2012/09/bombrobot.jpg

DARPA: Tactical Technology Office [Website]. (2015). Retrieved May 2, 2015, from http://www.darpa.mil/about-us/offices/tto

Davis, F. D. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Doctoral Dissertation, Massachusetts Institute of Technology. Retrieved from http://en.scientificcommons.org/7894517

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Science*, *35*(8), 982–1003. http://doi.org/10.1287/mnsc.35.8.982

de Visser, E., & Parasuraman, R. (2011). Adaptive Aiding of Human-Robot Teaming: Effects of Imperfect Automation on Performance, Trust, and Workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209–231. http://doi.org/10.1177/1555343411410160

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human Computer Studies*, *58*(6),

719–735. http://doi.org/10.1016/S1071-5819(03)00039-9

Deely, S., Dodman, D., Hardoy, J., Johnson, C., Satterthwaite, D., Serafin, A., & Waddington, R. (2010). *World Disasters Report 2010: Focus on Urban Risk*. (D. McClean, Ed.). Lyons, France: Imprimerie Chirat.

Dehais, F., Sisbot, E. A., Alami, R., & Causse, M. (2011). Physiological and subjective evaluation of a human-robot object hand-over task. *Applied Ergonomics*, *42*(6), 785–91. http://doi.org/10.1016/j.apergo.2010.12.005

Desai, M. (2012). *Modeling trust to improve Human-Robot Interaction*. (Doctoral Dissertation, University of Massachusetts Lowell, USA).

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. A. (2013). Impact of robot failures and feedback on real-time trust. *8th ACM/IEEE International Conference on Human-Robot Interaction*, 251–258. http://doi.org/10.1109/HRI.2013.6483596

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., … Yanco, H. A. (2012). Effects of changing reliability on trust of robot systems. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 73–80). New York, USA: ACM Press. http://doi.org/10.1145/2157689.2157702

Desai, M., & Yanco, H. A. (2005). Blending human and robot inputs for sliding scale autonomy. In *ROMAN 2005 IEEE International Workshop on Robot and Human Interactive Communication* (pp. 537–542). http://doi.org/10.1109/ROMAN.2005.1513835

Doroodgar, B., Ficocelli, M., Mobedi, B., & Nejat, G. (2010). The search for survivors: Cooperative human-robot interaction in search and rescue environments using semi-autonomous robots. In *IEEE International Conference on Robotics and Automation* (pp. 2858–2863). Anchorage, Alaska, USA: IEEE. http://doi.org/10.1109/ROBOT.2010.5509530

Dragon Runner 10 [Image]. (2015). QinetiQ. Retrieved from http://www.qinetiq.com/services-products/survivability/UGV/bomb-

disposal-eod/PublishingImages/dragon-runner-10.png

Driewer, F., Schilling, K., & Baier, H. (2005). Human-computer interaction in the PeLoTe rescue system. In *Safety, Security and Rescue Robotics, Workshop, 2005 IEEE International* (pp. 224–229). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1501253

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human Computer Studies*, *58*(6), 697–718. http://doi.org/10.1016/S1071-5819(03)00038-7

Eliav, A., Lavie, T., Parmet, Y., Stern, H., & Edan, Y. (2011). Advanced methods for displays and remote control of robots. *Applied Ergonomics*, *42*(6), 820–829. http://doi.org/10.1016/j.apergo.2011.01.004

Ellis, C. (1993). "There are survivors": Telling a story of a sudden death. *The Sociological Quarterly*, *34*(4), 711–730. Retrieved from http://www.jstor.org/stable/4121376?seq=3

Ellis, C., Adams, T., & Bochner, A. (2011). Autoethnography: An overview. *Historical Social Research/Historische Sozialforschung*, *12*(1), 273–290. Retrieved from http://www.jstor.org/stable/23032294

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In *Proceedings of the National Aerospace and Electronics Conference* (pp. 789–795). New York: IEEE. http://doi.org/10.1109/NAECON.1988.195097

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492. http://doi.org/10.1080/001401399185595

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–91. http://doi.org/10.3758/BF03193146

Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics* (4th ed.). SAGE Publications. Retrieved from https://uk.sagepub.com/en-gb/eur/discovering-statistics-using-ibm-spss-statistics/book238032

Fincannon, T., Barnes, L. E., Murphy, R. R., & Riddle, D. L. (2004). Evidence of the need for social intelligence in rescue robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vol. 2, pp. 1089–1095). IEEE. http://doi.org/10.1109/IROS.2004.1389542

Finomore, V., Satterfield, K., Sitz, A., Castle, C., Funke, G., Shaw, T., & Funke, M. (2012). Effects of the Multi-Modal Communication tool on Communication and Change Detection for Command & Control Operators. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, pp. 1461–1465). SAGE Publications. http://doi.org/10.1177/1071181312561410

Fiore, S., Jentsch, F., Salas, E., & Finkelstein, N. (2005). *Cognition, teams, and augmenting team cognition: understanding memory failures in distributed human-agent teams*. University of Central Florida, Orlando, USA.

Fireservice Recruitment [Website]. (2014). Retrieved July 7, 2014, from http://www.fireservice.co.uk/recruitment/faq#22

Fischer, K. (2014). Alignment or collaboration? How implicit views of communication influence robot design. In *International Conference on Collaboration Technologies and Systems (CTS)* (pp. 115–122). http://doi.org/10.1109/CTS.2014.6867552

Fong, T. W., Kaber, D., Scholtz, J., & Schultz, A. (2004). Common Metrics for Human-Robot Interaction. In *IEEE 2004 International Conference on Intelligent Robots and Systems*. Sendai, Japan.

Foster-Miller Solem [Image]. (2010). Army Guide. Retrieved from http://www.army-guide.com/images/solem000dof.jpg

Freedy, A., & de Visser, E. (2007). Measurement of trust in human-robot collaboration. In *IEEE Proceedings of the 2007 International*

Conference on Collaborative Technologies and Systems (pp. 106–114).

Gao, F., Clare, A. S., Macbeth, J. C., & Cummings, M. L. (2013). Modeling the Impact of Operator Trust on Performance in Multiple Robot Control. In *AAAI Spring Symposium: Trust and Autonomous Systems*. Retrieved from http://dspace.mit.edu/handle/1721.1/90334

Gao, F., Cummings, M. L., & Bertuccelli, L. F. (2012). Teamwork in controlling multiple robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 81–88). New York, USA: ACM Press. http://doi.org/10.1145/2157689.2157703

Glas, D. F., Kanda, T., Ishiguro, H., & Hagita, N. (2012). Teleoperation of Multiple Social Robots. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *42*(3), 530–544. http://doi.org/10.1109/TSMCA.2011.2164243

Goldberg, L. R. L. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. Retrieved from http://psycnet.apa.org/journals/pas/4/1/26/

Gow, A. J., Whiteman, M. C., Pattie, A., & Deary, I. J. (2005). Goldberg's "IPIP" Big-Five factor markers: Internal consistency and concurrent validation in Scotland. *Personality and Individual Differences*, *39*(2), 317–329. http://doi.org/10.1016/j.paid.2005.01.011

Green, S. A., Billinghurst, M., Chen, X., & Chase, J. G. (2008). Human-Robot Collaboration - A Literature Review and Augumented reality approach. *International Journal of Advanced Robotic Systems*, *5*(1), 1–18.

Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. *Interaction Studies*, *8*(3), 483–500.

Groom, V., Takayama, L., Ochi, P., & Nass, C. (2009). I Am My Robot : The Impact of Robot-building and Robot Form on Operators. In *Proceedings of the 4th ACM/IEEE international conference on Human*

robot interaction (pp. 31–36). ACM.

Guha-Sapir, D., Below, R., & Hoyois, P. (2015). EM-DAT: International Disaster Database – www.emdat.be. Université Catholique de Louvain – Brussels – Belgium.

Hamp, Q., Gorgis, O., Labenda, P., Neumann, M., Predki, T., Heckes, L., … Reindl, L. M. (2013). Study of efficiency of USAR operations with assistive technologies. *Advanced Robotics*, *27*(5), 337–350. http://doi.org/10.1080/01691864.2013.763723

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can You Trust Your Robot? *Ergonomics in Design: The Quarterly of Human Factors Applications*, *19*(3), 24–29. http://doi.org/10.1177/1064804611415045

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527. http://doi.org/10.1177/0018720811417254

Harriott, C. E., Buford, G. L., Zhang, T., & Adams, J. A. (2012). Assessing workload in human-robot peer-based teams. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12* (pp. 141–142). New York, USA: ACM Press. http://doi.org/10.1145/2157689.2157725

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam, Netherlands: North Holland Press.

Heger, F. W., & Singh, S. (2006). Sliding autonomy for complex coordinated multi-robot tasks: Analysis & experiments. In *Proceedings, Robotics: Systems and Science, Philadelphia*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.148.9892&

amp;rep=rep1&type=pdf

Helldin, T. (2014). *Transparency for Future Semi-Automated Systems - Effects of transparency on operator performance, workload and trust*. (Doctoral dissertation, Örebro University, Sweden).

Hendy, K. C., Liao, J., & Milgram, P. (1997). Combining time and intensity effects in assessing operator information-processing load. *Human Factors*, *39*, 30–47. http://doi.org/10.1518/001872097778940597

HM Government. (2008). Fire and Rescue Manual - Volume 2 Fire Service Operations - Incident Command. London, UK: The Stationary Office. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/7643/incidentcommand.pdf

Hoff, K. A., & Bashir, M. (2014). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. http://doi.org/10.1177/0018720814547570

Hoffman, G., & Breazeal, C. (2004). Collaboration in Human-Robot Teams. *AIAA 1st Intelligent Systems Technical Conference*, 1–18. http://doi.org/10.2514/6.2004-6434

Hoffman, G., & Breazeal, C. (2007). Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *Proceeding of the ACM/IEEE international conference on Human-robot interaction* (pp. 1–8). http://doi.org/10.1145/1228716.1228718

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, *28*(1), 84–88. http://doi.org/10.1109/MIS.2013.24

Holden, H., & Rada, R. (2011). Understanding the Influence of Perceived Usability and Technology Self-Efficacy on Teachers' Technology Acceptance. *Journal of Research on Technology in Education*, *43*(4), 343–367. http://doi.org/10.1080/15391523.2011.10782576

Hollnagel, E., & Woods, D. D. (1999). Cognitive systems engineering: new wine in new bottles. *International Journal of Human-Computer Studies*, *51*(2), 339–356.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. Retrieved from http://www.jstor.org/stable/10.2307/4615733

Horsch, C. H. G., Smets, N. J. J. M., Neerincx, M. A., & Cuijpers, R. H. (2013). Revealing unexpected effects of rescue robots' team-membership in a virtual environment. In *ISCRAM 2013 Conference Proceedings - 10th International Conference on Information Systems for Crisis Response and Management* (pp. 627–631). Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84905670746&partnerID=tZOtx3y1

Huang, H., Messina, E., & Albus, J. (2003). Toward a generic model for autonomy levels for unmanned systems (ALFUS). *Performance Metrics for Intelligent Systems (PerMIS) Workshop*. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA515323

Institute for Economic and Peace. (2014). Global Terrorism Index 2014, 1–94.

