Explaining Robot Behaviour

Beliefs, Desires, and Emotions in Explanations of Robot Action

Kaptein, F.C.A.

DOI
10.4233/uuid:1d92d61c-c124-4e7b-903e-bce246410bba

Publication date
2020

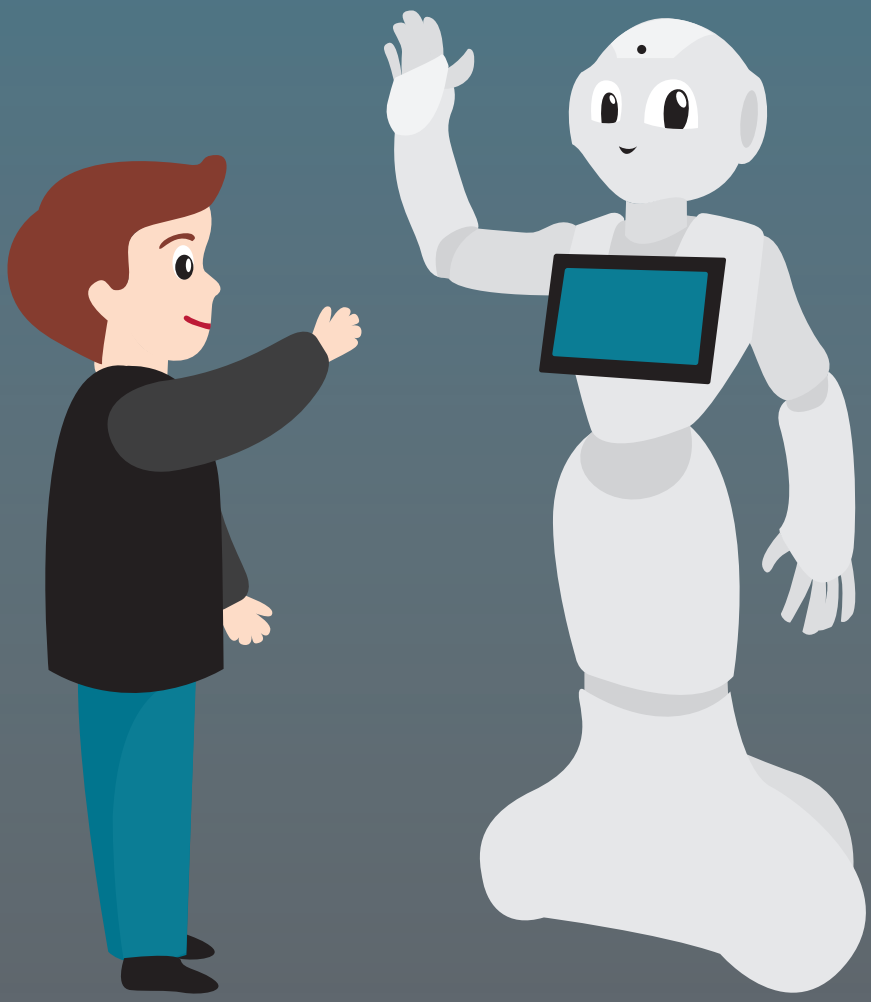
Document Version
Final published version

Citation (APA)
Kaptein, F. C. A. (2020). *Explaining Robot Behaviour: Beliefs, Desires, and Emotions in Explanations of Robot Action*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:1d92d61c-c124-4e7b-903e-bce246410bba

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Explaining Robot Behaviour

Beliefs, Desires, and Emotions in Explanations of Robot Action



Frank Kaptein

Explaining Robot Behaviour

Beliefs, Desires, and Emotions in Explanations of Robot Action

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof. dr. ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op woensdag 11 november 2020 om 10:00 uur

door

Frank Cornelis Adriaan KAPTEIN

Master of Science in Computer Science
Technische Universiteit Delft, Netherland,
geboren te Zoetermeer, Nederland.

Dit proefschrift is goedgekeurd door de promotoren

Samenstelling promotiecommissie bestaat uit:

Rector Magnificus,	voorzitter
Prof.dr. M.A. Neerincx	Technische Universiteit Delft, promotor
Prof.dr. K.V. Hindriks	Vrije Universiteit Amsterdam, promotor
Dr. J. Broekens	Universiteit Leiden, copromotor

Onafhankelijke leden:

Prof.dr. T. Belpaeme	Ghent University & University of Plymouth
Prof.dr.ir. D.A. Abbink	Technische Universiteit Delft
Prof.dr. C.M. Jonker	Technische Universiteit Delft
Dr. M.M.A. de Graaf	Universiteit Utrecht

TU Delft Delft University of Technology

Front & Back: Bregje Jaspers, proefschriftontwerp.nl

Printed by: ProefschriftMaken, proefschriftmaken.nl

ISBN 978-94-6423-040-6

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

”It’s a dangerous business..., going out of your door,” ... ”You step into the Road, and if you don’t keep your feet, there is no knowing where you might be swept off to.”

J.R.R. Tolkien
(Bilbo Baggins in *The Lord of the Rings*)

Contents

Summary xi

Samenvatting xv

1 Introduction 1
- 1.1 Self-Explanations by Robots . . . 2
- 1.2 Background: Explanations and Folk Psychology . . . 2
- 1.3 Related Work . . . 3
  - 1.3.1 Emotions Simulation for Intelligent Agents . . . 4
- 1.4 Definitions and Terminology . . . 5
- 1.5 The PAL project . . . 5
- 1.6 Research Questions And Thesis Structure . . . 6

References . . . 8

2 Cloud-based Social Robots for Health Education & Care 13
- 2.1 Introduction . . . 15
- 2.2 Related Work and Context . . . 16
  - 2.2.1 Related Work . . . 16
  - 2.2.2 Context: a Personal Assistant for a Healthy Lifestyle . . . 16
- 2.3 Principles for a Social Robot System for Long-term Interaction . . . 17
  - 2.3.1 Principle 1: Cloud-based Robots . . . 18
  - 2.3.2 Principle 2: Modular System . . . 18
  - 2.3.3 Principle 3: Common Knowledge-base and Terminology . . . 19
  - 2.3.4 Principle 4: Hybrid Artificial Intelligence . . . 19
- 2.4 System Implementation for a Social Robot in Health Education & Care . . . 19
  - 2.4.1 The Ontology . . . 19
  - 2.4.2 The Database . . . 22
  - 2.4.3 PAL Control & Inform . . . 24
  - 2.4.4 Activity Centre . . . 26
  - 2.4.5 Communication Between Modules . . . 26
  - 2.4.6 Multimodal Behaviour Manager . . . 27
  - 2.4.7 'The Hybrid Brain' . . . 30
- 2.5 Development and Test procedures . . . 37
- 2.6 Analyses of Performance . . . 38
- 2.7 Future Extensions . . . 39
- 2.8 Lessons Learned and Discussion . . . 40
- 2.9 Conclusion . . . 42

References . . . 42

3 Personalised Self-Explanation by Robots: The Role of Goals versus Beliefs in Robot-Action Explanation for Children and Adults **49**
3.1 Introduction . . . 51
3.2 Motivation for Research Conducted . . . 52
3.2.1 Goal-based and Belief-based Explanations . . . 52
3.2.2 Hypothesis . . . 52
3.3 Goal Hierarchy Trees . . . 53
3.3.1 The Structure of a Goal Hierarchy Tree . . . 54
3.3.2 Goal-based and Belief-based Agent-action Explanations 54
3.4 User Study . . . 54
3.4.1 Participants . . . 55
3.4.2 Designing a Goal Hierarchy Tree . . . 55
3.4.3 Set-up & Materials . . . 56
3.4.4 Variables & Design . . . 58
3.4.5 Procedure . . . 58
3.5 Results . . . 59
3.6 Discussion . . . 60
3.7 Conclusion . . . 61
References . . . 61

4 Evidence for the Use of Emotion in Human Explanations of Robot and Human Behaviour **65**
4.1 Introduction . . . 67
4.2 Background and Related Work . . . 68
4.2.1 Explanations and Folk Psychology . . . 68
4.2.2 Emotions and Coping Styles . . . 69
4.3 Research Questions . . . 69
4.4 Experiment . . . 72
4.4.1 Designing Conversations in Coping Styles . . . 72
4.4.2 Participants . . . 73
4.4.3 Experimental Design . . . 73
4.4.4 Materials and Measures . . . 73
4.4.5 Procedure . . . 75
4.5 Results . . . 76
4.5.1 Recognition of the Coping Styles . . . 76
4.5.2 Emotionality of Explanations . . . 79
4.5.3 Perception of Coping Styles . . . 81
4.6 Discussion . . . 84
4.6.1 Limitations . . . 86
4.6.2 Implications for Robot Self-Explanations . . . 86
4.7 Conclusion . . . 87
4.8 Acknowledgements . . . 87
References . . . 88

5 CAAF: A Cognitive Affective Agent Programming Framework 93
5.1 Introduction . . . 95
5.2 Motivation & Related Work . . . 96
5.3 A Model of Emotion for Cognitive Agent Programming Frameworks . . . 98
5.3.1 Semantics for a Basic Knowledge Representation & BDTE . . . 98
5.3.2 Closing the Semantic Gap between BDTE and BDI . . . 100
5.3.3 Querying the Emotion Base . . . 102
5.4 Proof of Consistency when Minimizing the (Re)Appraisal of Emotions . . . 103
5.5 Discussion . . . 104
5.6 Conclusion . . . 105
References . . . 106

6 Evaluating Cognitive and Affective Intelligent Agent Self-Explanations for Long-Term Health-Support 109
6.1 Introduction . . . 111
6.2 Motivation, Related Work, and Hypothesis . . . 112
6.3 Implementation of a Model for Explainable AI . . . 113
6.3.1 Explainable Actions . . . 113
6.3.2 Content of explanations . . . 115
6.3.3 Presentation of explanations . . . 115
6.4 Method . . . 116
6.4.1 Participants . . . 116
6.4.2 Experimental Design . . . 116
6.4.3 Measures and Variables . . . 117
6.4.4 Material & Set-Up . . . 117
6.4.5 Procedure . . . 117
6.5 results . . . 117
6.6 Discussion . . . 119
6.7 Conclusion . . . 120
References . . . 121

7 Conclusion 125
7.1 Findings . . . 126
7.2 Limitations . . . 130
7.3 Future Work . . . 131
7.4 Overall Contribution . . . 132
References . . . 132

Epilogue 135

A Adjusted Ways of Coping Questionnaire 137

B Filmed Conversations of Coping Styles 139
B.0.1 Conversations in Coping Styles . . . 139
B.0.2 Making Videos of the Conversations . . . 141

References 143
C T-values for Coping Style Recognition **145**
List of Publications **147**

Summary

Social humanoid robots are complex intelligent systems that in the near future will operate in domains including healthcare and education. Transparency of what robots intend during interaction is important. This helps the users trust them and increases a user's motivation for, e.g., behaviour change (health) or learning (education). Trust and motivation for treatment are of particular importance in these *consequential domains*, i.e., domains where the consequences of misuse of the system are *significant*. For example, rejecting treatment can have a negative impact on the user's health. Transparency can be enhanced by having the robot explain its behaviour to its users (i.e., when the robot provides *self-explanations*). Self-explanations help the user to assess to what extent he or she should trust the decision or action of the system.

Self-explanations of humanoid robots are typically based on how people explain their own and each other's behaviour amongst each other (i.e., *human behaviour explanations*). When people explain a person's (their own or someone else's) behaviour then they do so by referring to that person's beliefs, desires, and emotions. Humans make intuitive, split second decisions to decide what elements are best suited to explain behaviour in a situation to a particular receiver. In contrast, work on self-explanations by robots has mainly focused on referring to desires and sometimes beliefs, and in a non-personalised manner. The main question of this thesis is: 'Which aspects of human behaviour explanation can be used in the construction of social humanoid robot self-explanations and how should we generate such explanations?' In this thesis, we focus on two aspects of this question: 1) attuning explanations to the receiver; and 2) using emotions as part of the explanations.

In the introduction we give an overview of social robots, discuss how humans amongst each other explain behaviour, and how this inspired the design of explanations of autonomous agent behaviour (like social robots or virtual artificial characters). Furthermore, we discuss the European project affiliated with this thesis.

In chapter two, we discuss design principles and a resulting implementation for a system with a social humanoid robot. Issues were reaching long-term, personalised interaction, for different groups of users, in complex consequential and real-world application domains. We implemented a cloud-based, modular, social-robot system which provides personalised behaviour change support. The system is developed to autonomously interact with its users for a prolonged period of time (2 periods of 2.5 – 3 months). The context within which the system is developed is supporting diabetes management of children (aged 6-14). However, the system's architecture and principles are designed to provide health-support and education in a more general way. This chapter discusses the type of social robot system that serves as context for which we develop the explanations.

In chapter three, we aim to get a better understanding of whether and how robot self-explanations should be attuned to the receiver of the explanation. We look at user preferences and the differences between children and adults who receive explanations from a robot. We implemented a humanoid robot as a belief-desire-intention (BDI)-based agent and explained its actions using two different explanation styles. One based on the robot's beliefs that give context information on why the agent performed the action. The other based on the robot's goals that inform the user of the agent's desired state when performing the action. We investigated the preference of children and adults for goal- versus belief-based action explanations. From this, we learned that adults have a significantly higher tendency to prefer goal-based action explanations. Providing insight in preferences for BDI elements in explanations is an important preliminary step in the challenge of providing more personalised explanations in human-robot and human-agent interaction.

In chapter four, we address whether and how humans use emotions in their explanations of robot behaviour. Answering this question is important for two main reasons. First, it helps us design ways in which social robots can explain their own actions. Second, it gives insight into human attribution of mental states to robots. To study this, we presented filmed behaviours of a social humanoid robot coping with a distressing situation to MTurk participants. Coping was done in several styles drawn from literature. As a between subject control, we also presented all behaviours performed by a human actor. We asked participants to rate their recognition of these coping styles and how they would explain the behaviour (by typing this in an open text box). Results show that overall participants recognised the coping styles and used emotions in their explanations for both the robot and the human actor. Participants used significantly less emotions when explaining robot behaviour; however, with a very small effect size. Finally, for participants that were shown videos of human behaviour, we found that the recognised coping style correlated with the emotionality used in the explanations. We did *not* see this for participants that were shown videos of robot behaviour. We discuss implications of our findings for our understanding of human perception of robot behaviour. Finally, our analysis shows *that* emotions are often a part of the explanations; however, it is still unclear *when* emotions are a part of the explanations. We found that this is different for robots versus humans. The recognition of certain coping styles correlates with emotionality of the explanation when explaining human behaviour, but not when explaining robot behaviour. With this we identify an important line of future work. The main conclusion of this study is: if we intend to explain robot behaviour like a human would have, then we often need emotions as part of the explanation.

In the previous chapter, we looked at human explanations of robot behaviour. In chapter five, we look at the simulation of intelligent agent (e.g., robot) emotions. This is important because if the robot must use emotions in the explanations then it must be able to represent and generate them. Furthermore, we argue this should be done in such a way that the simulation stays close to emotion theory of how people understand and use emotions because people must understand the meaning of the emotion as used in the explanation. There are many computa-

tional models of emotion, all with their own specific value. However, these models typically simulate emotions based on cognitive appraisal theory. Which introduces a large set of appraisal processes not specified in enough detail for unambiguous implementation. This is particularly difficult for belief-desire-intention based (i.e., cognitive) agent programming. We present a framework based on the belief-desire theory of emotions (BDTE). This framework enables the computation of emotions for cognitive agents. In this paper, we bridge the remaining gap between BDTE and cognitive agent programming frameworks.

Chapter six presents two styles of robot self-explanations in our social robot system tested in a long-term in the wild study. Research in e-health support systems and human-robot interaction stresses the need for studying long-term interaction with users. We propose the effects of robot self-explanations should thus *also* be tested in prolonged interaction. We report on an experiment in which we tested the effect of cognitive, affective and lack of explanations on children's motivation to use an e-health support system. Children (aged 6-14) suffering from type 1 diabetes mellitus interacted with our system over a period of 2.5 - 3 months. Children alternated between the three conditions. Agent behaviours that were explained to the children included why 1) the agent asks a certain quiz question; 2) the agent provides a specific tip (a short instruction) about diabetes; and, 3) the agent provides a task suggestion, e.g., play a quiz, or, watch a video about diabetes. Their motivation was measured by counting how often children would accept the agent's suggestion, how often they would continue to play the quiz or ask for an additional tip, and how often they would request an explanation from the system. Surprisingly, children proved to follow task suggestions more often when no explanation was given, while other explanation effects did not appear. This is not in line with literature on related work and pedagogy and serves as an important lesson learned for developing explanations in long-term interaction. This is (to our knowledge) the first long-term study to report empirical evidence for an agent explanation effect, challenging future studies to uncover the underlying mechanism.

The work in this thesis shows that self-explanation algorithms should indeed consider more aspects of how humans amongst each other explain behaviour. (1) We show explanations must take the receiver of the explanation into account. Context like user type is essential. Furthermore, (2) we show people indeed use emotions themselves when explaining robot behaviour. Future work includes analysing how such personalised and emotion laden explanations would influence trust in the system. Furthermore, chapter six shows that an explanation effect on motivation occurred in long-term interaction. However, these effects were not in line with the expectations based on literature, showing the need for also more work on this. In this thesis, we designed and tested the explanations in a real-world ('in the wild') system in a consequential domain (helping children aged 6-14 to become more self-manageable with regards to their illness). Our research already shows that it is possible to address research questions in complex consequential domains, even with limited groups of users and over prolonged periods of interaction time. Overall, we conclude that work in explainable artificial intelligence, both in the social sciences as well as in human computer interaction, should consider individual

preferences and should consider emotions in addition to beliefs and desires when explaining robot or avatar behaviour.

Samenvatting

Sociale humanoïde robots zijn complexe intelligente systemen die in de nabije toekomst zullen opereren in domeinen zoals zorg en onderwijs. Transparantie van wat de robots nastreven tijdens de interactie is belangrijk. Dit maakt dat men ze eerder zal vertrouwen en verhoogt daarmee de gebruiker zijn motivatie tot, bijvoorbeeld, gedragsverandering (zorg) of leren (educatie). Vertrouwen en motivatie zijn inderdaad belangrijke onderwerpen in deze domeinen. Transparantie kunnen we versterken door de robot zijn gedrag uit te laten leggen aan de gebruiker (dit noemen wij hier '*zelf-verklaringen*'). Zelf-verklaringen helpen de gebruiker om in te schatten in welke mate hij/ zij beslissingen en gedragingen van het systeem moet vertrouwen.

Zelf-verklaringen van humanoïde robots zijn typisch gebaseerd op hoe mensen onderling hun eigen en elkaars gedrag verklaren (dit noemen we hier *mens-op-mens gedragsverklaringen*). Wanneer mensen het gedrag van een persoon (zichzelf of iemand anders) verklaren doen ze dit door te refereren naar de persoon zijn gedachtes, verlangens, en emoties. Mensen maken binnen een fractie van een seconde, intuïtieve beslissingen om te bepalen welke elementen het best passen om gedrag in een specifieke situatie uit te leggen aan een specifiek persoon. Daarentegen is onderzoek aangaande zelf-verklaringen van robots tot nu toe voornamelijk gefocust op het gebruik van verlangens, en soms gedachtes, op een niet-gepersonaliseerde wijze. De hoofdvraag van deze thesis is: 'Welke aspecten van mens-op-mens gedragsverklaringen kunnen gebruikt worden in het ontwikkelen van sociale humanoïde robot zelf-verklaringen en hoe kunnen we zulke verklaringen genereren?' In deze thesis focussen we op twee aspecten van deze vraag: 1) verklaringen afstemmen op de ontvanger van de verklaring, en 2) het gebruik van emoties als onderdeel van de verklaringen.

In de introductie geven we een overzicht van sociale robots, bespreken we mens-op-mens verklaringen, en bespreken we hoe zulke verklaringen het ontwikkelen van verklaringen in autonome agent systemen (zoals sociale robots of virtuele artificiële karakters) hebben geïnspireerd. Ten slotte bespreken we het Europese project geaffilieerd met deze thesis.

In hoofdstuk twee bespreken we de ontwikkelprincipes en een implementatie van een systeem met een sociale robot. De uitdaging was om lange-termijns-, gepersonaliseerde interactie te bewerkstelligen voor verschillende gebruikers groepen en in een complex zwaarwegend domein uit de samenleving ('real-world' in plaats van een verzonnen 'lab' domein). We hebben een 'cloud-based' (over het internet), modulair systeem ontwikkeld dat gedragsverandering en ondersteuning bied. Het systeem is ontwikkeld om autonoom met zijn gebruikers te interacteren over een langdurige periode (2.5 - 3 maanden). De context van het systeem is het ondersteunen van diabetes management van kinderen (leeftijd 6-14). Maar de

ontwikkelprincipes en de architectuur van het systeem zijn dusdanig opgezet dat het gedragsondersteuning op een generieke wijze kan ondersteunen. Het systeem besproken in dit hoofdstuk is ook het type sociale humanoïde robot systeem waar wij de uitleggingen voor ontwikkelen in deze thesis. Het dient dus ook als context voor de hierop volgende hoofdstukken.

In hoofdstuk drie onderzoeken we of en hoe robot zelf-uitleggingen aan de gebruiker moeten worden afgestemd. We kijken naar gebruikers voorkeuren voor verschillende type uitleggingen en testen op het verschil in voorkeur tussen volwassenen en kinderen. We hebben een robot geïmplementeerd als een BDI-based agent (dit is een term voor systemen die redeneren op basis van hun 'gedachtes', 'verlangens', en 'intenties'; of in het Engels 'beliefs', 'desires', en 'intentions'; BDI). De robot gaf zelf-verklaringen voor zijn gedrag in twee verschillende stijlen. Eén gebaseerd op zijn 'gedachtes' welke contextuele informatie omvatten over waarom de robot het gedrag vertoonde. En één gebaseerd op zijn verlangens welke tonen wat de robot wilde bereiken met het gedrag. We onderzochten de voorkeuren van kinderen en volwassenen voor deze verklaringen. We hebben hiervan geleerd dat volwassenen een sterkere voorkeur hebben voor verlangen-gebaseerde uitleggingen dan kinderen. Inzicht verkrijgen in voorkeuren voor uitlegstijlen is een belangrijke stap om gepersonaliseerde zelf-verklaringen te kunnen bieden.

In hoofdstuk vier onderzoeken we of en hoe mensen emoties gebruiken in *hun* uitleggingen van robot gedrag. Deze vraag is belangrijk om twee hoofdredenen. Ten eerste helpt het ons voor het ontwikkelen van robot-zelfverklaringen. Ten tweede verschaft het ons inzicht over hoe mensen denken over robot gedrag en welke mentale concepten (zoals bijvoorbeeld verlangens en emoties) ze attribueren aan het gedrag. Om dit te onderzoeken hebben we participanten van een MTurk studie gefilmde gedragingen laten zien van een sociale humanoïde robot welke omgaat (met de term uit het Engels: 'coping') met een stressvolle situatie. Coping werd in verschillende stijlen gedaan, gebaseerd op de literatuur. Ter controle werden anderen participanten een menselijke acteur getoond welke de coping stijlen vertoonde. Participanten gaven aan welke stijlen ze herkenden in het gedrag en we vroegen participanten om een uitleg te geven voor het gedrag (door deze te typen in een open tekstvak). Resultaten tonen dat de participanten in het algemeen, voor zowel de menselijke acteurs als voor de robot, de coping stijlen konden herkennen en dat ze emoties gebruikten in hun uitleggingen. Participanten gebruikten wel significant minder emoties bij het uitleggen van robot gedrag, maar met een zeer kleine effect grootte. We vonden dat voor onze set gedragingen 80% van de uitleggingen van menselijk gedrag emoties bevatte, en 75% van de uitleggingen van robot gedrag emoties bevatte. In dit hoofdstuk bespreken we de implicaties van onze resultaten voor ons begrip van hoe mensen robot gedrag waarnemen. Ten slotte toont onze analyse *dat* emoties vaak een deel zijn van uitleggingen, maar, het is nog steeds onduidelijk *wanneer* emoties een deel van de uitleg moeten zijn. Onze resultaten laten zien dat dit verschild voor mensen en robots. Onze resultaten tonen dat bij verklaringen van menselijk gedrag dit correleert met het toedichten van bepaalde coping stijlen aan het gedrag, maar bij robots niet. Hiermee identificeren wij een belangrijke vraag voor toekomstige studies. De hoofdconclusie

van deze studie is: als we robot gedrag willen verklaren zoals een mens dat doet, moeten we regelmatig emoties gebruiken als onderdeel van de uitleg.

In het hoofdstuk vier bekeken we menselijke uitleggingen van robot gedrag. In hoofdstuk vijf kijken we naar de simulatie van emoties van intelligente artificiële agenten (zoals robots). Als de robot emoties moet gebruiken in uitleggingen dan moet de robot deze emoties kunnen representeren en genereren. Verder beargumenteren we dat dit dusdanig moet dat de simulatie overeenkomt met emoties theoriën over hoe mensen emoties gebruiken en begrijpen zodat mensen de emotie in de uitleg ook inderdaad kunnen begrijpen. Er zijn vele computationele modellen van emotie, allen met hun eigen specifieke waarde. Maar deze modellen zijn typisch gebaseerd op 'appraisal theory'. Dit introduceert een grote set aan processen welke in onvoldoende detail zijn gedefinieerd om ze ondubbelzinnig te implementeren. Dit is met name lastig wanneer we een BDI-based agent programmeer taal gebruiken. Wij presenteren daarom een framewerk gebaseerd op de 'gedachte'-'verlangen' theorie (BDTE) van emotie. Dit framewerk maakt het mogelijk om emoties voor deze programmeer talen te berekenen. In deze paper sluiten we de kloof tussen BDTE en BDI-based agent programmeer framewerken.

Hoofdstuk zes presenteert twee stijlen van robot zelf-verklaringen in ons sociale robot systeem getest in een lange-termijn studie. Onderzoek aangaande e-health support systemen en mens-robot interactie benoemt vaak dat het belangrijk is om *lange-termijn* studies te doen. Wij argumenteren daarom dat onderzoek aangaande uitleggingen *ook* gedaan moet worden in lange-term studies. We rapporteren hier een experiment dat het effect van cognitieve, affectieve, en geen uitleggingen test op de motivatie van kinderen om een e-health support systeem te gebruiken. Kinderen (leeftijd 6-14) met diabetes type 1 hebben 2.5 tot 3 maanden geïnteracteerd met ons systeem. Kinderen alterneerden tussen de drie condities. De gedragingen van de artificiële agent die werden verklaard waren: 1) waarom de agent een specifieke quizvraag stelt; 2) waarom de agent een specifieke tip (een korte informatieve instructie aangaande diabetes) geeft; en 3) waarom de agent een taakvoorstel doet zoals bijvoorbeeld een quiz spelen, of een video over diabetes kijken. De motivatie van de kinderen werd gemeten door te tellen hoe vaak kinderen de taaksuggestie opvolgen, hoe lang ze de quiz blijven spelen dan wel volgende 'tips' vragen, en hoe vaak kinderen zelfstandig om een uitleg vragen. Tegen de verwachting in volgde kinderen taaksuggesties vaker op wanneer er geen uitleg was gegeven. We vonden geen verdere effecten van uitleggingen. Dit is niet in lijn met literatuur aangaande gerelateerd werk en pedagogie en dit dient als een belangrijke les voor het ontwikkelen van uitleggingen in lange-termijnsinteractie. Dit is bij ons weten de eerste lange-termijn studie die een empirisch bewijs opvoert dat uitleggingen inderdaad een effect hebben op de interactie. Het is nu aan toekomstige studies om te achterhalen wat het onderliggende mechanisme is.

Het werk in deze thesis toont aan dat zelf-verklaringen van robots inderdaad meer aspecten moeten meenemen van hoe mensen onderling gedrag verklaren. (1) Uitleggingen moeten de ontvanger van de uitleg in beschouwing nemen. Context zoals gebruikers type is van belang. (2) we tonen dat mensen emoties gebruiken bij het verklaren van robot gedrag. Toekomstige studies moeten analyseren hoe per-

sonalisatie en emoties vertrouwen in het systeem beïnvloeden. Verder toont ons zesde hoofdstuk dat uitleggingen een effect hebben op motivatie in lange-termijn interactie. Deze effecten waren alleen niet zoals verwacht gegeven literatuur op het onderwerp. Wat toont dat meer werk nodig is in dit gebied. In deze thesis hebben we verklaringen ontworpen en getest in een 'real-world' systeem in een zwaarwegend domein (kinderen met diabetes type 1 helpen om zelfstandig met hun ziekte om te kunnen gaan). Ons onderzoek toont dat het mogelijk is om onderzoeksvragen te adresseren in complexe domeinen, met een relatief kleine groep gebruikers, en over een lange interactie periode. We concluderen dat onderzoek naar zelf-verklarende artificiële agenten, zowel in de sociale wetenschappen als in mens-computer interactie, moet kijken naar individuele voorkeuren en moet kijken naar het gebruik van emoties als onderdeel van uitleggingen bij het verklaren van robot of avatar gedrag.

1

Introduction

1

1.1. Self-Explanations by Robots

Transparency of intelligent systems helps users to assess whether to trust decisions or actions of the system, to prevent misuse, and to increase motivation to use the system. Social robots are complex intelligent systems that in the near future will operate in domains including healthcare and education where trust in the system, understanding of the system, motivation to use the system and misuse of the system are important issues [2, 3]. As a result, transparency of robot behaviour is getting increasing attention [4].

EXplainable Artificial Intelligence (XAI) is a field that studies developing comprehensive and trustworthy systems [4–8]. This is studied by explaining the Artificial Intelligence (AI) algorithms themselves (a pressing topic also in the machine learning community [9]), by focusing on the human computer interaction, and analysing explanations in human communication [10]. In the present work, our main focus is on humanoid robots and avatars thereof that *self-explain* why they do the things they do.

Self-explanations of these robots are typically based on how humans amongst each other explain behaviour [4]. Humans typically explain behaviour based on the person's beliefs, desires, and emotions that caused the person to choose to act [11, 12]. Furthermore, human intuitively decide what beliefs, desires, and emotions to communicate in a particular situation and to a particular receiver. In contrast, work on self-explanations by robots has mainly focused on referring to desires and sometimes beliefs, and in a non-personalised manner. It seems there are aspects of how humans explain behaviour that are so far not thoroughly considered for designing robot self-explanations. The main question addressed in this thesis therefore is: 'Which aspects of human behaviour explanation can be used in the construction of social humanoid robot self-explanations and how should we generate such explanations?'. Where for *human behaviour explanations*, we consider both how humans explain their own behaviour as how humans explain someone else's behaviour. In particular, we focus on two aspects of this question: 1) attuning explanations to the receiver of the explanation; and 2) using emotions as content of the explanations.

In this introduction, we first discuss how humans explain behaviour amongst each other, i.e., folk psychology. Folk psychology is the most commonly used framework underpinning robot self-explanations [4] and also the framework we adopt for identifying and generating types of explanations in this thesis. Second, we discuss related work in XAI. Then, we aim to formulate a definition for what we mean with an explanation of agent behaviour, which we will use throughout the thesis. Finally, we discuss the thesis outline and research questions addressed.

1.2. Background: Explanations and Folk Psychology

People explain their behaviour to find meaning and to manage interactions [13]. When someone observes behaviour and attempts to explain that behaviour, the

Parts of this chapter have been published in [1]

observer might take the *intentional stance*. Which means the observer makes the assumption that the agent intended the action and rationally *chose* to do it [14]. Resulting explanations are then based on *folk psychology* [14–16]. In this section, we discuss the concept of folk psychology and how it relates to explanations.

Churchland [15] divides folk psychology in two classes: 1) fully intentional concepts like beliefs and desires; and 2) quasi-intentional concepts like, e.g., emotions, hunger and pain. He mentions that these quasi-intentional concepts regularly support simple explanations, of a more causal character (e.g., I was trembling because I was scared).

Malle [13, 16] calls fully intentional concepts *reasons*, and identifies a third type of reasons (besides beliefs and desires), which he calls *valuings*. In his own words: "Valuings directly indicate the positive or negative affect towards the action or its outcome" (p. 94 [13]). Examples of valuings are: like, enjoy, fear, or thrilling (one might recognize these as emotions, moods, and attitudes). Valuings are not beliefs (one can not have a false valuing), nor are they desires (desires are always directed at unachieved states, valuings can also be directed at already achieved states, e.g., one can value having a roof over ones head). Valuings combine features of both beliefs and desires, but can be subsumed under neither [13].

Döring [17] states that beliefs and desires are often unsatisfying when explaining an action; emotions are required. She divides actions in two subtypes, expressive actions (e.g., kicking a chair at home because you are angry about something that happened at work) and rational actions (e.g., crossing the street to get to the other side).

Expressive actions often require emotions for satisfactory explanations. Kicking the chair is intelligible by explaining you were angry. However, rational actions can also *require* emotions to satisfactory explain (rationalize) the action [17]. For example, quickly crossing the street can be explained by mentioning that you were scared of a dangerous looking person that was staring at you.

When provided in a social setting, emotions and motivations increase the acceptance of human action explanations [18]. They make actions more intelligible because they explain underlying values of the agent [19]. Humans use emotions to communicate intentions [20]. Emotions are an integral part of folk psychology.

1.3. Related Work

EXplainable Artificial Intelligence (XAI) is a sub-field of human agent interaction. It has its roots in Artificial Intelligence (AI), human-computer interaction (HCI), and the social sciences [10]. Much knowledge has already been accumulated with the study of expert systems [21]. From there, we can already consistently find that explanations are vital for acceptance and trust in the system's decisions, particularly in domains where decisions are judgemental and consequential (e.g., health-care) [22, 23]. Results that were later again verified by studies involving more modern intelligent systems [2, 24, 25]. In the present age, it has again become a pressing topic for the human-agent interaction community [4] and for the machine learning community [9]. This direction is further strengthened by political and societal awareness, for example, shown by the appearance of the new General Data Regu-

lation Law (GDPR) which underlines that users have the right to explanation when they are subjected to automatic decision-making [26]. In this thesis, we focus on agent self-explanations in human-agent interaction.

Current work in EXplainable AI (XAI) typically focuses on giving users some notion of the AI's reasoning in a reduced complexity form. Common approaches in human-agent interaction are to query a system's reasoning process [18, 27]. That information is then presented to the user. Most approaches applied to intelligent agents focus on the use of cognitive constructs such as beliefs, desires, intentions and goals. Which naturally links to the reasoning and decision making of the intelligent agents since this is often implemented using a BDI (belief-desire-intention)-based structure [7, 27]. These constructs are used to explain the actions of the agent in natural language [7, 11, 28–30].