Italy quake homeless in emergency shelters. (2012, May 21). *BBC News*. Retrieved from http://www.bbc.com/news/world-europe-18140543

Jian, J., Bisantz, A. M., Drury, C. G., & Llinas, J. (2000). Fondations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71.

Jonker, C. M., & Treur, J. (1999). Formal Analysis of Models for the Dynamics of Trust Based on Experiences. *Multi-Agent System Engineering*, *1647*, 221–231. http://doi.org/10.1007/3-540-48437-X_18

Jung, M., & Lee, J. (2013). Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 16th ACM*

Conference on Computer Supported Cooperative Work and Social Computing (pp. 1555–1566). San Antonio, Texas, USA.

Kaniarasu, P., Steinfeld, A., Desai, M., & Yanco, H. (2012). Potential measures for detecting trust changes. In *Proceedings of the 7th annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12* (pp. 241–242). http://doi.org/10.1145/2157689.2157775

Kaniarasu, P., Steinfeld, A., Desai, M., & Yanco, H. A. (2013). Robot confidence and trust alignment. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction* (pp. 155–156). IEEE. http://doi.org/10.1109/HRI.2013.6483548

Kaniarasu, P., & Steinfeld, A. M. (2014). Effects of blame on trust in human robot interaction. *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 850–855. http://doi.org/10.1109/ROMAN.2014.6926359

Kim, Y. C., Yoon, W. C., Kwon, H. T., Yoon, Y. S., & Kim, H. J. (2007). A Cognitive Approach to Enhancing Human-Robot Interaction for Service Robots. In *Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design* (pp. 858–867).

Klein, G., Bradshaw, J. M., Feltovich, P. J., & Woods, D. D. (2005). Common ground and coordination in joint activity. In W. B. Rouse & K. R. Boff (Eds.), *Organizational simulation* (pp. 1–42). Hoboken, N.J., USA: John Wiley & Sons.

KOHGA3 ground robot [Image]. (2011). Retrieved from http://spectrum.ieee.org/image/1812274

Komatsu, T., & Yamada, S. (2011). Adaptation gap hypothesis: How differences between users' expected and perceived agent functions affect their subjective impression. *Journal of Systemics, Cybernetics and Informatics*, *9*(1), 67–74. Retrieved from http://scholar.google.co.uk/scholar?cluster=10592265979395302519&hl=en&as_sdt=0,5#2

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. *Multiple-Task Performance*, (June), 279–328. Retrieved from http://hdl.handle.net/2060/19900020461

Kramer, R. M. (1994). The sinister attribution error: Paranoid cognition and collective distrust in organizations. *Motivation and Emotion*, *18*(2), 199–230. http://doi.org/10.1007/BF02249399

Kruijff, G.-J. M., Janíček, M., Keshavdas, S., Larochelle, B., Zender, H., Smets, N. J. J. M., … Sulk, M. (2014). Experience in system design for human-robot teaming in urban search and rescue. *Springer Tracts in Advanced Robotics*, *92*, 111–125. http://doi.org/10.1007/978-3-642-40686-7_8

Kruijff, G.-J. M., Pirri, F., Gianni, M., Papadakis, P., Pizzoli, M., Sinha, A., … Angeletti, S. (2012). Rescue robots at earthquake-hit Mirandola, Italy: A field report. In *2012 IEEE International Symposium on Safety, Security, and Rescue Robotics, SSRR 2012*. http://doi.org/10.1109/SSRR.2012.6523866

Kwok, K.-W., Sun, L.-W., Mylonas, G. P., James, D. R. C., Orihuela-Espina, F., & Yang, G.-Z. (2012). Collaborative gaze channelling for improved cooperation during robotic assisted surgery. *Annals of Biomedical Engineering*, *40*(10), 2156–67. http://doi.org/10.1007/s10439-012-0578-4

Larochelle, B., & Kruijff, G.-J. M. (2012). Multi-view operator control unit to improve situation awareness in USAR missions. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication* (pp. 1103–1108). Paris. http://doi.org/10.1109/ROMAN.2012.6343896

Larochelle, B., Kruijff, G.-J. M., & Van Diggelen, J. (2013a). Usage of Autonomy Features in USAR Human-Robot Teams. *International Journal of Robotics and Automation*, *4*(1), 19–30.

Larochelle, B., Kruijff, G.-J. M., & Van Diggelen, J. (2013b). Usage of Autonomy Features in USAR Human-Robot Teams. *International*

Journal of Robotics and Automation, *4*(1), 19–30.

Lee, J. D. (2008). Review of a Pivotal Human Factors Article: "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 404–410. http://doi.org/10.1518/001872008X288547

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. Retrieved from http://www.tandfonline.com/doi/full/10.1080/00140139208967392

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*, 153–184. http://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. http://doi.org/10.1518/hfes.46.1.50.30392

Lewis, L. (2014). *Investigating the ways in which virtual environments could influence aircraft passengers' comfort and experiences*. (Doctoral dissertation, Human Factors Research Group, University of Nottingham, United Kingdom).

Lewis, M., Wang, H., Chien, S. Y., Velagapudi, P., Scerri, P., & Sycara, K. (2011). Process and Performance in Human-Robot Teams. *Journal of Cognitive Engineering and Decision Making, SPECIAL ISSUE: Improving Human-Robot Interaction, Part II*, *5*(2), 186–208. http://doi.org/10.1177/1555343411409323.

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, *42*(6), 553–568. http://doi.org/10.1016/j.ergon.2012.09.001

Liu, Y., & Nejat, G. (2013). Robotic Urban Search and Rescue: A Survey from the Control Perspective. *Journal of Intelligent & Robotic Systems*, *72*(2), 147–165. http://doi.org/10.1007/s10846-013-9822-x

Lyons, J. B. (2013). Being Transparent about Transparency : A Model for Human-Robot Interaction. *Trust and Autonomous Systems: Papers from the 2013 AAAI Spring Symposium*, 48–53.

Lyons, J. B., & 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* (pp. 181–190). Springer International Publishing. http://doi.org/10.1007/978-3-319-07458-0_18

Martin, J. A. (2012). Bomb Squad Talon Robot [Inter-office memorandum]. Miami, FL: Special Investigations Section.

Matsuno, F., Sato, N., Kon, K., Igarashi, H., Kimura, T., & Murphy, R. R. (2014). Utilization of Robot Systems in Disaster Sites of the Great Eastern Japan Earthquake. In K. Yoshida & S. Tadokoro (Eds.), *Field and Service Robotics* (Vol. 92). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-40686-7

Mcknight, D. H., & Chervany, N. L. (1996). *The meanings of trust*. *Technical Report MISRC Working Paper Series 96-04*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.155.1213

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial Trust Formation in New Organizational Relationships. *The Academy of Management Review*, *23*(3), 473. http://doi.org/10.2307/259290

Menard, M. (2011). *Game Development with Unity* (1st ed.). Cengage Learning PTR.

Merritt, S. M. (2011). Affective processes in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(4), 356–370. http://doi.org/10.1177/0018720811411912

Merritt, S. M., & Ilgen, D. R. (2008). Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210.

http://doi.org/10.1518/001872008X288574

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2014). Are Well-Calibrated Users Effective Users? Associations Between Calibration of Trust and Performance on an Automation-Aided Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(1), 34–47. http://doi.org/10.1177/0018720814561675

Messina, E., & Jacoff, A. (2006). Performance standards for urban search and rescue robots. In *Proceedings of the SPIE Defense and Security Symposium*. Orlando, FL. http://doi.org/10.1117/12.663320

Mioch, T., Smets, N. J. J. M., & Neerincx, M. A. (2012). Assessing human-robot performances in complex situations with unit task tests. *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 621–626. http://doi.org/10.1109/ROMAN.2012.6343820

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control*, *21*(4-5), 203–211. http://doi.org/10.1177/014233129902100408

Muir, B. M. (1989). *Operators' Trust in and Use of Automatic Controllers Supervisory Process Control Task*. (Doctoral dissertation, University of Toronto, USA).

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(March 2015), 429–460. http://doi.org/10.1080/00140139608964474

Munich Re. (2015a). Loss events worldwide 1980 – 2014 - costliest events [Fact sheet]. Geo Risks Research, NatCatSERVICE.

Munich Re. (2015b). Loss events worldwide 1980 – 2014 - deadliest [Fact sheet]. Geo Risks Research, NatCatSERVICE.

Munro, A. (2011). Autoethnography as a reserach method in design reserach at Universities. *20/20 Design Vision*, 156–163. Retrieved from http://www.defsa.org.za/?q=system/files/2011conference/DEFSA Conference Proceedings 2011.pdf&download=1#page=163

Murphy, R. R. (2004). Human–Robot Interaction in Rescue Robotics. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, *34*(2), 138–153. http://doi.org/10.1109/TSMCC.2004.826267

Murphy, R. R. (2014). *Disaster Robotics*. Massachusetts, USA: MIT Press.

Murphy, R. R., Duncan, B. A., Collins, T., Kendrick, J., Lohman, P., Palmer, T., & Sanbirn, F. (2015). Use of a Small Unmanned Aerial System for the SR-530 Mudslide Incident near Oso, Washington. *Journal of Field Robotics*, 1556–4967. http://doi.org/10.1002/rob

Murphy, R. R., Rice, A., Rashidi, N., Henkel, Z., & Srinivasan, V. (2011). A multi-disciplinary design process for affective robots: Case study of Survivor Buddy 2.0. *2011 IEEE International Conference on Robotics and Automation*, 701–706. http://doi.org/10.1109/ICRA.2011.5979977

Murphy, R. R., & Schreckenghost, D. (2013). Survey of metrics for human-robot interaction. In *ACM/IEEE International Conference on Human-Robot Interaction* (pp. 197–198). http://doi.org/10.1109/HRI.2013.6483569

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*, *15*(6), 1044–1045. http://doi.org/10.1093/beheco/arh107

National Audit Office. (2008). *New Dimension – Enhancing the Fire and Rescue Services ' capacity to respond to terrorist and other large-scale incidents*. London, UK.

Neale, H., & Nichols, S. (2001). Theme-based content analysis: a flexible method for virtual environment evaluation. *International Journal of Human-Computer Studies*, *55*(2), 167–189. http://doi.org/10.1006/ijhc.2001.0475

Neerincx, M. A., & Grant, T. (2010). Evolution of Electronic Partners: Human-Automation Operations and ePartners During Planetary Missions. *Journal of Cosmology*, *12*, 3825–3833.

NIFTi [Website]. (2014). Retrieved January 1, 2015, from http://www.nifti.eu/

NIFTi UGV [Image]. (2013). Retrieved from http://www.nifti.eu/news/DSC_5511.jpg/image_preview

Nuttall, I. (2008). *Urban Search and Rescue (USAR) Canines within the U.K. Fire Service*. United Kingdom.

Oleson, K. E., Billings, D. R., Kocsis, V., Chen, J. Y. C., & Hancock, P. A. (2011). Antecedents of trust in human-robot collaborations. *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA 2011*, 175–178. http://doi.org/10.1109/COGSIMA.2011.5753439

Olsen, D. R., & Wood, S. B. (2004). Fan-out: Measuring human control of multiple robots. In *Proceedings of SIGCHI Conf. Human Factors Computer Systems* (pp. 231–238). Vienna, Austria.

Oriz, E., Fiorella, L., & Vogel-Walcutt, J. (2010). Teaming with a Robot : Effects on Teamwork Quality and Human-Robot Trust. In *Interservice/Industry Training, Simulation, and Education Conference* (pp. 1–8).

Ososky, S., Sanders, T., Jentsch, F., Hancock, P. A., & Chen, J. Y. C. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Proceedings of the SPIE Defense and Security, Unmanned Systems Technology XVI*. http://doi.org/10.1117/12.2050622

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. http://doi.org/10.1518/001872097778543886

Park, E., Jenkins, Q., & Jiang, X. (2008). Measuring trust of human operators in new generation rescue robots. In *7th JFPS International Symposium on Fluid Power*. Toyama, Japan. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Measuring+trust+of+human+operators+in+new+generation+rescue+robots#0

Partnership for Public Service. (2015). John Price: Applying space technology to rescue earthquake victims buried beneath the rubble. *The Washington Post*. Retrieved from http://www.washingtonpost.com/politics/federal_government/john-price-applying-space-technology-to-rescue-earthquake-victims-buried-beneath-the-rubble/2015/05/14/8e7b34d6-fa4b-11e4-9030-b4732caefe81_story.html

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, *316*(7139), 1236–1238. http://doi.org/10.1136/bmj.316.7139.1236

Peschel, J. M. (2012). Study and Development of the Rescue Robot to Accommodate Victims under Earthquake Disasters. In *Informatics in Control, Automation and Robotics* (pp. 89–100). Springer Berlin Heidelberg.

Pipe crawler Versatrax 150 [Image]. (2015). Inuktun. Retrieved from http://www.mswmag.com/images/uploads/gallery/30421/inuktun_services_versatrax_150__large.jpg

Preece, J., Rogers, Y., & Sharp, H. (2002). *Interaction design: Beyond human-computer interaction*. New York: Wiley.

Prewett, M. S., Johnson, R. C., Saboe, K. N., Elliott, L. R., & Coovert, M. D. (2010). Managing workload in human–robot interaction: A review of empirical studies. *Computers in Human Behavior*, *26*(5), 840–856.

http://doi.org/10.1016/j.chb.2010.03.010

QinetiQ. (2009). QinetiQ's Dragon Runner robots are sent to Afghanistan to support British troops. [Press release]. Retrieved from http://www.qinetiq.com/media/news/releases/Pages/dragon-runner-robots.aspx

R2i2 Delta Micro [Image]. (2014). RECCE robotics. Retrieved from http://www.recce-robotics.com/images/microVGTV2.jpg

R2i2 Extreme [Image]. (2015). RECCE robotics. Retrieved from http://www.recce-robotics.com/images/page4_img1.jpg

Recon Scout - Throwbot LE [Image]. (2015). Recon Robotics. Retrieved from http://www.recon-scout.com/graphics/Throwbot_LE_OCUII.jpg

Recon Scout Throwbot LE [Website]. (2015). Retrieved December 29, 2015, from http://www.reconrobotics.com/products/recon-scout_throwbot_LE.cfm

ReconRobotics. (2010). Recon Scout ThrowBot LE [Fact sheet]. Retrieved February 10, 2015, from https://www.simmonsle.com/uploads/c47f0e6b12d6706.Throwbot.LE.Aimee.pdf

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112. http://doi.org/10.1037/0022-3514.49.1.95

Robinette, P., Wagner, A. R., & Howard, A. M. (2015). *The effect of robot performance on human-robot trust in time-critical situations*. (GT--IRIM--HumAns-2015--001). Georgia Institute of Technology.

Ross, J. M., Szalma, J. L., & Hancock, P. A. (2007). Empirical Examination of Trust in Automation Across Multiple Agents in a Search and Rescue Operation. In *Human Factors and Ergonomics society 51th annual meeting* (pp. 1501–1505). http://doi.org/10.1037/e578042012-007

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–65. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/4865583

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*(1), 76–87. http://doi.org/10.1518/001872007779598082

Roy, D. (1959). "Banana Time": Job Satisfaction and Informal Interaction. *Human Organization*, 158–168. Retrieved from http://sfaa.metapress.com/index/07J88HR1P4074605.pdf

Rule, A., & Forlizzi, J. (2012). Designing interfaces for multi-user, multi-robot systems. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 97–104. http://doi.org/10.1145/2157689.2157705

Salas, E., Sims, D. E., & Burke, S. (2005). Is there a "Big Five" in Teamwork? *Small Group Research*, *36*(5), 555–599. http://doi.org/10.1177/1046496405277134

Sanders, T., Harpold, B., Kessler, T., & Hancock, P. A. (2015). Interpersonal distance effects on trust relationships in human-robot interaction. In *Proceedings 19th Triennial Congress of the IEA*. Melbourne.

Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y. C., & Hancock, P. A. (2011). A Model of Human-Robot Trust: Theoretical Model Development. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 55, pp. 1432–1436). http://doi.org/10.1177/1071181311551298

Sanders, T., Wixon, T., Schaefer, K. E., Chen, J. Y. C., & Hancock, P. A. (2014). The Influence of Modality and Transparency on Trust in Human - Robot Interaction. In *IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (pp. 456–459).

Schaefer, K. E. (2013). *The Perception and Measurement of Human-robot Trust*. (Doctoral dissertation, University of Central Florida, USA).

Retrieved from http://etd.fcla.edu/CF/CFE0004931/Schaefer_Kristin_E_201308_PhD.pdf

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T., Chen, J. Y. C., & Hancock, P. A. (2013). *A Meta-Analysis of Factors Influencing the Development of Trust in Automation : Implications for Human-Robot Interaction*. (No. ARL-TR-6984). Army Research Laboratory.

Schaerer, E., Kelley, R., & Nicolescu, M. (2009). Robots as animals: A framework for liability and responsibility in human-robot interactions. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 72–77). IEEE. http://doi.org/10.1109/ROMAN.2009.5326244

Scheutz, M., Schermerhorn, P., & Kramer, J. (2006). The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (pp. 226–233).

Selkowitz, A., Lakhmani, S., Chen, J. Y. C., & Boyce, M. (2015). The Effects of Agent Transparency on Human Interaction with an. In *59th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 806–810).

Sellner, B., Heger, F. W., Hiatt, L. M., Simmons, R., & Singh, S. (2006). Coordinated Multiagent Teams and Sliding Autonomy for Large-Scale Assembly. *Proceedings of the IEEE*, *94*(7), 1425–1444. http://doi.org/10.1109/JPROC.2006.876966

Seppelt, B. D., & Lee, J. D. (2007). Making adaptive cruise control (ACC) limits visible. *International Journal of Human Computer Studies*, *65*(3), 192–205. http://doi.org/10.1016/j.ijhcs.2006.10.001

Shah, B., & Choset, H. (2004). Survey on Urban Search and Rescue Robotics. *Journal of the Robotics Society of Japan*, *22*(5), 40–44.

Shapiro, S. P. (1987). The Social Control of Impersonal Trust. *American*

Journal of Sociology, *93*(3), 623. http://doi.org/10.1086/228791

Sharpe, D. (2015). Your Chi-Square Test is Statistically Significant: Now What? *Practical Assessment, Research & Evaluation*, *20*(8), 1–10.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators*. Massachusetts Institute of Technology, Man-Machine Systems Laboratory.

Shiotani, A., Tomonaka, T., Kemmotsu, K., Asano, S., Oonishi, K., & Hiura, R. (2006). World ' s First Full-fledged Communication Robot "wakamaru" Capable of Living with Family and Supporting persons. *Mitsubishi Heavy Industries, Ltd. Technical Review*, *43*(1), 1–2.

Simons, T. L., & Peterson, R. S. (2000). Task conflict and relationship conflict in top management teams: the pivotal role of intragroup trust. *The Journal of Applied Psychology*, *85*(1), 102–11.

Simpson, J. A. (2007). Foundations of interpersonal trust. In *Social psychology: Handbook of basic principles* (2nd ed., pp. 587–607). The Guilford Press.