In many AI applications involving intelligent agents, users require insight into the motivations behind a system's decisions [2, 31]. For example, in scenario-based training (e.g. disaster or military training), the agents in the training should be able to explain the rationale for their actions so that students can understand why the training unfolds as it does [6]. In tutor and pedagogical systems, natural dialog between the user and system has been shown to increase the training effect of such systems [32]. Debugging tools for BDI agent programs might benefit from a natural way of interaction involving asking why agents perform certain actions instead of looking at execution traces and internal mental states [33]. In human-agent teamwork [34, 35], explanations help to inform the other about the relevant individual and shared goals and intentions so that actions can be coordinated properly. In gaming and interactive storytelling [36, 37], having mechanisms to generate explanations of agent actions (the "story") could enhance the flexibility and appeal of the storyline.

XAI systems often use question lists, allowing the user a limited set of questions to ask [6, 8]. Such a question list then contains different types of questions. Simpler questions that require short factual answers, but also more nuanced questions that aim to find underlying motives of an AI system's decisions. Another approach focusses on the generation of explanations from beliefs and desires [7, 38]. One should then take special care in designing the reasoning of the agent [11]. If a good design is in place, then the XAI system can automatically choose the best explanation, based on the structure of the agent design, and characteristics of the user [11, 29].

1.3.1. Emotions Simulation for Intelligent Agents

In this thesis, we discuss robot self-explanations. One element of this is using emotions as content of the explanations. If the robot must use emotions in the explanations then it must be able to represent and generate them. Here, we briefly introduce emotion simulation for intelligent artificial agents.

Intelligent agents can simulate emotions via a computational model of emotion. A computational model of emotion describes the eliciting conditions for emotions, often including corresponding intensity of the emotions. They are typically based on cognitive theories of emotion [39]. A cognitive theory suggests that your emo-

tions are the result of thoughts and mental activity. For example, seeing emotions as consequences of cognitive evaluations (*appraisals*), relating the event to an individual's desires. For example, one is happy because one believes something to be true, and desires this to be true. Such models can be used in intelligent agent simulation to allow the agent to simulate and express emotions [39–43].

1.4. Definitions and Terminology

Here we provide definitions for the concepts used in this work. We are *explaining* the *behaviour* of *intelligent agents*. For *intelligent agent* we adopt the definition of Russel and Norvig [44].

Definition 1. (Intelligent Agent)
An intelligent agent is an entity that perceives its environment through sensors, autonomously decides how to act upon that environment, and then does so using actuators.

An intelligent agent, in our work, is embodied as a humanoid robot or a virtual avatar thereof. It chooses its actions based on its *mental state*, and updates its mental state based on what it perceives. A mental state can consist of, e.g., beliefs, desires, and emotions. An *event* is anything that happens in, and changes the state of the environment that the agent is situated in. An event can influence the agent's mental state when *perceived* by the agent's sensors. An action is a special type of *event*, directly caused by an agent by means of its actuators. If the action is performed by the agent itself, then it can *perceive* this by simply monitoring its own decision making.

When we talk about agent *behaviour*, then we mean one or more agent actions and/or reasons. Where a *reason* is a single belief, desire, or emotion, present in the mental state of the agent. We can now provide a definition for *explanation of agent behaviour* in our context.

Definition 2. (Explanation of Agent Behaviour)
Any number of reasons and events (but at least one of either) formulated in natural language, with the aim of communicating the agent's underlying intentions.

1.5. The PAL project

The context of our thesis is the PAL (a Personal Assistant for a Healthy Lifestyle) project. The PAL project helps children (aged 7-14) to cope with Type 1 Diabetes Mellitus (T1DM). The amount of children suffering from type 1 diabetes mellitus (T1DM) has doubled in less than 20 years. The growing burden of chronic illnesses on health and health-care has led to health policy responses increasingly referring to self-management. Becoming self-manageable requires long-term motivation for change. Which is especially difficult when the patient is a child.

There are several challenges. The child needs to learn to deal with medical issues like the proper use of an insulin pump, or eating regularly, but also with psychological issues like feeling different from one's classmates. The caregivers and

Figure 1.1: An example of an explanation given by a (mobile) avatar of the NAO robot in an application that children can play at home.

parents cannot always be there to help the child and will always have a different relationship with the child than that of a peer.

In the PAL project, there is human-robot interaction in hospitals and camps with scientists present, and continued long-term interaction with the children at home. We developed an elaborate system to educate the child on- and support the child with his/her diabetes whilst continuing to be a peer of the child (a *pal*). The system consists of a social robot, its (mobile) avatar, an expandable set of (mobile) health applications (diabetes diary, educational quizzes, sorting games, etc.) for interaction with the children. Additionally, there is a monitor app that allows parents to oversee the child's progress and a control app that allows caregivers to oversee and adjust how the system is configured.

In this complex AI system it is vital that the users understand and trust the system. For example, if the application keeps asking questions about hypos to the child, then it should be able to explain its underlying motivations. E.g., the system might explain that its aim is to educate the child, and it beliefs that playing a quiz about hypos is currently the best way to do so; or, the system might say it hopes that the child will increase its knowledge on hypos by answering quiz questions about hypos. We are developing an XAI module capable of generating such explanations.

1.6. Research Questions And Thesis Structure

Our main research question is:

Main Research Question
Which aspects of human behaviour explanation can be used in the construction of social humanoid robot self-explanations and how should we generate such explanations?

We focused on two aspects of this question: 1) attuning explanations to the receiver; and 2) using emotions in the explanations. We derived five research questions from this main question and addressed these in the respective chapters.

Before we study explanations themselves, in chapter 2, we specify the type of system and interaction that we are designing self-explanations for. We focus on social humanoid robots that interact with their users over prolonged periods of time. Challenges were reaching long-term, personalised interaction, for different groups of users, in complex consequential and real-world application domains. This system is used to support the children with their diabetes in the PAL-project. For designing and implementing this system, we addressed the following question:

Research Question 1; Chapter 2
What are the design principles for a social robot system that must autonomously run for several months?

In chapter 3, we work towards attuning explanations to the receiver of the explanation. Two common explanations styles in folk psychology are goal-based and belief-based explanations [13–16]. However, explanations based on folk psychology change as humans mature [13, 18]. For example, young children (4 years old) have trouble realising someone may have a belief that is false [45]. Second, children and adults alike are inclined to belief that others have similar beliefs and knowledge as they do [18]. However, adults have accumulated a vast amount of knowledge to which they can link new information [46]. Third, adults strongly desire (more than children) to know the goals you are pursuing when educating them [46, 47]. Our second research question is:

Research Question 2; Chapter 3
What are the differences in preference for goal-based versus belief-based social robot explanations between adults and children?

In the previous question, we address an important element of making explanations more attuned to the end-users. However, is still very much in line with traditional work in XAI which primarily focuses on beliefs, goals, and desires for explanations [4]. However, our discussion of the literature pointed out that emotions might play a role as well for explaining robot behaviour. In chapter 4, we study human explanations of robot behaviour and whether humans use emotions when explaining robot behaviour. Self-explanations and other person explanations are both typically based on folk psychology [13]. If people use emotions when explaining robot behaviour themselves, then this is a strong indicator that robot self-explanations benefit from the use of emotions as well. Our third research question is:

Research Question 3; Chapter 4
To what extent and in what way do humans use emotions in their explanations of robot behaviour?

Addressing this question, we found that people indeed use emotions in their explanations of robot behaviour. This is strong motivation to model emotions for the robot's behaviour and explanations thereof. First we must model the emotions themselves. Our social robot system uses a BDI-based agent programming for its high-level decision making. In chapter 5, we address the following question:

Research Question 4; Chapter 5
How can we incorporate emotion theory into BDI-based agent programming?

Finally, we argued that emotions may play a role in robot self-explanations. From literature, we found that humans often use emotions in their explanations [15, 17]. They increase the acceptance of explanations [18]. Citing only beliefs and desires in action-explanations is often insufficient, emotions can be *required* for constructing an explanation that is perceived as satisfying by the receiver of the explanation [17]. In addition, our own work concerning research question 2 shows people themselves indeed use emotions when explaining robot behaviour. In chapter 6, we address the following question:

Research Question 5; Chapter 6
What are the effects of cognitive and affective explanations on motivation to use a social robot/ avatar system during long-term interaction?

Finally in chapter 7, we present overall conclusions. We discuss the limitations of our work and potential directions for future continued work. Finally, we discuss some more general contributions from the thesis as a whole.

References

[1] F. Kaptein, J. Broekens, K. V. Hindriks, and M. Neerincx, *Caaf: A cognitive affective agent programming framework,* in *Intelligent Virtual Agents* (2016) pp. 317–330.

[2] S. R. Haynes, M. A. Cohen, and F. E. Ritter, *Designs for explaining intelligent agents,* International Journal of Human-Computer Studies **67**, 90 (2009).

[3] I. Leite, C. Martinho, and A. Paiva, *Social robots for long-term interaction: a survey,* International Journal of Social Robotics **5**, 291 (2013).

[4] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, *Explainable agents and robots: Results from a systematic literature review,* in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*

(International Foundation for Autonomous Agents and Multiagent Systems, 2019) pp. 1078–1088.

[5] M. Van Lent, W. Fisher, and M. Mancuso, *An explainable artificial intelligence system for small-unit tactical behavior,* in *National Conference on Artificial Intelligence* (2004) pp. 900–907.

[6] M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, *Building explainable artificial intelligence systems,* in *Innovative Applications of Artificial Intelligence* (2006) pp. 1766–1773.

[7] J. Broekens, M. Harbers, K. Hindriks, K. Van Den Bosch, C. Jonker, and J.-J. Meyer, *Do you get it? user-evaluated explainable bdi agents,* in *Multiagent System Technologies* (Springer, 2010) pp. 28–39.

[8] G. Taylor, K. Knudsen, and L. S. Holt, *Explaining agent behavior,* in *Behavior Representation in Modeling and Simulation* (2006).

[9] O. Biran and C. Cotton, *Explanation and justification in machine learning: A survey,* in *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8 (2017) p. 1.

[10] T. Miller, *Explanation in artificial intelligence: Insights from the social sciences,* Artificial Intelligence (2018).

[11] M. Harbers, J. Broekens, K. Van Den Bosch, and J.-J. Meyer, *Guidelines for developing explainable cognitive models,* in *International Conference on Cognitive Modeling* (2010) pp. 85–90.

[12] M. De Graaf and B. Malle, *People's explanations of robot behavior subtly reveal mental state inferences.* in *Human-Robot Interaction (HRI), 2019 11th ACM/IEEE International Conference on, in press* (ACM, 2019).

[13] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction.* (MIT Press, 2004).

[14] D. C. Dennett, *Three kinds of intentional psychology,* in *Reduction, Time and Reality*, edited by R. Healey (Cambridge University Press, Cambridge, 1981) pp. 37–61.

[15] P. M. Churchland, *Folk psychology and the explanation of human behavior,* The future of folk psychology: Intentionality and cognitive science , 51 (1991).

[16] B. F. Malle, *How people explain behavior: A new theoretical framework,* Personality and social psychology review **3**, 23 (1999).

[17] S. A. Döring, *Explaining action by emotion,* The Philosophical Quarterly **53**, 214 (2003).

[18] F. C. Keil, *Explanation and understanding,* Annual Review of Psychology **57**, 227 (2006).

[19] C. Tappolet, *Emotions and the intelligibility of akratic action,* in *Weakness of Will and Practical Irrationality*, edited by S. Stroud and C. Tappolet (Oxford: Clarendon Press, 2003) pp. 97–120.

[20] E. Hudlicka, *To feel or not to feel: The role of affect in human–computer interaction,* International journal of human-computer studies **59**, 1 (2003).

[21] W. Swartout, C. Paris, and J. Moore, *Explanations in knowledge systems: Design for explainable expert systems,* IEEE Expert **6**, 58 (1991).

[22] B. M. Muir, *Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems,* Ergonomics **37**, 1905 (1994).

[23] L. R. Ye and P. E. Johnson, *The impact of explanation facilities on user acceptance of expert systems advice,* Mis Quarterly , 157 (1995).

[24] J. D. Lee and K. A. See, *Trust in automation: Designing for appropriate reliance,* Human factors **46**, 50 (2004).

[25] B. Y. Lim, A. K. Dey, and D. Avrahami, *Why and why not explanations improve the intelligibility of context-aware intelligent systems,* in *Human Factors in Computing Systems* (2009) pp. 2119–2128.

[26] P. Carey, *Data protection: a practical guide to UK and EU law* (Oxford University Press, Inc., 2018).

[27] K. V. Hindriks, *Debugging is explaining,* in *International Conference on Principles and Practice of Multi-Agent Systems* (Springer, 2012) pp. 31–45.

[28] M. Harbers, K. van den Bosch, and J.-J. C. Meyer, *A study into preferred explanations of virtual agent behavior,* in *International Workshop on Intelligent Virtual Agents* (Springer, 2009) pp. 132–145.

[29] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, *Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults,* in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on* (IEEE, 2017) pp. 676–682.

[30] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, *Enabling robots to communicate their objectives,* Autonomous Robots , 1 (2017).

[31] H. C. Lane, M. G. Core, M. Van Lent, S. Solomon, and D. Gomboc, *Explainable artificial intelligence for training and tutoring*, Tech. Rep. (DTIC Document, 2005).

[32] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, *Autotutor: an intelligent tutoring system with mixed-initiative dialogue,* IEEE Transactions on Education **48**, 612 (2005).

[33] J. Broekens and D. DeGroot, *Formalizing cognitive appraisal: from theory to computation,* in *Cybernetics and Systems* (Vienna, 2006) pp. 595–600.

[34] M. Harbers, C. Jonker, and B. Van Riemsdijk, *Enhancing team performance through effective communication,* in *Proceedings of the 4th Annual Human-Agent-Robot Teamwork Workshop* (2012) pp. 1–2.

[35] F. Flemisch, D. Abbink, M. Itoh, M.-P. Pacaux-Lemoine, and G. Weßel, *Shared control is the sharp end of cooperation: Towards a common framework of joint action, shared control and human machine cooperation,* IFAC-PapersOnLine **49**, 72 (2016).

[36] M. Cavazza, F. Charles, and S. J. Mead, *Character-based interactive storytelling,* IEEE Intelligent Systems **17**, 17 (2002).

[37] M. Theune, S. Faas, D. K. J. Heylen, and A. Nijholt, *The virtual storyteller: Story creation by intelligent agents,* in *Technologies for Interactive Digital Storytelling and Entertainment* (2003) pp. 204–215.

[38] M. Harbers, K. Van den Bosch, and J.-J. Meyer, *Design and evaluation of explainable bdi agents,* in *Web Intelligence and Intelligent Agent Technology* (2010) pp. 125–132.

[39] S. Marsella, J. Gratch, and P. Petta, *Computational models of emotion,* A Blueprint for Affective Computing-A sourcebook and manual **11**, 21 (2010).

[40] W. S. Reilly, *Believable Social and Emotional Agents.*, Tech. Rep. (DTIC Document, 1996).

[41] M. S. El-Nasr, J. Yen, and T. R. Ioerger, *Flame—fuzzy logic adaptive model of emotions,* in *Autonomous Agents and Multi-agent systems* (Springer, 2000) pp. 219–257.

[42] A. Popescu, J. Broekens, and M. van Someren, *Gamygdala: An emotion engine for games,* IEEE Transactions on Affective Computing **5**, 32 (2014).

[43] J. Dias, S. Mascarenhas, and A. Paiva, *Fatima modular: Towards an agent architecture with a generic appraisal framework,* in *Emotion Modeling* (Springer, 2014) pp. 44–56.

[44] S. Russell, P. Norvig, and A. Intelligence, *A modern approach,* Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs **25**, 27 (1995).

[45] H. W. H. Mayringer, *False belief understanding in young children: Explanations do not develop before predictions,* International Journal of Behavioral Development **22**, 403 (1998).

[46] S. Lieb and J. Goodlad, *Principles of adult learning,* (2005).

[47] M. S. Knowles *et al.*, *The modern practice of adult education*, Vol. 41 (New York Association Press New York, 1970).

2

Cloud-based Social Robots for Health Education & Care

2

Making the transition to long-term interaction with social-robot systems has been identified as one of the main challenges in human-robot interaction. This paper identifies four design principles to address this challenge and applies them in a real-world implementation: cloud-based robot control, a modular design, one common knowledge base for all applications, and hybrid artificial intelligence for decision making and reasoning. The control architecture for this robot includes a common Knowledge-Base (ontologies), Data-Base, Hybrid Artificial Brain (dialogue manager, action selection and explainable AI), Activities Centre (Timeline, Quiz, Break & Sort, Memory, Tip of the Day, ...), Embodied Conversational Agent (ECA; i.e., robot and avatar), and Dashboards (for authoring and monitoring the interaction). Further, the ECA is integrated with an expandable set of (mobile) health applications. The resulting system is a Personal Assistant for a healthy Lifestyle (PAL) which supports diabetic children with self-management and educates them on health-related issues (48 children, aged 6-14, recruited via hospitals in the Netherlands and in Italy). It is capable of autonomous interaction 'in the wild' for prolonged periods of time without the need for a 'Wizard-of-Oz' (up until 6 months online). PAL is an exemplary system that provides personalised, stable and diverse, long-term human-robot interaction.

This paper is submitted to ACM Transactions on Human-Robot Interaction (THRI). The author of this thesis is the main author of this chapter and the corresponding submitted paper. However, in a number of sections different co-authors have taken the lead.

Specifically, different co-authors have taken the lead in writing sections: 2.4.1 - 2.4.4, 2.4.6, 2.4.7 except for the part about explainable AI, and section 2.7. In these sections, different co-authors were also the main developers of the corresponding software design and implementations.

The full list of authors for the corresponding submitted paper is: FRANK KAPTEIN (Delft University of Technology, Netherlands), BERND KIEFER (Deutsches Forschungszentrum für Künstliche Intelligenz, Germany), ANTOINE CULLY (Imperial College London, United Kingdom), OYA CELIKTUTAN (King's College London, United Kingdom), BERT BIERMAN (Produxi, Netherlands), RIFCA PETERS (Delft University of Technology, Netherlands), JOOST BROEKENS (Leiden University, Netherlands), WILLEKE VAN VUGHT (TNO, Netherlands), MICHAEL VAN BEKKUM (TNO, Netherlands), YIANNIS DEMIRIS (Imperial College London, United Kingdom), MARK A. NEERINCX (Delft University of Technology, Netherlands).

2.1. Introduction

There is an increasing interest in long-term human-robot interaction. Social robots are commonly applied to education, health-care, public spaces, work environments, and home environments [1]. These systems often need to interact with several users and user groups at the same time and require interaction over prolonged periods of time in order to achieve their individual goals [1].

Current social robot systems have their own specific value, but remain simple and scripted in nature and miss the required comprehensive, prolonged, and personalised support. For example, in EU project ALIZ-E (www.aliz-e.org) a social robot was developed for children to support them in the self-management of diabetes [2, 3]. However, much of the implemented functionality remained scripted and required a 'Wizard of Oz'. Furthermore, the children interacted with the robot 'only' in a limited number of subsequent sessions [4, 5].

To establish long-duration pervasive human-robot interaction, our approach is to develop a personalised social-robot *with* its avatar that allows the user to always and anywhere engage in a divers set of activities over a prolonged period of time (cf. [6]). We propose four principles for the implementation of such a system. It must: (1) have a connection to the cloud to delegate parts of the computational problems to external computers; (2) be modular to support parallel and incremental development of functionality; (3) have a common knowledge-base and vocabulary in the different parts of the system and for the human-agent interaction; and (4) have hybrid artificial intelligence solutions (e.g., agent-based and machine learning) that all have their own contribution to the problem. We discuss these principles separately and we discuss how they were incorporated in the system's development.

This paper presents the PAL system (a Personal Assistant for a healthy Lifestyle), an exemplary system of human-robot interaction that enables long-term support for health education and care. The robot autonomously interacts with children (aged 6-14) 'in the wild' over a period of several months. The PAL system is a fully integrated and autonomous system that interacts with the children, their parents, and the health-care professionals for prolonged periods of time. It is composed of a social robot, its (mobile) avatar, and an expandable set of (mobile) health applications (diabetes diary, educational quizzes, sorting games, etc.). The system allows for adaptation to the patient's condition and activities on the fly. It ran robustly during the duration of the experiment, i.e., more than half a year (single users interacted for 2.5 to 3 months but started at different moments)

First, we discuss the related work in human-robot long-term interaction systems and discuss the context of our work in section 2.2. Then, we discuss and motivate the four design principles of our system in section 2.3. The system architecture, as well as how the principles led to certain decisions is described in section 2.4. We also describe the *process* of development and testing (i.e., decisions we made to streamline development in such a large scale project with several project partners) in section 2.5. We analyse the performance of our system (usage statistics and stability) in section 2.6. Finally, we discuss and conclude upon our efforts in sections 2.8 and 2.9.

2.2. Related Work and Context

This section first describes the state-of-the-art for robot systems in long-term interaction. Then we discuss our context, the PAL (a Personal Assistant for a healthy Lifestyle) project. We argue that state-of-the-art robot systems all have their own specific value, but miss the required prolonged, comprehensive, and personalised support to successfully apply a robot system in health-care & education. Finally, we present important technical requirements for such a system.

2.2.1. Related Work

Leite et. al. [1] surveyed existing social robot systems, identifying four domains for such systems: Health Care, Education, Work Environments and Public Spaces, and Home. It is possible for a system to fit in multiple domains simultaneously, for example, a robot might have a health support function as well as a health education function. Only in more recent work, social robot systems have been investigated in long-term studies (in EU projects like PAL, ALIZ-E, LIREC, L2TOR, UPA4SAR, WYSIWYD and PATRICIA). This is because long-term interaction requires a degree of robustness, versatility, and autonomy. Something that technology only more recently is starting to provide.

Animal-like companion robots such as Pleo [7, 8], Paro [9, 10], and the AIBO robotic dog of SONY [11] have been used for some time in health-care and show potential with respect to treatment [12] and in maintaining adherence during prolonged interaction [13]. Such robots can provide comfort to their (elderly) users [9, 10], and develop social skills of the users (autistic children, 4-12 years old) [8]. However, such systems are limited in the richness and personalisation of the interaction because they lack dialogue capabilities and direct educational functions.

Humanoids may have a harder time in maintaining long-term interaction with users. The embodiment of a robot influences the expectations we have of the robot's capabilities [14]. For example, we might expect a humanoid robot to communicate using natural language. Managing those expectations is challenging when attempting to maintain interaction with (especially child) users [15].

Several long-term studies have taken place where a social robot attempts to educate and/or support a user's health [3, 16, 17]. To maintain a prolonged interaction with a robot it becomes vital that the robot truly has added value compared to the other technology available to the users, i.e., the robot must by functionally-relevant [18], or provide unique experiences to the user [19]. This seems quite challenging when the robot is applied to health-care and/or education. Still, a well designed robotic system can help in the execution of educational and health-related tasks [3, 20].

2.2.2. Context: a Personal Assistant for a Healthy Lifestyle

The context of our system is the PAL project. The PAL project helps children (48 children, aged 6-14, recruited via hospitals in the Netherlands and in Italy) to cope with Type 1 Diabetes Mellitus (T1DM). The amount of children suffering from T1DM has doubled in less than 20 years. The growing burden of chronic illnesses on

health and health-care has led to health policy responses increasingly referring to self-management. Becoming self-manageable requires long-term motivation for change, which is especially difficult when the patient is a child.

There are several challenges. The child needs to learn to deal with medical issues like the proper use of an insulin pump, or eating regularly, but also with psychological issues like feeling different from one's classmates. The caregivers and parents cannot always be there to support the child and their relationship will always be different than that of a peer.

In the PAL project, we developed a system to educate the child on- and support the child with his/her diabetes whilst continuing to be a peer of the child (a *pal*). The system consists of a social robot, its (mobile) avatar, an expandable set of (mobile) health applications (diabetes diary, educational quizzes, sorting games, etc.) for interaction with the children. Additionally, there is a monitoring dashboard that allows parents to oversee the child's progress and an authoring tool that allows caregivers to oversee and adjust how the system is configured. For example, a caregiver could increase the difficulty for a certain *learning goal* when the child shows good progress, or select a new goal to work on altogether.

There is interaction in hospitals and camps with scientists present, and continued long-term interaction at the children's homes. In the hospitals and camps, the PAL agent is a NAO robot. In the home interactions, there is an avatar impersonating the robot on a tablet screen. During all interactions, the PAL agent makes decisions and proposes activities to the child. It makes these proposals based on the configuration and progress of the child's personal learning goals.

In the PAL project we have both a robot and an avatar as possible embodiments of the PAL agent. Robots have been shown to have a positive impact on motivation and learning [21]. For example, the NAO robot developed by Softbank (formerly Aldebaran) has already been used successfully in ALIZ-E, where children learn and are supported by the (robot-based) health-care system [2, 3]. However, a pragmatic problem with any sufficiently advanced humanoid of good quality is that it is an expensive device. This means that it is not feasible to provide a large group of users with their own personal robot. However to make developing content for an interactive robot attractive, a large user base is necessary. Perhaps this is a problem that will become less relevant in the future if humanoids become more affordable. Still, it may be a long time before owning a robot is as common as owning a car. Meanwhile virtual avatars are needed to support the development of human-robot interaction. In our context, the children can have an avatar of the robot on the tablet where the mobile health applications are installed. In the hospitals and camps, the children can interact with the physical robot.

2.3. Principles for a Social Robot System for Long-term Interaction

In this section, we provide four main principles for developing a personalised long-term social robot system. Our vision is that such a system 1) should have a cloud-based implementation to distribute heavy computations and allow real-time adap-

tation of the system's functionality; 2) should be developed in a modular way to facilitate parallel development; 3) must contain a common knowledge-base and terminology for the different project partners, the different parts of the system, and the human-agent interaction; and 4) have hybrid artificial intelligence solutions (e.g., agent-based and machine learning) to contribute to the different (sub-) parts of the complexity. We discuss these four principles separately.

2.3.1. Principle 1: Cloud-based Robots

The first principle for a social robot system in long-term interaction is that the system should be *cloud-based*.

Cloud-based computing offers several advantages over stand-alone robot systems [22]. It allows the use of (1) *external libraries* for machine learning approaches to, for example, generate sentences for dialogue. It enables using (2) *external computers* to delegate complex computational tasks, e.g., a statistical analyses of previous behaviours and their outcomes. Enables the (3) *sharing of data and outcomes* of behaviours amongst different robots. So, when one robot learns that playing the quiz is a great way of teaching children to count carbohydrates, then it can share this knowledge with the other robots. Finally, Kehoe et. al. [22] also mention that cloud robots enable (4) *Human Computation*, i.e., using crowd-sourcing for analysing, e.g., images and error recovery. However, we have not investigated this in our context since we strove for a more autonomous system.

In addition to the advantages of cloud-based robots as stated in [22], we would specifically state that it facilitates (5) *Personalisation and Adaptability* of the system. The different users of the robot system can adapt parameters online and thereby steer the robot's behaviour in desired directions. In this way, the human expert (the health-care professional within the context of PAL) can personalise the robot to the specific patient. Finally, (6) *integration with internet services* has sparked interest in the development of social robot systems. For example recently (in 2018), AIBO was relaunched with improved artificial intelligence. It uses cloud-based techniques to apply deep learning for its reasoning and to develop a unique personality, depending on the behaviour of the owner. Another example is the ALEXA chatbot, which can be seen as a object shaped robot. ALEXA's main functionality is to easily provide internet services to the users.

There are risks associated with cloud-based computing that involved data security and privacy. However, cloud-based computing as a system design principle does not automatically exclude usage of this principle in health-care or education. For example in the PAL project, the servers doing the computation were managed by the hospitals or university and one can easily envision dedicated servers for cloud-based health applications with sophisticated data security and privacy management.

2.3.2. Principle 2: Modular System

The second principle is that the system must be *modular*. Handling complexity in software development is facilitated by developing (nearly) modular components that are responsible for providing particular aspect of such a system [23]. Different

techniques are built by software developers from different organisations at different locations (countries). For example, one German project partner provided a dialogue framework and another British partner an action selection framework. Setting up the architecture in a modular way allows to connect these frameworks and facilitates parallel development of them. The concerning functions need to be addressed as building blocks for intelligence, much like a 'society of mind' [24] and in line with recent virtual agent architectures [25].

2.3.3. Principle 3: Common Knowledge-base and Terminology

The third principle is that the system must have a common terminology and knowledge-base that: (1) provides an unambiguous vocabulary in communication between stakeholders; (2) supports system implementation of knowledge-based reasoning functionality; and (3) serves as a basis for interoperability in human-agent interaction.

A common way of defining a knowledge-base is by means of an *ontology*. An ontology clarifies the structure of knowledge [26]. It contains explicit, formal specifications of terms in the domain and of the relations among them [27]: it is used to represent real-world objects and concepts, and to specify properties of and relations between those objects.

2.3.4. Principle 4: Hybrid Artificial Intelligence

Finally, the fourth principle is that the system must be comprised of several artificial intelligence techniques that all have their own contribution to the interaction. For example, machine learning (ML) techniques excel at learning optimal policies when given large amounts of data. Within the context of PAL this may mean that ML can learn what activities the robot should propose to a child in order to teach the child something about a specific issue (like, measuring blood sugar levels). On the other hand, agent-based techniques allow to implement expert knowledge on a human-understandable manner. For example, when the definition of a good blood sugar level differs per hospital, then an agent system can easily change a single belief without having to change anything in the implementation and logic or having to re-train.

2.4. System Implementation for a Social Robot in Health Education & Care

Here, we provide a technical description of the architecture of the PAL system. Figure 2.1 shows an overview of the architecture. This section will discuss the sub-parts of this picture individually. First, we discuss the ontology with the knowledge and content. Second, we discuss the database. Then, we discuss the user interfaces. Finally, we discuss the brain and its individual elements.

2.4.1. The Ontology

This section specifically describes our work on developing a common knowledge-base and terminology (i.e., principle 3). The ontology represents concepts related

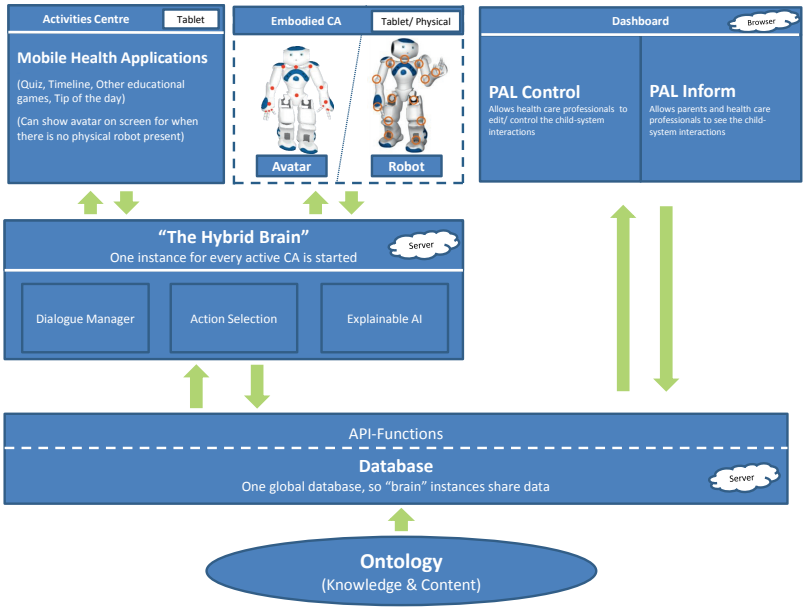


Figure 2.1: The high level architecture of the PAL system. On the top of the picture the user interfaces are shown. The child interface consists of a tablet application that connects to a physical nao robot or an avatar thereof (which is then shown on the screen of the tablet application). On this application several mobile health applications are connected, like, an educational quiz or a timeline where the child can keep track of his/ her blood sugar values, activities, and food regime. The health-care professional and parents can both see the child's progress on the learning objectives in the web-page interface. Only the health-care professional (in PAL control) can also adjust the robot's behaviour, i.e., steer the support to best fit the current state of the child's treatment. 'The brain' is responsible for making the actor's behaviour intelligent and lifelike. For every child that logs in the system, a "brain" instance is started on the cloud. This allows the complex computations for the actor (robot or avatar) to happen on a more powerful, external computer. All elements of the system connect to the common database and communicate to the database by means of API functions. The database is structured by the ontology, which is developed in cooperation with the health-care professionals.

to diabetes, actors and tasks involved in self-management, and emotions involved in human-agent interaction. It defines the definitions and relations between these concepts.

Developing such an ontology requires close collaboration with experts in the field, in this case the health-care professionals. This section gives an overview of some of the main decisions made and how they underpin the system's workings.

Ontology Frames

The entire ontology in the PAL project is constructed by integrating separate ontologies, linking them by means of a top-level ontology. These separate models function as high-level building blocks for smaller, more specific areas of interest (frames) [28].

- An ontology of human/machine roles and actors and locations involved in self-management.
- a generic ontology of tasks for actors (human and/or artificial), associated with goals and roles.
- An ontology of learning objectives, defining learning goals and tasks (activities) specific to the diabetes self-management domain, and child specific status of these learning objectives.
- Abstract ontologies that define notions of events and processes and various properties of time.
- A dialogue management ontology that contains dialogue acts and some semantic frames, based on the DIT++ taxonomy of dialogue acts (ISO standard 24617-2) and FrameNet, respectively
- An episodic memory ontology as a system responsible for capturing specific events, or episodes, in order for the PAL system to interact with a human user in a meaningful manner over prolonged periods of time.
- An ontology for storing and reasoning over the affective process and state of a child, that allows the PAL system to estimate the emotions experienced by the child.

We have reused existing ontologies to cover the various frames wherever possible. Although the frames of interest mentioned above are typically generic in nature, pre-existing models for these frames may differ (slightly) in scope and/or intention and may thus be a partial fit to the intended scope of the frame in the context of PAL. Whereas e.g. self-management activities of diabetes are a relevant topic, the entire professional medical diagnosis and treatment model of diabetes is out of scope. We have adapted some of the existing models by either extending them with additional concepts or by taking a profile (part) from the model whenever there are details/concepts in the model that are irrelevant to the scope of PAL. An example of reuse is displayed in the adoption of the well-known ontology for task world models [29] in the frame for tasks/goals and learning objectives. These objectives steer the behaviour of the robot and the treatment of the child.

PAL Objectives Model (POM) One important part of the ontology is the frame of **learning objectives** that covers most relevant aspects that the children (aged 7-14) with T1DM encounter in their daily lives and must learn to become self-manageable. These objectives steer the behaviour of the robot and the treatment of the child. Prior to (and during) the usage of the PAL system, the health-care professional, the parents, and the child together choose a subset of relevant learning objectives, based on the child's individual needs, interests and knowledge. The learning objectives consist of achievements, goals and tasks. Achievements define a set of goals that are required to enable the user to carry out an real world challenge represented by the achievement. An achievement does not define new

Figure 2.2: Child playing the Break & Sort game with the NAO robot using the integrated PAL system.

knowledge or skills but groups together goals with a similar level and related to a specific challenge to make goal setting easier. Vice versa; goals are specified that describe the end state of knowledge, skills or attitudes that a child should have to accomplish an achievement. Learning goals are hierarchically structured by difficulty level (i.e., novice to master) and level of complexity based on the Taxonomy of Educational Objectives of Bloom et. al. [30]. An example of an achievement within our context is: *'I can go to a sleepover'* which contains goals as: *'I know when to ask for help'*, *'I know what to take with me to a sleepover'*, and *'I know that I should take extra insulin when I eat extra carbohydrates'*. To attain these goals, children have to do activities (and thereby complete tasks) within the myPAL application. The robot has several tasks/activities in its database that educate the child on these subjects. For example, the child might do a quiz together with the social robot, play a memory game, or the robot may simply explain something about the subject (a *tip of the day*).

Learning objectives can be attuned to a child's developmental stage and the child's personal and environmental context. Learning objectives are labelled with a knowledge level, difficulty level and prerequisite knowledge. Additionally they can be linked to a device (pen, pump, sensor) or a hospital. In this way, the ontology facilitates intelligent personalisation of interaction and learning process which enhances motivation and learning gain [31].

2.4.2. The Database

The central data hub of the PAL system is based on an extended Resource Description Framework (RDF) storage component and reasoner (HFC) [32]. Its special features allow putting the terminological knowledge (the ontology of diabetes knowledge, the definition of user and agent model, and static knowledge for language and dialogue processing) as well as the dynamic knowledge that is produced and consumed in the running instances of the myPAL app into one data repository, thereby fostering principle 3 the need for a common knowledge-base. The database plays a pivot role for principle 1 *cloud computing*, because it helps the different computers to use the same data, have customised reasoning rules together

with a streaming reasoning approach, and allows to infer new data in real-time.

The database allows connecting different sub-ontologies (frames) by using equivalence statements. This fosters principle 2 (modularity) in the knowledge building process and facilitates re-usability of the work. Another advantage of structured databases, like RDF-based or graph databases, is the flexibility when it comes to adding or changing data structures. In general, this is much easier to achieve than for relational databases (RDBs), especially because the knowledge representation, i.e., the specification of the data, and the (dynamic) data itself are in the same format. This allows quite simple checking consistency by custom reasoning rules.

The PAL system is based on a very particular implementation for RDF storage and reasoning that allows using n-tuples instead of the usual triples. This makes it possible to directly attach time and confidence information to every data chunk in a more efficient way than with currently available RDF storage solutions [33–35]. As a consequence, the database can contain a flow of events and data that is susceptible to temporal and probabilistic reasoning.

One price to pay for this increased flexibility can be an increased resource footprint, especially when it comes to memory consumption. Every user that starts the tablet application starts a reactive system that frequently reads from and writes to the database. This, in turn, unconditionally triggered computations that need to be synchronised between the several instances. With many users simultaneously using the system, this can put a large burden on the server's CPU. We found out that many computations were unnecessarily triggered since their computation did not depend on the changed data. Therefore, HFC was extended with stream reasoning functionality that reduces computations using highly efficient filters.

Technical Description

The database itself is an enhanced version of the aforementioned HFC. HFC is an RDF in-memory storage with a forward chaining reasoning engine. The reasoner comes with predefined sets of reasoning rules for different OWL dialect, but it also permits to add custom reasoning rules for specific purposes, e.g., temporal or probabilistic reasoning. For the PAL project, a server / client communication layer, an object synchronisation layer, and a persistence layer were added.

The communication layer is based on the event-based middleware TECS. Event-based middleware developed at DFKI[1] which allows to easily extend the server API with special functionality, e.g., executing complicated or cascaded queries directly in the server and providing the result as a return value to improve performance. This is used by the modules in the PAL system and provides a clean separation between database functionality and the module specific requirements.

In addition, this layer is used to enforce access control rules, which are based on a user hierarchy specified in the ontology, and secret security tokens that are internally exchanged between the web proxy and the database when a user logs in. These tokens are also used to encrypt the data stream between the app and the server functionality, allowing for safe data exchange.

[1] http://www.dfki.de

The object synchronisation layer implements an efficient and correct exchange of data between the local user interfaces (the myPAL app and the PAL Control & Inform web application) and the central database on the cloud. Here, connected parts of the RDF storage are treated like data objects, e.g., the data for a child consists of a unique reference (a uniform resource identifier, i.e., URI) and the properties (and values) that are connected to that reference. When some (possibly embedded) value of such a data object changes, the app needs to know at least the relation down to the object reference to integrate this change into its internal program state.

To achieve that, the classes in the ontology that can be synchronised in this way are marked in the ontology, and local graph search in the ontology guarantees to deliver complete data chunks that the app can process. This process works in both directions, in the sense that the database gets complete chunks, but only stores those data bits that have actually changed.

Finally, the persistence layer guarantees that all changes to the database are efficiently reflected in background storage, such that the database can be backed up safely, and be restored to the current state in case of server shutdown or crash.

2.4.3. PAL Control & Inform

To be able to author the behaviour of the system, personalise the system towards the child and monitor how the child uses the system, we have developed a web-application with two dashboard modules and associated interfaces accessible depending on user role.

The monitoring dashboard, *palInform*, is the software module and associated interface that enables care takers to get an insight into how their children use the system, their progress, and what nutritional and medical values they fill in. It provides a timeline of the most important events, based on system activity and the data the child enters, in an aggregated manner. The monitor displays glycemic, insulin, nutritional, activity and emotion-related data for a child, as entered by the child via the timeline in the myPAL app. Further, it displays goal attainment data in relation to time. The monitor is available for both health-care professionals (HCP) and the parent(s). To ensure data is shared with parents in a way that respects the child's privacy, the child and parent can set agreements on what information will or will not be available. Agreement options are used to balance between the conflicting values of privacy and medical safety. Requirements were developed together with the hospitals in the PAL project.

The authoring dashboard, *palControl*, is a tool to enable health-care professionals to set learning objectives (i.e., achievements, goals and tasks) for children during or in between meetings. It further enables the HCP to enter child data including personal data and preferences such as sports and hobbies and whether the child uses a pen or a pump for insulin intake. Two issues were leading in its design: 1) how to formalise the learning objectives based on the medical protocols and informal expertise of HCPs and 2) how to design the interface (and mechanisms behind) to facilitate easy personal goal selection and progress monitoring during the intake meeting and further guidance of the education process ([31] and covered in the

(a) Welcome Screen (b) Task Suggestion

(c) Timeline (d) Quiz

(e) Sorting Game (f) Goal Progress Screen

Figure 2.3: Six screen-shots of the myPAL application used by the children.

ontology section 2.4.1, 2.4.1).

Both modules allow real-time changes to specified learning objectives (content & display) and personalised learning goal setting through the ontology, with the reasoning system. The modules thus most strongly link to principle 1 using cloud techniques and to principle 3 using a common knowledge-base, i.e., the modules facilitate real-time adaptation of the system's configuration and thereby its reasoning. It allows a HCP to adapt to system's functioning to the different phases of the treatment plan. The health-care professional can review the progress and tweak the systems behaviour accordingly. In line with the learning objectives discussed in section 2.4.1, this can simply mean that the health-care professional updates the set of learning objectives the child and robot should currently work on.

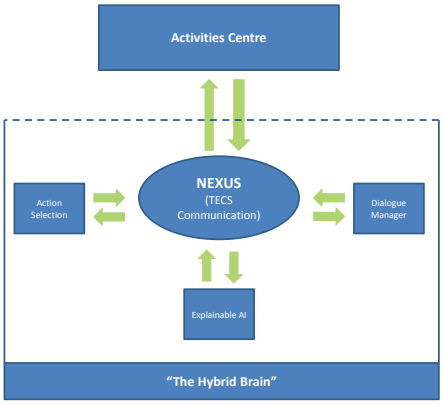


Figure 2.4: Communication between the modules in the brain and with the child interface. All messages go through a common messaging board called the *nexus*. Modules individually decide to *subscribe* to *types of messages* and can themselves send messages of a particular type. Modules in this way individually implement what to do when a particular message is sent.

2.4.4. Activity Centre

The activity centre is an application comprised of several activities that the Conversational Agent (CA) can perform with the user (shown on the top-left part of figure 2.1). In our context this is a tablet application (myPAL) connected to a physical (nao) robot or an avatar thereof. myPAL contains health-related activities that support the child's treatment and that educate the child on diabetes.

MyPAL contains several games, i.e., educational quizzes, sorting games, and memory games. The activities in the application have been set-up in a modular way (principle 2). Allowing to add new games by implementing the required interfaces. In addition to the educational games, the PAL actor can provide a 'Tip of the day', where it provides the child with some information concerning diabetes. myPAL also provides a list of videos about several diabetes-related topics and a list of real-world tasks that the child must perform his-/ herself. Finally, myPAL contains a timeline where the child can keep track of his/her blood sugar values, glycemic corrections, activities (sport and other), and food regime. All these activities are related to the child's current set of learning objectives. myPAL shows the child his/her progress on the learning objectives and provides 'task suggestions', i.e., proposes particular educational activities to work on the learning objectives in a targeted manner. Figure 2.3 shows some screen-shots of the different components of myPAL.

2.4.5. Communication Between Modules

The activity centre, conversational agent, and the separate modules in the hybrid brain all connect to a global communication platform, the 'nexus' (see figure 2.4).

These modules can all subscribe to *types of messages* and send those message types. In this way different modules can individually decide what information is relevant for them and how they should respond to new information.

For example, the child may click the quiz which is then send around on the nexus. The action selection module must respond to this message by choosing different quiz topics, the dialogue manager wants to know so that a dialogue act about starting a quiz can be initiated. The behaviour manager, on the other hand, might not need to know this at all. It will generate a movement only after the dialogue manager sends a high level 'speech-and-movement' message.

This greatly fosters modularity (principle 2) of the system. A new module can implement its own workings independently of the rest of the system. It can subscribe to messages that carry information relevant to the module. Of course, integration of the module still requires coordination amongst developers. The other modules should subscribe to new messages sent by the new module and should implement protocols on how to use this data for the human-agent interaction.

Modules can be independently maintained and improved as long as the interface contract is unchanged. The modularity also improves flexibility and reusability when requirements change. Even when the functionality of some module is extended, only modules that will profit from these enhancements have to be changed.

2.4.6. Multimodal Behaviour Manager

The robot-movement manager converts gesture specifications ('commands send over the nexus') into values about joints and times to be send to the robot or its avatar (i.e., it can communicate gesture & posture commands to both the robot and its avatar). The current framework meets two important challenges for cross-platform social human-agent interactions. First, the behaviours of the virtual and physical NAO have completely the same foundation and expression mechanisms, so that they can be perceived as really similar. Second, modulation of these gestures is possible to adapt affective (emotion, style) expressions conveyed in these behaviours.

The task of the robot-movement manager can be divided into two parts: 1, executing the multimodal utterances on the (virtual) NAO; and 2, making the (virtual) NAO appear lively.

Executing a multimodal utterance (1) is done when another module in the brain sends a specific message type over the nexus. Such a utterance consists of the name of the gestures to execute and the text to be spoken as well as an (optional) emotional (mood) modulation. From this content it constructs the necessary values to move the joints of the (virtual) NAO which results are then published to the messaging board. From the gesture name the robot-movement manager calculates the position of the joints of the (virtual) NAO and the time available for the joints to reach that position. When the execution of multimodal utterance is finished, the robot-movement manager sends this information the other modules.

Making the (virtual) NAO look lively (2) makes the interaction more engaging for the users. It is done by implementing continuous autonomous moves. These autonomous moves need to be carefully combined with the multimodal utterances

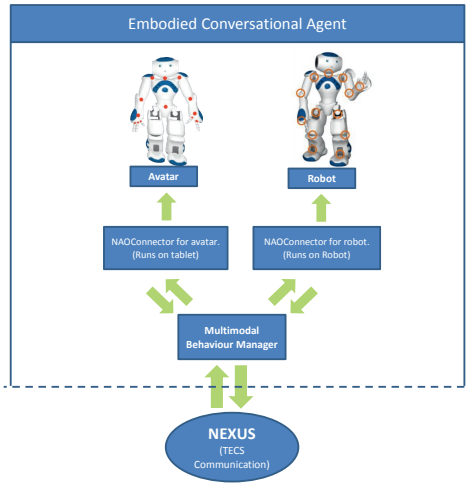


Figure 2.5: Graphical representation of relations between the robot-movement manager and the embodiments.

sent by other modules of the PAL system. Furthermore, style adaptation in autonomous move is challenging because these motions are not pre-designed and thus require real-time modulation which is less controllable and predictable.

The physical NAO The implementation for the real NAO is a wrapper which converts the data received (from the robot-movement manager) into the commands which can be send to the NAO's own execution system. The majority of the messages originate from the autonomous move system. The main issue to solve is handling all messages in parallel without blocking the processes.

The virtual NAO The implementation for the virtual NAO is embedded in the Unity environment in which the tablet application (the activity centre) is developed and the task is to execute the moving of the avatar according to the values received. Since there was no implementation available it needed to be developed.

A 3D Model was developed composed of virtual objects connected through a series of junctions, which are meant to take the place of real NAO's gears, as displayed in Figure 2.6. Next, an algorithm was developed to translate the rotations of the real NAO's gears in rotations of the NAO avatar's junctions around their three axis's (x, y, z), so that we can obtain the same movements using the same commands from the robot-movement manager.

Making the avatar's leg movement realistic and correct in regards to the message send, was the most difficult part of the implementation here. We developed a hybrid solution which allows calculating only a portion of the parameters required to apply a simulated gravity to some parts of the 3D model.

The implementation discussed in this section allows to have all complex com-

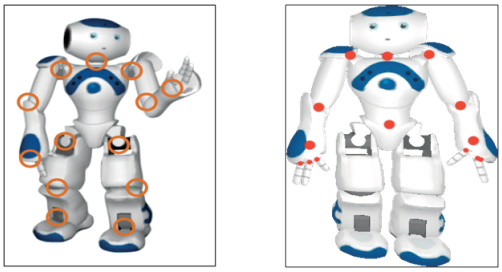


Figure 2.6: The real NAO on the left with the joint indicated and the avatar on the right with the junctions indicated which were implemented.

putations concerning what behaviours the robot *should* do on external computers (principle 1, using cloud techniques) while the embodiment specific computations (transferring joints and times to actual movement) are done on the device itself. In the PAL project, the NAO robot was the target platform. However, the system could easily be applied to different types of robots (again linking to the principle of modularity 2). When a new robot platform would be introduced then high-level gestures like 'wave arm' are still sent in the same way over the nexus. The behaviour manager and robot connectors, however, would require extension to support the new embodiment.

Stylised Behaviours Humans (often unconsciously) use social signals to inform others about their affective stance or attitude; based on observations we evaluate someone as, among other things, warm or cold, competent or incompetent, friendly or hostile, and dominant or submissive (e.g., [36, 37]). For artificial agents (both virtual and robotic), to engage in meaningful interactions with humans, the importance of social intelligence is widely acknowledged [38, 39]. On top of that, communication style is important in educational settings: teachers should use appropriate styles when interacting with students and this is not different for robots that teach [40]. Although previous work is available on the expression of affect by robots and agents [41–44]) and rapport building between agents and humans [45], there was no clear way of modulating expressive style of robots in an implicit way (with notable exceptions being some work on virtual agents [46, 47]). With implicit we mean that the base behaviours and scenario of the robot are the same, but the style of the robot differs. In several works, we have shown that it is possible to manipulate style in a subtle manner. We were able to manipulate children's perception of warmth and competence of a robot [40], the perception of warmth [48], and the perception of dominance [49]. In general we found that different gestures as well as parameter-based modulation enabled us to express style in a recognisable manner.

Based on the theories from educational sciences, we defined three interaction styles for the PAL Actor: *friendly*, *direct*, and *neutral*. These styles are defined by the factors warmth, competence, and dominance, for which we designed and evaluated non-verbal behavioural patterns [40, 49, 50]. The factors define the

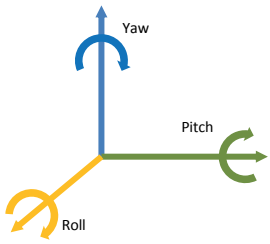


Figure 2.7: Definitions of Yaw, Roll, and Pitch as used here to define robot joint manipulation

Table 2.1: NAO joints and adjustment values for the stand posture per style

	Original	**Neutral**	**Friendly**	**Direct**
HeadYaw	-0.1	0.0	changing	changing
HeadPitch	-9.6	0.0	-10	-20
LShoulderPitch	80	80	60	70
LShoulderRoll	10.4	8.0	20	35

styles as follows: *friendly*, high warmth, low competence and low dominance; *direct*, low warmth, high competence and high dominance; *neutral*, low warmth, low competence and low dominance. Verifying how well these factors implement the style (as friendly, direct, and neutral) is future work. Definitions of each style, and the mapping of a specific style to each activity, have been stored in the common knowledge-base.

We defined a minimal set of NAO joint adjustments to express each style. Some parameters are directly mapped to a specific joint (e.g., head tilt vertical with `HeadPitch`). Other parameters require adjustment of multiple joints (e.g., gesture openness requires adjustment of `ElbowYaw` and `ElbowRoll` relative to `SholderRoll`). See figure 2.7 for the meaning of the words yaw, roll and pitch. These adjustment were applied to the original 'stand' posture as designed by SoftBank (Alderbaran) to create a start and end pose for each style (see Figure 2.8), and these joint adjustment values (see Table 2.1) were used to calculate the relative joint adjustment for each key frame of each gesture in each style. Additionally, manual adjustments have been applied to distinctive motions for specific gestures in each style to avoid exceeding joint limits and/or creating jerky, unnatural motions.

During child-PAL Actor interactions the PAL system selects the appropriate interaction style based on the child model and the ongoing activity. Whenever the Behaviour Manager receives a nexus behaviour message (see Section 2.4.6) it uses a 'mood' value to determine the selected style, and selects behaviours accordingly.

2.4.7. 'The Hybrid Brain'

The brain is responsible for the decision making and managing the content of the behaviour of the PAL actor (i.e., the NAO robot or avatar). The brain runs on the cloud (principle 1) which allows the use of more powerful computers for com-

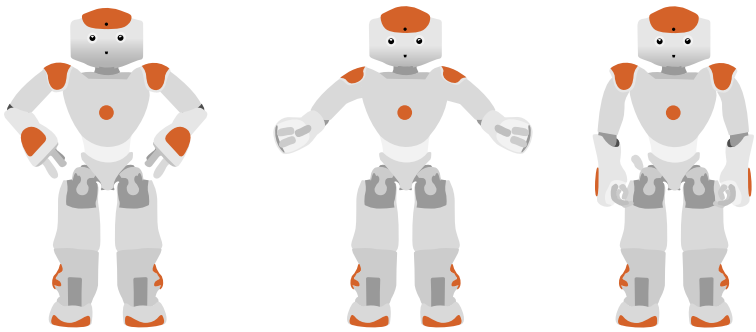


Figure 2.8: Example of a base-pose for the styles implemented in PALY3 resp. Direct, Friendly, and Neutral.

puting the context dependent optimal behavioural strategies. The brain uses the communication mechanism described in section 2.4.5 to support easily connecting additional (sub-)modules (principle 2). Finally, the brain consists of several artificial intelligence techniques (principle 4, hybrid). The action selection module uses mainly machine learning optimisation of protocols, while the explainable AI module uses BDI structures and ontology querying for its implementation, and the dialogue management uses a mixture of rule-based processing and statistical selection for the optimal strategy. In this section, we describe the different modules that support the human-agent interaction.

Action Selection and User Models

The PAL system uses an action selection module in conjunction with user models in order to personalise the behaviour of the system. Personalising the application to the particularities of each user is essential for two reasons: 1) it increases the engagement of the user to the PAL system, and 2) it allows the user to reach more effectively its personal goal(s) by adapting to his/her preferences.

The action selection is based on a hierarchical action selection architecture, HAMMER [51] [52], which uses multiple models to generate and evaluate multiple action possibilities. It takes place at two different levels in the PAL system. The first level is when multiple options are available during a dialogue. For instance, when the avatar can suggest to start one of the three games of the application, the action selection module will select the one that the user models predict as most beneficial for the user. An action is deemed beneficial if it increases the knowledge level of the user or if it increases its (predicted) happiness (typically, winning in a game increases happiness). The second level of decision making is within the quiz game in order to select the topic of the questions that are asked. Here again, a user model is used to predict the knowledge level of the users on the different topics

covered by the PAL system. This information is to select of topics that are not too difficult nor too easy, but just with the right level of difficulty. Such a topic selection approach allows the user to remain in his "zone of proximal development", which is known to provide optimal educational path[53] [54].

One of the main challenges in the action selection and user modelling is to produce accurate predictions from few data. Typically, a large amount of interactions is required to make an accurate estimation of the difficulty level of the child on one particular topic. This comes from the fact that each question returns only a limited amount of information: it only informs the system if the child managed to respond correctly to the question or not. This binary information is not enough to infer the actual knowledge level on a topic and the system needs to accumulate several dozens of responses in order to make an accurate assessment that is not biased by non-knowledge related factors (e.g., random guess, mistakes caused by distraction or ambiguous formulation of questions). For instance, 10 data points only provide a rough approximation of the user level, while 100 data points provide a more accurate estimation. This difficulty is amplified by the number of topics on which we would like an estimate of the knowledge level. For instance, if twenty questions are used to make an (potentially inaccurate) estimate on one topic, then more than 580 questions are required to form a global estimate of the user's knowledge level on each of the individual topics. The level estimations are made per topic to allow a child to excel in one topic, while progressing more slowly, or not at all, in others. This large amount of questions most of the time represents a limitation for intelligent tutoring systems, as they may be unable to provide a personalised educational path as long as the estimation of the knowledge level of the user is not completed. Designing a model that can account for sparse data and, therefore, provide accurate estimate with only a little amount of data is of crucial importance. In the PAL system this is partly overcome by the manual estimation of the child's entry level by the HCP via goal setting as described in section 2.4.3.

Technical Description

In the PAL system, we have introduced a novel user model that leverages data from previous users of the application in order to bootstrap the predictions made by the user model. This user model is also able to track, in real-time, the evolution of the children's difficulty levels. Tracking this is important to continuously provide adequate level of difficulty for the users.

The realisation of this model is centred around three main features: 1) the model relies on Gaussian Processes to track online the evolution of the student's knowledge level over time, 2) it uses collaborative filtering to rapidly provide long-term predictions by leveraging the information from previous users, and 3) it automatically generates abstract representations of knowledge components via automatic relevance determination of covariance matrices. The model has been evaluated on three datasets, including data from real users and the results demonstrate that the model converges to accurate predictions in average 4 times faster than the compared methods. A detailed evaluation of the model's technical characteristics can be found in [55].

Dialogue Management

The PAL actor needs verbal and non-verbal communication skills to support the child in his/her learning process and to become a real companion. Dialogue is present in almost all activities of the app, guiding the child, or giving feedback to current and past performance in games, or to her/his treatment, in which case the data is provided by entries in the timeline.

The dialogue manager (DM) is responsible for multimodal generation (language and gestures) and, currently to a lesser part, multimodal interpretation. In section 2.4.6, we discussed how a multimodal utterance should be processed causing the PAL actor (robot or avatar) to actually output the utterance. During idle phases, the robot-movement manager makes the actor to appear lively by constantly executing 'autonomous moves'. The DM, on the other hand, chooses the particular (deliberate) movements, like gesturing to something on the screen or waving to the child.

To be an interesting partner for conversation over a longer period of time, the communication strategy must be adaptive to the user. The DM takes a long-term perspective (using a user model) and takes short-term aspects into account, such as the current user mood or recent important events, as well as the parameters from the stylised behaviour model described earlier to modulate the interactions. This already points to the importance of a *world model* and a *memory* for the agent, which enables it to reason about past events and interactions, and subsequently uses its knowledge in the conversation. This, together with a high variability in dialogue strategies and a rich repertoire of verbal and non-verbal expressions to choose from, helps to make the artificial agent more appealing.

Dealing with children requires high reliability concerning the content and the way things are presented. This makes it difficult to use pure machine-learning based methods, because the results will never be fully predictable, not to mention the problem of collecting enough data for dialogue strategies in the first place. Together with the need for very flexible dialogue that was mentioned earlier, we decided to go for a rule-based approach with statistical selection, provided by the Action Selection unit described in section 2.4.7. The long-term dialogue memory is implemented by a specialised RDF/OWL knowledge-base (described in section 2.4.2), enabling us to use world knowledge and temporal reasoning in the dialogue management.

Rule-based dialogue systems are in between learning-based approaches and hierarchical state machines concerning flexibility and implementation complexity. For the PAL system, which already needs fairly complicated dialogue management, state machines would be unmanageable in size. Machine learning approaches on the other hand lack the predictability that is essential for an application in the health sector. Given this set of requirements, we decided to develop a rule-based dialogue management framework which is tightly connected to the specialised reasoning engine and the statistical action selection.

A central achievement of the PAL project concerning dialogue processing is the development of the dialogue framework VOnDA [56], which facilitates creating reactive dialogue management engines. In the case of PAL, the DM follows the Information State/Update tradition [57]. The framework is unique in that it uses a

specialised RDF reasoner which allows to attach temporal information to the RDF triples as implementation for the information state. This has the advantage that a long-term memory is directly build into the architecture of the dialogue system. Figure 2.9 shows a schematic view of the module.

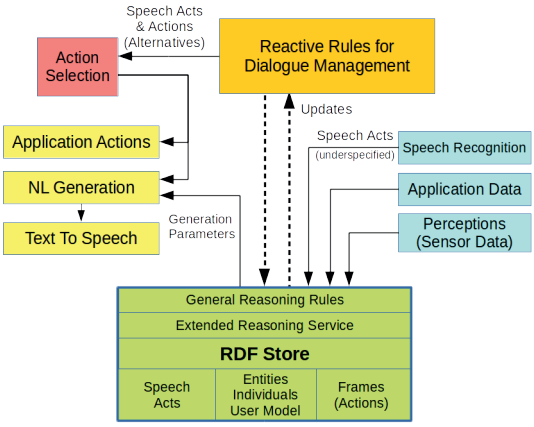


Figure 2.9: A schematic VOnDA agent

If information changes, the previous state is still kept inside the database, which allows to additionally use information from the past for dialogue strategies. Furthermore, the RDF store also provides a flexible specification layer for domain knowledge, including knowledge about natural language concepts, such as dialogue acts and semantic frames, but also for domain-specific data structures that are used in the rule base. Since RDF/OWL is a well-established standard, there is plenty of tool support to create the ontologies which serve as basis for the dialogue management. With VOnDA , we have created a framework that tackles the following design goals:

- Flexible and uniform specification of dialogue semantics, knowledge and data structures
- Scalable, efficient, and easily accessible storage of interaction history and other data, resulting in a large information state
- Readable and compact rule specifications, facilitating access to the underlying RDF database, with the full power of a programming language
- Transparent access to underlying programming code for simple integration with the host system

The dialogue management of PAL itself is implemented mostly as VOnDA rules, with supporting Java code to hook it up to the communication hub (the nexus), or when specialised functionality is needed, e.g., for complex data base queries and computations.

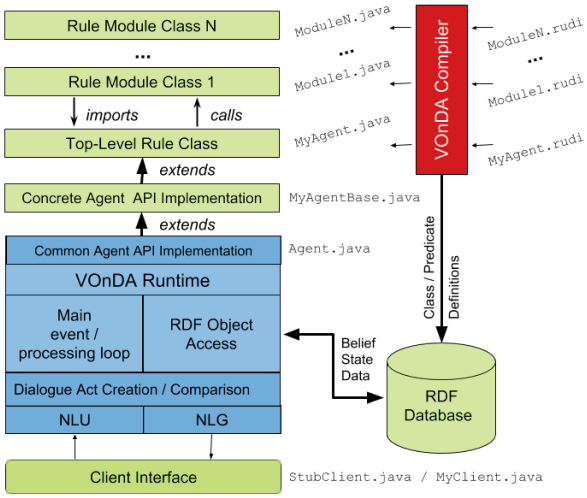


Figure 2.10: A schematic VOnDA agent

Technical Description

VOnDA consists of a compiler that turns rule descriptions into programming code, using data structure specifications from an ontology, and a run-time library that is used in the agent implementation and executes the rule code produced by the compiler. The compiler uses the class and property definitions in the ontology for type inference and type checking.

The rule language itself strongly resembles by Java/C++ if/then/else statements and expressions. Access to the database is modelled after field access, where RDF URIs are treated like references to objects, and properties like fields. Figure 2.11 shows a small example with a tiny part of the ontology and some code using objects from this class hierarchy.

```
user = new Animate;
user.name = "Joe";
set_age:
if (user.age <= 0) {
  user.age = 15;
}
```

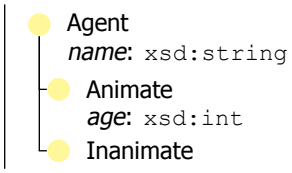


Figure 2.11: Ontology and VOnDA code

During run-time, the set of reactive rules is executed whenever there is a change in the information state. These changes are caused by incoming sensor or application data, intents from the speech recognition, or expired timers. The incoming data is put into the database, which hereupon triggers a notification event. For efficiency, the rules work on cached database content. Any change produced by rule applications is put back into the database, re-triggering the rules until a fixpoint is reached. The event-based nature of the system allows to create a very responsive

system that is able to react to external stimuli in real-time.

Apart from the main dialogues, there are interaction modules that are tightly connected to the dialogue manager and cover particular parts of the dialogue functionality, namely *targeted feedback* to timeline entries, an *episodic memory* reasoning about past events, and a module for *off-activity talk*, which uses self-disclosure to engage the child in a conversation whose goal is not to increase his or her knowledge, but to increase the bond to the virtual agent. This again shows the modularity (principle 2) of our implementation. The modules subscribe to and send messages over the nexus. They can be activated or deactivated without impairing the rest of the system's functionality. In a scientific project this has the additional benefit of allowing additional experiments and pilots to test the effects of specific parts of the system.

Targeted Feedback

The targeted feedback module is active during the timeline activity of the child interface. It first determines how close the child is following the routine of entering data into the timeline, and subsequently praises or encourages the child, depending on the frequency of the entries, and the data itself. If alarmingly high or low values are reported, the child is advised to talk to her/his parents to prevent a critical situation.

Episodic Memory

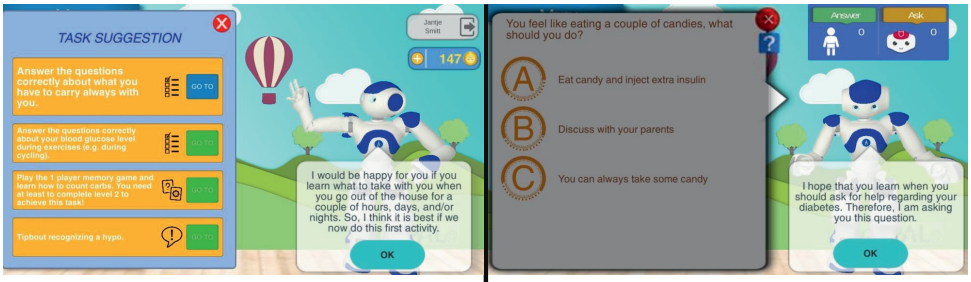
In contrast to the *targeted feedback*, the episodic memory module aggregates data from the past to detect events like the recent completion of a task or an achievement, exceptionally disciplined behaviour, and the like. These events are then subject to remarks and questions during the welcome phase directly after log-in. In this way, the agent encourages positive behaviour and shows that it takes interest in the child's daily activities.

Off-Activity Talk / Self-Disclosure

The goal of the off-activity talk (or social talk) module is to improve the virtual agent's social connection to the user, and make it more likeable. As a consequence, it should increase the user's inclination to follow the agent's advice and guidance. The current module is in large parts based on the study in [58] enhanced with additional introductory dialogue moves and more elaborate prompts. Based on an estimate of 'intimacy' between the user and the agent, which is based on general usage parameters, like the frequency of proper reactions to prompts by the agent, the agent reveals personal secrets of its habits or former encounters with other robots or persons.

Explainable Artificial Intelligence

EXplainable Artificial Intelligence (XAI) is the capability of a system to explain/ justify its behaviour to its users. The General Data Protection Regulation (GDPR) law states that users have the right to explanations [59]. It has been a significant topic during the development of the PAL system.



(a) Explanation during a Task Suggestion

(b) Explanation during the Quiz

Figure 2.12: Two screen-shots of explanations given in the PAL system. Screen-shot (a) shows an explanation for a task suggestion, screen-shot (b) shows the quiz.

XAI is of particular importance in applications where the system makes consequential decisions, like, in health-care [60, 61] and is of particular importance in long-term interaction since lack of trust in a behaviour change systems causes the users to not rely on the given advice [62]. Lack of trust can cause users to misuse or even abandon the system altogether [1, 63]. XAI has been shown to increase a user's trust in and understanding of the system's behaviour [60, 64–67].

Within the context of the PAL system, XAI is the capability of the actor to provide utterances that serve to make the system's behaviour intelligible. We implemented XAI for three different activities, quiz questions, tip of the day, and task proposals. These activities have already been discussed in the child interface section 2.4.4. The XAI module monitors the system state by monitoring the messaging board. The module can provide an explanation during one of these activities by monitoring what goal the action selection is currently pursuing. It generates the explanation partly by using annotations and partly by generating the sentences. A goal has a description annotated to it. For example, 'ask for help regarding your diabetes'. Then there is a set of preceding sentences and following sentences that can be put in front and behind this description. The XAI module works together with the dialogue manager to produce the complete explanations and integrate them into the dialogue. For example, 'I hope that you learn when you should ask for help concerning your diabetes. Therefore, I am asking you this (quiz) question'. See figure 2.12 for two examples of such explanations given during activities. Explanations could be given on the initiative of the PAL agent or of the child's own accord. In the later case, the child could request an explanation by pressing a question mark visible on the screen (see also figure 2.12 b showing the question mark during the quiz).

2.5. Development and Test procedures

Creating an integrated system requires far more solid software engineering approaches than commonly applied in scientific projects, i.e., it is required to program the modules 'defensively' so that they can continue (to their best of capabilities) when other modules break. Furthermore, it requires rigorous development cycles with plenty of time to test all the new features in systematic and thorough ways

using, e.g., unit testing [68], integration testing [69], system testing [70] and performance testing [71]. Research projects need to be aware of this and have the necessary resources and knowledge available. For example, have commercial companies involved in the project.