Singer, S., & Akin, D. (2011). A Survey of Quantitative Team Performance Metrics for Human-Robot Collaboration. In *41st International Conference on Environmental Systems* (pp. 1–19). Portland, Oregon. Retrieved from http://spacecraft.ssl.umd.edu/publications/2011/AIAA-2011-5248.Sharon.pdf

Sklar, E., Ozgelen, A. T., Munoz, J. P., Gonzalez, J., Manashirov, M., Epstein, S. L., & Parsons, S. (2011). Designing the HRTeam Framework : Lessons Learned from a Rough-and-Ready Human / Multi-Robot Team. In *The Autonomous Agents and MultiAgent Systems* (pp. 232–251). http://doi.org/10.1007/978-3-642-27216-5_15

Sobre-Denton, M. S. (2012). Stories from the Cage: Autoethnographic Sensemaking of Workplace Bullying, Gender Discrimination, and White Privilege. *Journal of Contemporary Ethnography*, *41*(2), 220–250.

http://doi.org/10.1177/0891241611429301

Stanton, N. A., Young, M. S., & Walker, G. H. (2007). The psychology of driving automation: a discussion with Professor Don Norman. *International Journal of Vehicle Design*, *45*(3), 289–306. http://doi.org/10.1504/IJVD.2007.014906

Steinbauer, G., Maurer, J., & Krajnz, H. (2014). R 3 : Request a Rescue Robot. In *Safety, Security, and Rescue Robotics (SSRR), 2014 IEEE International Symposium* (pp. 1–2). http://doi.org/10.1109/SSRR.2014.7017682

Stopforth, R., Bright, G., & Harley, R. (2010). Performance of the improvements of the CAESAR robot. *International Journal of Advanced Robotic Systems*, *7*(3), 217–226.

Stubbs, K., Hinds, P. J., & Wettergreen, D. (2007). Autonomy and Common Ground in Human-Robot Interaction: A Field Study. *IEEE Intelligent Systems*, *22*(2), 42–50.

TALON robot [Image]. (2015). QinetiQ. Retrieved from https://www.qinetiq.com/media/Image Library/talon-robot-large.jpg

The Fire Service College. (2014). Urban Search and Rescue Technician 2 (USART2) [Course material]. Moreton-in-Marsh.

The Personal Qualities and Attributes [Website]. (2014). Retrieved April 15, 2014, from http://www.fireservice.co.uk/recruitment/pqas

Tower, M. (2014). Live Coverage: Robo cop with $19,000 price tag proposed for Saginaw Police Department. [Press release]. Retrieved from http://www.mlive.com/news/saginaw/index.ssf/2014/02/police_robot_proposed_for_sagi.html

Tsui, K. M., Desai, M., Yanco, H. A., Cramer, H., & Kemper, N. (2010). Using the "negative attitude toward robots scale" with telepresence robots. In *Proceedings of the 10th Performance Metrics for Intelligent Systems Workshop on PerMIS 2010* (pp. 243–250). ACM Press.

http://doi.org/10.1145/2377576.2377621

United Nations. (2014). *World Urbanization Prospects: The 2014 Revision, Highlights*. Department of Economic and Social Affairs; Population Division (ST/ESA/SER.A/352).

Virk, G. S., Gatsoulis, Y., Parack, M., & Kherada, A. (2008). Mobile Robotic Issues for Urban Search and Rescue. In *Proceedings of the 17th World Congress, The International Federation of Automatic Control* (pp. 3098–3103). Seoul, Korea.

Wagner, A. R. (2015). Exploring human-robot trust: Insights from the first 1000 subjects. In *2015 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 485–486). http://doi.org/10.1109/CTS.2015.7210395

Wakamaru [Image]. (2013). Mitsubishi Heavy Industries. Retrieved from http://www.mhi.co.jp/en/products/detail/__icsFiles/artimage/2010/03/26/ce_pd_ts_re/wakamaru_about01.jpg

Wang, H., Chien, S. Y., Lewis, M., Velagapudi, P., Scerri, P., & Sycara, K. (2009). Human Teams for Large Scale Multirobot control. In *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA* (pp. 1269–1274).

Wang, H., Lewis, M., Velagapudi, P., Scerri, P., & Sycara, K. (2009). How search and its subtasks scale in N robots. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction* (pp. 141–147). La Jolla, CA. http://doi.org/10.1145/1514095.1514122

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and Reliance on an Automated Combat Identification System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(3), 281–291. http://doi.org/10.1177/0018720809338842

Wang, Y., Shi, Z., Wang, C., & Zhang, F. (2014). Human-Robot Mutual Trust in (Semi)autonomous Underwater Robots. In A. Koub & A. Khelil (Eds.), *Cooperative Robots and Sensor Networks* (pp. 115–137). Heidelberg: Springer-Verlag. http://doi.org/10.1007/978-3-642-

55029-4_6

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*(3), 433–441. http://doi.org/10.1518/001872008X312152

Wegner, R., & Anderson, J. (2004). Balancing robotic teleoperation and autonomy for urban search and rescue environments. *Advances in Artificial Intelligence*, 16–30. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-24840-8_2

West Midlands Fire Station. (2013). USAR - Equipment and Vehicles and Vehicles [Fact sheet]. West Midlands Fire Service. Retrieved from https://www.wmfs.net/sites/default/files/USAR Equipment and Vehicles_0.pdf

Wiethoff, C., & Lewick, R. J. (2000). Trust, Trust Development, and Trust Repair. In M. Deutsch & P. T. Coleman (Eds.), *The handbook of conflict resolution: Theory and practice* (pp. 86–107). San Francisco, CA: Jossey-Bass.

Wilkowska, W., & Ziefle, M. (2009). Which factors form older adults' acceptance of mobile information and communication technologies? *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5889*, 81–101. http://doi.org/10.1007/978-3-642-10308-7_6

Williamson, O. E. (1993). Calculativeness, trust, and economic organization. *Journal of Law and Economics*, *34*, 453–502.

Xu, A., & Dudek, G. (2013). Towards Modeling Real-Time Trust in Asymmetric Human-Robot Collaborations. In *Proceedings of the 16th International Symposium on Robotics Research (ISRR '13)*. Retrieved from http://www.cim.mcgill.ca/~mrl/pubs/anqixu/isrr2013_trust_study.pdf

Yagoda, R. E. (2011). *WHAT! You want me to trust a ROBOT? The development of a human robot interaction (HRI) trust scale*. Master thesis, North Carolina State University, USA).

Yagoda, R. E., & Gillan, D. J. (2012). You Want Me to Trust a ROBOT? The Development of a Human–Robot Interaction Trust Scale. *International Journal of Social Robotics*, *4*(3), 235–248. http://doi.org/10.1007/s12369-012-0144-0

Zaied, A. (2012). An Integrated Success Model for Evaluating Information System in Public Sectors [Image]. *Journal of Emerging Trends in Computing and Information Sciences*.

Appendix A - Study I: General Questionnaire

Human-Robot Collaboration

General questionnaire

---

1. What is your gender? Please tick the appropriate circle.

   ○ Female ○ Male

---

2. What age are you?

   _____ years

---

3. What is your occupational title?

   _____

   If you are a researcher/student, please state your area of research/course of study.

   _____

---

4. How frequently are you using a computer?

   ○ everyday ○ 1-2 times a week ○ 1-2 times a month ○ less than once a month

---

5. Do you play computer games, app games or console games?

○ Yes ○ No

If Yes, please state the frequency of playing computer games, app games or console games.

○ everyday ○ 1-2 times a week ○ 1-2 times a month ○ less than once a month

Please state the type (PC, app or console) and name of the games you are playing.

_____
_____
_____
_____

---

6. Do you have experience with robots?

○ Yes ○ No

If Yes, please state occasion and name of the robot.

_____
_____
_____
_____

Please rate the following statements about robots by ticking the appropriate circle.

In general…

I would feel uneasy if I was given a job where I had to use robots.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

1. The word "robot" means nothing to me.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

2. I would feel nervous operating a robot in front of other people.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

3. I would hate the idea that robots or artificial intelligences were making judgements about things.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

4. I would feel very nervous just standing in front of a robot.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

5. I would feel paranoid talking with a robot.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

6. I would feel uneasy if robots really had emotions.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

7. Something bad might happen if robots developed into living beings.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

8. I feel that if I depend on robots too much, something bad might happen.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

9. I am concerned that robots would be a bad influence on children.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

10. I feel that in the future society will be dominated by robots.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

11. I feel that in the future, robots will be commonplace in society.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

12. I would feel relaxed talking with robots.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

13. If robots had emotions, I would be able to make friends with them.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

14. I feel that I could make friends with robots.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

15. I feel comforted being with robots that have emotions.

○ Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

16. I feel comfortable being with robots.

o Strongly disagree	○ Disagree	○ Undecided	○ Agree	○ Strongly agree

Thank you very much for taking part in my study. Please hand this questionnaire back to the researcher.

Personality questions

(This was an online questionnaire and the system provided space behind each item to select one of the answers)

How I am in general? The following section lists a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please indicate the extent to which you *agree or disagree* with that statement.

- 1=disagree strongly
- 2=disagree a little
- 3=neutral

- 4=agree a little
- 5=agree strongly

I am someone who…

...tends to find fault with others

...does a thorough job

...can be somewhat careless

...is relaxed, handles stress well

...is a reliable worker

...tends to be disorganized

...worries a lot

...is generally trusting

...tends to be lazy

...perseveres until the task is finished

...does things efficiently

...remains calm in tense situations

...makes plans and follows through with them

…likes to cooperate with others

...trust in things other people say

Appendix B - Study II: General questionnaire

Human-Robot Collaboration

- General Questionnaire -

---

7. What is your gender? Please tick the appropriate circle.

○ Female ○ Male

---

8. What age are you?

_____ years

---

9. What is your occupational title?

_____

If you are a researcher/student, please state your area of research/course of study.

_____

---

10. How frequently are you using a computer?

○ everyday ○ 1-2 times a week ○ 1-2 times a month ○ less than once a month

---

11.Do you play computer games, app games or console games?

○ Yes ○ No

If Yes, please state the frequency of playing computer games, app games or console games.

○ everyday ○ 1-2 times a week ○ 1-2 times a month ○ less than once a month

Please state the type (PC, app or console) and name of the games you are playing.

_____

_____

_____

_____

12.Do you have experience with robots?

○ Yes ○ No

If Yes, please state occasion and name of the robot.

_____

_____

_____

_____

Please rate the following statements about robots by ticking the appropriate circle.

Personality questions

(This was an online questionnaire and the system provided space behind each item to select one of the answers)

How I am in general? The following section lists a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please indicate the extent to which you *agree or disagree* with that statement.

- 1=disagree strongly
- 2=disagree a little
- 3=neutral
- 4=agree a little
- 5=agree strongly

I am someone who…

...tends to find fault with others

...does a thorough job

...can be somewhat careless

...is relaxed, handles stress well

...is a reliable worker

...tends to be disorganized

...worries a lot

...is generally trusting

...tends to be lazy

...perseveres until the task is finished

...does things efficiently

...remains calm in tense situations

...makes plans and follows through with them

…likes to cooperate with others

...trust in things other people say

Appendix C - Study II: Post-task questionnaire

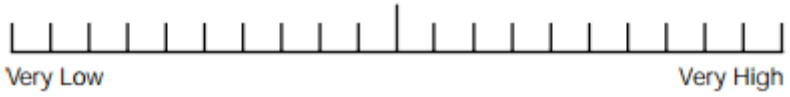
Post-task Questionnaire

Please rate the task according to these scales by circling the appropriate **vertical** line. Thank you.