We used `git` for code sharing and version management. The master branch was always running on the main server. For every feature, all modules had a branch with the features name to develop it. Usually, multiple features were simultaneously implemented. Each on had a lead developer. Additionally, a global lead developer oversaw the larger process. When a feature was implemented then the separate modules had unit tests as a first check of code validity and the feature's lead developer tested the system manually for face validity. Next, the feature was merged to a common developer branch. The goal of this was to merge in parallel developed features before they were merged to the master branch. Next, an independent developer tested the system, including rerunning the unit tests. Finally, a test-module was run that automatically acted as a user of the system. The goal of that test-module was to test all aspects of the system and see whether things run as they were supposed to. Then, finally, the master branch was updated. There the automatic tests were run again. Additionally, two more types of tests were done in this stage before the system was deployed. First, a stress test was done where 10 to >20 user accounts (controlled by the developers) logged into the system simultaneously and actively interacted with the conversational agent and did the activities. This was needed to find and repair potential synchronisation errors and to test whether our non commercial servers were able to handle the complexity. After the stress test non-developer researches had to use and test the system. If all those tests were successful, then a new short-term experiment or pilot could be done. These served also as a final test where our target (child) users interacted with the system under supervision of researchers. This development procedure, over a course of three years, resulted in the system presented in this paper and used by children for a period of 2.5 to 3 months. The socio-cognitive engineering activities that focused on continuous stakeholder involvement, domain analyses, integration of human factors theories and claims analyses to validate the design rationale (in formative and summative evaluations) are described in a separate paper [72].

2.6. Analyses of Performance

In total there were 48 (25 Dutch and 23 Italian) children with Type 1 Diabetes Mellitus (T1DM) aged 6-14 that used the final system. The children were recruited via two hospitals in the Netherlands and one in Italy. There were no consequences to dropping out intermediately. 47 children had an average of 19 log-ins (STD = 12.9, minimum = 1, maximum = 55). One child was excluded from analyses due to a glitch in the data caused by a system error. The Randomized Controlled Test that compared the knowledge, performance and health conditions of these children with children who had "casual care" will appear in a separate paper [73]).

There were a total of 756.667 lines of code with over 7500 commits. More than 95 branches. The system was up and running more than 95% of the time. One of the stress tests had more than 20 user accounts (controlled by the developers)

simultaneously log-in and actively use the system running on one small-to-mid size virtual machine. Neither the stress test nor the test with real users showed any signs of performance issues. Nevertheless, the system is still a research prototype and would have to be hardened to adhere to commercial standards and possibly host thousands of users simultaneously. This was beyond the scope of the current project and will require further development and testing.

2.7. Future Extensions

Several extensions are possible in this system. It would be beyond the scope of this paper to provide an exhaustive list. However, we briefly discuss a few modules that were developed during the project, but did not yet reach the status of full integration in the system.

GPS tracking

Behavioural adaption to individual differences (context awareness) is compulsory for achieving good outcomes in health education and care through social robotics. One aspect is modelling social context (e.g., home vs. hospital) through collecting and processing global positioning system (GPS) data. We have successfully implemented a module that tracks a user's location. The pending future work is to integrate this with dialogue management and action selection so that conversations and activities are shaped to this context.

GPS data was collected with a frequency of 1 sample per minute. To deal with potential inaccurate GPS estimations, we applied an averaging filter over time, using a window size of about 10 minutes. Prior to real-time interactions, we saved pre-defined fixed locations (e.g., hospital locations) in the database. During the interactions, once a location was estimated, we first compared the estimated location with the saved locations in the HFC database. To do so, we computed the distance between two points on the earth, where we set the earth radius to 6371.0 in km. If the distance between the estimated location and each of the saved locations was greater than a threshold (3 km, in our case), then we saved the estimated location in the database as a new location. Otherwise, the estimated location was not saved in the database. This procedure was repeated for each estimated location until the user logged off.

Eye-gaze tracking

A second form of context awareness was implemented through eye-gaze tracking. We attempted to infer aspects of a user's mental state using eye gaze, which has been shown to be useful in training contexts [74]. Interaction logs and verbal protocols are generally not adequate for genuine cognitive and social profiling [75]. However, nonverbal cues such as eye gaze have been frequently studied for inferring user internal states in this context, as eye movements directly reflect what is at the centre of an individual's visual attention, and are linked to cognitive processes in the mind [76].

To address this, we pushed the limits of the state-of-the-art, and implemented a Convolutional Neural Network based method for accurate gaze tracking, which

runs on a GPU-enabled tablet, and takes users' face images captured via the front facing camera as an input and predicts users' gaze fixations on the tablet screen in real-time (approximately 10 fps) [77]. We further developed a classification scheme to predict users' mental states, i.e., knowledgeability, from the estimated gaze fixations while they were playing an educational game with the robot/avatar [77]. Our ultimate goal was to benefit from mental state predictions to better personalise quiz questions and make the avatar exhibit user-aware behaviours. For example, if the avatar senses that a particular user is having difficulties answering the question, the avatar can offer a hint to support the user.

The eye-gaze tracking was not fully evaluated in the version of the PAL system used in the current round of hospital evaluations because it required real-time on-device processing (to preserve the privacy of the user) and the mobile devices used did not have sufficient computational power available for this.

Automatic Speech Recognition

We considered Automatic Speech Recognition (ASR) to enhance interaction with our users. ASR could be a way to add more convenient conversational capabilities to the system, or add a dictation mode to the timeline, and possibly improve real-time emotion estimation for an even more empathic companion ([78]). At that time, however, solutions based on openly available resources like software tools and pre-trained acoustic models turned out not to achieve the necessary recognition quality for such a project. On the other hand, cloud-based solutions posed a severe data and personality protection issue since speech data are classified as biometric data and therefore on the highest protection level of the European Union Data Protection Guidelines. Therefore, we used ASR only for prototypical demonstration, as an outlook towards future extensions. Still, with the current advent of large coverage embedded solutions, and cloud ASR that explicitely guarantees adherence to the EU guidelines, integrating ASR would be a beneficial future extension for these types of systems.

2.8. Lessons Learned and Discussion

This paper presented a social robot system architecture that facilitates long-term human robot interaction. The four primarily important aspects of our architecture were: using cloud based techniques (principle 1); having a modular architecture (principle 2); having a common knowledge-base to support development and support the human-agent interaction (principle 3); and, using hybrid artificial intelligence (principle 4). This section discusses the global lessons learned concerning the development of such a system and concerning the four principles.

Building an integrated system in a research project poses unique challenges, which are usually avoided implementing small, isolated prototypes that exhibit the behaviour which is under investigation. In an integrated system, the functionality of the modules must be synchronised to achieve the intended overall performance. This needs more consultation and arrangements between the implementing groups, joint definition of interfaces, down to the functional level, and thorough development cycles and testing procedures ([70]). For these requirements, resources have

to be allocated that are usually avoided in projects conducting basic research. Considering this, PAL is among the few research projects that created an integrated system that was used successfully by real, untrained users over a longer time and also has the potential to be commercially exploited in an extended and consolidated form.

The cloud-based system architecture (principle 1) we presented facilitates long-term interaction. Firstly, it enables access to external libraries (Big Data, for example used in the dialogue of this system); secondly, it allows for more complex computations by using external computational power (cloud-computing). In this way the robot can provide a more stable and believable interaction with more diverse behaviours. Thirdly, it facilitates version management which is particularly important for research where functionality is incrementally added. Fourthly, it helps the health-care professional to seamlessly personalise and adapt the system's behaviour based on the patient's individual development.

Furthermore, modularity (principle 2) reduces complexity, both for the development and application phase. However, we found that there are limits to the amount of modularity one should strive for. For example, it seems attractive to separate the agent's non-communicative behaviour (the application logic, so to say) and the (possibly multi-modal) communication layer. This, however, turns out to be challenging, because both parts heavily depend on each other to create a believable and consistent persona for the agent. For example, an ongoing dialogue can not simply be stopped at any point to perform some task, but has to be properly brought to an end for the robot not to be perceived as impolite or untrustworthy. On the other hand, the urgency of the task might require a more or less elaborate form of shutting down an ongoing conversation. We found that there is a delicate balance between modularity and coordination.

Another important aspect is that the system allows to plug in a virtual avatar of the robot. Modularity is maintained since other modules do not need to know whether the avatar or the physical robot is connected. This implementation could be extended by implementing controllers for other embodiments like a (physical and/or virtual) Pepper robot. Having a avatar of the robot is important because many robots (like the nao robot) are used for educational and/or health-care related support, but are too expensive for commercial use. The conversational agent (CA) is meant to 'bind' the user to the application. Having a robot present during meetings in the hospital and an avatar when using the application at home allows to make use of the motivational gains of a CA without having to buy a robot for every child with diabetes.

For a common knowledge-base (principle 3), we chose to develop an ontology in the context domain of diabetes type 1. By working closely together with health-care professionals, we believe to have created a reusable knowledge-base that can be profitable for further research on combining health-care and artificial intelligence. For example, the formalised T1DM self-management learning goals and achievements are available as ontology as well as in an online co-creation tool[2] to allow for further specification and usage. We believe that reasoning over this expert

[2] https://confluence.ewi.tudelft.nl/display/PO1/PAL+Objectives

knowledge is best done by applying agent-based or expert system implementations. The symbolic AI proposes several possible learning paths to support the children to become self-manageable. However, machine learning can then optimise the possible learning paths to the specific children. We show in our architecture how such hybrid techniques (principle 4) can work together successfully. It is beyond the scope of this paper to discuss the performance of the system on the children their diabetes-control and well-being. For this we refer to [72] and [73].

2.9. Conclusion

The PAL system architecture is a cloud-based, modular, long-term human-robot interaction framework that uses hybrid artificial intelligence and a common knowledge-base to shape the interaction. This system has run with over 40 users for two periods of two and a half to three months. The system remained stable and continued to show (more and more) behaviours and support in health education & care. This work can serve as a blueprint for future long-term human-avatar and human-robot interaction studies and thereby facilitate incremental research.

References

[1] I. Leite, C. Martinho, and A. Paiva, *Social robots for long-term interaction: a survey,* International Journal of Social Robotics **5**, 291 (2013).

[2] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, *et al.*, *Multimodal child-robot interaction: Building social bonds,* Journal of Human-Robot Interaction **1**, 33 (2013).

[3] A. Coninx, P. Baxter, E. Oleari, S. Bellini, B. Bierman, O. B. Henkemans, L. Cañamero, P. Cosi, V. Enescu, R. R. Espinoza, *et al.*, *Towards long-term social child-robot interaction: using multi-activity switching to engage young users,* Journal of Human-Robot Interaction **5**, 32 (2016).

[4] O. A. B. Henkemans, B. P. Bierman, J. Janssen, R. Looije, M. A. Neerincx, M. M. van Dooren, J. L. de Vries, G. J. van der Burg, and S. D. Huisman, *Design and evaluation of a personal robot playing a self-management education game with children with diabetes type 1,* International Journal of Human-Computer Studies **106**, 63 (2017).

[5] R. Looije, M. A. Neerincx, J. K. Peters, and O. A. Blanson Henkemans, *Integrating robot support functions into varied activities at returning hospital visits,* International Journal of Social Robotics **8**, 483 (2016).

[6] J. van der Drift Esther, R.-J. Beun, R. Looije, O. A. B. Henkemans, and M. A. Neerincx, *A remote social robot to motivate and support diabetic children in keeping a diary,* in *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, 2014) pp. 463–470.

[7] Y. Fernaeus, M. Håkansson, M. Jacobsson, and S. Ljungblad, *How do you play with a robotic toy animal?: a long-term study of pleo,* in *Proceedings of the 9th international Conference on interaction Design and Children* (ACM, 2010) pp. 39–48.

[8] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul, and B. Scassellati, *Social robots as embedded reinforcers of social behavior in children with autism,* Journal of autism and developmental disorders **43**, 1038 (2013).

[9] T. Shibata and K. Tanie, *Physical and affective interaction between human and mental commit robot,* in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, Vol. 3 (IEEE, 2001) pp. 2572–2577.

[10] S. Sabanovic, C. C. Bennett, W.-L. Chang, and L. Huber, *Paro robot affects diverse interaction modalities in group sensory therapy for older adults with dementia,* in *Rehabilitation Robotics (ICORR), 2013 IEEE International Conference on* (IEEE, 2013) pp. 1–6.

[11] M. R. Banks, L. M. Willoughby, and W. A. Banks, *Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs,* Journal of the American Medical Directors Association **9**, 173 (2008).

[12] J. Broekens, M. Heerink, and H. Rosendal, *Assistive social robots in elderly care: a review,* Gerontechnology **8**, 94 (2009).

[13] C. Kertész and M. Turunen, *What can we learn from the long-term users of a social robot?* in *International Conference on Social Robotics* (Springer, 2017) pp. 657–665.

[14] M. Kwon, M. F. Jung, and R. A. Knepper, *Human expectations of social robots,* in *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on* (IEEE, 2016) pp. 463–464.

[15] M. Ligthart, O. B. Henkemans, K. Hindriks, and M. A. Neerincx, *Expectation management in child-robot interaction,* in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on* (IEEE, 2017) pp. 916–921.

[16] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, *Interactive robots as social partners and peer tutors for children: A field trial,* Human–Computer Interaction **19**, 61 (2004).

[17] C. D. Kidd and C. Breazeal, *Robots at home: Understanding long-term human-robot interaction,* in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on* (IEEE, 2008) pp. 3230–3235.

[18] M. De Graaf, S. Ben Allouch, and J. Van Dijk, *Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study,* in

Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (ACM, 2017) pp. 224–233.

[19] F. Kaplan, *Everyday robotics: robots as everyday objects,* in *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies* (ACM, 2005) pp. 59–64.

[20] J. Li, R. Kizilcec, J. Bailenson, and W. Ju, *Social robots and virtual agents as lecturers for video instruction,* Computers in Human Behavior **55**, 1222 (2016).

[21] R. Looije, M. A. Neerincx, and V. d. Lange, *Children's responses and opinion on three bots that motivate, educate and play,* (2008).

[22] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, *A survey of research on cloud robotics and automation.* IEEE Trans. Automation Science and Engineering **12**, 398 (2015).

[23] N. R. Jennings and M. Wooldridge, *Applications of intelligent agents,* in *Agent technology* (Springer, 1998) pp. 3–28.

[24] M. Minsky, *Society of mind* (Simon and Schuster, 1988).

[25] J. Gratch, A. Hartholt, M. Dehghani, and S. Marsella, *Virtual humans: a new toolkit for cognitive science research,* in *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35 (2013).

[26] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, *What are ontologies, and why do we need them?* IEEE Intelligent systems , 20 (1999).

[27] T. R. Gruber, *A translation approach to portable ontology specifications,* Knowledge Acquisition 5(2) , 199 (1993).

[28] M. A. van Bekkum, H.-U. Krieger, M. A. Neerincx, F. Kaptein, B. Kiefer, R. Peters, and S. Racioppa, *Ontology engineering for the design and implementation of personal pervasive lifestyle support.* in *SEMANTiCS (Posters, Demos, SuCCESS)* (2016) pp. 5–8.

[29] M. V. Welie, G. C. V. D. Veer, and A. Eliëns, *An ontology for task world models,* Design, Specification and Verification of Interactive Systems '98 , 57 (1998).

[30] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, D. R. Krathwohl, *et al.*, *Taxonomy of educational objetives: the classification of educational goals: handbook I: cognitive domain*, Tech. Rep. (New York, US: D. Mckay, 1956).

[31] R. Peters, J. Broekens, and M. A. Neerincx, *Guidelines for tree-based learning goal structuring,* in *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17 (ACM, 2017) pp. 401–405.

[32] H.-U. Krieger and C. Willms, *Extending OWL ontologies by cartesian types to represent n-ary relations in natural language,* in *Language and Ontologies 2015* (2015).

[33] H.-U. Krieger, *Capturing graded knowledge and uncertainty in a modalized fragment of owl,* in *Proceedings of the 8th International Conference on Agents and Artificial Intelligence. International Conference on Agents and Artificial Intelligence (ICAART-2016), February 24-26, Rome, Italy*, INSTICC (INSTICC, 2016) conference Website: http://www.icaart.org.

[34] H.-U. Krieger, *Integrating graded knowledge and temporal change in a modal fragment of owl,* in *Jaap van den Herik; Joaquim Filipe;: Agents and Artificial Intelligence. Revised selected papers from the 8th International Conference, ICAART 2016*, Lecture Notes in Computer Science, LNCS (Springer-Verlag, 2016) pp. xxx–yyy.

[35] H.-U. Krieger and S. Schulz, *A modal representation of graded medical statements,* in *Annie Foret; Glyn Morell; Reinhard Muskens; Rainer Oswald; Sylvain Pogodalla: Formal Grammar 2015/2016*, Lecture Notes in Computer Science, LNCS, Vol. 9804 (Springer-Verlag, Berlin Heidelberg, 2016) pp. 1–17.

[36] S. T. Fiske, A. J. Cuddy, and P. Glick, *Universal dimensions of social cognition: Warmth and competence,* Trends in cogn. sci. **11**, 77 (2007).

[37] S. L. Ellyson and J. F. Dovidio, *Power, dominance, and nonverbal behavior: Basic concepts and issues,* in *Power, dominance, and nonverbal behavior* (Springer, 1985) pp. 1–27.

[38] T. Fong, I. Nourbakhsh, and K. Dautenhahn, *A survey of socially interactive robots,* Robot. and autonomous syst. **42**, 143 (2003).

[39] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, *Bridging the Gap Between Social Animal and Unsocial Machine: A Survey of Social Signal Processing,* Affect. Comput. **3**, 69 (2011).

[40] R. Peters, J. Broekens, and M. A. Neerincx, *Robots educate in style: The effect of context and non-verbal behaviour on children's perceptions of warmth and competence,* in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on* (IEEE, 2017) pp. 449–455.

[41] G. Castellano, M. Mancini, C. Peters, and P. W. McOwan, *Expressive copying behavior for social agents: A perceptual analysis,* Trans. on Syst., Man, and Cybern. **42**, 776 (2012).

[42] C. Clavel, J. Plessier, J.-C. Martin, L. Ach, and B. Morel, *Combining facial and postural expressions of emotions in a virtual character,* in *Int. Workshop on Intell. Virtual Agents* (Springer, 2009) pp. 287–300.

[43] M. Destephe, A. Henning, M. Zecca, K. Hashimoto, and A. Takanishi, *Perception of emotion and emotional intensity in humanoid robots gait,* in *Proc. 2013 Int. Conf. on Robot. and Biomimetics* (IEEE, 2013) pp. 1276–1281.

[44] C. Breazeal, *Emotion and sociable humanoid robots,* Int. j. of human-comput. stud. **59**, 119 (2003).

[45] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, *Creating rapport with virtual agents,* in *Intelligent Virtual Agents*, edited by C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007) pp. 125–138.

[46] T.-H. D. Nguyen, E. Carstensdottir, N. Ngo, M. S. El-Nasr, M. Gray, D. Isaacowitz, and D. Desteno, *Modeling warmth and competence in virtual characters,* in *Proc. Int. Conf. on Intell. Virtual Agents* (Springer, Delft, The Netherlands, 2015) pp. 167–180.

[47] B. Ravenet, M. Ochs, and C. Pelachaud, *From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes,* in *Proc. Int. Conf. on Intell. Virtual Agents*, Vol. 8108 LNAI (Springer, 2013) pp. 263–274.

[48] P. Prajod, M. Al Owayyed, T. Rietveld, J.-J. van der Steeg, and J. Broekens, *The effect of virtual agent warmth on human-agent negotiation,* in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19 (International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2019) pp. 71–76.

[49] R. Peters, J. Broekens, K. Li, and M. Neerincx, *Robot dominance expression through parameter-based behaviour modulation,* in *Proc. Int. Conf. on Intell. Virtual Agents* (ACM, 2019) p. in press.

[50] R. Peters, J. Broekens, K. Li, and M. A. Neerincx, *Robot dominance expression through parameter-based behaviour modulation,* in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (ACM, 2019) pp. 224–226.

[51] Y. Demiris and B. Khadhouri, *Hierarchical attentive multiple models for execution and recognition,* Robotics and Autonomous Systems **54**, 361 (2006).

[52] Y. Demiris, L. Aziz-Zadeh, and J. Bonaiuto, *Information processing in the mirror neuron system in primates and machines,* Neuroinformatics **12**, 63 (2014).

[53] L. S. Vygotsky, *Mind in society: The development of higher psychological processes* (Harvard university press, 1980).

[54] Y. Demiris, *Knowing when to assist: Developmental issues in lifelong assistive robotics,* in *International Conference of the IEEE engineering in medicine and biology society (EMBC)* (IEEE, 2009) pp. 3357–3360.

[55] A. Cully and Y. Demiris, *Online knowledge level tracking with data-driven student models and collaborative filtering,* (under review at) Transactions on Knowledge and Data Engineering (2018).

[56] B. Kiefer, A. Welker, and C. Biwer, *VOnDA: A Framework for Ontology-Based Dialogue Management,* in *International Workshop on Spoken Dialogue Systems Technology (IWSDS)* (Springer, 2019).

[57] D. R. Traum and S. Larsson, *The information state approach to dialogue management,* in *Current and new directions in discourse and dialogue* (Springer, 2003) pp. 325–353.

[58] F. Burger, J. Broekens, and M. A. Neerincx, *Fostering relatedness between children and virtual agents through reciprocal self-disclosure,* in *BNAIC 2016: Artificial Intelligence*, edited by T. Bosse and B. Bredeweg (Springer International Publishing, Cham, 2017) pp. 137–154.

[59] P. Carey, *Data protection: a practical guide to UK and EU law* (Oxford University Press, Inc., 2018).

[60] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, *Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system,* Computers and biomedical research **8**, 303 (1975).

[61] S. R. Haynes, M. A. Cohen, and F. E. Ritter, *Designs for explaining intelligent agents,* International Journal of Human-Computer Studies **67**, 90 (2009).

[62] J. D. Lee and K. A. See, *Trust in automation: Designing for appropriate reliance,* Human factors **46**, 50 (2004).

[63] B. M. Muir, *Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems,* Ergonomics **37**, 1905 (1994).

[64] D. N. Lam and K. S. Barber, *Comprehending agent software,* in *Autonomous Agents and Multiagent Systems* (2005) pp. 586–593.

[65] C. Conati and K. VanLehn, *Providing adaptive support to the understanding of instructional material,* in *Proceedings of the 6th international conference on Intelligent user interfaces* (ACM, 2001) pp. 41–47.

[66] L. R. Ye and P. E. Johnson, *The impact of explanation facilities on user acceptance of expert systems advice,* Mis Quarterly , 157 (1995).

[67] B. Y. Lim, A. K. Dey, and D. Avrahami, *Why and why not explanations improve the intelligibility of context-aware intelligent systems,* in *Human Factors in Computing Systems* (2009) pp. 2119–2128.

[68] Y. Cheon and G. T. Leavens, *A simple and practical approach to unit testing: The jml and junit way,* in *European Conference on Object-Oriented Programming* (Springer, 2002) pp. 231–255.

[69] P. C. Jorgensen and C. Erickson, *Object-oriented integration testing,* Communications of the ACM **37**, 30 (1994).

[70] B. Beizer, *Software system testing and quality assurance* (Van Nostrand Reinhold Co., 1984).

[71] E. J. Weyuker and F. I. Vokolos, *Experience with performance testing of software systems: issues, an approach, and case study,* IEEE transactions on software engineering **26**, 1147 (2000).

[72] M. A. Neerincx, W. van Vught, O. Blanson Henkemans, E. Oleari, J. Broekens, R. Peters, F. Kaptein, Y. Demiris, B. Kiefer, D. Fumagalli, and B. Bierman, *Socio-cognitive engineering of a robotic partner for child's diabetes self-management,* Frontiers in Robotics and AI **(in press)** (2019).

[73] O. Blanson Henkemans, E. Oleari, C. Pozzi, D. Baranzini, S. Pal, M. Bakker, W. van Vught, M. A. Neerincx, J. Broekens, R. Peters, F. Kaptein, Y. Demiris, B. Kiefer, D. Fumagalli, B. Bierman, R. Bonfanti, A. Rigamonti, M. Schouten, and G. J. van der Burg, *Robotic playmate to develop diabetes self-management competency and performance during childhood: A randomized controlled trial,* (to appear).

[74] T. Georgiou and Y. Demiris, *Adaptive user modelling in car racing games using behavioural and physiological data,* User Modelling and User-Adapted Interaction **27**, 267 (2017).

[75] V. Surakka, M. Illi, and P. Isokoski, *Chapter 22 - voluntary eye movements in human—computer interaction,* in *The Mind's Eye*, edited by J. Hyönä, R. Radach, and H. Deubel (North-Holland, Amsterdam, 2003) pp. 473 – 491.

[76] G. Underwood, *Cognitive Processes in Eye Guidance* (Oxford University Press, 2005).

[77] O. Celiktutan and Y. Demiris, *Inferring human knowledgeability from eye gaze in mobile learning environments,* in *European Conference on Computer Vision Workshops (ECCVW* (2018).

[78] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, *Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds,* in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2019) pp. 5866–5870.

3

Personalised Self-Explanation by Robots: The Role of Goals versus Beliefs in Robot-Action Explanation for Children and Adults

A good explanation takes the user who is receiving the explanation into account. We aim to get a better understanding of user preferences and the differences between children and adults who receive explanations from a robot. We implemented a Nao-robot as a belief-desire-intention (BDI)-based agent and explained its actions using two different explanation styles. Both are based on how humans explain and justify their actions to each other. One explanation style communicates the beliefs that give context information on why the agent performed the action. The other explanation style communicates the goals that inform the user of the agent's desired state when performing the action. We conducted a user study (19 children, 19 adults) in which a Nao-robot performed actions to support type 1 diabetes mellitus management. We investigated the preference of children and adults for goal- versus belief-based action explanations. From this, we learned that adults have a significantly higher tendency to prefer goal-based action explanations. This work is a necessary step in addressing the challenge of providing personalised explanations in human-robot and human-agent interaction.

Parts of this chapter have been published in [1]

3.1. Introduction

Explainable Artificial Intelligence (XAI) is the capability of a system to explain its own behaviour. XAI is known to have a positive influence on user trust in and understanding of the intelligent systems [2–4]. Intelligent systems are becoming increasingly complex, which makes it difficult for the users to understand the system's actions [5]. XAI is important in areas such as medical support [2], fire-fighting [6, 7], and education [4].

A theoretical approach towards explaining actions is the *intentional stance*. When adopting the intentional stance, one assumes that actions result from intentions of the *actor* (i.e., the human or agent performing the action) [8]. In everyday human communication, two common explanation styles for intentional actions are goal-based and belief-based explanations [9]. A goal-based explanation communicates the actor's desired outcome of the action. A belief-based explanation provides information about the context and the circumstances that caused the actor to choose one action over another. How humans explain an actor's actions in everyday communication is referred to as *folk psychology* [10, 11], and has been used in developing XAI for BDI-based (belief, desire intention) agents [7, 12–14].

A good explanation is *personalised*, i.e., it takes the user that is receiving the explanation into account. As we mature, we develop our capabilities to create and understand explanations for someone else's actions [9, 15, 16]. Furthermore, different educational strategies are required for adults and children [17, 18]. Therefore, self-explaining robots and agents that educate their users are likely to need different explanation strategies for children and adults.

To address this challenge of personalising explanations, we compared two different explanation styles on two different user groups. We constructed goal-based and belief-based robot-action explanations. We then asked both children and adults what explanation best helped them to understand the different actions. We tested whether these different explanation styles significantly differ on this metric, taking user group into account as a factor.

The context of this thesis and of this chapter is the PAL (a Personal Assistant for a Healthy Lifestyle) project. This project helps children (aged 7-14) to cope with type 1 diabetes mellitus. In this project, we develop an agent controlling a Nao-robot or its virtual avatar. The system autonomously interacts with the children and their parents for prolonged periods of time. It helps the children to cope with their medical health issues. Therefore, it is important that the different users trust and understand the actions of the PAL-agent. To facilitate this, we develop the capability to explain these actions to the different users. Consider the following robot action: '*the PAL-robot tells the child that one has a hypo when one's blood glucose level is below 4.0 millimoles per litre*'. We may explain this by saying the robot wants to teach the child how to detect and treat a hypo (goal); or, that it thinks the child does not know what blood measurements indicate that one has a hypo (belief).

We will first review related work in the field of XAI, and, in particular, work that focuses on self-explaining agents in Section 3.2. Then, in Section 3.3, we describe a generally applicable representation of a BDI-based agent's decision making and how we derive belief-based and goal-based explanations from this. We explain the

set-up of our experiment in Section 3.4. Finally, we present the results and discuss them.

3.2. Motivation for Research Conducted

When developing intelligent agents, one should consider enhancing them with self-explaining capabilities. Previous studies have shown that XAI enhances user trust in and understanding of intelligent systems [2–4, 19, 20]. This is especially important for intelligent agents because they are often designed to operate semi-autonomously, and they often operate in consequential domains like medicine or military [21].

Recent work on XAI for intelligent agents used automatically generated folk psychology based explanations [7, 12, 13]. A folk psychology based explanation communicates the beliefs and goals that led to the agent's behaviour. One adopts the *intentional stance*, meaning one explains the agent's action by explaining the *reasons* (beliefs and goals) for the agent's intention [8]. Example folk psychology based explanations are, 'I proposed to play a sorting game together because I thought that you liked that game' (belief); or, 'I proposed to play a sorting game together because I wanted to play a game with you' (goal). Folk psychology based explanations provide concise, human-like explanations of an agent's actions. They are mainly directed at the end-users of the intelligent system (see, e.g., [14, 22] for explanations more directed towards agent developers).

Previous work on XAI took user knowledge into account [23, 24]. These studies classified a user as a beginner or expert and used this to provide explanations that better fit the individual user's preferences. However, we belief more elaborate user models are required for good personalised explanations. In this chapter, we compare two common explanation styles for two distinct user groups: children and adults.

3.2.1. Goal-based and Belief-based Explanations

Folk psychology is how humans in everyday communication explain and predict intentional actions [9, 10]. Two common explanations styles in folk psychology are goal-based and belief-based explanations [8–11]. A goal-based explanation communicates the actor's desired outcome of the action. It provides an answer to the questions, 'To what end?' or 'For what purpose?' A belief-based explanation provides information on why the actor chose a certain action over another. It provides information about the context and the circumstances. Goals are easier to infer from the action itself, whereas beliefs provide information specific to the particular actor that performed the action and context in which the action was performed. Malle [9] writes that to infer an actor's belief, one needs to take the *perspective* of this particular actor.

3.2.2. Hypothesis

In this chapter, we define two explanation algorithms: one that always provides the triggering condition (belief) that caused the agent to perform the action and

one that always provides the parent goal that the agent is trying to achieve. We test which explanation algorithm is preferred by adults and children by presenting them example actions explained by these algorithms. Our hypothesis is:

Hypothesis. *Adults have a stronger preference than children for goal-based over belief-based social robot explanations.*

There is psychological support for this hypothesis. First, a difference in itself is likely because explanations based on folk psychology change as humans mature [9, 16]. For example, young children (4 years old) have trouble realising someone may have a belief that is false [15]. Second, children and adults alike are inclined to belief that others have similar beliefs and knowledge as they do [16]. However, adults have accumulated a vast amount of knowledge to which they can link new information [18]. Third, adults strongly desire (more than children) to know the goals you are pursuing when educating them [17, 18].

3.3. Goal Hierarchy Trees

Previous work on XAI showed how one can use an agent's beliefs and goals for generating action explanations [13, 25]. However, this means that one should take special care in designing and formulating the beliefs and goals of the agent [13, 26]. Previous work proposed the use of a goal hierarchy tree (GHT) to develop a high level design of the agent's reasoning [7, 12] and provides guidelines for their development [13]. GHTs are based on hierarchical task analysis, a technique from cognitive psychology used to specify complex human tasks [27]. In this section, we describe the structure of a GHT, and how we can construct explanations from it. We adopt GHTs as agent design and use these to test how one can personalise explanations for different users.

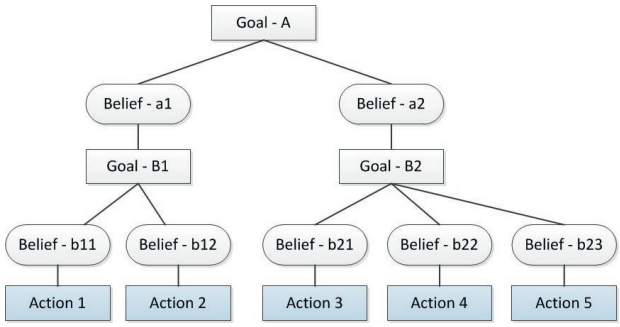


Figure 3.1: The square nodes are goals the agent adopts. The top node is the agent's main goal. When following the edges, sub-goals are represented that the agent adopts to achieve the main goal. Actions to achieve (sub-)goals are shown as shaded square nodes (the leaves) of the tree. Conditions (beliefs) that cause an agent to adopt a sub-goal or perform a particular action are shown as rounded nodes.

3.3.1. The Structure of a Goal Hierarchy Tree

Figure 3.1 shows the structure of a GHT. A BDI-based agent that runs in accordance with a GHT chooses actions as follows. Based on the agent's current goals and beliefs, it chooses an action to perform. If multiple actions are applicable, then the agent randomly chooses one. When no actions are applicable then the agent remains idle until its beliefs change, which can cause it to adopt new goals and can make new actions become applicable.

A GHT does not model what external events can occur and how events and agent-actions cause the agent to update its beliefs. Rather, the GHT shows a high-level design of the agent's *reasoning*, i.e., what action it should perform given a current state of beliefs and goals. This is sufficient for our purpose of generating explanations based on an agent's reasoning. However, if one wants to run a BDI-based agent that acts in accordance with the GHT, then this additional modelling is required.

3.3.2. Goal-based and Belief-based Agent-action Explanations

One can explain an action by means of the goal one aims to achieve, or by explaining why it was possible to perform the action (belief) [9, 11]. Explanations should not be too long [13, 16]. We need to be selective in what beliefs and goals we communicate. In this chapter, we use the following explanation algorithms.

The **belief-based explanation algorithm** selects the belief directly above the action (triggering condition). For example, action-2 is explained by Belief-b12 and action-3 by Belief-b21 (Figure 3.1).

The **goal-based explanation algorithm** selects the goal directly above the action (parent goal). For example, action-2 is explained by Goal-B1 and action-3 by Goal-B2 (Figure 3.1).

We use these algorithms for the sake of testing the hypothesis of Section 3.2.2. Both explanation algorithms consist of one element (one belief or one goal) in the goal hierarchy tree. The goal-based explanation provides the most direct response to the question, 'What is your purpose?' The belief-based explanation provides the most direct answer to why the agent chose a particular action over another. Thus, they closely resemble how humans explain their actions as discussed in Section 3.2.1. Furthermore, they are short and thus unlikely to flood the receiver of the explanation with too much information [13, 16]. Therefore, we can use these algorithms to determine a preference for belief-based or goal-based agent-action explanations in a general way.

3.4. User Study

We developed a GHT within the context of the PAL-project and set up an experiment using the explanation algorithms from Section 3.3.2. We tested whether goal-based or belief-based action explanations are better received by the participants (children and adults). We tested for a significant difference in preference within and between these user groups.

We belief the PAL-project is a good domain for testing XAI. Firstly, it provides

Table 3.1: Distribution of children and adults over the 4 conditions

	Random Seed of Scenarios			
	Normal Order of Explanations		Reversed Order of Explanations	
	Normal Order of Scenarios	Reversed Order of Scenarios	Normal Order of Scenarios	Reversed Order of Scenarios
Children	6	5	4	4
Adults	5	4	5	5

us with both children and adults that interact with the agent; and secondly, it is an exemplary, consequential domain were the agent interacts with its users for prolonged periods of time (the type of domains where XAI is especially important [21]).

3.4.1. Participants

The participants were recruited from a diabetes camp for children. Children diagnosed with type 1 diabetes mellitus were recruited by the Dutch Diabetes Association DVN. These children and their parents were invited to participate in our experiment. Rejecting this had no influence on participation in other activities during the camp.

Participants with diabetes were a good choice for testing our hypothesis. First, diabetes is a use case within the health domain which is of particular relevance to explainable AI [21]. Second, the transparency provided by personalised explanations can help children trust and understand how the system tries to help them [2–4, 19, 20], which is in line with the societal- and research goals within the PAL project.

In total, there were 21 children and 20 adults (parents of the children) present at the camp. One child did not participate in the experiment. One child participated, but was looking over his friend's shoulder while filling in answers. One adult did not fill in the initial sheet asking for data like age, gender, and education. This left us with 19 children (12 male, aged 8-11) and 19 adults (8 male, aged 35-48).

3.4.2. Designing a Goal Hierarchy Tree

Figure 3.3 presents our design of a GHT for our agent. This GHT is a translation from its Dutch counterpart, since the experiment was performed in the Netherlands. Based on this GHT, the agent chose different actions to support a child in diabetes management.

The GHT specifies two styles of support. The agents aims to educate the child when the child is in a good mood to learn new things. The agent aims to cheer up the child when the child is sad. We call this *cognitive support*, and *affective support*, respectively. The way that this agent provides these types of support is defined by ontologies that were developed in cooperation with health-care professionals [28]. The here developed GHT resembles the treatment plan provided by these experts.

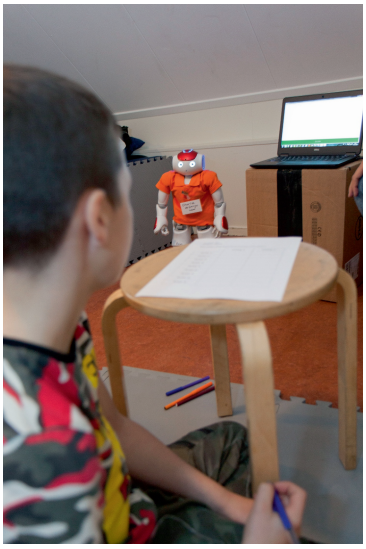


Figure 3.2: Set-up of the experiment. The Nao-robot verbally presents example scenarios and provides two explanations for each of these. The screen textually shows what the Nao-robot is saying, so the child can always read-back on the screen what happened. The child then puts a mark at the most preferred explanation.

3.4.3. Set-up & Materials

The GHT shown in Section 3.4.2 has nine different robot actions. These actions can all be explained by using the belief-based explanation algorithm or by using the goal-based explanation algorithm. A Nao-robot presented all the actions to the participants. For each action, it proposed two explanations obtained from the algorithms. The participants had a forced choice to prefer either one of the proposed explanations.

The robot was located in front of the participants and next to a laptop screen (Figure 3.2). For every action and corresponding explanations presented by the Nao-robot, the laptop screen showed the action performed and explanations provided. In this way, the participants could read back what the robot said. For example, the screen could look like:

Action: 'I tell Jimmy to take dextrose when he is having a hypo.'
Explanation 1: 'Jimmy does not yet know how to correct his blood-sugar when he has a hypo.'
Explanation 2: 'I want to teach Jimmy how to successfully cope with hypos.'

By using the robot, the children experience the experiment as a fun activity rather than a chore. Robots have been shown to have a positive impact on motivation and learning [29]. We chose to also use a screen, since Nao-robots do not always pronounce words very well. By using a screen, the participants can always read back what the robot said.

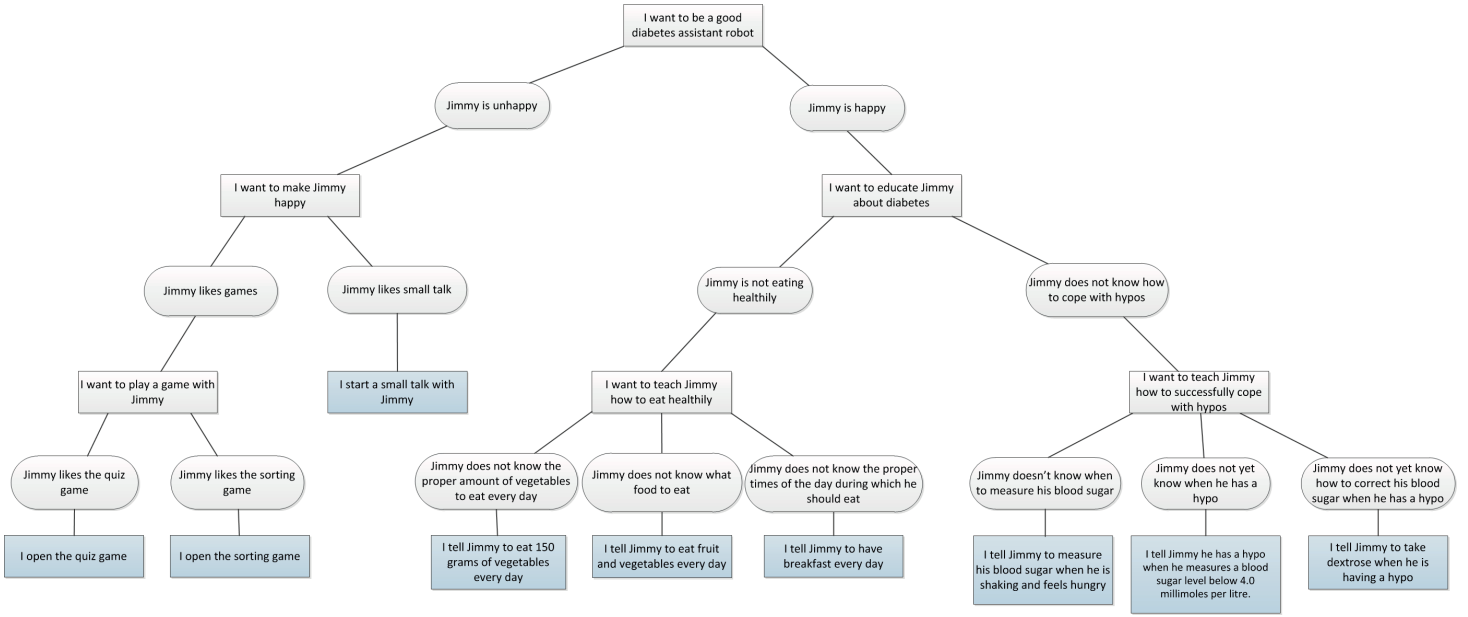


Figure 3.3: In this figure, the goal hierarchy tree of the PAL agent is shown. The rectangular nodes are goals the agent adopts. The top node is the agent's main goal. When following the edges, sub-goals are represented that the agent adopts in order to achieve the main goal. The triggering conditions (beliefs) that determine whether the agent should adopt a sub-goal are represented in rounded nodes. The agent's actual actions are the shaded nodes (the leaves) of the tree.

3.4.4. Variables & Design

The presentation of an action including the two explanations is henceforth referred to as a *scenario*. A scenario starts with the robot saying: 'I performed action a'. With a being one of the actions in the GHT and phrased exactly as shown in the GHT (Figure 3.3). Then, the robot says: 'How should I explain this? One: $explanation$-1; or two, $explanation$-2'. Where explanation 1 and 2 are the belief-based explanation and the goal-based explanation as explained in section 3.3.2. The exact texts are thus also depicted in the GHT.

A participant is shown nine scenarios, one for each action, in random order. Whether the participant was shown the belief-based explanation first or the goal-based explanation first was also chosen randomly for every scenario. After the experiment, we counted the *percentage of scenarios where the participant preferred a goal-based explanation*. So, if the participant preferred the belief-based explanation in six scenarios and the goal-based explanation in the other three scenarios, then this variable is 33% for the participant.

Due to the camp setting, we were not able to do the experiment with every user separately. We were forced to have the participants do the experiment in small groups (group size of 2-3 for the children, 4-5 for the adults). The individuals in the groups were not allowed to discuss amongst each other nor look at each other's answers before the experiment was over. However, a consequence of having groups was that participants in the same group also saw the same order of actions and the same order of explanations. Thus, we counterbalanced the conditions. We produced a single random seed of scenarios. The actions were first put in random order. For every action separately, the system then randomly chose explanation 1 to be belief-based and explanation 2 goal-based, or vice versa. We counterbalanced this among the participants. I.e., the participants saw the actions in this randomly chosen order, or they saw the actions in reversed order. Furthermore, they saw the order of the explanations in this randomly chosen order, or they saw the explanations for these actions in reversed order. The participants were evenly distributed over these 4 conditions (see Table 3.1).

3.4.5. Procedure

In small groups, the participants were asked to enter the room and were seated in front of the robot and laptop. The researcher informed the participants that he would remain present during the experiment but that the robot would guide the experiment. Additional questions could be directed to the researcher.

The Nao-robot started the experiment with a small presentation. Here, it told the participants that it wants to learn how to explain its behaviour to them and that it needs their input. It explained that it will provide example scenarios where it helped a fictional child 'Jimmy' to deal with diabetes. In this starting presentation, the robot said it sometimes plays a game with Jimmy and sometimes tries to educate Jimmy concerning diabetes management. The robot said that in all the example scenarios it wants to explain its action and always considers two possible explanations. The participants were then asked to select the explanation that best helped them to understand why the PAL-robot performed that action.

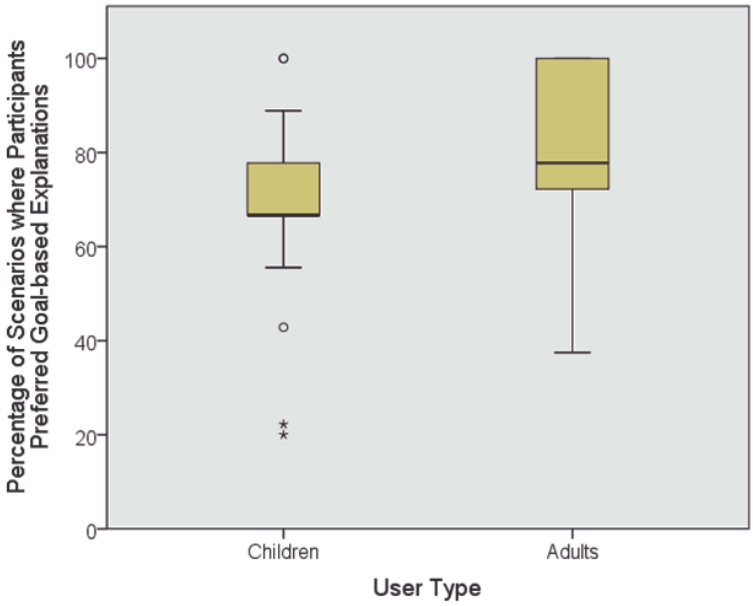


Figure 3.4: A box plot showing the distribution of preferring goal-based explanations over belief-based explanations. On the x-axis the two user types (children and adults) are depicted. The y-axis shows the percentage of scenarios where the subject preferred the goal-based explanation over the belief-based explanation. Adults have a significantly higher preference for goal-based action explanations (Median = 0.778) than children (Median = 0.667).

After the presentation the robot verbally presented the nine scenarios, one for each action, and the screen showed the scenarios in text. With the children, the researcher paid special attention to prevent them from looking over each other's shoulder while choosing the best explanation. Once the experiment was finished the robot and researcher thanked the participants for their help.

3.5. Results

To test the preference towards the different explanation styles, we counted the percentage of scenarios where the participants preferred a goal-based explanation. A one-sample Wilcoxon signed rank test shows that the median of preferring goal-based explanations, rather than belief-based explanations, is significantly above 50% for children ($med = 0.667, 95\%\ CI = [0.667, 0.778],\ p = .007$) and adults ($med = 0.778, 95\%\ CI = [0.667, 1.0],\ p < .000$). So, both user groups significantly prefer goal-based explanations over belief-based explanations. Figure 3.4 shows the distributions of preferring goal-based explanations for children and adults.

Furthermore, a Mann-Whitney test indicated that the preference towards goal-based explanations, rather than belief-based explanations, was greater for adults ($med = 0.778$) than for children ($med = 0.667$), U = 112.5, p = .042, r = .33. Adults prefer goal-based explanations significantly more than children.

3.6. Discussion

The results in the previous section show that there is a significant preference for goal-based explanations in both user groups. However, it would be premature to state that goal-based explanations are always preferable. Previous studies have shown contradicting findings on this subject. The work in [13] analyses three studies that provide explanations based on a goal hierarchy tree. Two studies in a firefighting domain [6, 7], (one with experts and one with laymen) and one in a cooking domain [12] (where users were perceived as experts, since they all knew how to cook). In the non-expert domain, the participants showed a preference for belief-based explanations. In the expert domains, goal-based explanations were preferred. It is hypothesized in [13] that this difference may be due to the expert level. Since both the children (who have diabetes mellitus themselves) and the adults (their parents) are familiar with the domain, it can be expected that these users would prefer goal-based explanations. Future work should further explore this by testing this domain on layman (adult & child) users and comparing the results with the here presented findings.

Another finding in the results is that adults significantly prefer goal-based explanations more than their children. This has two possible explanations. First, according to adult learning psychology, adults need to know the objectives of the instruction [17, 18]. They are goal-oriented learners that rely on their vast personal experience. They prefer to know how instructions help them to enhance their existing abilities, rather than children who learn under the assumption that all instructions will help them sometime in the future [17]. Adults thus prefer knowing the objectives (goals) that the robot is pursuing when performing actions to educate its user.

The second explanation is that children are more motivated to understand a robot character. To infer an agent's belief, one needs to take the *perspective* of this particular agent [9]. Children and adults alike are better at perspective taking when their motivation to do this is high [30]. A higher motivation then correlates with a better understanding of belief-based explanations. On the other hand, adults are better adapted to perspective taking than children. Adults are faster at adjusting when they learn their initial perspective is incorrect [30]. Being an adult then also suggests a better understanding of belief-based explanations, since the adult is more flexible at adjusting her perspective to match that of the agent. In conclusion, if perspective taking is the explanation for these results, then motivation of the children must have had a stronger influence than the flexibility of the adults.

There are three sources that potentially limit the generalisability of our results. First, the chosen scenarios depicted in the GHT (Figure 3.3) potentially have traits that we are unaware of but that influence the preference for an explanation style. Second, the PAL-agent aims to educate its users on type 1 diabetes mellitus and maintain a positive mood in the user. A domain that resembles this type of system behaviour may be more likely to find similar preferences for explanation styles. Third, the children and adults may also have traits that are not representative for the entire population (e.g., culture) and that influence the preference for an explanation style. There are, however, many similarities between our user groups (i.e., they

both face the problem of managing diabetes, they work with the same caregivers at the same hospitals, they even share the same genes). Within our sample space, we therefore belief that child/ adult is the only factor responsible for this difference in preferred explanation style. However, to address this issue of generalisability, future work includes replicating our study with more divers scenarios, contexts, and users.

The presented experiment was a start. In future work, we will systematically expand the design space. For example, individual user preferences and differences in the context in which the agent performed the action can also have an influence on how one should construct the explanation. Furthermore, one can combine goal-based and belief-based explanations providing the user with more information. However, explanations should not become too long [16]. The explainer should thus be careful with when and how to add further information to an explanation. Finally, we tested the subjective preference towards the explanation styles. A next step is to test how this influences user behaviour and trust in the system.

3.7. Conclusion

In this chapter, we compared the preference for goal-based versus belief-based social robot action explanations between two user groups. We presented children and adults with a set of example robot actions and provided two possible explanations for these. Belief-based explanations communicated the context (a belief) preceding the decision to perform an action. Goal-based explanations provided the agent's purpose (a goal) of the action. The users were asked to choose the explanation that *best helped them to understand why the Nao-robot performed this action*.

We found that adults have a significantly *higher* preference for goal-based explanations than children. This is the first evidence that self-explanations of intelligent agents are perceived differently by children and adults. This work is a necessary step towards providing *personalised explanations* in human-robot and human-agent interaction.

References

[1] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, *Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults,* in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on* (IEEE, 2017) pp. 676–682.

[2] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, *Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system,* Computers and biomedical research **8**, 303 (1975).

[3] D. N. Lam and K. S. Barber, *Comprehending agent software,* in *Autonomous Agents and Multiagent Systems* (2005) pp. 586–593.

[4] C. Conati and K. VanLehn, *Providing adaptive support to the understanding of instructional material,* in *Proceedings of the 6th international conference on Intelligent user interfaces* (ACM, 2001) pp. 41–47.

[5] M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, *Building explainable artificial intelligence systems,* in *Innovative Applications of Artificial Intelligence* (2006) pp. 1766–1773.

[6] M. Harbers, K. van den Bosch, and J.-J. C. Meyer, *A study into preferred explanations of virtual agent behavior,* in *International Workshop on Intelligent Virtual Agents* (Springer, 2009) pp. 132–145.

[7] M. Harbers, K. Van den Bosch, and J.-J. Meyer, *Design and evaluation of explainable bdi agents,* in *Web Intelligence and Intelligent Agent Technology* (2010) pp. 125–132.

[8] D. C. Dennett, *Three kinds of intentional psychology,* in *Reduction, Time and Reality*, edited by R. Healey (Cambridge University Press, Cambridge, 1981) pp. 37–61.

[9] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction.* (MIT Press, 2004).

[10] P. M. Churchland, *Folk psychology and the explanation of human behavior,* The future of folk psychology: Intentionality and cognitive science , 51 (1991).

[11] B. F. Malle, *How people explain behavior: A new theoretical framework,* Personality and social psychology review **3**, 23 (1999).

[12] J. Broekens, M. Harbers, K. Hindriks, K. Van Den Bosch, C. Jonker, and J.-J. Meyer, *Do you get it? user-evaluated explainable bdi agents,* in *Multiagent System Technologies* (Springer, 2010) pp. 28–39.

[13] M. Harbers, J. Broekens, K. Van Den Bosch, and J.-J. Meyer, *Guidelines for developing explainable cognitive models,* in *International Conference on Cognitive Modeling* (2010) pp. 85–90.

[14] K. V. Hindriks, *Debugging is explaining,* in *International Conference on Principles and Practice of Multi-Agent Systems* (Springer, 2012) pp. 31–45.

[15] H. W. H. Mayringer, *False belief understanding in young children: Explanations do not develop before predictions,* International Journal of Behavioral Development **22**, 403 (1998).

[16] F. C. Keil, *Explanation and understanding,* Annual Review of Psychology **57**, 227 (2006).

[17] M. S. Knowles *et al.*, *The modern practice of adult education*, Vol. 41 (New York Association Press New York, 1970).

[18] S. Lieb and J. Goodlad, *Principles of adult learning,* (2005).

[19] L. R. Ye and P. E. Johnson, *The impact of explanation facilities on user acceptance of expert systems advice,* Mis Quarterly , 157 (1995).

[20] B. Y. Lim, A. K. Dey, and D. Avrahami, *Why and why not explanations improve the intelligibility of context-aware intelligent systems,* in *Human Factors in Computing Systems* (2009) pp. 2119–2128.

[21] S. R. Haynes, M. A. Cohen, and F. E. Ritter, *Designs for explaining intelligent agents,* International Journal of Human-Computer Studies **67**, 90 (2009).

[22] A. Hedhili, W. L. Chaari, and K. Ghédira, *Causal maps for explanation in multi-agent system,* in *Intelligent Informatics* (Springer, 2013) pp. 183–191.

[23] S. Gregor and I. Benbasat, *Explanations from intelligent systems: Theoretical foundations and implications for practice,* MIS quarterly , 497 (1999).

[24] G. Milliez, R. Lallement, M. Fiore, and R. Alami, *Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring,* in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, 2016) pp. 43–50.

[25] J. Broekens, D. DeGroot, and W. A. Kosters, *Formal models of appraisal: Theory, specification, and computational model,* Cognitive Systems Research **9**, 173 (2008).

[26] G. Taylor, K. Knudsen, and L. S. Holt, *Explaining agent behavior,* in *Behavior Representation in Modeling and Simulation* (2006).

[27] J. M. Schraagen, S. F. Chipman, and V. L. Shalin, *Cognitive task analysis* (Psychology Press, 2000).

[28] M. A. Neerincx, F. Kaptein, M. A. van Bekkum, H.-U. Krieger, B. Kiefer, R. Peters, J. Broekens, Y. Demiris, and M. Sapelli, *Ontologies for social, cognitive and affective agent-based support of child's diabetes self-management,* Artificial Intelligence for Diabetes , 35 (2016).

[29] R. Looije, M. A. Neerincx, and V. d. Lange, *Children's responses and opinion on three bots that motivate, educate and play,* (2008).

[30] N. Epley, C. K. Morewedge, and B. Keysar, *Perspective taking in children and adults: Equivalent egocentrism but differential correction,* Journal of Experimental Social Psychology **40**, 760 (2004).

4

Evidence for the Use of Emotion in Human Explanations of Robot and Human Behaviour

The main question addressed in this chapter is whether and how humans use emotions in their explanations of humanoid robot behaviour. Addressing this is important because it: (1) helps us design how robots can explain their own actions; and (2) gives insight into human attribution of mental states to robots. We presented filmed behaviours of a human or a (humanoid) robot coping with a distressing situation to MTurk participants. Between-subjects, behaviours were shown in different coping styles (based on literature), performed by the human or robot actor type, and in a health or museum scenario. Participants rated their recognition of these coping styles and provided a textual explanation for the behaviour. Results show that participants recognised most coping styles (two were not recognised properly, one style was recognised in humans but not in robots). Furthermore, participants used emotions in their explanations for both the robot and human actor, and the recognition of the coping style correlated with the emotionality of the explanations for the human actor type, but not for the robot. These results, for the first time, show that emotions are used in the explanation of robot and human behaviour, and that coping styles are recognised in robot behaviour.

This paper is submitted to ACM Transactions on Human-Robot Interaction (THRI), special issue on explainable AI

4.1. Introduction

In the near future robots will perform support functions that involve interaction with humans in domains including education, health care and hospitality [1]. The behaviour and decision making of these robots will become complex. A key problem related to this complexity is that if a user does not understand how a system or robot makes a decision, that user's trust in the system is impacted as well, which can lead them to misuse and abandonment of the interaction [1, 2]. A common approach to increase a robot's intelligibility and user's trust in the robot is by providing users with explanations of the behaviour [3–9]. This is referred to as eXplainable AI (XAI).

A common way to develop XAI for intelligent agents (like humanoid robots and avatars), is by basing it on how humans explain behaviour amongst each other [9], (i.e., folk psychology [10, 11]). For example, you see a man running to cross a busy street. If you are asked to answer the question "why?" then you tend to explain that by "the man does not *want* to be run over by a car", explaining the behaviour with the goal of staying alive. Folk psychology can be used in human-agent/human-robot interaction in two ways. First, one can use folk psychology as a basis to generate *self-explanations* of behaviour by humanoid robots [3, 5, 7]. Second, one can use folk psychology as a framework to analyse *people's explanations* of robot behaviour as a proxy for the user's perception of the underlying intentional structure [8]. Both approaches can provide insight into how self-explanations by robots should, in principle, be designed. The former by testing the effects of explanations on the users (e.g., the effect on trust); the latter by analysing the structure of people's explanations and using that as input to generate more human-like self-explanations.

The majority of the research in self-explanation and people's explanation ignores the role of emotion in the explanation of behaviour and instead focuses on beliefs, desires, goals and other more "cognitive" constructs [3, 6–8, 12, 13]. However, it has been argued that emotions play an essential role in *human* explanations [14]. Also for robots and virtual agents, it has been argued that emotions might play an important role in the generation of self-explanations for robot behaviour [15].

The main question addressed in this chapter is whether and how humans use emotions in their explanations of robot behaviour and how these explanations differ for similar human behaviour. To study this question, we focus on behaviour resulting from coping strategies for the following reason. Coping strategies are triggered by emotion and are aimed at emotion regulation [16]. The use of emotion in the explanation of the resulting behaviours can serve as existence proof. If people *do* explain robot behaviour using emotions, then this should be observed in behaviours resulting from coping strategies. If people *do not* use emotions for explanations of this behaviour then that is a strong indicator that they also won't use it when the robot shows other types of (semi-)intentional behaviour.

We presented filmed behaviours of either a robot or a human actor coping with a distressing situation to Amazon Mechanical Turk participants. The actors applied a coping style in their behaviour from the set of styles that the literature distinguishes, alternating the style over participants. We asked participants to rate their recognition of coping styles in this behaviour and how they would explain the behaviour.

We further investigated the extend to which emotions are used in the participant's explanations of the behaviour, and whether this depends on the coping style and actor type (robot versus human).

4.2. Background and Related Work

EXplainable Artificial Intelligence (XAI) is a sub-field of human-agent interaction. It has its roots in Artificial Intelligence (AI), human-computer interaction (HCI), and the social sciences [17]. Early studies in expert systems suggest that explanations are important for acceptance of, and trust in the system's decisions, particularly in domains where decisions are judgemental and consequential (e.g., health-care) [2, 18, 19]. These findings have been replicated in studies with modern intelligent systems [4, 20, 21]. With the introduction of even more advanced artificial intelligence, transparency and explanation have again become increasingly important topics in human-agent interaction [9] and machine learning [22]. This importance is further emphasised by the General Data Regulation Law (GDPR) [23].

Current work in EXplainable AI (XAI) in artificial agents and robots typically presents the AI's reasoning in a reduced complexity form. First, the system's reasoning process is queried, then that information is presented to the user[3, 24, 25]. Most approaches in human-agent interaction use cognitive constructs such as beliefs, desires, intentions and goals to explain the actions of the agent [6, 7, 12, 13, 26].

4.2.1. Explanations and Folk Psychology

Most of the work in XAI for robots and (other) agents is based on folk-psychology [9]. Folk-psychology proposes that humans explain each others' behaviour in terms of mental constructs like beliefs and goals [10, 11, 27]. People use folk psychology to explain behaviour when they assume the behaviour comes from an intentional agent [10]. For example, your colleague took some days off from work (the action to be explained) because he **thought** (belief) there were no immediate deadlines and he **wanted** (goal) to take some time to relax and recuperate. XAI based on folk psychology typically explains behaviour based on the system's beliefs and goals.

Previous work in XAI confirms that folk-psychology provides a solid basis for agent explanations. Several early studies showed that humans are able to identify which beliefs and goals should be used by an intelligent agent [13], and that humans use beliefs and goals to explain robot behaviour [8, 28]. In addition, several studies have investigated people's mental state ascriptions to robots [8, 17, 28–32]. For example, in [8] participants were shown textual descriptions of robot behaviour. Participants used beliefs and goals when explaining the described behaviour; and, participants sometimes referred to the robot as a programmed machine. Usage of emotions in the explanations was not reported. Wortham et. al. [32] showed people video's of human-robot interaction. They found that robot explanations indeed improved the user's mental model's accuracy, but did not test the occurrence of mental constructs (beliefs, goals, emotions) in the participant's explanations.

4.2.2. Emotions and Coping Styles

In affective computing, the field that uses emotion in technology [33], there are three main views on emotion that are used most frequently: a categorical view, a dimensional view and an appraisal view. The categorical view proposes to organise emotions into groups of distinct emotion types (e.g., [34]). The dimensional view proposes that all emotions share a common base called *core affect*, usually expressed in one or more continuous scales, for example valence, arousal, and dominance [35]. Appraisal theory proposes that emotions are the result of an assessment of the situation in terms of consequences for the individual [36–38].

Emotions play an important role in coping. In essence, coping is a response to deal with situations that are appraised as personally relevant and taxing, aimed at reducing the resulting distress [16, 39]. Coping is a response to emotions and can mitigate emotions [16, 40]. A prominent view on coping styles including their link to emotions is proposed by Folkman and Lazarus [16, 41]. They define 8 styles of coping in the 'Ways of Coping' (WoC) and designed a questionnaire to measure their prevalence in ones behaviour [41]. Table 4.1 shows the questions per style in that questionnaire.

Because we use these coping styles in the construction of the videos that participants view in our study, we explain these in more detail here. When using the **confrontive** style (C), one tries to solve the problem by confronting the responsible agent. For example, one tries to change the responsible agent's mind and/or expresses anger. When using **distancing** (D), one tries to deal with the distress on a more internal level. One tries to mentally distance oneself from what happened. When using **self-controlling** (S-C), one tries to solve the problem by first of all keeping ones feelings to oneself and by thinking before acting. When using *seeking social support* (Soc-S), one tries to seek help from an external party. When using **accepting responsibility** (A-R), one seeks the fault by oneself and tries to make up for this. When using **escape avoidance** (E-A), one wishes that the situation would simply go away and tries to get away from it. When using **planfull problem solving** (P-S), one aims to solve the underlying problem by making changes in the situation, When using **positive reappraisal** (P-R), one changes ones thinking about the situation and reappraises it more positively.

4.3. Research Questions

In the related work section we showed that many of the agent-based explanation approaches use beliefs and goals to construct explanations. However, it has been argued that humans use emotions when explaining behaviour amongst each other ('I quickly crossed the street out of **fear** for the **angry** looking man') [14]. It has also been proposed that expressing emotions can play a role in robot transparency [42]. Beliefs and goals may be insufficient for generating explanations. Agent self-explanations may also need to consider emotions of the agent and of the agent's user [15, 43].

Because coping strategies are triggered by emotion and are aimed at emotion regulation, the resulting behaviours are a good candidate to investigate whether and how people use emotions for the explanations thereof. If people *do* explain robot

Table 4.1: The Ways of Coping questionnaire [41]

Coping Style	**Description**
Confrontive (C)	1. Stood my ground and fought for what I wanted. 2. Tried to get the person responsible to change his or her mind. 3. I expressed anger to the person(s) who caused the problem 4. I let my feelings out somehow. 5. Took a big chance or did something very risky. 6. I did something which I didn't think would work, but at least I was doing something
Distancing (D)	1. Made light of the situation; refused to get too serious about it. 2. Went on as if nothing had happened. 3. Didn't let it get to me; refused to think too much about it. 4. Tried to forget the whole thing. 5. Looked for the silver lining, so to speak; tried to look on the bright side of things. 6. Went along with fate; sometimes I just have bad luck.
Self-Controlling (S-C)	1. I tried to keep my feelings to myself. 2. Kept others from knowing how bad things were. 3. Tried not to burn my bridges, but leave things open somewhat. 4. I tried not to act too hastily or follow my first hunch. 5. I tried to keep my feelings from interfering with other things too much. 6. I thought about how a person I admire would handle this situation and used that as a model. 7. I tried to see things from the other person's point of view.
Seeking social-support (Soc-S)	1. Talked to someone to find out more about the situation. 2. Talked to someone who could do something concrete about the problem. 3. I asked a relative or friend I respected for advice. 4. Talked to someone about how I was feeling. 5. Accepted sympathy and understanding from someone. 6. I got professional help.
Accepting-responsibility (A-R)	1. Criticized or lectured myself. 2. Realized I brought the problem on myself. 3. I made a promise to myself that things would be different next time. 4. I apologized or did something to make up.
Escape-Avoidance (E-A)	1. Wished that the situation would go away or somehow be over with. 2. Hoped a miracle would happen. 3. Had fantasies or wishes about how things might turn out. 4. Tried to make myself feel better by eating, drinking, smoking, using drugs or medication, etc. 5. Avoided being with people in general. 6. Refused to believe that it had happened. 7. Took it out on other people. 8. Slept more than usual.
Planful problem-solving (P-S)	1. I knew what had to be done, so I doubled my efforts to make things work. 2. I made a plan of action and followed it. 3. Just concentrated on what I had to do next – the next step. 4. Changed something so things would turn out all right. 5. Drew on my past experiences; I was in a similar situation before. 6. Came up with a couple of different solutions to the problem.
Positive-reappraisal (P-R)	1. Changed or grew as a person in a good way. 2. I came out of the experience better than when I went in. 3. Found new faith. 4. Rediscovered what is important in life. 5. I prayed. 6. I changed something about myself. 7. I was inspired to do something creative.

behaviour using emotions, then this should be observed in behaviours resulting from coping strategies. If people *do not* use emotions for explanations of this behaviour then that is a strong indicator that they also won't use it when the robot shows other types of (semi-)intentional behaviour.

The main question addressed in this chapter is therefore whether and how humans use emotions in their explanations of robot behaviour, and how these explanations differ for similar human behaviour. To study this question, we presented M-Turkers filmed behaviours of a (humanoid) robot or human actor, coping with a distressing situation using different coping styles. The construction of these videos and styles is explained in the material section (4.4.4) and more thoroughly in Appendix B.

We are studying emotionality of explanations for human and robot behaviour, where behaviour is in one of several coping styles. If people do not recognise different coping styles then we should not expect coping style to influence the explanations.

It might be that people recognise coping styles in human behaviour but not in robot behaviour. Literature suggests that people do often ascribe human characteristics to nonhuman agents (like, robots) [30]. However, people might not be able to recognise some of the more complex characteristics [29]. The mental model people have of the robot may differ from the mental model they have of the human (even if the behaviour is similar) [29, 44, 45]. Furthermore, people often ascribe *multiple coping styles simultaneously* to behaviour [16, 39, 41] (which could be the case for both human and robot behaviour). Our first sub-question is:

Research Question 1. *Do people recognise coping styles in the behaviours of the actors, and what are the similarities and differences for human and robot behaviours?*

Secondly, we study the *emotionality* of people's (natural language) explanations of robot behaviour. Studying people's explanations has been argued to provide subtle insights in people's mental models of the robots [8]. In our case, ascribing the human characteristic to have emotions as motivation for behaviour. The different coping styles discussed in this section differ in the way the one deals with the distressful situation. Some are very much objective and problem oriented, others aim more specifically at dealing with the negative emotions one experiences [16, 39, 41]. Our second research question is:

Research Question 2. *Does coping style influence emotionality of the explanation, and what are the similarities and differences for human and robot behaviours?*

Finally, we expect that certain coping styles may be natural and appropriate when humans portrait them, but might seem very strange when a robot does it (or the other way around). For example, people may experience more discomfort with a confrontive *robot* than with a confrontive *human*. Or, people might find escape-avoidance unnatural for a human health coach, but appropriate for a robot. Our third question is:

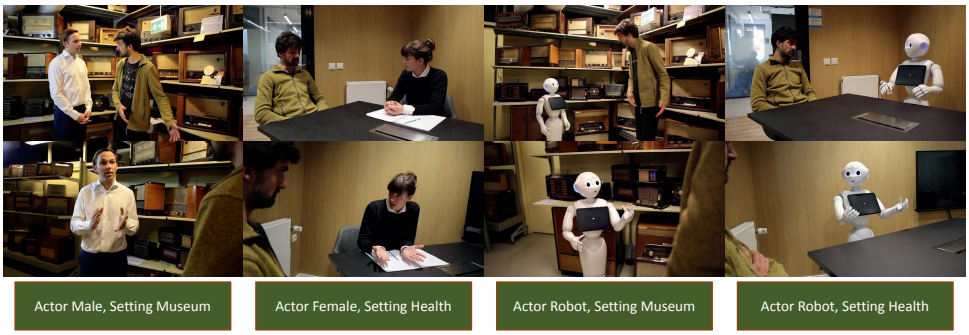


Figure 4.1: Snapshots of the videos of the conversations. The four on the top show snapshots of the initial part of the conversation. After the actor playing Bob distresses Robin, Robin copes with that behaviour in one of the four styles. The four pictures below show snapshots of the coping response by the male actor, female actress, and robot that take the role of Robin.