Mental Demand – How mentally demanding was this task?

Physical Demand – How physically demanding was this task?

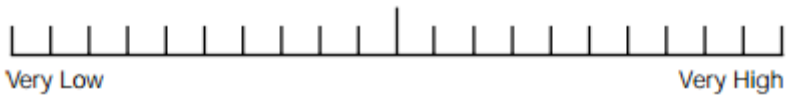
Temporal Demand – How hurried or rushed was the pace of this task?



Performance – How successful were you in accomplishing what you were asked to do?

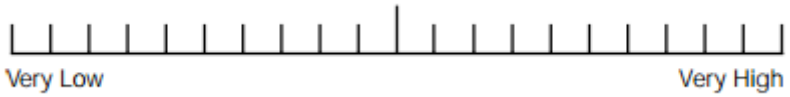
Effort – How hard did you have to work to accomplish your level of performance?

Frustration – How insecure, discouraged, irritated, stressed and annoyed were you?



Enjoyment – How pleased, entertained, satisfied and happy were you?

Engagement – How involved, thrilled or immersed were you?



Human-Robot Collaboration
- Robot perception -

Do you think the amount of information given to you by the robot was appropriate?

Too little	O	O	O	O	O	Too much
	1	2	3	4	5	

Do you think information given to you by the robot was detailed enough?

Not detailed enough	O	O	O	O	O	Too detailed
	1	2	3	4	5	

Do you think the information given to you by the robot was helpful?

Not at all helpful	O	O	O	O	O	Very helpful
	1	2	3	4	5	

To what extent did the robot contribute to the success of the task performance? Please tick the appropriate circle.

O 10%	O 20%	O 30%	O 40%	O 50%
O 60%	O 70%	O 80%	O 90%	O 100%

Please indicate to what extent you agree/disagree with the following statements.

Our team accomplished the task efficiently.

Strongly disagree	O	O	O	O	O	O	O	Strongly agree
	1	2	3	4	5	6	7	

I felt I was working with an intelligent being.

Strongly disagree	O	O	O	O	O	O	O	Strongly agree
	1	2	3	4	5	6	7	

The robot was very competent.

Strongly disagree	O	O	O	O	O	O	O	Strongly agree
	1	2	3	4	5	6	7	

I can trust the robot.

Strongly disagree	O	O	O	O	O	O	O	Strongly agree
	1	2	3	4	5	6	7	

I would like to operate this robot again.

Strongly disagree	O	O	O	O	O	O	O	Strongly agree
	1	2	3	4	5	6	7	

I think the robot malfunctioned.

Strongly disagree	O	O	O	O	O	O	O	Strongly agree
	1	2	3	4	5	6	7	

I am wary of the robot.

Strongly disagree O 1 O 2 O 3 O 4 O 5 O 6 O 7 Strongly agree

I am confident in the robot.

Strongly disagree O 1 O 2 O 3 O 4 O 5 O 6 O 7 Strongly agree

The robot is dependable.

Strongly disagree O 1 O 2 O 3 O 4 O 5 O 6 O 7 Strongly agree

The robot is reliable.

Strongly disagree O 1 O 2 O 3 O 4 O 5 O 6 O 7 Strongly agree

I felt competent operating the robot.

Strongly disagree O 1 O 2 O 3 O 4 O 5 O 6 O 7 Strongly agree

How familiar are you with the robot?

Not familiar at all O 1 O 2 O 3 O 4 O 5 O 6 O 7 O 8 O 9 O 10 very familiar

---

In your opinion, the way the robot communicated was:

Confusing O 1 O 2 O 3 O 4 O 5 Clear

Inconsistent O 1 O 2 O 3 O 4 O 5 Consistent

Hard to understand O 1 O 2 O 3 O 4 O 5 Easy to understand

Unfriendly O 1 O 2 O 3 O 4 O 5 Friendly

Unnatural O 1 O 2 O 3 O 4 O 5 Natural

Machinelike O 1 O 2 O 3 O 4 O 5 Humanlike

Thank you.

Appendix D - Study II: RVP analysis; in-between events

1. Event: In-between events

These events occurred independent from the robot's actions (e.g. finding a target, indicating reliability, etc.) and were recorded throughout the scenario.

1.1. Attention allocation

Attention allocation means the attention participants allocated towards the robot or the secondary task. Participants were free to choose to do a secondary task which involved to count boxes on the screen and press the appropriate number on the keyboard. Their incentive to actually do the task was that they could gain points for each correct answer. Therefore, participants used different strategies of allocating their time between the robot and the secondary task. From the retrospective verbal protocol the comments were collected and categorised under different themes (see Table 45).

Event	In between events: Attention allocation	
Theme	Task switching [152]	
Sub-theme	**From R to secondary [89]**	**From secondary to R [63]**
Raw data theme	• **Space/clear area in front [21]** • **R focus/advancing towards obvious point [10]** • R is slow/not turning [9] • R succeeded [9] • **More experience [7]** • Clear environment [6] • Illuminated areas [3] • Just panning, no focus [6] • **P identified target [5]** • More trust [3] • Low workload [4] • R got stuck [4] • P presumes no targets at start [1] • objects only on boxes/in-between boxes free time [1]	• **Cluttered/complex environment [10]** • **R moves/turns to new area [14]** • **Bad picture quality [10]** • Less trust in R [6] • R focusses on sth. [4] • Lower light [6] • R too fast [2] • **R stopped/slowed down [6]** • Too much secondary task/might miss something [1] • Difficult secondary task [1] • P anticipated important area [1] • Waiting for R to identify target [1] • Two missed [1]

Table 45 - TBCA overview of sub-event: Attention allocation (robot/secondary task)

When participants mentioned they switched between the tasks they could either do it away from the robot or *towards the secondary task* [89] or the other way round [63].

Task switching – From robot to secondary task [89]

They mostly switched from the robot to the secondary task when they *could see a clear and wide area in front* of the robot [21]:

- "Again familiarise with, looking at the scene, see there is nothing there, switching back to the secondary task." (P14; Parker)
- "[…] I guess I waited for the wide open, [to start secondary task] [hesitation sound] yeah there we go. I am playing the game [secondary task]. So I just wanted to see what was round the corner I guess and make sure it is still like wide open, so definitely I was trusting the robot in wide open areas, when visibility was high and both I could see that is nothing there." (P16; Roy)
- "[…], there is now objects here, and you can probably hear me, blitzing a lot at the secondary task, because nothing new is coming into view." (P10; Parker)

This goes hand in hand with the statements which commented that they were switching to the secondary task when the *robot was advancing towards an obvious point* [10]:

- "[…] it was clearly visible that the robot is advancing on this particular position, so unless something popped up in front of it, […], I can just, just leave it to it – whatever it is doing." (P02; Parker)
- "[…] it's obviously when he is travelling to a place, then there are not gonna be any items of importance. That was the time to get clicking." (P04; Roy)

It also showed that the *more familiar and experienced* participants felt the more they switched to the secondary task [7]. This is in agreement with the finding that the second performed task had significantly higher performance scores (cf. 5.4.2.1 Performance, p.134):

- "[…].So the more I probably, the more experienced I get with it the more am I try the other task [secondary], but it's still, I didn't want to miss a thing on the other one [robot]." (P24; Roy)
- "Yeah I am doing the secondary task at this point quite, it's maybe down to getting comfortable with the task as much as anything else. I doubt I was entirely consistent throughout, you get more comfortable with doing something here." (P10; Roy)

Interestingly, as soon as some *participants identified a target* (not yet the robot) [5] they were doing the secondary task and at the same time waiting that the robot identified/not identifies the target:

- "So as soon I picked something up, I thought: Okay it's not incorrectly identified that and then have a go at that [secondary task]." (P17; Parker)
- "When I knew what it [target] was, I clicked [secondary task]." (P01; Roy)

Generally, participants switched to the secondary task when the *picture was clear and not cluttered* [21]. Experience with the task led to more secondary task performance.

Task switching – From secondary task to robot [63]

The most stated reason for switching back towards the robot was the *robot moving/turning towards a new area* [14]:

- "You can hear me clicking some questions as it going straight forward again, fairly slowly and then as it begins to turn, I go back [supervising the robot]." (P09; Parker)
- "Again wide space, [hesitation sound], starts to pivot – that's when I look." (P02; Roy)

Also, c*luttered and complex environments* [10] made participants switch back to the robot:

- "I paid attention every time he might change camera angles. Like another field of view. And then he moved slow." (P05; Parker).

- "Again we are going to a bit more of a busier area here, so I have stopped doing the secondary task and the reliability is high, so." (P14; Roy)
- "Here I am waiting because it's like the corner so I am not doing anything on the secondary task. And there is a lot of stuff here, so I am thinking: ok." (P23; Roy)

Another theme that came up when switching back to the robot was *bad picture quality* [10] (bad video stream):

- "[hesitation sound] and I was struggling make out, [hesitation sound], images and things, so I really really really slowed down on the secondary task quite quickly and virtually neglected it, I did a bit but not very much." (P12; Roy).
- "Yes there so again, poor picture quality, certainty warrant more attention at this point." (P10; Parker)

Also when the *robot slowed down, stopped* or behaved uncharacteristically [6] participants allocated their attention to it:

- "Because it slowed down, I went back to look at it [robot]. And then what it looked at, - I didn't feel it was one of those [required objects to find]." (P02; Parker)
- "Here it seemed to kind of freeze or didn't seem to be doing much. That why I stopped [secondary task] for a few seconds and then got back to the task. Because the speed it was taking seemed to be kind of uncharacteristic for what is doing for." (P02; Roy)
- "Here, when it stopped I stop [secondary task]. When it slows down you can't hear me click. Oh because I thought: Why did he stopped?" (P08; Parker)

It can be assumed that the robot's movements have a huge influence on attention allocation. Not only what the robot sees, also how fast or slow it moves. As expected, busy or cluttered environments with potential targets and bad picture quality led participants to watch the robot more. Other literature reported that if a robot worked too slow the trust towards the robot decreased (Robinette et al., 2015).

The next three themes in the attention allocation event are *flicking between tasks* [8], *no switching* [22], and *distractions* [9] (see Table 46).