Research Question 3. *3. Does coping style influence the perception of the behaviour in terms of naturalness, appropriateness, warmth, competence, and discomfort, and what are the similarities and differences for human and robot behaviours?*

4.4. Experiment

4.4.1. Designing Conversations in Coping Styles

We designed several conversations between two individuals. The content of the conversations is such that in all of them one does something that is distressing for the other. The other then copes with that in one of the styles discussed in the previous section. For convenience, we refer to the person causing the distress as 'Bob', and the person coping with that as 'Robin'.

We chose to exclude the 'seek social support' style in our study because we focus on a conversation between two individuals to avoid additional complexity in perspective taking for the participants. Seeking social support would require an additional actor in the scenario. Furthermore, the behaviour is framed within a health-care or public museum scenario (two common human-robot interaction scenarios [1]). In appendix B, we discuss the design and validation of these conversations at length.

Using the validated conversations, we continued to translate them into videos where they were played out by actors. We wanted videos of the conversations rather than textual descriptions to make sure that people have the same mental image of the robot. A textual description like 'the Robot did X' might invoke different mental state images for different participants. Furthermore, non-verbal characteristics and movements might have an influence on the perception. We chose a Pepper robot from Softbank as embodiment. Pepper is a humanoid robot which makes having dialogue with it seem natural. In addition, its size makes it reasonable to have it in a museum or hospital giving guidance to users.

Professional actors (one male, one female) played out the coping styles in the (validated) conversations. Then, we animated a Pepper robot to move similar to the actor. Figure 4.1 shows eight snapshots of the resulting videos. The initial part of the conversation shows both actors in the conversation (see the upper four pictures). The part where the character 'Robin' copes is filmed as close-up (see the lower four pictures). In appendix B, we discuss the design of the videos in more detail.

4.4.2. Participants

We recruited participants via Amazon Mechanical Turk. We required them to have a 95% or higher acceptance rate and to have some number of previous studies completed in the past. Because participants had to understand the content for the conversations, and because participants had to write down an explanation in English, we only accepted participants that were located in the US. Participants got a 1.1 Dollar monetary compensation for their time. In total, 577 participants participated in our study. Twelve were excluded because they did not pay sufficient attention to the videos (see below). Eight participants were excluded due to giving out of context answers to the explanation question such as referring to the actor 'Robin Williams'. Finally, ten more participants were excluded from analyses because their answers were written in incomprehensible English, or because they did not really answer the question (e.g., if they answered 'I don't know'). (These participants *did* get a monetary compensation for their time.) This left us with a total of **547 participants** (244 female, 299 male) used for analyses. Of these, 532 were native speakers and the remaining 15 indicated their English was at a good professional level.

4.4.3. Experimental Design

We had a between-subjects design. Participants were randomly assigned to one of the 42 different stimuli (two scenarios, i.e., health/ museum; three actors, i.e., male/ female/ robot; and seven coping styles (from ways of coping as discussed in section 4.4.1). Note that with actor, we refer to the agent that played the role of 'Robin' in our videos. The two human actors were chosen to control for effects induced by *human actor*. The two scenarios were used to control for a context effect imposed by scenario. Actor type (human versus robot) and coping style (7 styles) are the main independent variables, resulting in a 2x7 between subject setup. Participants were equally distributed over these 2x7 conditions and randomly distributed over the control variables context and male-female actor.

4.4.4. Materials and Measures

The experiment involved viewing the video and clipped video as explained above, and some questionnaires to measures the participant's perception of Robin's behaviour.

In the full video, the entire conversation (see section 4.4.1) was shown. In the clipped video only the close-up part of the video was shown where Robin shows the coping response itself. The full videos lasted for about one minute. The clipped

Table 4.2: Example participants' explanations of Robin's behaviour analysed by the LIWC sentiment miner

Emotionality by LIWC	Text
0	He wants him to develop a healthier way.
0	She was diffusing the issue
0	She was programmed to respond in that way.
5.9	Maybe she felt that she had overstepped purposing this
8.3	To try and calm the man and retain him as a customer.
14.3	Robin is programmed to respond this way to customers frustrations when they are lost.
15.4	He was trying to calm down the customer and ensure he was happy
25	She was frustrated with him for being stubborn

videos lasted for about 20 seconds.

For the experiment, we use several questionnaires. First, we ask the participant to give an explanation in natural language for Robin's behaviour by answering the question: 'Why did Robin respond in that way?'. Participants were shown an open text-box in which they had to type their explanation. They had to type at least three words. We refer to this as the participant's *explanation* of Robin's behaviour.

We measured the *emotionality* of a participant's explanation of Robin's behaviour using a state of the art sentiment miner LIWC [46] on the full text explanation given by the participants. LIWC counts words in psychologically meaningful categories (for example, emotions) [46]. We use the affect outcome of the algorithm. Which is a function of the count of different emotion words in the explanation.

Furthermore, we measure what coping styles participants recognised in the behaviour. For this purpose, we designed an adapted version of the ways of coping (WoC) questionnaire (i.e., the *adjusted WoC*). The original questionnaire is somewhat long for an Amazon Mechanical Turk study and was initially designed for recognising coping styles in ones own behaviour. The questions are not all referring to attributes recognisable in someone else. To account for this, we developed an adjusted Ways of Coping questionnaire specifically for our purposes. For every coping style, there are three questions on a 4-point Likert scale (coded as [0-3]) that measure to what extent a participant perceived the style in the behaviour. We count the values per style to get seven variables ($coping_{styleType}Count$), where style type is one of the seven coping styles). We use these variables as a measure for the recognition of the coping styles. Please note the difference between the dependent measure (*recognised coping style*) and the independent variable coping style as intended, modelled, and validated by the authors (*modelled coping style*). We will use these terms in the remainder of this chapter. For a full description of the content and design of this questionnaire, we refer to appendix A.

Naturalness and *appropriateness* of the coping behaviour is measured as two

separate items both on a 5-point Likert scale.

To measure *Warmth*, *competence*, and *discomfort* we use the RoSAS questionnaire [47]. This is an 18-item, 9-point Likert scale which measures participants' judgements of social attributes of Robin with three underlying dimensions: warmth, competence, and discomfort (6 questions per dimension). We average over the 6 items per scale.

Finally, we administered an affinity/ attitude-towards-robots questionnaire which was only shown to people in robot conditions. This was done via nine multiple choice questions (5-point Likert scale) in total. This last questionnaire is not used in the analyses reported in this chapter, so we do not discuss it further here.

4.4.5. Procedure

Participants had to finish the survey within 3 hours after accepting the hit. Average completion time of the survey was about 8 minutes. Participants were *not* in any way informed on what the different coping styles looked like or even that the behaviour of the actor was in a style. Nor were they informed that there were different possible responses versions of the scenario. We consciously primed them as little as possible. We asked them to explain the behaviour before asking any other questions so that the explanations gotten are unbiased as we could make them. The complete procedure was as follows.

When participants accepted the hit on Amazon Mechanical Turk they were directed to the online experimentation environment. There, participants gave their consent to use their data and were then directed to the demographics form (age, gender, language proficiency, education, and employment). Language proficiency was the only required field. The other questions were requested but not obligatory. Still, all participants answered all demographic questions.

Next, participants were shown a message asking them to unmute their speakers/ headphones. If they declared that they had done so, they were shown the full video corresponding to their assigned condition. Every time a video was shown, the display was automatically put to full-screen. Participants could not fast forward. If participants closed the full screen then they were shown a message asking them to not do that and they had to watch the video from the start. Participants *could* pause the video.

After watching the full video, participants replied to two simple questions about the content to check if they paid attention to the video. If the participants answered incorrectly, then they were forced to re-watch the video and were given a second attempt to answer the test-questions. If they answered incorrectly again, then they were excluded from the experiment.

After correctly answering the test questions, participants were shown the clipped video. After watching the clipped video, they had to provide their *explanation* for Robin's behaviour and answer the questions about the naturalness and appropriateness of the behaviour.

Following this, the participants had to watch the clipped video again and answer the questions to the adjusted WoC questionnaire.

Then, participants had to watch the clipped video for the third and final time,

followed by the RoSAS questionnaire [47]. For participants assigned to a human actor condition, this was the end of the questionnaire. Participants in the robot conditions were additionally shown the affinity and attitude questions after the RoSAS questions.

4.5. Results

4.5.1. Recognition of the Coping Styles

In this section, we analyse what coping styles participants actually perceived/ recognised within our modelled styles. An initial 7x2x2 MANOVA examined the recognition of the coping styles as dependent variables (the variables $coping_{styleType}Count$). Independent variables were: (1) the seven modelled coping styles, (2) the actor type (human versus robot) and (3) the scenario (to check for a potential influence of settings health versus museum).

Significant main effects were found for modelled coping style (Wilks' lamba = $p < 0.0005$), actor type (Wilks' lamba = $p < 0.0005$), and scenario (Wilks' lamba = $p < 0.0005$). There was also a significant effect for coping style x actor type (Wilks' lamba = $p < 0.0005$) and coping style x scenario (Wilks' lamba = $p < 0.0005$).

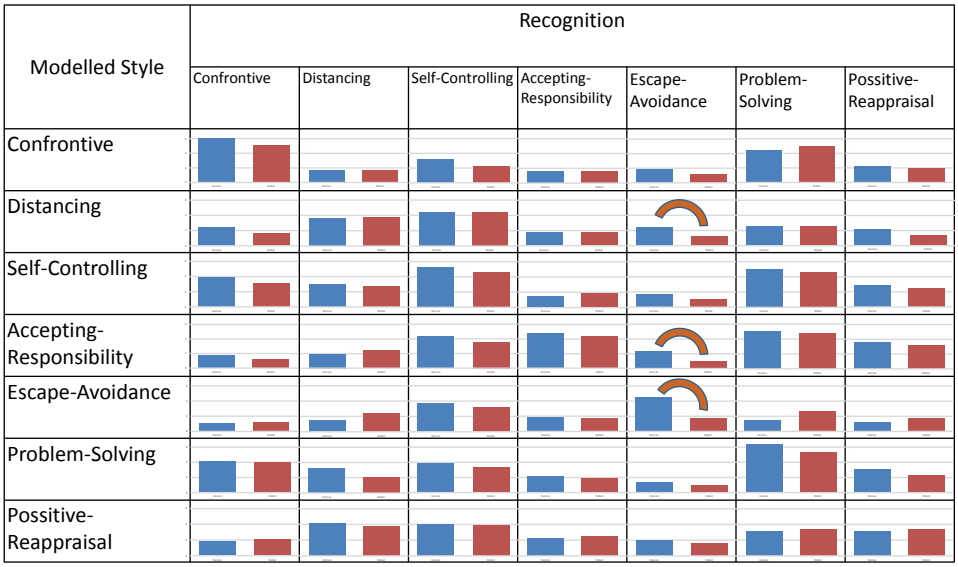
We put the significance level of the following between-subject effects at p=$0.05/7 \sim 0.007$ using a Bonferonni correction for the seven dependent variables (one for the recognition of each style, i.e., the variables $coping_{styleType}Count$). Modelled coping style significantly influenced the recognition of all seven coping styles ($p < 0.0005$ for all seven dependent variables). Actor type significantly influenced the recognition of escape avoidance (p<0.0005) and of self-controlling (p=0.006). Scenario significantly influenced the recognition of confrontive (p=0.006) and of problem-solving (p<0.0005). Finally, there was an interaction effect of actor type and modelled coping style for the recognition of escape-avoidance (p=0.002); and, there was an interaction effect of modelled coping style and scenario on the recognition of confrontive (p<0.0005), self-controlling (p=0.003), accepting-responsibility (p<0.0005) and problem-solving (p=0.006). There were no further significant effects.

For the remainder of this section, we first provide an extensive post-hoc analyses of how the modelled styles were recognised across the actor types. Then, we discuss how the different *human actors* influenced the recognition as a test for the reliability of our findings across different actors. Then, we discuss the differences between the different scenarios (health versus museum).

Coping Style Recognition Across the Modelled Coping Styles

For every *modelled coping style*, we gather the fingerprint it has with regards to the recognition ($coping_{styleType}Count$). For every modelled style, and for every style type (7x7), we test whether the perception of the style type is significantly higher than the average perception of the other 6 style types. In addition, we do this for both actor types (human/ robot) individually. This results in tables 4.3.4.3a (human actor) and 4.3.4.3b (robot actor). We use a paired samples t-test to compare the perception of a particular style with the average perception of the other 6 styles as baseline. Table 4.3 shows the results of this analyses. Dark green coloured cells are

Figure 4.2: Histograms of the means of the recognition of the different coping styles in the differently modelled styles.



Note: Rows show the differently modelled styles. Columns show the extent to which participants recognised a style in the modelled style. The blue bars show the extent to which the style was recognised for the human behaviour, the red one for the robot behaviour. The orange/ red symbol (⌒) means there is a significant difference between the actor types. (I.e., the recognition of escape-avoidance in the styles modelled as distancing, accepting-responsibility, and escape-avoidance is significantly higher for the human versus the robot actor type.)

significantly above, and dark red cells are significantly below the respective baselines. Light green and red cells are not significant when considering a Bonferonni correction (significance level at $p = 0.05/7 \sim 0.007$), but would be when testing as LSD (significance level at $p = 0.05$) [1]. Grey cells are not significant. Tables showing the exact t and p values can be found in appendix C of this chapter.

When looking at table 4.3.4.3a, we see that five of the seven modelled styles are properly recognised. In addition, participants often perceived one additional style simultaneously (i.e., recognised another style in the modelled style). For example, the confrontive style is significantly recognised as such. In addition, it is also recognised as problem-solving. The table further shows that the positive-reappraisal style is *not* recognised and distancing is *poorly* recognised (borderline *not* significant for human conditions and borderline significant for robot conditions). When we look at table 4.3.4.3b, then we see that for the robot actor conditions four of the seven styles are significantly recognised as such. Besides positive-reappraisal and

[1] Throughout the chapter, we report when an analyses was significant at the LSD level; however, we only consider correlations significant when they are so also after a Bonferonni correction.

distancing, participants were not able to recognise escape-avoidance in the robot conditions.

Table 4.3: Means and Standard Deviations for style recognition.

(a) Human Actor: Means and Standard Deviations

Modelled Style	Recognition C	D	S-C	A-R	E-A	P-S	P-R
Confrontive (C)	M=5.9, STD=1.6	M=1.7, STD=2.2	M=3.2, STD=2.4	M=1.5, STD=2.2	M=1.9, STD=2.1	M=4.5, STD=2.1	M=2.3, STD=2.3
Distancing (D)	M=2.5, STD=2.0	M=3.6, STD=1.9	M=4.4, STD=2.0	M=1.7, STD=2.0	M=2.5, STD=2.1	M=2.6, STD=2.5	M=2.2, STD=2.2
Self-Controlling (S-C)	M=3.8, STD=1.7	M=3.0, STD=2.0	M=5.2, STD=2.0	M=1.6, STD=1.5	M=1.7, STD=1.9	M=5.1, STD=2.0	M=2.9, STD=2.1
Accepting-Responsibility (A-R)	M=1.6, STD=1.8	M=1.8, STD=2.3	M=4.2, STD=2.5	M=4.9, STD=1.8	M=2.4, STD=2.3	M=4.9, STD=2.4	M=3.5, STD=2.3
Escape-Avoidance (E-A)	M=1.1, STD=1.6	M=1.6, STD=1.5	M=3.8, STD=2.1	M=1.9, STD=2.0	M=4.6, STD=1.8	M=1.6, STD=2.3	M=1.3, STD=2.0
Problem-Solving (P-S)	M=4.2, STD=1.4	M=3.2, STD=1.9	M=3.8, STD=2.2	M=2.3, STD=2.3	M=1.4, STD=1.9	M=6.4, STD=1.7	M=3.2, STD=2.5
Possitive-Reappraisal (P-R)	M=1.9, STD=2.1	M=4.1, STD=2.0	M=4.0, STD=2.0	M=2.2, STD=1.9	M=1.9, STD=1.9	M=3.2, STD=2.6	M=2.1, STD=2.1

(b) Robot Actor: Mean and Standard Deviation

Modelled Style	Recognition C	D	S-C	A-R	E-A	P-S	P-R
Confrontive (C)	M=5.1, STD=2.2	M=1.7, STD=2.0	M=2.2, STD=2.0	M=1.5, STD=1.9	M=1.1, STD=1.7	M=4.9, STD=2.5	M=2.0, STD=2.1
Distancing (D)	M=1.6, STD=1.8	M=3.7, STD=1.8	M=4.3, STD=2.1	M=1.8, STD=1.7	M=1.2, STD=1.7	M=2.6, STD=2.1	M=1.5, STD=1.9
Self-Controlling (S-C)	M=3.1, STD=1.7	M=2.8, STD=1.8	M=4.6, STD=2.2	M=1.9, STD=1.9	M=1.0, STD=1.3	M=4.5, STD=2.0	M=2.4, STD=1.7
Accepting-Responsibility (A-R)	M=1.1, STD=2.0	M=2.3, STD=1.7	M=3.5, STD=2.2	M=4.6, STD=1.7	M=.9, STD=1.7	M=4.5, STD=2.0	M=3.2, STD=2.2
Escape-Avoidance (E-A)	M=1.3, STD=2.0	M=2.5, STD=1.9	M=3.3, STD=2.2	M=1.7, STD=1.8	M=1.8, STD=2.0	M=2.7, STD=2.3	M=1.7, STD=2.6
Problem-Solving (P-S)	M=4.1, STD=1.5	M=2.0, STD=1.9	M=3.3, STD=2.1	M=1.9, STD=1.5	M=1.0, STD=1.9	M=5.3, STD=1.9	M=2.2, STD=2.2
Possitive-Reappraisal (P-R)	M=2.1, STD=2.2	M=3.8, STD=1.8	M=4.0, STD=2.1	M=2.4, STD=2.1	M=1.7, STD=2.2	M=3.3, STD=2.5	M=3.5, STD=2.3

Note: grey cells are not significant, light red and green cells are significant as LSD ($p < 0.05$) but not after a Bonferonni correction ($p <\sim 0.007$), dark red and green cells are significant. Green cells show a mean above-, and red cells show a mean below the baseline. Where baseline is the average of the *other* styles. E.g., the confrontive style is significantly recognised as *confrontive* compared to the average of the other six styles.

To get a clearer view on the specific interaction effects of coping style and actor type (human versus robot), we did some final t-tests. For every modelled style, we conducted seven independent-samples t-tests to compare $coping_{styleType}Count$ for the human actor conditions and the robot actor conditions. The recognition of escape-avoidance was influenced by actor type in the distancing (D), accepting-responsibility (A-R), and escape-avoidance (E-A) conditions. These styles were all perceived to have less escape-avoidance in them in the robot compared to the human conditions. These results indicate that escape-avoidance as coping style was not perceived in the robot behaviour. Or in other words, robot behaviour was significantly less attributed with escape-avoidance as underlying behavioural motivator.

Moderating Variable; Human Actor

In our following test, we ignore the participants assigned to a robot condition and focus on those assigned to one of the two types of human actor conditions. We test whether the *human actor* variable influences the recognition. We did a 7x2x2 MANOVA using the seven $coping_{styleType}Count$ as dependent variables; and, coping style, *human actor*, and scenario as independent variables. Human actor showed no significant main effect nor a significant interaction effect with one of the other variables. This gives us confidence that results regarding coping style recognition from our study are stable across different human actors and thus reliable.

Moderating Variable; Scenario

Next, we tested the effects of scenario (health versus museum; see table 4.4). We do not go through all differences individually but rather focus on the differences that we believe most noteworthy. Scenario has an effect on the results in many of the conditions (C, A-R, E-A, P-S). It is clear that context needs to be considered when testing the subjective recognition of the different styles. Arguably the most critical biases are when scenario influences the extent to which the modelled style itself is recognised (e.g., style C recognised as C or style A-R recognised as A-R). The style C and A-R were less well recognised in the museum conditions.

Table 4.4: Scenario Influences for Coping Style Recognition

Modelled Style	Recognition						
	C	D	S-C	A-R	E-A	P-S	P-R
Confrontive (C)	Health	-	-	-	-	-	-
Distancing (D)	-	-	-	-	-	-	-
Self-Controlling (S-C)	-	-	-	-	-	-	-
Accepting-Responsibility (A-R)	Museum	-	-	Health	-	Museum	-
Escape-Avoidance (E-A)	-	-	-	Museum	-	-	-
Problem-Solving (P-S)	-	-	-	Museum	-	-	-
Possitive-Reappraisal (P-R)	-	-	-	-	-	-	-

Note: grey cells are not significant, light red cells are significant as LSD ($p < 0.05$) but not after a Bonferonni correction ($p <\sim 0.007$), dark red cells are significant and show what conditions score a *higher* recognition

4.5.2. Emotionality of Explanations

In the previous section, we discussed the recognition of coping styles in the behaviours. In this section, we discuss the participant's explanation of the behaviour. Particularly, we discuss the emotionality of the explanations given (research question 2). We did a 7x2x2 ANOVA using emotionality as dependent variable; and, coping style, actor type and scenario as independent variables. The main effect for modelled coping style was not significant. The main effect for actor type was significant ($F(1, 519) = 10.134, p = 0.002$). The main effect for scenario was significant ($F(1, 519) = 11.477, p = 0.001$). Furthermore, the interaction of modelled coping

Table 4.5: Emotionality of Explanations per Actor Type

Emotionality score (by LIWC)	Percentage of explanations	
	Human actor type	Robot actor type
0	21.3%	27.6%
>0, <=8	27.6%	29.7%
>8, <=15	29.1%	33.7%
>15	22%	9%

style and scenario was significant ($F(6, 519) = 2.795, p = 0.011$). First, we discuss the main effect of Actor Type more thoroughly. Then, we discuss the coping styles where scenario had an significant effect.

An independent-samples t-test was conducted to compare emotionality of explanation in the human actor type conditions and the robot actor type conditions. There was a significant difference in the means for human actor type ($M = 9.44, SD = 8.08$) and robot actor type ($M = 7.37, SD = 6.78$); $t(545) = 3.256, p = 0.001$. In general, participants explained the behaviour of human actor types with significantly more emotionality than the behaviour of the robot actor type. The Cohen's d effect size measure is 0.278, i.e., humans are explained with ~ 0.28 deviations more emotionality than robots. Table 4.5 shows some further descriptives. These statistics can be compared with the example explanations in table 4.2 for interpretation.

Secondly, several independent-samples t-tests were conducted to compare emotionality of explanation in the health scenario conditions and the museum scenario conditions across the coping styles. The effect of scenario was significant in the confrontive, escape-avoidance, and problem-solving conditions. In the confrontive conditions, there was a significant difference in the means for health scenario ($M = 5.95, SD = 6.97$) and museum scenario ($M = 10.90, SD = 7.10$); $t(79) = -3.158, p = 0.002$. In the escape-avoidance conditions, there was a significant difference in the means for health scenario ($M = 6.26, SD = 7.18$) and museum scenario ($M = 11.37, SD = 8.80$); $t(69) = -2.692, p = 0.009$. In the problem-solving conditions, there was a significant difference in the means for health scenario ($M = 3.89, SD = 4.71$) and museum scenario ($M = 9.10, SD = 7.98$); $t(76) = -3.465, p = 0.001$. For these three modelled coping styles, participants explained the behaviour in the museum scenario with significantly more emotionality than the behaviour in the health scenario. The Cohen's d effect size measure is 0.278, i.e., museum scenarios were explained with ~ -0.29 deviations more emotionality than health scenarios.

Moderating Variable; Human Actor

In our following test, we ignore the participants assigned to a robot condition and focus on those assigned to one of the two types of human actor conditions. We test whether the *human actor* variable influences the recognition. We did a 7x2x2 ANOVA using emotionality as dependent variable; and, coping style, *human actor*, and scenario as independent variables. Human actor showed no significant main

effect nor a significant interaction effect with one of the other variables. This gives us confidence that results regarding emotionality from our study are stable across different human actors and thus reliable.

Correlation Between Recognised Coping Styles and Emotionality

There was no main effect of modelled coping style on emotionality of explanations. However, from our findings in the previous section 4.5.1, we can conclude that the modelled styles generally are perceived as several coping styles simultaneously. Note also that people indeed often cope in several styles simultaneously [16, 39, 41] as discussed in the related work section 4.2.2. For example, the conversation modelled to have a confrontive coping response is also perceived to have a problem-solving coping response (see tables 4.3a and 4.3b). In this subsection, we show that perceived coping style *does* significantly *correlate* with emotionality of explanation.

A simple linear regression was calculated to predict emotionality of explanations based on $coping_{styleType}Count$ (with all seven coping style types). We used the backward elimination method. A significant regression was found ($F(2, 544) = 7.590, p = .001$, with an $R^2 of 0.027$). Participant's emotionality of explanation increases when $coping_{E-A}Count$ increases and decreases when $coping_{P-S}Count$ increases. Note that R^2 is only 0.027. This means that only 2.7% of the variance in emotionality can be explained by this model. This is not much. However, we will discuss in section 4.6 that we can not expect too large effect sizes for this test.

Next, we did two separate linear regressions to predict emotionality of explanations based on $coping_{styleType}Count$ controlling for actor type, i.e., one for the human- and one for the robot actor type. A simple linear regression considering only cases with a *human* actor type was calculated to predict emotionality of explanations based on $coping_{styleType}Count$. Just like before, we used the backward elimination method. A significant regression was found ($F(3, 264) = 7.071, p < .000$, with an $R^2 of 0.074$). Table 4.6 shows the model. Participant's emotionality of explanation increases when $coping_{E-A}Count$ increases and when $coping_{A-R}Count$ increases. Participant's emotionality of explanation decreases when $coping_{P-R}Count$ increases. We can see that when focusing solely on human actor type conditions, the model could explain more variation in the emotionality of explanations (i.e., the value of R^2 which was 7.4%). In contrast, a simple linear regression (considering only cases with a *robot* actor type) was calculated to predict emotionality of explanations based on $coping_{styleType}Count$. No significant regression was found, the model explained 0% of the variance in the participant's emotionality of explanation.

These results indicate that coping style indeed correlates with emotionality of explanation when explaining *human* behaviour. However, for explaining *robot* behaviour, we found no correlation.

4.5.3. Perception of Coping Styles

Finally, for research question 3.3, We test the effect of actor type and (modelled) coping style on the perception in terms of naturalness, appropriateness, warmth, competence, and discomfort. We conducted 5, $7x2x2$ MANOVA's (5 dependent variables, 7 modelled coping styles, 2 actor types). A Bonferonni correction puts

Table 4.6: Linear regression model to predict emotionality of explanation based on perceived coping style for the human actor type only

	Beta	95% Confidence Interval for Beta		t-value	p-value
		Lower Bound	Upper Bound		
Intercept	7.57	4.41	10.74	4.710	<0.0005
Accepting-Responsibility	0.57	0.07	1.07	2.228	0.027
Escape-Avoidance	0.60	0.13	1.06	2.507	0.013
Problem-Solving	-0.76	-1.22	-0.30	-3.259	0.001

the significance level at $p <= .01$.

Coping style had a significant influence on all metrics (Wilks' Lamba p=< 0.0005): naturalness (F(6,519)=11.998, p<.0005), appropriateness (F(6,519)=4.243, p<.0005), warmth (F(6,519)=3.880, p<.001), competence (F(6,519)=11.973, p<.0005), and discomfort (F(6,519)=5.596, p<.0005).

Actor type had a significant influence (Wilks' Lamba p=< 0.0005). With specific effects on naturalness (F(1,519)=14.831, p<.0005), competence (F(1,519)=8.976, p=.003), and discomfort (F(1,519)=16.485, p<.0005), but not on appropriateness, and warmth.

Scenario had a significant influence (Wilks' Lamba p=< 0.0005). With specific effects on naturalness (F(1,519)=28.384, p<.0005), competence (F(1,519)=15.673, p<.0005), and discomfort (F(1,519)=14.418, p<.0005), but not on appropriateness and warmth.

Furthermore, there were some **interaction effects**. The interaction between actor type and coping style was significant (Wilks' Lamba p=< 0.009). There was an effect on warmth (F(6,519)=2.867, p=.009) and competence (F(6,519)=3.630, p=.002), but not on naturalness, appropriateness, and discomfort. The interaction between scenario and coping style was significant (Wilks' Lamba p=< 0.0005). There was an effect on naturalness (F(6,519)=5.583, p<.0005), appropriateness (F(6,519)=4.764, p<.0005), and competence (F(6,519)=4.169, p<.0005), but not on warmth and discomfort. The interaction between scenario and actor type was significant (Wilks' Lamba p=< 0.021). However, there were no specific effects found here. Finally, there was *no* effect of the interaction for coping style times actor type times scenario.

Figure 4.3 shows box-plots of the perception of the coping styles. Robots were perceived more positively on three metrics (naturalness of the behaviours, competence of the actor, and discomfort (opposite effect) imposed by the actor). Additionally, we can see that accepting-responsibility and escape-avoidance are less positively perceived than the other coping styles in general. However, these styles are *much* more positive in the robot conditions.

Moderating Variable; Human Actor

In our following test, we ignore the participants assigned to a robot condition and focus on those assigned to one of the two types of human actor conditions. We test whether the *human actor* variable influences the recognition. We did a 7x2x2 MANOVA using the perception in terms of naturalness, appropriateness, warmth,

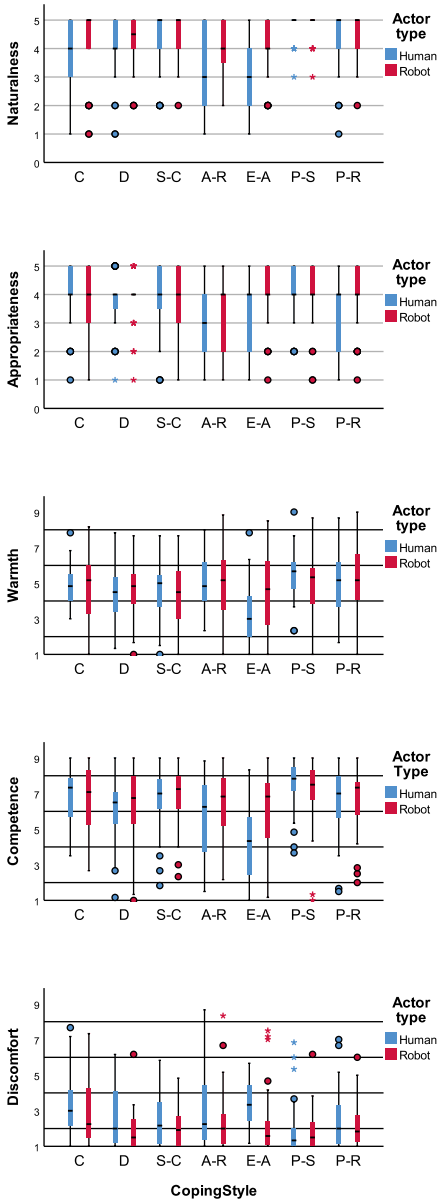


Figure 4.3: Perception of the coping styles in terms of naturaleness, appropriateness, warmth, competence, and discomfort across the coping styles and actor types

competence, and discomfort as dependent variable; and, coping style, *human actor*, and scenario as independent variables. Human actor showed no significant main

effect nor a significant interaction effect with one of the other variables. This gives us confidence that results regarding perception from our study are stable across different human actors and thus reliable.

4.6. Discussion

In the previous section, we looked at (1) the recognition of coping styles; (2) the emotionality of the participants' explanations; and (3) the perception of the coping styles in terms naturalness, appropriateness, warmth, competence, and discomfort. In this section, we discuss these findings.

Coping Style Recognition

We found that most coping styles were recognised in Robin's behaviour regardless of actor type. The distancing style was not strongly recognised for either humans or robots. The positive-reappraisal style was not recognised at all. In general, coping styles can be divided in emotion-based coping and problem-based coping [16]. Emotion-based coping (D, A-R, E-A, P-R) focuses primarily on regulating the negative feeling of distress caused by the situation. Problem-based coping (C, S-C, P-S) focuses on dealing with the situation to address the *cause* of the distress. Especially distancing and positive-reappraisal focus strongly on the more internal processes, which makes it understandable that these were harder to recognise. The problem-based coping styles were recognised more clearly as can be seen in table 4.3.

Interestingly, escape-avoidance was recognised for the human actor type but not at all for the robot actor type. Looking specifically at the biases imposed by actor type (figure 4.2, we can see that people perceive significantly less escape-avoidance in robot behaviour than in human behaviour). This has implications for human-robot interaction. We simply do not expect robots to deal with problems by trying to get away from them. Programming a robot to behave in that way might not be understood by people. When developing interaction models for when the robot is unable to solve a problem, then we must adopt different strategies.

Emotionality of Explanations

Besides coping style recognition, we gathered unrestricted spontaneous people's explanations of the behaviours. Such explanations provide insights in people's mental state ascription imposed on behaviour [8].

Our results indicate that people use emotions when explaining the behaviours. We tested the emotionality depending on coping style and on actor type. Actor type had an influence on the emotionality of the explanations. People did use emotions when explaining robot behaviour, but they used less emotions than for explaining the human behaviour.

Modelled coping style *did not* predict emotionality. However, we *did* find a correlation between *perceived* coping style and emotionality of explanations. This correlation occurred for the human actor type, but not for the robot actor type. We propose a possible explanation for this correlation in our findings. Perhaps there is some aspect of the behaviour, that causes one to recognise escape-avoidance and

that causes one to ascribe emotions to the behaviour, that people did not ascribe to the robot but that they *did* ascribe to the humans. For example, escape-avoidance is associated with having 'hopes and whishes' [16, 41]. It is possible that this is what impacts recognition of escape-avoidance as well as emotionality of the explanations. A follow-up study could be to test the influence of hopes and wishes (rather than coping style) on emotionality using a similar setup as used this study. One can define and validate behaviours that correspond with a low versus high amount of hopes and wishes ascribed to it, have a human versus robot actor type as between-subjects control, and measure the emotionality of participants' explanations of the behaviours.

There was a significant correlation between perceived coping style and emotionality of the explanations for the *human actor type*. However, the R^2 was quite small (7.4% of the variance could be explained with the model). Still, we should not expect this value to be much higher. First, there was no main effect of modelled coping style on emotionality and there *was* an effect of modelled coping style on perceived/ recognised coping style. This already implies that perceived coping style is unlikely to have a strong correlation with emotionality. Furthermore, 'when people use emotions in explanations' is very much an unsolved problem in social sciences as well [11, 14, 27]. If there were strong straightforward correlations then earlier work would have already found and reported these.