Event	In between events: Attention allocation		
Theme	**Flicking between tasks [8]**	**No switching (stay with R) [22]**	**Distractions [9]**
Sub-theme	**General [8]**	**General [22]**	**General [9]**
Raw data theme	• General mentioning [3] • **Bad picture quality [1]** • **R slowed down [3]** • **make sure [1]**	• **Unfamiliar with task [6]** • **Bad picture quality [5]** • **Less trust in R [4]** • R more important [2] • **Forgot secondary task [3]** • Unsure about low or high reliability area [1] • no attention for two tasks [1]	• **secondary task mistake [4]** • **Checking target list [2]** • **Switching between tasks [1]** • Bad picture quality [1] • Unknown object [1]

Table 46 - TBCA overview of sub-event: Attention allocation (switching)

Flicking between tasks [8]

Additionally a few participants reported that they were flick*ing between the tasks* [8], the reasons were similar to the listed above: *Bad picture quality* [1], *Robot was slow/slowed down* [3] or just *making sure* the robot does the right thing [1].

No task switching (stay with robot) [22]

Three participants totally *forgot about the secondary task*. Other reasons for not switching and staying on the robot (mostly at the beginning of the task) were that participants felt *unfamiliar with the task* [6]:

- "At this stage I was just trying to allocate as much attention as I could. Because I didn't know, what I am gonna be coming up against." (P08; Roy; 00:01)
- "I wanted to the secondary task again, but I was thinking: It is a new robot. I need to familiarise myself with its competence before I concentrate on that [secondary task]." (P17; Parker; 00:01)

- "I didn't look at the other task, because I had to get my eye in." (P07; Parker; 01:09)

The *bad picture quality* [5]:

- "I gave it a high level of attention, to know what was going on. The picture quality coming in and out as well, I didn't feel I could stop looking at it. And I don't think at any point through the whole exercise I felt that I could stop looking at it, may I should stop looking at it." (P06; Parker; 04:32)
- "So [hesitation sound] again it's just-I don't know- the image I think flickers a lot or moves a lot, and I found it very, - I was really really concentrating, really hard on this […]." (P12; Roy; 02:03)

Low trust in the robot [4]:

- "The same as before I didn't even try to look at the secondary task, at all. I still feel that, as that why could trust, - or felt I should trust the robot enough to pick it up." (P06, Roy)
- "No, I just literally started I think towards the end but [hesitation sound] I was thinking I actually [unclear] was wrong and I didn't want to miss anything." (P13; Roy)

To sum up the main reasons for not switching were when people felt unfamiliar with the task, experienced bad picture quality, had less trust in the robot or they forgot about the secondary task.

Distractions [9]

It distracted participants when they *made a mistake at the secondary task* [4]. As soon as the secondary task feedback became red (false answer), participants were looking back to the secondary task and were distracted from supervising the robot. Other distractions were *checking the target against the target list* [2] or switch*ing too much between tasks* [1].

1.2. Robot/Interface characteristics

Participants stated positive, neutral and negative comments about the robot's characteristics and the interface, as shown in Table 47.

Event	In-between events: Robot/Interface characteristics			
Theme	**Positive [14]**	**Neutral [15]**		**Negative [13]**
Sub-theme	**General [14]**	General [3]	**Questions [12]**	**General [13]**
Raw data theme	• **More feedback [2]** • **Good distance to objects [3]** • **Good speed [2]** • **Understandable search strategy of R [2]** • R's speed adjusts to environment [1] • P likes humanoid voice [1] • 360 degree view [1] • R looks thoroughly [1] • Accurate reliability assessment [1]	• R faster [1] • R stopped to look at objects [1] • R gave less information [1]	• Can the camera turn? Or the whole robot? [1] • Can R identify targets while moving? [1] • Does it identify when it pans? [1] • **What are the dimensions of R? [4]** • What does the R feedback means to me? [1] • Where the delay come from? [1] • Does the robot try to identify when stopped? [1] • How much is the reliability affected? [1] • Does it need to get close to identify? [1]	• **No feedback at the beginning [3]** • **R too fast [3]** • No building plan [1] • R can't access where humans could [1] • R speed does not adjust to environment [1] • No robot sound [1] • No feedback while R stopped [1] • too close to objects = light is in the way [1] • Too much feedback [1]

Table 47 - TBCA overview of sub-event: Robot interface characteristics (positive, neutral, negative)

Positive [14]

The *additional feedback* from Parker was positive [2].

- "It was giving me a lot more information this time. I thought it was a lot better, I liked being more informed and finding out why was it thought that it was in low identification. It gave me an understanding how I have to look harder. So when it was low lighting, I knew I have to kind of really [unclear: crouch] forward [towards screen] and yeah I thought the higher amount of information was much better. I almost maybe trust it more, I think." (P08; Parker)
- "Yeah but it gave more information. So I was happy with that." (P19; Parker)

Further the robot had a *good distance to objects* [3], *good speed* [2] and an *understandable/comprehensible search strategy* [2].

Neutral [15]

Of importance in the *neutral* theme were the *questions* people asked/inferred about the robot. These questions can give an insight to what is missing regarding information flow between robot and operator.

Some questions that the participants asked or inferred were about the *physical dimensions of the robot* [4]:

- "I was just thinking here: Is it getting stuck or is it gonna crashing into something […]." (P21; Roy)
- "At this point I also thought if the robot could squeeze under a gap that small, and I thought that is a bit of a stupid. Unless it's collapsible." (P20; Roy)

The current study setup did not allow the participant to see the robot before performing the tasks, because the robot design should not influence the recorded data.

The robot's process of identifying a target was questioned in many different ways, it seems that a more detailed explanation is necessary:

- Yeah it had two errors there. Yes so I guess it was [hesitation sound] yeah it needs to get close to the objects in order to pick it up. I assume. (P15; Parker); *Does it need to get close to identify?*

This missing information could be provided during training: A representation of the decision making process of the robot and the mechanism of identifying targets (iteration of planes, points, heat pattern, etc.) could be beneficial.

Negative [13]

Overall a huge negative factor for participants was that they had to wait for the robot to identify a target. This issue was not directly stated as a negative trait of the robot but became visible when looking at the previous analysed events and how often participants mentioned that they had to wait [31+].

Three participants criticised *not having any feedback at the beginning*:

- "So here I couldn't really hear anything from the robot for a while, so I was wondering if the mic, headset was still working. […]." (P16; Roy)
- "To begin with, I was a little bit, [hesitation sound], confused, because I thought, I didn't know whether I was having to speak about everything, or whether the robot was gonna tell me straight away whether it found anything, […] that was why I didn't know it was gonna tell me it found something, so I was getting as bit anxious." (P13; Roy)
- "I was a little bit worried, because I didn't know whether this was gonna talk to me or not. So I was a little, I wanted it to talk to me. I want it to say: Okay I am off. You know something like that." (P07; Parker)

This is similar to the comments of "*1st target found*" (see 5.4.5.1.7), where it was useful for the participants to know that the system works and they could familiarise with the voice.

Again, it is important to provide continuously the status of the robot and giving a starting message, so that people hear the voice and know the robot works properly. In addition, the robot motor sound could positively contribute to the overall understanding of the robots state. For instance, when we are driving a car uphill the motor needs more power and will sound differently, the same could be applied to the robot, if the robot drives over rubble (not obviously visible for the operator) and is therefore uncharacteristic slow, the operator can hear a changed motor sound and can more easily infer that the surface might be challenging for the robot.

In some cases the robot was *too fast* for the participants [3]. Even though the robot is autonomous it should be possible to slow it down to be adaptable to different skilled operators.

Ideas [13]

Ideas for improvements mentioned by the participants is shown in Table 48.

Event	In-between events: Robot/Interface characteristics		
Theme	**Ideas [13]**		
Sub-theme	**Camera and Movement [5]**	**Screen/Display [4]**	**General [4]**

Raw data theme	**• R speed adjust to P skills [1]** **• Camera not fixed on R better [1]** **• Get surround view [3]**	**• Screen feedback of reliability [1]** **• Showing reliability bars on screen constantly [1]** **• Indicate percentage of how sure R is about target [1]** **• Diagram of robot indicating status [1]**	**• Feedback if there is no target [1]** **• Command: Go back/check error [1]** **• more chatty/human [1]** **• Explanation what R is telling [1]**

Table 48 - TBCA overview of sub-event: Robot interface characteristics (participant ideas)

Participants' ideas that could improve the usability and behaviour of the robot were collected in this theme. The presented ideas incorporated giving *feedback if there is no target* [1], a *command to go back/check again* [1], a *more chatty human voice* [1], and an *explanation of what the robot is actually telling the operator* [1]:

- "I did think that it might be good, it might update you on other things, like you were just waiting for it to go around and find something, before it spoke to you. So it didn't say anything when it saw something that you saw, okay is that important. It is not what it is looking for but shouldn't it say something, is where." (P18; Parker); *feedback if there is no target*
- "[…] I felt it should be more: Oh look, - more chatty, more human I guess. Because that [the robot] was talking in such a way [succinct and standardised procedure protocol] I felt inhibited to do that [talking natural towards robot]." (P07; Roy); *more chatty human*
- "So better understanding on what it is telling me, would be needed." (P24; Parker); *explanation of what the robot is telling the operator*

Regarding the camera and robot movement participants' whished that they would be provided with a *surround view* [3] of the area when entering a new section:

- "[Hesitation sound], I think if I were operating it, as I say, I would have stopped in various positions and if the robot was the right size I would have done a 180 (shows 180 grad circle with hands). To get every surface as it were." (P09; Parker)

- "But it would have been, if I would have been in control, like to turn left now, because I wanna see if there is something left." (P03; Parker)
- "[…] I would have liked it, so it was going through this path and there things in both sides, eventually turned on this side [right] the robot that I was thinking that I would like if it had stopped at this point and make a whole turn, because I was thinking: Okay I haven't seen any on this wall. And I was wondering if it's gonna turn around that way or is it gonna go straight." (P23; Parker)

Other ideas consisted of *adjusting the robots speed* to the operator's skills:

- "[…] once you got into it after like three or four minutes, it was probably a bit slow. Once I skilled up. It could have probably gone a bit quicker. […]. So it was almost like, [hesitation sound], I lost a little bit of focus, I guess, or I could have done, if he would have gone on much longer on that speed." P07; Roy)

And a rotatable camera (*camera not fixed on R better*) [1]:

- ("Yeah and the camera only is pointing forwards. The robot is looking forwards, I figured, if I would have a separate camera I could have turned otherwise." P03; Parker).

With respect to the *screen/display* participants were not sure if they were still in low or high reliability areas, which inferred *visible feedback (status) of reliability* [1]:

- "There was one point where it said, yeah, I think it was here, where it said: reliability high, where I thought: Does that mean that through this period reliability has been low? Because I thought it only been in that little bit in the beginning, where it going into the shading, and I was like: Oh okay. I thought it was light before this." (P18; Parker)

Moreover, *showing reliability bars on screen constantly* (e.g. light, heat, accessibility and overall reliability) [1]:

- "[..] but I find it quite helpful, to have a thing of lighting levels [draw bar on interface], there is a bar and I don't know, [hesitation sound] the thing about heat affecting the sensors at one point, so you could

have that. And then be able to look at that and say okay well the overall reliability is high, because everything else is high. [Hesitation sound]. While it was good for the robot to say things, cause that meant that they bring it to my attention better, rather than just having a bar or something that would flash up. It also meant that, I only knew stuff when it chose to tell me. So if I was just driving along, and I was curious about darkness and what is the reliability level? It wouldn't... It only tell me when it changed. And I had to remember what the last change was." (P03; Parker)

The same participant also whished a *percentage value of how sure the robot is about a target* [1]:

- "[…] I felt like we were kind of competing, like the robot was trying to see stuff and I tried to see stuff, when there was two people competing at it. And I felt like collaborating if the robot would kind of say, here is something I am 20% sure it is a person. Here is something I am 60% sure, and then I can kind of filter that in that way the robot can spot stuff and I can seeing the robot flanks. And I don't know how much was this one? 20%? And 20% is a low enough threshold that it would capture stuff that I see and it doesn't." (P03; Parker)

Another beneficial statement was made by this participant about the *visualisation (diagram) of the robots status* [1]:

- "[..] And it might get stuck and you didn't know why. And that might take both time and a very complicated vocabulary for the robot to say that rear left tire was experiencing lower than usual traction, - whatever. Whereas you can just see a little diagram, tyre, tyre, tyre... low traction [drawing robot shape in screen corner]. Could be like that. And that's the problem [pointing and drawing in the air]." (P03; Parker)

Recapping these ideas the movement of the robot should be adjustable by the operator. If the robot turns towards new areas a surround view should be given to the operator. If possible, the reliability (low or high) should be

visualised on the screen, as well as the light levels, heat levels and accessibility levels (e.g. bars in percent). Furthermore, if the robot identifies a target it could also provide a percentage of how sure it is about the identification. If there is no target the robot would indicate that, too. Another feature could be a top view of the robot indicating any faults and the specific location of the fault.

Appendix E - Study III: General questionnaire

Virtual Robot Rescue Study

- General Questionnaire -

13.What is your gender? Please tick the appropriate circle.

O Female O Male

14.What age are you?

_____ years

15.What is your occupational title?

_____

If you are a researcher/student, please state your area of research/course of study.

_____

16.How frequently are you using a computer?

O daily O more than once a week O more than once a month O less than once a month O never

---

17.Do you play computer games?

O daily	O more than once a week	O more than once a month	O less than once a month	O never

If applicable, please state the type and name of the games you are playing most frequent.

_____
_____
_____

18.Do you have experience with robots?

O Yes O No

If Yes, please state occasion and name of the robot.

_____
_____
_____

In general… I would say that I trust robots

O Strongly disagree	O Disagree	O Undecided	O Agree	O Strongly agree

---

Personality test

How Accurately Can You Describe Yourself?

Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute

confidence. Indicate for each statement whether it is “Very Inaccurate”, “Moderately Inaccurate”, “Neither Accurate/Nor Inaccurate”, “Moderately Accurate”, or “Very Accurate” as a description of you. Please rate the following 50 items.

Participants could tick a circle behind each of the questions indicating how accurate the statement describes themselves. The questions asked were:

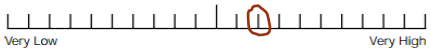
1. I am the life of the party.
2. Feel little concern for others.
3. I am always prepared.
4. Get stressed out easily.
5. Have a rich vocabulary.
6. Don't talk a lot.
7. I am interested in people.
8. Leave my belongings around.
9. I am relaxed most of the time.
10. Have difficulty understanding abstract ideas.
11. Feel comfortable around people.
12. Insult people.
13. Pay attention to details.
14. Worry about things.
15. Have a vivid imagination.
16. Keep in the background.
17. Sympathize with others' feelings.
18. Make a mess of things.
19. Seldom feel blue.
20. I am not interested in abstract ideas.
21. Start conversations.
22. I am not interested in other people's problems.
23. Get chores done right away.
24. I am easily disturbed.
25. Have excellent ideas.
26. Have little to say.
27. Have a soft heart.
28. Often forget to put things back in their proper place.

29. Get upset easily.
30. Do not have a good imagination.
31. Talk to a lot of different people at parties.
32. I am not really interested in others.
33. Like order.
34. Change my mood a lot.
35. I am quick to understand things.
36. Don't like to draw attention to myself.
37. Take time out for others.
38. Shirk my duties.
39. Have frequent mood swings.
40. Use difficult words.
41. Don't mind being the center of attention.
42. Feel others' emotions.
43. Follow a schedule.
44. Get irritated easily.
45. Spend time reflecting on things.
46. I am quiet around strangers.
47. Make people feel at ease.
48. I am exacting in my work.
49. Often feel blue.
50. I am full of ideas.

Appendix F - Study III: Post-task questionnaire

Virtual Robot Rescue Study

- Post-task Questionnaire -

Please rate the task according to these scales by circling the appropriate **vertical** line. Here is an example: 

Mental Demand – How mentally demanding was this task?

Physical Demand – How physically demanding was this task?

Temporal Demand – How hurried or rushed was the pace of this task?

Performance – How successful were you in accomplishing what you were asked to do?

Effort – How hard did you have to work to accomplish your level of performance?

Frustration – How insecure, discouraged, irritated, stressed and annoyed were you?

Enjoyment – How pleased, entertained, satisfied and happy were you?

Engagement – How involved, thrilled or immersed were you?

---

1. Please rate your performance for the last scenario and tick the appropriate circle.

Poor					Excellent
1	2	3	4	5	6
O	O	O	O	O	O

Please explain your answer:

_____

_____

_____

_____

2. Please rate the robot's overall performance for the last scenario.

Poor					Excellent
1	2	3	4	5	6
O	O	O	O	O	O

Please explain your answer:

_____

_____

_____

_____

3. Which mode would you prefer to use?

O Manual mode O Auto mode O No preference

4. Please indicate which mode had the better performance?

O Manual mode O Auto mode O No preference

5. How difficult did you perceive the task?

Extremely difficult O	O	O	O	O	Not at all difficult O

6. How complex do you rate the task?
   (Complexity means the simultaneous occurrence of several task components that influence your performance.)

Extremely complex O	O	O	O	O	Not at all complex O

7. How confident were you in performing the task?

Extremely confident O	O	O	O	O	Not at all confident O

Trust questionnaire for Human robot interaction (Schaefer, 2013)

Please tick the appropriate percentage value. If you think something is not applicable select N/A.

Participants could answer the statements in 10% intervals from 0% to 100% or tick not applicable (N/A).

What % of the time did this robot…

1. Act consistently
2. Protect people
3. Act as part of the team
4. Function successfully
5. Malfunction

6. Clearly communicate
7. Require frequent maintenance
8. Openly communicate
9. Have errors
10. Perform a task better than a novice human user
11. Know the difference between friend and foe
12. Provide Feedback
13. Possess adequate decision- making capability
14. Warn people of potential risks in the environment
15. Meet the needs of the mission
16. Provide appropriate information
17. Communicate with people
18. Work best with a team
19. Keep classified information secure
20. Perform exactly as instructed
21. Make sensible decisions
22. Work in close proximity with people
23. Tell the truth
24. Perform many functions at one time
25. Follow directions
26. Considered part of the team

What % of the time was this robot…

27. Responsible
28. Supportive
29. Incompetent
30. Dependable
31. Friendly
32. Reliable
33. Pleasant
34. Unresponsive
35. Autonomous
36. Predictable
37. Conscious
38. Lifelike
39. A good teammate

40. Led astray by unexpected changes in the environment

The Muir (1989) trust questionnaire

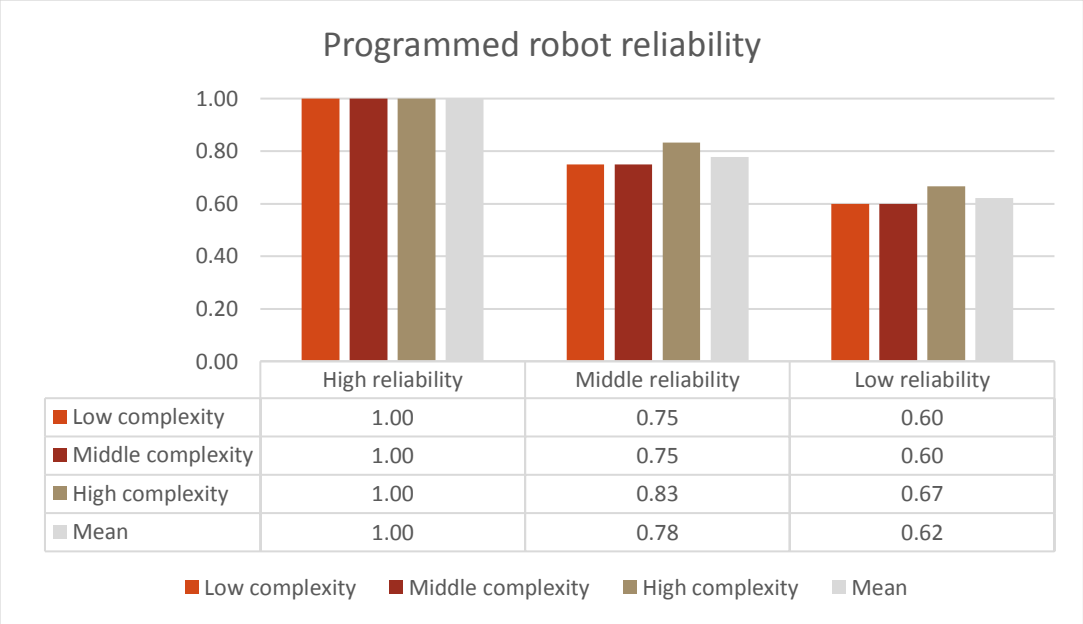
Please select a value from 1 to 10 regarding the following questions.