In summary, participants recognised the coping styles within the robots, and participants used emotions to explain the robot behaviour. However, the perception of the coping style in the robot behaviour did *not* correlate with the emotionality of the resulting explanation of the behaviour. People are able to recognise these emotional behaviours in robots, and people attribute emotions to a robot's underlying motivations. However, the recognition did not impact the emotionality of the explanations. These results imply that people attribute mental states and motivations differently to robots than they do for humans. Similar to nuance differences found in related studies on anthropomorphism [29, 30, 44]. Our results indicate that people's estimate of what emotions cause a robot's behaviour are independent of the robot's coping style for stressful situations.

Perception of the Behaviour

We found that the behaviours were perceived more positively for the robots than for the humans in terms of naturalness of the behaviours, competence of the actor, and discomfort (opposite effect) imposed by the actor. This was not what we expected. We provide several possible explanations for this finding. It might be a novelty effect. People are initially more positive about the robot because they find it an interesting new type of entity [1, 48, 49]. In that case, the novelty effect will gradually wear off. A second explanation is that people have lower expectations of the robot in these scenarios. This would cause them to be more positive about the robot's behaviour [50]. Finally, maybe people might feel more comfortable with a robotic support in these specific scenarios. For example, it has been proposed people might be more comfortable sharing health-related issues with a robotic health coach [49].

There were also some interaction effects. Escape-avoidance was perceived positively for robots but not for humans. So the modelled behaviour worked in the stressful encounter. However, this does not imply that escape-avoidance should be modelled in robots. In fact, our results imply the opposite. Designers of humanoid robots must take the user's mental models into account as well as the social cues that robots emit [45]. If escape-avoidance is not recognised, then modelling behaviour in this style can cause unexpected results. Participant's assessment of the behaviour was independent of the robots motivations. However, we should strive for transparency in a robot's intentions [9]. People predict robot behaviour and frame their interaction based on their mental models [29, 45]. We must ensure that these expectations are aligned.

4.6.1. Limitations

Participants of our study were citizens from the united states of America. Our results our therefore tested within and for the English language. There might be differences across languages for the emotionality of the explanations. There might even be cultural differences, for example, British people might differ somewhat from our population. In addition, we recruited participants via Amazon Mechanical Turk. Many people do mTurk studies for some additional income which biases the distribution of people that would have done our study. Another limiting factor is that we chose a particular humanoid robot (the Pepper robot from Softbank). There might be subtle differences on all our measures when testing this on a different platform. Finally, we chose to have participants type the explanations rather than speak them out loud. Typing made gathering and analysing the data much less error prone. However, there might be differences in the explanations when they are given in another way. For example, spoken explanations might be longer than types explanations if the participants finds typing less convinient than talking.

4.6.2. Implications for Robot Self-Explanations

Next, in this discussion we want to address the possibility to use people's explanations of robot behaviour as input for eXplainable Artificial Intelligence (XAI). This is an important current topic in human robot interaction given the importance of transparency on trust and comprehension [4, 20, 32, 51], and the appearance of the recent GDPR [23]. For humans, explanations by third parties are not the same as self-explanations; however, there are similarities in the types and frequencies of mental constructs (for example, beliefs, goals, emotions) used [11, 52]. The usage of beliefs and especially goals as explanations has been widely tested in the XAI community [3, 6, 7, 12, 13]. However, it has been proposed that considering emotions might be required for explainable agents like robots [15]. When considering our findings as input for XAI, then our results confirm that the use of emotions must be considered for explaining robot behaviour. Whether this is directly as part of self-explanations as in [43], or the expression as part of a transparency mechanism as in [42], or both, remains an open question though.

Our results indicated the use of emotions in explanations for robot behaviour is lower than that for human behaviour. Furthermore, our results show that the emo-

tionality is stable across different styles to cope with distressful encounters. This implies the task is at least slightly easier for robot designers. Some emotionality in robot self-explanations seems necessary, but in a less involved manner than for humans. An important follow-up study would be to let the robot self-explain the behaviours modelled in this study using the exact explanations as given by the participants of this study and then measure the effect (naturalness/ appropriateness/ warmth/ competence/ discomfort) on other participants.

4.7. Conclusion

In this chapter, we investigated people's perception of robot versus human behaviour. The behaviours were modelled to represent several coping styles from literature [41]. We measured (1) whether people could recognise the coping styles; (2) what spontanious unrestricted explanations people give for the behaviour; and (3) how positive and accepting people were towards the behaviour (i.e., naturalness, appropriateness, warmth, competence, and discomfort). For all these outcomes we considered the influence of the actor type (human versus robot) and scenario (health versus museum).

We found that people did not recognise escape-avoidance in robot behaviour. Robot designers must take this into account. Even though the behaviour was perceived as positive, we conclude that designers of robot behaviour should take extra care when implementing escape-avoidance like behaviours because people do not recognise them in robots. When the robot shows behaviours that do not match a user's expectations then users might become increasingly aware that they are dealing with a programmed machine rather than an actual intentional agent (i.e., breaking the illusion of life [53]). Misaligned expectations can cause users to quit the interaction [50].

XAI is indeed often based on how people explain behaviour amongst each other [9, 12]. How people explain behaviour can serve as input for eXplainable AI (XAI) [8, 17] as well. Our work shows that emotions in robot self-explanations seems necessary, but in a less involved manner than for humans, and verifies that further research on this topic is needed.

Our main finding is that we show that, and shed light on what way, people use emotions when explaining robot behaviour. They do so with less frequency then they do for human behaviour. Still, only about $1/4th$ of the explanations is devoid of any emotionality when analysing with the LIWC sentiment miner (about $1/5th$ for the human behaviour explanations). Furthermore, we found a difference in what causes people to explain *human behaviour* with emotions and what causes people to explain *robot behaviour* with emotions. The perception of coping style correlates with the emotionality of explanations when people explain human behaviour, but not when people explain robot behaviour.

4.8. Acknowledgements

We want to thank Bart Vastenhouw for the great work with filming, editing, and distributing of the videos. We also want to thank Gerben Tuin, Gina Lamprell, and

Pietro Pasotti for their terrific acting performance. This work has been supported by the PAL project, Horizon2020 grant nr. 643783-RIA.

References

[1] I. Leite, C. Martinho, and A. Paiva, *Social robots for long-term interaction: a survey,* International Journal of Social Robotics **5**, 291 (2013).

[2] B. M. Muir, *Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems,* Ergonomics **37**, 1905 (1994).

[3] M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, *Building explainable artificial intelligence systems,* in *Innovative Applications of Artificial Intelligence* (2006) pp. 1766–1773.

[4] B. Y. Lim, A. K. Dey, and D. Avrahami, *Why and why not explanations improve the intelligibility of context-aware intelligent systems,* in *Human Factors in Computing Systems* (2009) pp. 2119–2128.

[5] M. Harbers, K. Van den Bosch, and J.-J. Meyer, *Design and evaluation of explainable bdi agents,* in *Web Intelligence and Intelligent Agent Technology* (2010) pp. 125–132.

[6] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, *Enabling robots to communicate their objectives,* Autonomous Robots , 1 (2017).

[7] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, *Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults,* in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on* (IEEE, 2017) pp. 676–682.

[8] M. De Graaf and B. Malle, *People's explanations of robot behavior subtly reveal mental state inferences.* in *Human-Robot Interaction (HRI), 2019 11th ACM/IEEE International Conference on, in press* (ACM, 2019).

[9] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, *Explainable agents and robots: Results from a systematic literature review,* in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, 2019) pp. 1078–1088.

[10] D. C. Dennett, *Three kinds of intentional psychology,* in *Reduction, Time and Reality*, edited by R. Healey (Cambridge University Press, Cambridge, 1981) pp. 37–61.

[11] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction.* (MIT Press, 2004).

[12] M. Harbers, J. Broekens, K. Van Den Bosch, and J.-J. Meyer, *Guidelines for developing explainable cognitive models,* in *International Conference on Cognitive Modeling* (2010) pp. 85–90.

[13] J. Broekens, M. Harbers, K. Hindriks, K. Van Den Bosch, C. Jonker, and J.-J. Meyer, *Do you get it? user-evaluated explainable bdi agents,* in *Multiagent System Technologies* (Springer, 2010) pp. 28–39.

[14] S. A. Döring, *Explaining action by emotion,* The Philosophical Quarterly **53**, 214 (2003).

[15] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, *The role of emotion in self-explanations by cognitive agents,* in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (IEEE, 2017) pp. 88–93.

[16] R. S. Lazarus, *Emotion and adaptation.* (Oxford University Press, 1991).

[17] T. Miller, *Explanation in artificial intelligence: Insights from the social sciences,* Artificial Intelligence (2018).

[18] W. Swartout, C. Paris, and J. Moore, *Explanations in knowledge systems: Design for explainable expert systems,* IEEE Expert **6**, 58 (1991).

[19] L. R. Ye and P. E. Johnson, *The impact of explanation facilities on user acceptance of expert systems advice,* Mis Quarterly , 157 (1995).

[20] S. R. Haynes, M. A. Cohen, and F. E. Ritter, *Designs for explaining intelligent agents,* International Journal of Human-Computer Studies **67**, 90 (2009).

[21] J. D. Lee and K. A. See, *Trust in automation: Designing for appropriate reliance,* Human factors **46**, 50 (2004).

[22] O. Biran and C. Cotton, *Explanation and justification in machine learning: A survey,* in *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8 (2017) p. 1.

[23] P. Carey, *Data protection: a practical guide to UK and EU law* (Oxford University Press, Inc., 2018).

[24] F. C. Keil, *Explanation and understanding,* Annual Review of Psychology **57**, 227 (2006).

[25] K. V. Hindriks, *Debugging is explaining,* in *International Conference on Principles and Practice of Multi-Agent Systems* (Springer, 2012) pp. 31–45.

[26] M. Harbers, K. van den Bosch, and J.-J. C. Meyer, *A study into preferred explanations of virtual agent behavior,* in *International Workshop on Intelligent Virtual Agents* (Springer, 2009) pp. 132–145.

[27] P. M. Churchland, *Folk psychology and the explanation of human behavior,* The future of folk psychology: Intentionality and cognitive science , 51 (1991).

[28] S. Thellman, A. Silvervarg, and T. Ziemke, *Folk-psychological interpretation of human vs. humanoid robot behavior: exploring the intentional stance toward robots,* Frontiers in psychology **8**, 1962 (2017).

[29] S. Kiesler and J. Goetz, *Mental models of robotic assistants,* in *CHI'02 extended abstracts on Human Factors in Computing Systems* (ACM, 2002) pp. 576–577.

[30] N. Epley, A. Waytz, and J. T. Cacioppo, *On seeing human: a three-factor theory of anthropomorphism.* Psychological review **114**, 864 (2007).

[31] A. Sciutti, A. Bisio, F. Nori, G. Metta, L. Fadiga, and G. Sandini, *Robots can be perceived as goal-oriented agents,* Interaction Studies **14**, 329 (2013).

[32] R. H. Wortham, A. Theodorou, and J. J. Bryson, *What does the robot think? transparency as a fundamental design requirement for intelligent systems,* in *Ijcai-2016 ethics for artificial intelligence workshop* (2016).

[33] R. W. Picard *et al.*, *Affective computing,* (1995).

[34] P. Ekman, *Basic emotions,* Handbook of cognition and emotion **98**, 16 (1999).

[35] A. Mehrabian, *Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,* Current Psychology **14**, 261 (1996).

[36] K. R. Scherer, *Appraisal considered as a process of multilevel sequential checking,* in *Appraisal processes in emotion: Theory, methods, research*, edited by K. R. Scherer, A. Schorr, and T. Johnstone (Oxford University Press, 2001) pp. 92–120.

[37] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions* (Cambridge university press, 1990).

[38] R. Reisenzein, *Appraisal processes conceptualized from a schema-theoretic perspective: Contributions to a process analysis of emotions.* in *Appraisal processes in emotion: Theory, Methods, Research*, edited by K. R. Scherer, A. Schorr, and T. Johnstone (Oxford University Press, 2001) pp. 187–201.

[39] S. Folkman, *Stress: appraisal and coping,* in *Encyclopedia of behavioral medicine* (Springer, 2013) pp. 1913–1915.

[40] S. C. Marsella and J. Gratch, *Ema: A process model of appraisal dynamics,* Cognitive Systems Research **10**, 70 (2009).

[41] S. Folkman and R. S. Lazarus, *Ways of coping questionnaire* (Consulting Psychologists Press, 1988).

[42] J. Broekens and M. Chetouani, *Towards transparent robot learning through tdrl-based emotional expressions,* IEEE Transactions on Affective Computing (2019).

[43] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, *Evaluating cognitive and affective intelligent agent explanations in a long-term health-support application for children with type 1 diabetes,* in *2019 Eighth International Conference on Affective Computing and Intelligent Interaction (in press)* (IEEE, 2019).

[44] N. Shechtman and L. M. Horowitz, *Media inequality in conversation: how people behave differently when interacting with computers and people,* in *Proceedings of the SIGCHI conference on Human factors in computing systems* (ACM, 2003) pp. 281–288.

[45] S.-l. Lee, I. Y.-m. Lau, S. Kiesler, and C.-Y. Chiu, *Human mental models of humanoid robots,* in *Proceedings of the 2005 IEEE international conference on robotics and automation* (IEEE, 2005) pp. 2767–2772.

[46] Y. R. Tausczik and J. W. Pennebaker, *The psychological meaning of words: Liwc and computerized text analysis methods,* Journal of language and social psychology **29**, 24 (2010).

[47] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, *The robotic social attributes scale (rosas): Development and validation,* in *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction* (ACM, 2017) pp. 254–262.

[48] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn, *Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study,* in *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication* (IEEE, 2007) pp. 564–569.

[49] O. A. B. Henkemans, B. P. Bierman, J. Janssen, R. Looije, M. A. Neerincx, M. M. van Dooren, J. L. de Vries, G. J. van der Burg, and S. D. Huisman, *Design and evaluation of a personal robot playing a self-management education game with children with diabetes type 1,* International Journal of Human-Computer Studies **106**, 63 (2017).

[50] M. Ligthart, O. B. Henkemans, K. Hindriks, and M. A. Neerincx, *Expectation management in child-robot interaction,* in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on* (IEEE, 2017) pp. 916–921.

[51] N. Wang, D. V. Pynadath, and S. G. Hill, *Trust calibration within a human-robot team: Comparing automatically generated explanations,* in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, 2016) pp. 109–116.

[52] B. F. Malle, *How people explain behavior: A new theoretical framework,* Personality and social psychology review **3**, 23 (1999).

[53] J. Bates *et al.*, *The role of emotion in believable agents,* Communications of the ACM **37**, 122 (1994).

5

CAAF: A Cognitive Affective Agent Programming Framework

Cognitive agent programming frameworks facilitate the development of intelligent agents like robots and avatars. By adding a computational model of emotion to such a framework, one can program agents capable of using and reasoning over emotions. Computational models of emotion are generally based on cognitive appraisal theory; however, these theories introduce a large set of appraisal processes, which are not specified in enough detail for unambiguous implementation in cognitive agent programming frameworks. We present CAAF (Cognitive Affective Agent programming Framework), a framework based on the belief-desire theory of emotions (BDTE), that enables the computation of emotions for cognitive agents (i.e., making them cognitive affective agents). In this chapter we bridge the remaining gap between BDTE and cognitive agent programming frameworks. We conclude that CAAF models consistent, domain independent emotions for cognitive agent programming.

5

Parts of this chapter have been published in International Conference on Intelligent Virtual Agents 2019 [1]

5.1. Introduction

Interaction with intelligent agents is facilitated by providing such agents with affective abilities. For example, affective abilities in intelligent agents have been applied to facilitate *entertainment* [2, 3], to make an agent more *likable* for the user [4], to get *empathic* reactions from the user [5], and to create the so-called *the illusion of life* [6, 7], where characters are modelled to appear more life-like.

Cognitive agents can be programmed in frameworks like, e.g., GOAL [8], Jadex [9], or Jason [10]. A cognitive agent is an autonomous agent that perceives its environment through sensors and acts upon that environment with actuators [11]. It does so based on its *beliefs*, *desires* and *intentions*. Cognitive agents have a *mental state* and a *reasoning cycle* (see Figure 5.1). The mental state consists of *beliefs* and *desires*. Beliefs are the agent's representation of its environment. The agent can believe it is walking down the street, or that it is raining outside. Desires are things the agent *wants* to be true. For example, the agent can want to have an umbrella. The *intention* to get an umbrella reflects the agent's commitment to achieve that desire. After sensing *percepts* from the environment, the agent updates its mental state. Based on its beliefs, desires, and intentions, the agent reasons about its next action. The environment can change by itself, in response to an action of the agent, or actions from other agents that are situated in the same environment; thus, the agent may not always be *certain* of the exact *state of affairs* in its environment.

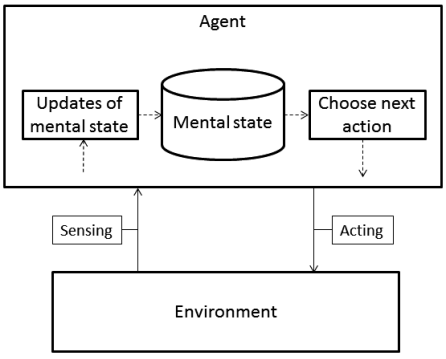


Figure 5.1: The reasoning cycle of a cognitive agent.

By adding a computational model of emotion to cognitive agent programming frameworks, one can program intelligent agents capable of using and reasoning over emotions. Computational models of emotion are usually based on cognitive appraisal theories [12]. Cognitive appraisal theory proposes that emotions are consequences of cognitive evaluations (*appraisals*), relating the event to an individual's desires. For example, one is happy because one believes something to be true, and desires this to be true.

However, cognitive appraisal theories [13–15] typically introduce a large set of appraisal processes, which are not specified in enough detail for unambiguous im-

plementation in cognitive agent programming frameworks. Psychological theories are developed to explain emotions for humans. These theories are thus not obligated to provide worked out computational specifications for the appraisals. Here we address this problem by integrating a computational model of the belief-desire theory of emotions (BDTE) [16, 17] with a BDI (belief-desire-intention)-based, cognitive agent programming framework. We present CAAF, a Cognitive Affective Agent programming Framework. Emotions are computed based on BDTE for two reasons: 1) because it is conceptually close to the BDI agent framework; and 2) it does not introduce a large set of appraisals that are difficult to describe in a computational manner.

The two main contributions of this work are: 1) We define semantics for the programming constructs of cognitive agents, formalizing how an agent updates its *mental state*, and how emotions are computed. 2) We show when the agent should minimally (re)appraise, by proving that, under some circumstances, the computation of emotions stays consistent when reducing the frequency with which the agent's emotions are recomputed, thereby increasing the efficiency of the computation.

5.2. Motivation & Related Work

In this article, we focus on computational models of emotion based on cognitive appraisal theory. A computational model of emotion describes the eliciting conditions for emotions, often including corresponding intensity. A popular appraisal theory among computer scientists, is the OCC-model [14, 18, 19]. The appraisal theory by Lazarus [13], and the sequential check theory (SCT) by Scherer [15, 20] have also found some attention among computer scientists. For example, the computational model EMA [21, 22] is mainly based on the appraisal theory by Lazarus [13], where the link between appraisal and coping is emphasized. EMA models how emotions develop and influence each other. For example, sadness can turn into anger at the responsible source. In [23] a formal notation for the declarative semantics of the structure of appraisal is proposed. Using this, a computational model of emotion is developed based on SCT.

The OCC model is the most implemented cognitive appraisal theory. Computational models based on the OCC model include AR [24], EM [7], FLAME[25], FearNot! [5], FAtiMA [26], and GAMYGDALA [2]. In AR [24] agents judge events based on their pleasantness, and whether they are confirmed, unconfirmed, or disconfirmed. For example, sadness is achieved when an agent confirms an unpleasant event. In EM [7] the aim is to build 'believable agents', agents that appear emotional and engage in social interactions. The EM architecture facilitates artists to model emotional agents in their applications. In FLAME the desirability of an event is modelled with fuzzy sets. For example, they define a fuzzy set 'undesirable event'. Individual events are then partly a member of this set, the amount of membership is adaptively learned over time. FearNot! is an application that helps children to cope with bullying. The agents use planning and expected utility to derive proper emotional responses. Currently the emotional responses in FearNot are triggered with a more enhanced model FAtiMA. FAtiMA divides the appraisal

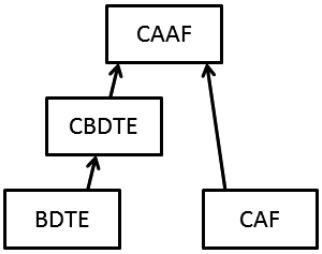


Figure 5.2: CAAF is build upon CBDTE [17] and CAFs (Cognitive Agent programming Frameworks). With CAAF, we close the gap between CBDTE and CAFs, and provide a fully worked out, computational account of BDTE.

into different modules, all responsible for a separate part of the computation. This enables implementing such modules independently. GAMYGDALA is an emotion engine that can be added to games by annotating events with their influence on the beliefs and desires of different characters.

An underlying problem with many appraisal theories is that cognitive agent programming frameworks lack the required knowledge representations to compute most appraisal processes. For example, a computational model of emotion that aims to describe the OCC-model in total [14], including emotion intensities, needs to model 12 different appraisals. For many of these appraisals it is unclear *how* they should be implemented, e.g., *deservingness*, *sense of reality*, or *proximity*. Other appraisals, e.g., praiseworthiness, require complex constructs like norms and values to be represented by the agent. SCT [15, 20] additionally introduces multiple layers in the appraisal process. An event is first analysed in a reactive, bodily responsive, type of way, and later analysed with increasingly nuanced cognitive processes. The computational model of emotion, EMA [22], is mainly based on the appraisal theory by Lazarus [13]. EMA [22] aims to simplify the appraisal processes, introduced by the underlying appraisal theories, and models them from a knowledge representation consisting of beliefs, desires, intentions, and (decision-theoretic) plans. This is conceptually closer to cognitive agent programming frameworks; however, though these frameworks are suited for programming decision-theoretic plans, they do not always do so. This would thus put constraints on the agent programming frameworks for which we want to compute emotions.

The appraisals and knowledge representation proposed by the belief-desire theory of emotion (BDTE) [16, 17] are more compatible with cognitive agent programming frameworks. In BDTE, emotions are derived only from beliefs and desires. In its minimal form BDTE requires only two appraisals. This makes BDTE more suitable as a basis for simulated emotions for such frameworks.

In this chapter, we integrate a computational model of BDTE with a cognitive agent programming framework (CAF), hence developing CAAF. In [17], Reisenzein extended BDTE to a computational form (CBDTE). CBDTE has been referred to as a computational model of emotion [12]; however, Reisenzein acknowledges that the motivation behind developing CBDTE was not to develop a worked-out com-

putational model, but rather to clarify aspects of BDTE [17]. Here, we build upon CBDTE, and close the gap between CAFs and CBDTE (see figure 5.2). Thus, this chapter presents a *full* computational account of BDTE, and formalizes how a cognitive agent should (efficiently) compute emotions.

5.3. A Model of Emotion for Cognitive Agent Programming Frameworks

In this Section we present CAAF. We present the formal semantics needed to integrate BDTE with cognitive agent programming. Further, based on this formal system we show in Section 4 that emotions can be computed in an efficient way using the model presented here.

5.3.1. Semantics for a Basic Knowledge Representation & BDTE

The mental state of an agent requires a *knowledge representation*. The agent needs to *represent* states of affairs, to *store* these representations, and to *change* the stored representations.

Representing the states of affairs is achieved with a *language*. This language needs to define a syntax of *well-formed formulae*. We write $\varphi \in \mathcal{L}$ to denote that φ is a formula of language $\mathcal{L}$. Here, a formula is a single proposition that contains information about a *state of affairs*, i.e., it is a sentence that *expresses whether a state of affairs is true (or not)*. We do not define how *logical connectives* work in this language, i.e., symbols that connect propositions such that the sense of the compound proposition depends only on the original sentences (for example, φ_1 *and* φ_2). The contribution of this chapter is to define semantics for the programming constructs of cognitive agents, formalizing how an agent updates its *mental state*, and how emotions are computed.

Storing states of affairs is done with a *set*. The belief, desire and emotion base are represented in the semantics as a set of formulae, mapped to a value $[0,1]$. These bases are a subset of some language $\mathcal{L}$, but contain further information as well. A belief base has the form: $\Sigma : \langle C : \mathcal{L} \to [0,1] \rangle$, where C is mapping of a formula φ to (exactly one) certainty value between $[0,1]$. We denote $b\{\varphi \to c\} \in \Sigma$ for 'the agent believes φ with certainty c'. Furthermore, we add the constraint that if C contains the mappings $b\{\varphi \to c\}$ and $b\{\neg\varphi \to c'\}$, then $c = 1 - c'$. A desire base has the form $\Gamma : \langle U : \mathcal{L} \to [0,1] \rangle$, where U is mapping that maps formula φ to a utility value between $[0,1]$. We denote $d\{\varphi \to c\} \in \Gamma$ for 'the agent desires φ with utility (strength of desire) u'. Finally an emotion base has the form $\Upsilon : \langle I : \mathcal{L} \times \Theta \to [0,1] \rangle$, where $\theta \in \Theta$ is an emotion label (happy, unhappy, hope, fear, surprise, relieve, or disappointment), and I maps formula $\varphi \in \mathcal{L}$ and label $\theta \in \Theta$ to an intensity value between $[0,1]$. We denote $e\{\varphi \times \theta \to i\} \in \Upsilon$ for 'the agent has emotion θ (concerning formula φ) with intensity i'. Note that traditional boolean propositional logic (where formulae are either true or false, rather than mapped to a value between $[0,1]$) would be sufficient for programming cognitive (BDI-based) agents [8]. However, for the computation of many emotions in BDTE we need values between $[0,1]$. For

example, an agent that applies for a new job cannot feel hope (according to BDTE) when it only knows if it got the job afterwards. It should reason over the certainty of getting this job. For example, after having a good job interview. Also note that the emotions in Υ contain a formula, rather than just a label and intensity. With this we model the apparent directedness of emotions, in line with BDTE [17]. One is happy *about* some formula, e.g., φ = 'I will get a new job'.

Changing the knowledge representation is denoted with a combine operator $\oplus$. Given some set S and some set T containing a number of formulae, $S \oplus T$ denotes an update of S with T. $\oplus$ is a simple set join, with elements in set T taking priority over elements in set S, to allow updating of c, u and i in S. For all formulae $\varphi \in S$ and $\varphi \in T$, the mapping $\varphi \to n$ in the resulting set is taken from the set T. Thus, $\oplus$ is not symmetric, i.e., $S \oplus T \neq T \oplus S$.

Definition 3. (Combine $\oplus$)
Given some sets S, and T, which contain a number of elements $e = \{\varphi \to n\}$, where φ is a formula $\varphi \in \mathcal{L}$, and n a value $n \in [0,1]$. $S \oplus T$ is defined as follows:

$$e \in S \oplus T \qquad \textit{iff} \qquad e \in T, \ \textit{or} \ (e \in S \ \textit{and} \ e \notin T)$$

A knowledge representation is a pair $\langle \mathcal{L}, \oplus \rangle$, where $\mathcal{L}$ is a language to represent states of affairs, and $\oplus$ defines how a set of formulae is updated with another set of formula. Using our definition of a knowledge representation, we can now formally define what a *mental state* of an agent is. We call this initial definition a 'Simple Mental State' because we will expand it later in the chapter.

Definition 4. (Simple Mental State)
A mental state is a pair $\langle \Sigma, \Gamma \rangle$ where Σ is called a belief base, and Γ is a desire base.

The aim of the work presented here is to add *emotional reasoning* to these agent programming frameworks. The belief-desire theory of emotion (BDTE) [16, 17] provides a method for computing emotional responses based solely on ones beliefs and desires. For BDTE we need only the beliefs and desires, before and after an agent's update of its mental state. We could imagine that a computation of an agent program is a sequence of mental states $m_0, m_1, m_2, \ldots$. BDTE then enables the computation of an agent's emotions in a mental state m_i by using the belief- and desire base corresponding to mental states m_{i-1} and m_i. Based on BDTE we can define the inner workings of this function [17].

Definition 5 describes BDTE in a computational manner. This is based on CBDTE [17]. In function $R(\Sigma, \Sigma', \Gamma, \Gamma') \to \Upsilon$ (R for Reisenzein's appraisal [17]), we denote Σ as the belief base of mental state m_{i-1}, Γ as the desire base of mental state m_{i-1}, Σ' as the belief base in mental state m_i, and Γ' as the desire base of mental state m_i. The function $R(\Sigma, \Sigma', \Gamma, \Gamma')$ computes all new emotions resulting from changes in the mental state.

Definition 5. (BDTE R)
Given function $R(\Sigma, \Sigma', \Gamma, \Gamma') \to \Upsilon$. Let S be the set containing all φ such that $b\{\varphi \to c\} \in \Sigma$, $b\{\varphi \to c'\} \in \Sigma'$, $d\{\varphi \to u\} \in \Gamma$, and $d\{\varphi \to u'\} \in \Gamma'$, with $c \neq c'$, or $u \neq u'$. $S = \{\varphi_1, \ldots, \varphi_n\}$. If we iterate through S with $i = 1..n$, add the following emotions as follows: $\Upsilon = E_1 \oplus E_2 \oplus \ldots \oplus E_n$, such that:

$e\{\varphi_i \times happy \to u\} \in E_i$	*iff*	$c' = 1 \;\&\; u > 0$
$e\{\varphi_i \times unhappy \to u\} \in E_i$	*iff*	$c' = 0 \;\&\; u > 0$
$e\{\varphi_i \times hope \to c' \times u\} \in E_i$	*iff*	$0 < c' < 1 \;\&\; u > 0$
$e\{\varphi_i \times fear \to (1 - c') \times u\} \in E_i$	*iff*	$0 < c' < 1 \;\&\; u > 0$
$e\{\varphi_i \times surprise \to 1 - c\} \in E_i$	*iff*	$c' = 1$
$e\{\varphi_i \times surprise \to c\} \in E_i$	*iff*	$c' = 0$
$e\{\varphi_i \times relief \to 1 - c\} \in E_i$	*iff*	$c' = 1 \;\&\; u > 0$
$e\{\varphi_i \times disappointment \to c\} \in E_i$	*iff*	$c' = 0 \;\&\; u > 0$

For example, let $\varphi_1 =$*'I got a new job'*, $b\{\varphi_1 \to 1\} \in \Sigma'$ (i.e., the agent beliefs to have gotten a new job), and $d\{\varphi_1 \to 0.9\} \in \Gamma$ (i.e., the agent strongly desires to have gotten a new job), then Definition 5 prescribes $e\{\varphi_1 \times happy \to 0.9\} \in \Upsilon$ (i.e., the agent is very happy that it got a new job).

With these definitions we already have a framework to implement emotions, which basically works as proposed in previous work [17]. We might imagine that the computation of an agent program results in a sequence of mental states $m_0, m_1, m_2, \ldots$. Computing emotions can then be done by computing Υ over two consecutive mental states. However, this approach does not take into account that emotion intensities decay over time, how to deal with multiple appraisals of the same emotion label (θ), or the fact that you might want to store emotions for reasoning purposes. Furthermore, computation based on BDTE gives a large set containing multiple emotions for every formula φ the agent has in its mental state, meaning we need a method to abstract useful information from it.

5.3.2. Closing the Semantic Gap between BDTE and BDI

In this Section we expand the model such that BDTE can be used for agent programming in an efficient way, including decay, repeated appraisals, and querying the emotions. We start with expanding the mental state of an agent with an emotion base. With this we can store the current emotional state of an agent, and query this when needed.

Definition 6. (Mental State)
A mental state is a triple $\langle \Sigma, \Gamma, \Upsilon \rangle$ *where* Σ *is called a belief base,* Γ *is a desire base, and* Υ *is an emotion base.*

With an emotion base storing the emotional responses we can now define a function that gradually decays the intensities of the stored emotions. Function $d(\Upsilon, \Delta t)$ is responsible for decaying the emotional state Υ over time Δt. For the consistency of our model (see Section 5.4) we define Δt to be zero within one reasoning cycle of an agent. Between reasoning cycles, Δt is a function over the actual system time passed between the start of the previous and current reasoning cycle. Function decay is a mapping $d : \Upsilon \to \Upsilon'$, that decreases the intensity $i \in [0, 1]$ for all elements $e\{\varphi \times \theta \to i\} \in \Upsilon$.

Definition 7. (Decay Function d)
Let $e\{\varphi \times \theta \to i\} \in \Upsilon$. d *is a function* $d(\Upsilon, \Delta t) \to \Upsilon'$ *defined as:*

$$e\{\varphi \times \theta \to f(\theta, i, \Delta t)\} \in d(\Upsilon, \Delta t) \qquad \textit{iff} \qquad e\{\varphi \times \theta \to i\} \in \Upsilon$$

Where $f(\theta, i, \Delta t)$ is a function that decreases the intensity i, and for all emotions $e \in \Upsilon$ the emotion also exists in Υ' with a decayed intensity. The function can be initialized differently for every emotion label $\theta \in \Theta$. An example of exponential decay for happy would be: $f(happy, i, \Delta t) = i - i \times \Delta t$.

We adopt the view in [7] that decay may need different instantiations for different emotions, depending on the corresponding emotion label $\theta \in \Theta$. For example, hope and fear may decay slower than surprise. In our model an agent programmer can adjust the default decay function, for every emotion label independently.

The above defined functions come together in (i.e., are sub-functions of) function **EM**. This function is a mapping: **EM**$(\Sigma \times \Sigma \times \Gamma \times \Gamma \times \Upsilon) \rightarrow \Upsilon$.

Definition 8. (Emotion Base Transformer **EM**)
*Let Σ, Γ, and Υ be a belief base, desire base, and emotion base in some mental state m. Further, let Σ', and $dbase'$ be the belief base and desire base after some update on this mental state. Function **EM**$(\Sigma \times \Sigma' \times \Gamma \times \Gamma' \times \Upsilon) \rightarrow \Upsilon'$ computes the emotion base in this updated mental state as follows:*

$$\Upsilon' = d(\Upsilon, \Delta t) \oplus R(\Sigma, \Sigma', \Gamma, \Gamma')$$

This function is called when the belief base or desire base of an agent change. This happens through *updates*. There is a set of build-in updates that act on the mental state bases of the agent. Updates change the belief and desire bases of the agent. Whilst performing these updates, the agent will automatically add emotions to its emotion base Υ.