	Not at all 1	2	3	4	5	6	7	8	9	Com-pletely 10
To what extent can the system's behaviour be predicted from moment to moment?	O	O	O	O	O	O	O	O	O	O
To what extent can you count on the system to do its job?	O	O	O	O	O	O	O	O	O	O
How confident are you that the system will be able to cope with all situations in the future?	O	O	O	O	O	O	O	O	O	O
Overall how much do you trust the system?	O	O	O	O	O	O	O	O	O	O

Appendix G - Study III: Programmed robot reliability

Programmed reliability across reliability levels

For a better understanding Figure 128 will illustrate the programmed robot reliabilities across the conditions. The values represent the ratio between the programmed number of targets missed and the programmed number of targets found. The high complexity values are slightly higher for middle and low reliability due to the different number of targets present in the trial. In the high reliability conditions the robot is programmed not to make any mistakes. In the middle robot reliability condition the robot is programmed to make one mistake and be unreliable for a certain amount of time, which is visible by not inspecting all required areas during the low reliability time as shown in the reliability profiles in Figure 128, p. 396. During low reliability the robot will have two low reliability sections and makes in each of them a mistake.



	High reliability	Middle reliability	Low reliability
Low complexity	1.00	0.75	0.60
Middle complexity	1.00	0.75	0.60
High complexity	1.00	0.83	0.67
Mean	1.00	0.78	0.62

Figure 128 - Overview of the programmed reliability levels of the robot

Programmed reliabilities across task complexity

Figure 129 gives an overview of the task complexity conditions and the related mean programmed robot performance. Unfortunately due to the

different number of targets the high complexity condition's mean of 83% is 5% higher than the low or middle complexity means. This needs to be taken into account when interpreting results across the task complexity levels.

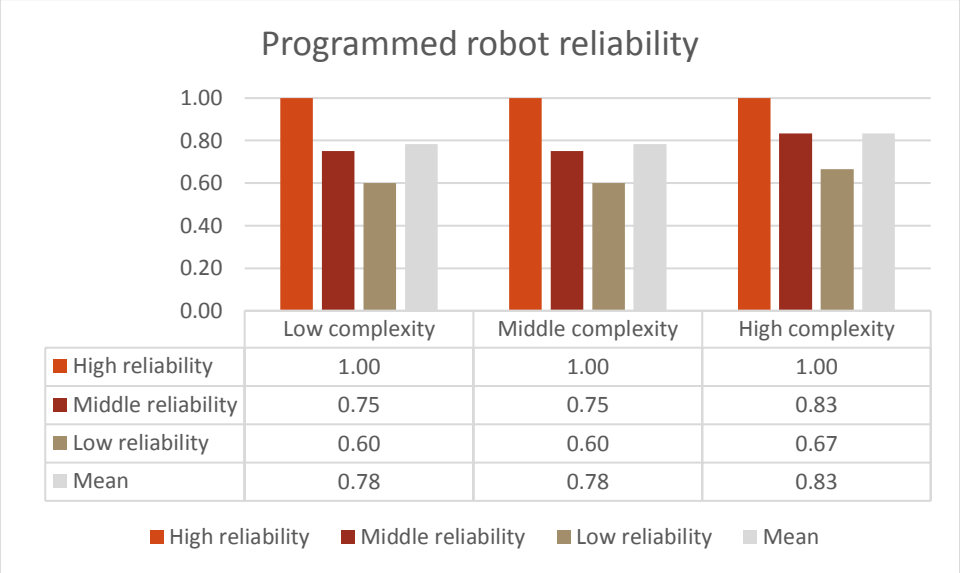


Figure 129 - Overview of programmed robot reliability levels across task complexity

Although the programmed reliability is fixed in each condition and not a dependent variable, the fact that participants are free to choose between manual and auto mode gives a different reliability for each run for each participant.

Appendix H - Study IV: General questionnaire

Virtual Robot Rescue Study 2

- General Questionnaire -

---

19. What is your gender? Please tick the appropriate circle.

O Female O Male

20. What age are you?

_____ years

21. What is your occupational title?

_____

If you are a researcher/student, please state your area of research/course of study.

_____

22. Do you have prior experience with robots? (E.g. took part in one of the previous robot studies, worked with robots, etc.)

_____

_____

Please read the following statements and tick the appropriate circle.

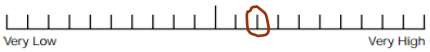
Trust propensity scale	**Strongly disagree**				**Strongly agree**
1. I usually trust robots until there is a reason not to.	O 1	O 2	O 3	O 4	O 5
2. For the most part, I distrust robots.	O 1	O 2	O 3	O 4	O 5
3. In general, I would rely on a robot to assist me.	O 1	O 2	O 3	O 4	O 5
4. My tendency to trust robots is high.	O 1	O 2	O 3	O 4	O 5
5. It is easy for me to trust robots to do their job.	O 1	O 2	O 3	O 4	O 5
6. I am likely to trust a robot even when I have little knowledge about it.	O 1	O 2	O 3	O 4	O 5

Appendix I - Study IV: Post-task questionnaire

Virtual Robot Rescue Study 2

- Post-task Questionnaire – Condition: …….

Please rate the task according to these scales by circling the appropriate **vertical** line. Here is an example:



Mental Demand – How mentally demanding was this task?

Physical Demand – How physically demanding was this task?

Temporal Demand – How hurried or rushed was the pace of this task?

Performance – How successful were you in accomplishing what you were asked to do?

Effort – How hard did you have to work to accomplish your level of performance?

Frustration – How insecure, discouraged, irritated, stressed and annoyed were you?

Please rate the following statements with respect to the scenario you just performed.

8. How complex do you rate the task?
   (Complexity means the simultaneous occurrence of several task components that influence your performance.)

Not at all complex O	O	O	O	O	Very complex O

---

9. How difficult did you perceive the task?

Not at all difficult O	O	O	O	O	Very difficult O

10. Please rate **<u>your</u>** performance for the last scenario and tick the appropriate circle.

Poor 1 O	2 O	3 O	4 O	5 O	6 O	Excellent 7 O

11. Please rate the **<u>robot's</u>** overall performance for the last scenario.

Poor 1 O	2 O	3 O	4 O	5 O	6 O	Excellent 7 O

12. How confident were you in performing the task?

Not at all confident	O	O	O	O	Very confident

O					O

Trust questionnaire – short (Schaefer, 2013)

Please tick the appropriate percentage value.

Participants could answer the statements in 10% intervals from 0% to 100% or tick not applicable (N/A).

What % of the time did/was this robot…

1. Act consistently
2. Function successfully
3. Malfunction
4. Have errors
5. Provide Feedback
6. Meet the needs of the mission
7. Provide appropriate information
8. Communicate with people
9. Follow directions
10. Dependable
11. Reliable
12. Unresponsive
13. Predictable
14. Perform exactly as instructed

Appendix J - Study IV: Analysis of trust questionnaire

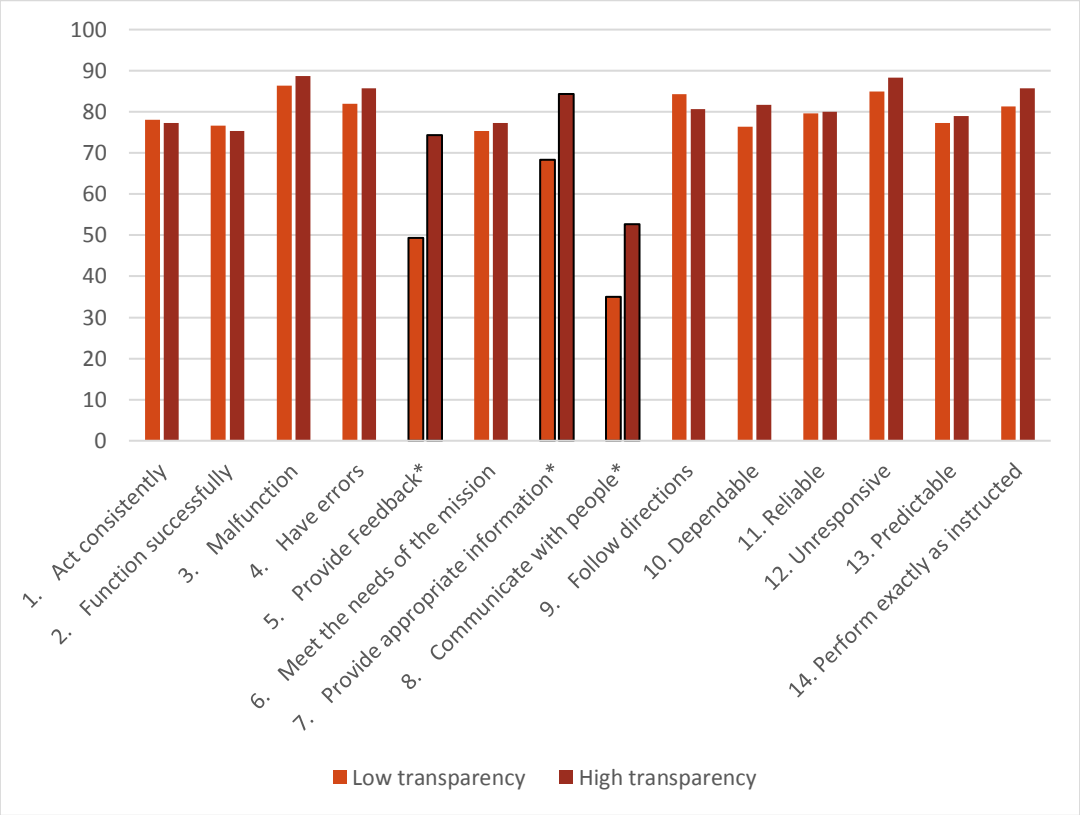


Figure 130 - Detail analysis of short trust questionnaire (* indicates the biggest changes between conditions)

Appendix K - Digital Appendix

Instructions:

1. To access the Digital Appendix, please click on the following link: https://www.dropbox.com/sh/8y1cy1kid7hhl4v/AAC1AIuZXV6Sy96-buZ-QWlda?dl=0
2. Password for sensitive datasets needs to be requested via e-mail: Katharina.Gabrecht@nottingham.ac.uk
3. If there are any problems accessing the data, please write to the e-mail above.

..