Definition 9. (Mental State Transformer $\mathcal{M}$)
Let $\varphi \in \mathcal{L}$, and $n \in [0, 1]$. The mental state transformer function $\mathcal{M}(update, m) \rightarrow m'$ *is a mapping from built-in updates (update = [insert, adopt, drop]) and mental states $m = \langle \Sigma, \Gamma, \Upsilon \rangle$ to mental states as follows:*

$$\begin{array}{lcl} \mathcal{M}(\textbf{\textit{insert}}(\varphi, n), m) & = & \langle \Sigma \oplus \{\varphi \rightarrow n\}, \Gamma, \Upsilon' \rangle \\ \mathcal{M}(\textbf{\textit{adopt}}(\varphi, n), m) & = & \langle \Sigma, \Gamma \oplus \{\varphi \rightarrow n\}, \Upsilon' \rangle \\ \mathcal{M}(\textbf{\textit{drop}}(\varphi), m) & = & \langle \Sigma, \Gamma \oplus \{\varphi \rightarrow 0\}, \Upsilon' \rangle \end{array}$$

*with Υ' = **EM**$(\Sigma, \Sigma', \Gamma, \Gamma', \Upsilon)$, where Σ' is the belief base, and Γ' is the desire base in the resulting mental state m'.*

Mental state bases are defined as sets, thus, if a previous mapping $\{\varphi \rightarrow n\}$ exists in the mental state, then the updates defined above overwrite the previous mapping. In BDTE the claim is made that emotions are subconscious meta-representations of ones beliefs and desires [17]. In the definition above, we model this with function **EM**, which automatically updates the emotions when updating the beliefs, and desires in the mental state.

Definition 10. (Transition rule)
Let m be a mental state, and u be an update ([insert, adopt, drop]) performed in

mental state m. The transition relation $\xrightarrow{\mathrm{u}}$ is the smallest relation induced by the following transition rule.

$$\frac{\mathcal{M}(\mathrm{u}, m) \text{ is defined}}{m \xrightarrow{\mathrm{u}} \mathcal{M}(\mathrm{u}, m)}$$

The execution of an agent as explicated above, results in a *computation*. A computation in this context is a list of mental states and corresponding updates, performed by the agent. The new mental state is derived from the transition rule in Definition 10. The agent chooses its next update from the set of possible updates in the current state, this set is filled through the rules defined by the programmer. The computation starts in the initial mental state of the agent.

Definition 11. (Mental Computation)
A mental computation is a sequence of mental states $m_0, \mathrm{u}_0, m_1, \mathrm{u}_1, m_2, \mathrm{u}_2, \ldots$ such that for each i we have that $m_i \xrightarrow{\mathrm{u}_i} m_{i+1}$ can be derived using the transition rule of Definition 10.

The emotion update function **EM** is triggered as part of the Mental State Transformer (Definition 9). It is a part of the mapping from $m_i \xrightarrow{\mathrm{u}_i} m_{i+1}$. Emotions are thus computed after every mental state change of an agent.

Figure 5.1 showed the reasoning cycle of an agent. The mental computation, defined in Definition 11, operates solely in the 'updates of mental state' box. This means that in the model presented here, an agent senses its environment and starts updating its mental state based on these observations. With these mental state updates, we now defined how emotions are automatically changed accordingly. After updating its mental state, the agent can choose a new action to perform in the environment, which in turn changes the environment. The agent then again senses the changes in the environment, and the cycle starts anew.

5.3.3. Querying the Emotion Base

Querying the emotion base of an agent is useful. For example, if one wants to know if the agent is happy then one should inspect the emotion base for formulae about which the agent is happy. However, a computation based on BDTE gives a large set containing multiple emotions for every formula φ the agent has in its mental state. We therefore need a function that abstracts over these formulae.

To model this, we define an overall *affective state*, which summarizes the agent's emotions. We compute this affective state with function A. This function computes abstractions from the emotion base that enable a programmer to, for example, query the overall happiness of an agent. It summarizes the emotions in some emotion base Υ. It does so by taking all formulae in the emotion base Υ, for all emotion labels $\theta \in \Theta$, and computing a single intensity from these emotions in Υ concerning the emotion label θ.

Besides the computational argumentation there is also a psychological argumentation to define the affective state. In [27] Reisenzein argues that emotions have a hedonic tone, different than that of beliefs and desires. It *feels* a certain

way to have an emotion, which is essentially different from how a belief or desire feels. In his own words: "To account for the hedonic tone of emotions in BDTE, one must assume that 'emotional' belief-desire configurations cause a separate mental state that carries the hedonic tone. [27]" By means of an affective state we model this hedonic tone of emotions.

Definition 12. (Affective State Ω)
Ω is a function, that computes a generalized affective state which summarizes the emotions $e\{\varphi \times \theta \rightarrow i\} \in \Upsilon$ for some emotion label $\theta \in \Theta$.

$$\Omega(\theta, \Upsilon) = \log_2(\sum_{e\{\varphi\times\theta\rightarrow i\}\in\Upsilon} 2^{i\times 10})/10$$

In our model we have implemented $\Omega(\theta, \Upsilon)$ with a logarithmic function ($Log_2(\sum 2^{i\times 10})/10$), where we sum over all emotions $e\{\varphi \times \theta \rightarrow i\} \in \Upsilon$ corresponding to label θ. Other possible functions might be normal combine: $i' = I/(I+1)$, with I the summation of all intensities concerning θ), or a simple MAX function (taking the highest intensity emotion corresponding to θ.

From these functions the logarithmic is computationally speaking slightly less efficient; however, the function forces the resulting intensity to be as least as large as the highest value, but takes other values into account. For example, happiness about three different propositions: φ_1 = 'Getting a new job', φ_2 = 'Buying a new car', and φ_3 = 'Going out for dinner', with corresponding intensities: $[0.7, 0.6, 0.3]$, will compute to an overall happiness of 0.76 with logarithmic combine, to 0.62 with normal combine, and to 0.7 with the MAX function.

We do not claim that this is the only correct way to compute the overall affective state, but rather that an agent programmer *requires* a summary to efficiently query the emotion base, and that the here proposed approach will thus help the programmer.

5.4. Proof of Consistency when Minimizing the (Re)Appraisal of Emotions

In Section 5.3, we defined the (re)computation of an agent's emotions to occur after every mental state update. However, this is not a computationally optimal approach. In this Section we show how one can optimize this by showing when an agent should minimally (re)compute its emotions (i.e., when the agent should *(re)appraise*).

There are three conditions that should trigger a reappraisal: 1, An agent should reappraise before querying its emotion base, if it has updated its mental state since the last reappraisal, since otherwise it would query an outdated emotional state. 2, An agent should reappraise before a mental state update if the last reappraisal was in a previous reasoning cycle, otherwise the emotions are not correctly decayed. 3, An agent should reappraise when it performs a mental state update on a formula that had already been updated after the last reappraisal, otherwise the previous update will be lost. Since 1 and 2 directly follow from the formal semantics, we need only to show that 3 is true. We do so by proving that if we assume that

updates refer to different formulae, appraisal can be postponed to the last update. From this one can infer point 3.

Theorem 1. *Consistency For Delayed Appraisal*
Let $\mathfrak{u}_1, \mathfrak{u}_2, .., \mathfrak{u}_n$ *be different mental state updates, with* $\varphi_1, \varphi_2, \ldots, \varphi_n$ *the formulae these updates refer to respectively. Furthermore, let* $\mathfrak{u}'_1, \mathfrak{u}'_2, .., \mathfrak{u}'_n$ *be the same mental state updates; however, for these mental state updates we define the Mental State Transformer (Definition 9) to delay updating the emotion base until* $\mathfrak{u}'_n$*. Furthermore let* $\varphi_1 \neq \varphi_2 \neq \ldots \neq \varphi_n$*. Consider the following two possible reasoning cycles:*

$$rc_1: \quad m_0 \xrightarrow{\mathfrak{u}_1} m_1 \xrightarrow{\mathfrak{u}_2} \ldots \xrightarrow{\mathfrak{u}_n} m_n$$
$$rc_2: \quad m_0 \xrightarrow{\mathfrak{u}'_1} m'_1 \xrightarrow{\mathfrak{u}'_2} \ldots \xrightarrow{\mathfrak{u}'_n} m'_n$$

where rc_2 *delays updating the emotion base until update* $\mathfrak{u}'_n$*. Under the constraint that* $\varphi_1 \neq \varphi_2 \neq \ldots \neq \varphi_n$*, we can derive that* $m_n = m'_n$*.*

5

To show the truth of this claim, let the knowledge bases corresponding to mental state m_i be denoted with, $m_i = \langle \Sigma_i, \Gamma_i, \Upsilon_i \rangle$. Since Σ and Γ are updated normally we need only to show that $\Upsilon_n = \Upsilon'_n$. To this end, we first need to define a property of the definitions. We defined Δt in function d (decay) to be zero within one reasoning cycle. Furthermore, $d(\Upsilon, 0) = \Upsilon$. Due to this, we can ignore decay when comparing reasoning cycles rc_1 and rc_2. If we denote E_i to be the set of emotions resulting from function R in transition $m_{i-1} \xrightarrow{\mathfrak{u}_i} m_i$, then we can write:

$$\begin{aligned}
\Upsilon_1 &= d(\Upsilon_0, 0) \oplus E_1 \\
&= \Upsilon_0 \oplus E_1 \\
\Upsilon_2 &= d(\Upsilon_0 \oplus E_1, 0) \oplus E_2 \\
&= \Upsilon_0 \oplus E_1 \oplus E_2 \\
\Upsilon_n &= \Upsilon_0 \oplus E_1 \oplus E_2 \oplus \ldots \oplus E_n.
\end{aligned}$$

The emotion base resulting from reasoning cycle 2 can be found with the same definitions. Since the update of the emotion base is delayed, the emotion base $\Upsilon'_{n-1} = \Upsilon_0$. Furthermore, the computation of new emotions (Definition 5) will consider all updated formulae:

$$\begin{aligned}
\Upsilon'_n &= d(\Upsilon_0, 0) \oplus \{E_1 \oplus E_2 \oplus \ldots \oplus E_n\} \\
&= \Upsilon_0 \oplus E_1 \oplus E_2 \oplus \ldots \oplus E_n.
\end{aligned}$$

If $\varphi_1 \neq \varphi_2 \neq \ldots \neq \varphi_n$, then the emotions in sets $E_1, \ldots, E_n$ do not overwrite each other when added to the emotion bases. Therefore, we can conclude that $\Upsilon_n = \Upsilon'_n$. Together we can now also conclude $m_n = m'_n$.

5.5. Discussion

In this section we discuss some drawbacks of using BDTE as psychological background. BDTE models a limited range of emotions compared to other theories

(BDTE models 7 emotions, while, for example, OCC models over 20 different emotions). Should an agent programmer want to use the emotions in the agent's decision making, then a smaller set of emotions might be more conceivable; however, there can also be domains in which the set of emotions modelled by BDTE is too limited. For example, when a programmer needs the agent to properly reason over empathic emotions like gratitude and remorse, then BDTE is inadequate in its current form.

Future work could thus complement this framework by modelling social emotions. In [28], Reisenzein discusses possible extensions of BDTE to take social emotions into account. For example, he proposes introducing *altruistic desires*. For example, pity is then explained as a form of displeasure following from the frustration of an *altruistic desire* (desiring something good for someone else). However, this does not provide explanations for all social emotions (e.g., anger). When adding social emotions, one might need to complement the presented framework with additional concepts such as norms.

5.6. Conclusion

In this chapter, we presented CAAF (a Cognitive Affective Agent programming Framework), a framework where emotions are computed automatically when agents update their mental states. We presented semantics showing the programming constructs of these agents in a domain-independent manner. With these constructs, a programmer can build an agent program with cognitive agents that automatically compute emotions during runs. We chose BDTE to compute new emotions because it is conceptually close to the BDI architecture and therefore allowed us to embed emotions without introducing many additional concepts in the mental states of the agents.

Our semantics facilitate incremental work. For example, if it is desirable to change the affective state (Definition 12) with a global mood, then one could change the function that computes the affective state (function A), without being forced to adjust the entire framework. One might also want to enable programmers to adjust the emotion base without changing the belief base. Definition 9 defined functions to update the agent's mental state. We could simply complement this definition to contain function $Appraise$, capable of inserting emotions in the emotion base (Υ), similar to the update $insert$ for the belief base (Σ). This fits well in the modular approach suggested by Marsella et. al. [12], where models can implement parts of a complete cycle of emotional reasoning. For example, one could add a module capable of using emotions to guide the agent's decision making (e.g., what action to perform in the environment, or when to decrease the utility of a desire as a type of coping behaviour). The framework presented in this chapter provides a modular, domain-independent, and consistent implementation for the computation of emotions for cognitive agent programming frameworks, thereby facilitating the development of intelligent virtual agents with affective abilities.

References

[1] F. Kaptein, J. Broekens, K. V. Hindriks, and M. Neerincx, *Caaf: A cognitive affective agent programming framework,* in *Intelligent Virtual Agents* (2016) pp. 317–330.

[2] A. Popescu, J. Broekens, and M. van Someren, *Gamygdala: An emotion engine for games,* IEEE Transactions on Affective Computing **5**, 32 (2014).

[3] P. Rizzo, *Why should agents be emotional for entertaining users? a critical analysis,* in *Affective interactions* (Springer, 2000) pp. 166–181.

[4] R. Beale and C. Creed, *Affective interaction: How emotional agents affect users,* International journal of human-computer studies **67**, 755 (2009).

[5] J. Dias and A. Paiva, *Feeling and reasoning: A computational model for emotional characters,* (2005).

[6] J. Bates *et al.*, *The role of emotion in believable agents,* Communications of the ACM **37**, 122 (1994).

[7] W. S. Reilly, *Believable Social and Emotional Agents.*, Tech. Rep. (DTIC Document, 1996).

[8] K. V. Hindriks, *Programming rational agents in goal,* in *Multi-Agent Programming:* (Springer, 2009) pp. 119–157.

[9] A. Pokahr, L. Braubach, and W. Lamersdorf, *Jadex: A bdi reasoning engine,* in *Multi-agent programming* (Springer, 2005) pp. 149–174.

[10] R. H. Bordini, J. F. Hübner, and M. Wooldridge, *Programming multi-agent systems in AgentSpeak using Jason*, Vol. 8 (John Wiley & Sons, 2007).

[11] S. Russell, P. Norvig, and A. Intelligence, *A modern approach,* Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs **25**, 27 (1995).

[12] S. Marsella, J. Gratch, and P. Petta, *Computational models of emotion,* A Blueprint for Affective Computing-A sourcebook and manual **11**, 21 (2010).

[13] R. S. Lazarus, *Emotion and adaptation.* (Oxford University Press, 1991).

[14] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions* (Cambridge university press, 1990).

[15] K. R. Scherer, *Appraisal theory,* in *Handbook of cognition and emotion*, edited by T. Dalgleish and M. J. Power (1999) pp. 637–663.

[16] R. Reisenzein, *Appraisal processes conceptualized from a schema-theoretic perspective: Contributions to a process analysis of emotions.* in *Appraisal processes in emotion: Theory, Methods, Research*, edited by K. R. Scherer, A. Schorr, and T. Johnstone (Oxford University Press, 2001) pp. 187–201.

[17] R. Reisenzein, *Emotions as metarepresentational states of mind: Naturalizing the belief–desire theory of emotion,* Cognitive Systems Research **10**, 6 (2009).

[18] C. Adam, A. Herzig, and D. Longin, *A logical formalization of the occ theory of emotions,* Synthese **168**, 201 (2009).

[19] B. R. Steunebrink, M. Dastani, and J.-J. C. Meyer, *The occ model revisited,* in *Proc. of the 4th Workshop on Emotion and Computing* (2009).

[20] K. R. Scherer, *Appraisal considered as a process of multilevel sequential checking,* in *Appraisal processes in emotion: Theory, methods, research*, edited by K. R. Scherer, A. Schorr, and T. Johnstone (Oxford University Press, 2001) pp. 92–120.

[21] J. Gratch and S. Marsella, *A domain-independent framework for modeling emotion,* Cognitive Systems Research **5**, 269 (2004).

[22] S. C. Marsella and J. Gratch, *Ema: A process model of appraisal dynamics,* Cognitive Systems Research **10**, 70 (2009).

[23] J. Broekens, D. DeGroot, and W. A. Kosters, *Formal models of appraisal: Theory, specification, and computational model,* Cognitive Systems Research **9**, 173 (2008).

[24] C. D. Elliott, *The affective reasoner: A process model of emotions in a multi-agent system,* (1992).

[25] M. S. El-Nasr, J. Yen, and T. R. Ioerger, *Flame—fuzzy logic adaptive model of emotions,* in *Autonomous Agents and Multi-agent systems* (Springer, 2000) pp. 219–257.

[26] J. Dias, S. Mascarenhas, and A. Paiva, *Fatima modular: Towards an agent architecture with a generic appraisal framework,* in *Emotion Modeling* (Springer, 2014) pp. 44–56.

[27] R. Reisenzein, *What is an emotion in the belief-desire theory of emotion?* (2012).

[28] R. Reisenzein, *Social emotions from the perspective of the computational belief-desire theory of emotion,* in *The Cognitive Foundations of Group Attitudes and Social Interaction* (Springer, 2015) pp. 153–176.

6

Evaluating Cognitive and Affective Intelligent Agent Self-Explanations for Long-Term Health-Support

Explanation of actions is important for transparency of-, and trust in the decisions of smart systems. Literature suggests that emotions and emotion words - in addition to beliefs and goals - are used in human explanations of behaviour. Furthermore, research in e-health support systems and human-robot interaction stresses the need for studying long-term interaction with users. However, state of the art explainable artificial intelligence for intelligent agents focuses mainly on explaining an agent's behaviour based on the underlying beliefs and goals *in* short-term experiments*. In this chapter, we report on a long-term experiment in which we tested the effect of cognitive, affective and lack of explanations on children's motivation to use an e-health support system. Children (48 children aged 6-14) suffering from type 1 diabetes mellitus interacted with a virtual robot as part of the e-health system over a period of 2.5 - 3 months. Children alternated between the three conditions. Agent behaviours that were explained to the children included why 1) the agent asks a certain quiz question; 2) the agent provides a specific tip (a short instruction) about diabetes; or, 3) the agent provides a task suggestion, e.g., play a quiz, or, watch a video about diabetes. Their motivation was measured by counting how often children would follow the agent's suggestion, how often they would continue to play the quiz or ask for an additional tip, and how often they would request an explanation from the system. Surprisingly, children proved to follow task suggestions more often when* no explanation *was given, while other explanation effects did not appear. This is to our knowledge the first long-term study to report empirical evidence for an agent explanation effect, challenging the next studies to uncover the underlying mechanism.*

Parts of this chapter have been published in 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019) [1]

6.1. Introduction

Humans are increasingly supported by Artificial Intelligence (AI), for example, at home using virtual assistants, in health care settings, and in education [2]. Transparency of why such systems provide particular advice or choose certain actions, as well as user trust in such systems, is important [3–5]. Therefore, the ability to provide explanations to motivate the reasoning behind the AI's decisions, i.e., eXplainable AI (XAI), becomes increasingly important. This trend is supported by the recent General Data Protection Regulation (GDPR) law, which states that users have the right to explanations [6].

Current XAI for agents is often based on folk psychology, i.e., how humans in their everyday lives explain their decisions amongst each other [7]. Such explanations are based on the beliefs and goals of the system. For example, 'I suggest you watch this video about diabetes because I **think** (a system belief) it contains valid information about proper blood sugar levels, and I **want** (a system goal) you to learn when your blood sugar level would be too low'. Using beliefs and/or goals for explaining intentional behaviour is common in both human-human communication and in XAI [8–12]. We refer to this as providing *cognitive explanations*.

Literature suggests that emotions and emotion words - in addition to beliefs and goals - are used in human explanations of behaviour [13–15]. Humans explain their decisions also based on their emotions. For example, 'I called the hospital because I was **scared** (emotion) that I might have a hypo (too low blood sugar level)'. As such, explanations of agents based on beliefs and/or goals may not always be sufficient and emotions may be required as part of the explanations in human-agent interaction [16].

Furthermore, research in e-health support systems and human-robot interaction stresses the need for studying long-term interaction with users [2, 17–19]. However, state of the art of XAI for intelligent agents has focussed mainly on explaining an agent's behaviour based on the underlying *beliefs and/or goals* in *short-term experiments* [5, 9, 12, 20].

In this chapter, we report on a long-term experiment in which we tested the effect of cognitive, affective and lack of explanations on children's motivation to use an e-health support system. Children (aged 6-14) suffering from Type 1 Diabetes Mellitus (T1DM) interacted with a virtual robot as part of the e-health system over a period of 2.5 to 3 months. Children alternated between the three conditions. Agent behaviours that were explained to the children included why 1) the agent asks a certain quiz question; 2) the agent provides a specific tip (a short instruction) about diabetes; or, 3) the agent provides a task suggestion, e.g., play a quiz, or, watch a video about diabetes. Their motivation was measured by counting how often children would follow the agent's suggestion, how often they would continue to play the quiz or ask for an additional tip, and how often they would request an explanation from the system.

6.2. Motivation, Related Work, and Hypothesis

First, we motivate why intelligent agents in consequential domains, such as healthcare, must be able to explain their behaviour. As computer systems become more powerful, more complexity is introduced in their decision making [5]. To maintain trust in a system in the long-term, the system must be clear about the task it is trying to achieve [2]. Lack of trust in a behaviour change system causes users to not rely on the given advice [21], and can cause them to misuse or even abandon the system [22]. XAI has been shown to have a positive impact on a user's trust in several studies [3, 4, 23, 24]. Indeed, such consequential domains often include explainable AI for transparency and intelligibility [7].

Now we motivate why emotions need to be considered in the generation of explanations. XAI is typically based on how humans explain their behaviour amongst each other, i.e., on *folk psychology* [13, 14, 25]. This refers to the use of beliefs, goals and emotions to explain behaviour [14, 15]. Explanations using beliefs and goals (which we call *cognitive explanations*) are often used in XAI [8–12]. However, using emotions and emotion words for explanations (which we call *affective explanations*) has not yet been properly tested in XAI. Still, synthetic emotions expressed by agents have the potential to influence user attitudes and behaviour [26], and explanations of agents based on beliefs and goals may not always be sufficient, emotions may be required as part of the explanations in human-agent interaction [16].

Finally, we motivate why long-term experiments are essential. Explanations are typically done by using the agent's beliefs and/or goals and in short-term experiments [5, 7, 9, 12, 20]. However, the importance of testing long-term effects has been stressed in human-robot interaction and e-health [2, 17–19]. Long-term interaction typically has more repetition of information and interaction patterns, and such systems need to overcome novelty effects. Related work shows that reasons to stop using a robot change over time [19]. In the short-term, the robot must be enjoyable and easy to use, in the long-term it must be functionally relevant.

The context of this work is the PAL project (a Personal Assistant for a healthy Lifestyle). Here we develop a support application with a (Nao) robot and virtual avatar thereof that helps children (aged 6-14) with T1DM to cope with their illness. The child sets personal learning goals with the caregiver, such as, 'recognise hypo and correct blood sugar accordingly'. The PAL agent then shapes the activities to support the child to achieve these goals. For example, during the quiz PAL might ask the child what the child should do when (s)he suddenly starts shaking and is feeling very hungry. The child can then ask PAL why the agent asks the child this question. The XAI module developed and reported upon here enables the system to respond along the lines of: 'I would be happy for you if you learn how to recognise that you have a hypo, and learn what you should then do'.

Because agent explanation of action is important for trust and motivation, because emotions need to be considered as part of the explanations, and because XAI needs to be evaluated in such long-term experiments, we address the following question:

What is the effect of cognitive, affective and lack of explanations on the motivation of children to use an e-health support system in long-term interaction?

We look at several motivational effects of explanation style and split our research question into four hypotheses. First, we want to know if children appreciate and use explanations. We assess this by measuring the total number of requested explanations.

Hypothesis 1. *There is a difference in total number of requested explanations induced by explanation style (cognitive versus affective explanations).*

Second, we expect explanations to have an effect on the usage of the system. People desire to know the goals they are pursuing when being educated [27, 28]. Explanations may help a user to better understand why an action is proposed, thereby understanding the learning goal. In previous work it was found that adults, more than children, prefer goal-based over belief-based explanations [12]. Here we are interested in the effect of cognitive versus affective explanations.

Hypothesis 2. *There is a difference in the average number of questions in a quiz before children close it given the explanation style (cognitive versus affective versus lack of explanation).*

Hypothesis 3. *There is a difference in how often children request an additional tip given the explanation style (cognitive versus affective versus lack of explanation).*

Finally, to directly assess the motivational value of explanations, we look at how often a task suggestion by the system is followed. We expect such task suggestions to be followed more often when they are explained because in general people are more motivated to learn something when they know why they should learn it [27, 28].

Hypothesis 4. *There is a difference in how often children follow a task suggestion after they received an explanation, induced by explanation style (cognitive versus affective versus lack of explanation).*

6.3. Implementation of a Model for Explainable AI

In our model, explanations consist of some raw *content* and a *presentation* of the content. The content of the explanation is the goal that the agent is pursuing with its behaviour. The presentation is the resulting set of sentences generated. We consider two different *styles* in which these sentences can be formulated, (i.e., *cognitive* and *affective* explanations).

6.3.1. Explainable Actions

We explain three different types of actions shown by the PAL agent. 1) Asking the user a quiz question (e.g., 'What should you do when you are experiencing a hypo whilst doing sports?'). 2) Giving the user a *tip of the day* or shortly a *tip* (e.g., 'When

your blood sugar level is below 4.0 mmol/L then you have a hypo'). 3) Suggesting a task to do (e.g., 'play the quiz' or 'watch this video').

Quiz questions and tips are activities that the child can do within the system. A child can play a quiz as often and for as many questions as they like. When a child requests a tip, then (s)he can request *next tips* as often as (s)he likes. Suggesting an activity happens when the child is shown a list of four possible activities ('tasks') to do in the system. This always happens when the application starts. Additionally, the child can request a (new) list of possible tasks at any moment. The PAL agent then always suggests that the top-most task would currently be the best task to do. It can potentially motivate this suggestion further by explaining *why* it thinks the child should do that activity (See also figure 6.1). The text used to suggest a task is chosen randomly from a set of pre-made sentences. For example, in figure 6.1.a the text is 'Let's do the first activity' and in figure 6.1.c the text is 'I think you should do this first activity'. In the cognitive and affective styles, the explanation and task suggestion texts are concatenated in a single text balloon.

Figure 6.1: Four screen-shots of the PAL system. Screen-shots (a-c) show task suggestions, screen-shot (d) shows the quiz. During a task suggestion the user is always shown a list of four possible tasks. The top-most task is then suggested by the PAL agent as being the 'best' to do at the current time. In screen-shot (a) the PAL agent explains why it is a good task to do by providing a cognitive explanation, in (b) it provides an affective explanation, and in (c) it provides no explanation for its suggestion. Finally, screen-shot (d) shows an example of an affective explanation given during the quiz.

6.3.2. Content of explanations

The content of the explanations is the goal that the agent is trying to pursue. Which is a common approach in XAI [8–12]. However, an action often pursues multiple goals. For example, (a proposal for) watching a video can be valuable for a large list of unrelated learning goals (like, 'recognise hypo', 'be able to talk with friends about diabetes', and 'start eating more vegetables'). This is a complicating factor since an explanation loses its value when it becomes too long [29], so we should not mention all the goals in an explanation.

Within PAL we chose a simple solution for this problem. We pick a random goal as content for the explanation to show the child why an activity is beneficial for the child's self-management. Clearly, we are not claiming that this is the best way of selecting content for the explanation. However, we do believe that this is a valid way that fits our purposes (i.e., measure the effect of explanations on a child's motivation to use the system in long-term interaction).

6.3.3. Presentation of explanations

With the content of the explanation being a single goal, we still need to share this information with the child. So, we need a way to transform it into some natural language sentence. We do this partly by automation and partly by annotation. The learning goals are annotated with a natural language sentence that describes them, e.g., 'how to recognise that your blood sugar level might be too high (hyper), and what you should then do'. We can then automatically put a sentence in front of that that completes the explanation, e.g., 'I want you to learn..'. And, we can add a sentence behind to refer to the explained action, like, 'That is why I ask you this question', or 'And that is why I gave you this tip (of the day)'. So a full explanation can be: 'I want you to learn how to recognise that your blood sugar level might be too high (hyper), and what you should then do. That is why I ask you this question.'

We differentiate the sentences before and after the description to prevent repetitiveness in sentences. For example, 'I want you to learn' can be interchanged with 'my aim is that you learn', and 'That is why I ask you this question' can be interchanged with 'So, remember the answer to this question well!'. We have 3 different sentences to precede the goal description and 5 different sentences for every explainable action to follow it. We implemented the explanations in three languages (English, Dutch, Italian), which is a strong proof of concept that similar implementation is possible in at least a large set of languages.

In addition, this implementation allowed us to differentiate the *style* of the explanation. We consider *cognitive explanations* and *affective explanations*. The cognitive explanations are phrased like above, affective explanations use emotion words in the phrasing of the explanations. For example, we can exchange the sentence 'I want to' with 'It would make me happy if you'. In that way, the full explanation becomes: 'It would make me happy if you learn how to recognise that your blood sugar level might be too high (hyper), and what you should then do. That is why I ask you this question.' This shows that this implementation enables providing, with a very simple manipulation of the sentence generation, explanations in different styles.

6.4. Method

We evaluated the different explanation styles in a long-term (2.5 - 3 months) experiment.

6.4.1. Participants

In total there were 48 (25 Dutch and 23 Italian) children with T1DM aged 6-14. The children were recruited via hospitals in the Netherlands and in Italy. There were no consequences to dropping out intermediately.

6.4.2. Experimental Design

When a child logs into the system (s)he is set to an initial experimental condition randomly. There are three possible conditions, *Cognitive Explanations*, *Affective Explanations*, and *No Explanations*. The children rotate between the three conditions (within-subjects testing).

It was not possible to test our hypothesis between subjects in this particular experiment. This experiment is part of a larger project where multiple experiments have been tested simultaneously. A requirement was therefore that all children would see the same content in the system. This meant that it was not possible to distribute the conditions randomly over the children and then keep them in that condition.

There were two phases of the experiment. The system had some small differences in the two phases. Task suggestions are only given in the second phase. Quizzes and tips were given in both phases. Furthermore, there were minor changes between the phases in activities without explanations. The experimental conditions switched per week in the first phase and per log-in in the second phase. We changed this in the second phase because many children used the system actively for only one or two weeks, which causes them to not have enough exposure to the different conditions. Children that participated in the first phase were allowed to do so again in the second phase. 4 children (Dutch) and 9 children (Italian) did both phases.

Finally, both cognitive and affective explanations can be offered to the children in two different ways. 1) On the initiative of the PAL agent. Meaning the PAL agent simply gives the explanation for its behaviour. 2) On the initiative of the child. Meaning the system shows a question mark. The child can choose to press the question mark of his/her own accord. See figure 6.1.d for an example during the quiz.

Task suggestions are *always* explained when the child is in the cognitive or affective condition, and they are always explained on the initiative of the PAL agent. For the quiz and the tips the PAL agent provides explanations automatically 20% of the time. The other cases the child is shown a question mark. There is an exception to this. When the quiz is opened through the task suggestions rather than manually, then all questions in the quiz are for the same underlying goal which has already been mentioned during (the explanation for) the task suggestion itself. The explanations for the questions would always have the exact same content. Questions during a quiz opened in this way always only show a question mark.

6.4.3. Measures and Variables

For hypothesis 1, we test how often children request explanations of their own accord. We count how often children press the question marks (visible during the quiz and the tips) given an explanation condition (cognitive or affective). There is no measure in the no explanation condition since children cannot request explanations in that condition.

For hypothesis 2, we count the number of questions a child answered before closing the quiz. We then compute the average quiz length in the different styles for that child.

For hypothesis 3, the number of times the child manually request a 'next tip'. When the child receives a tip of the day, then the child can choose to either close the screen or press the 'next tip' button. We compute the average of next tip presses in the different styles for that child.

For hypothesis 4, we test whether children are more inclined to follow task suggestions in the different conditions (cognitive, affective, and lack of explanations). When presented with a task suggestion, the child can accept the suggestion by pressing the top-most task in the screen (see figure 6.1), or the child can reject the suggestion by either closing the screen or choosing another task in the list. We log the child's decision and measure the percentage of times the child actually chooses the suggested task given the explanation condition.

6.4.4. Material & Set-Up

There are two main locations where children interact with the PAL system, at home and at the hospital. At the hospital, the children interact with a physical Nao robot from Aldebaran and the PAL system. There they interact with a Health-Care Professional (HCP) and a researcher present. At home, they get a tablet with a virtual avatar of the robot and the same health-care applications (quiz, sorting game, etc.). At home, they interact with the system individually.

6.4.5. Procedure

Children were first invited to come to a hospital. There they were introduced to the PAL agent and system. Together with the HCP, they set some specific goals to advance their self-management of their diabetes (e.g., 'learn to recognise when you might have a hypo'). The system shapes the activities and task suggestions to work towards those goals. At the end, the children were given the tablet with the avatar to take with them to their houses. For 2.5 to 3 months they could play with the PAL system as often and long as they wanted. At the end of the period, they were invited to the hospital again.

6.5. results

One child (out of 48) was excluded from analyses due to a glitch in the data caused by a system error. The remaining 47 children had an average of 19 log-ins (STD = 12.9, minimum = 1, maximum = 55). Only three children requested an explanation in both the cognitive and the affective style. In section 6.6, we discuss possible

improvements on our method for addressing the first hypothesis in future work.

For the second hypothesis, a one-way within subjects (or repeated measures) ANOVA was conducted to compare the effect of (IV) explanation style (cognitive, affective, and no explanations) on (DV) the average length of the quiz measured by the number of questions. There was no significant effect of the IV explanation style, Wilks' Lambda = 0.88, F (2,19) = 1.319, p = .291.

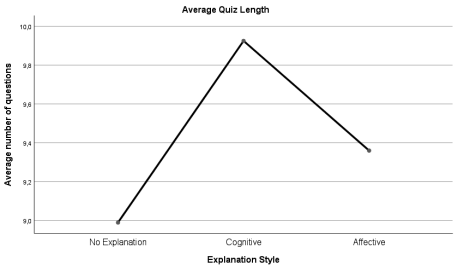


Figure 6.2: The average number of questions per child and per style before children close the quiz in the different explanation styles.

For the third hypothesis, a one-way within subjects (or repeated measures) ANOVA was conducted to compare the effect of (IV) explanation style on (DV) how often children request another tip. There was no significant effect of the IV explanation style, Wilks' Lambda = 0.93, F (2,45) = 1.772, p = .182.

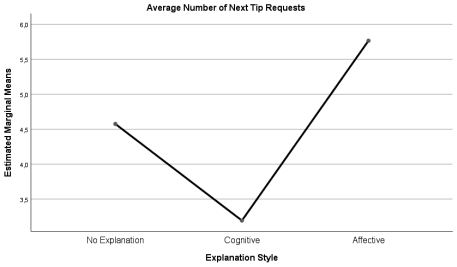


Figure 6.3: The average number of times per child and per style that children requested a next tip in the different explanation styles.

Finally for the fourth hypothesis, a one-way within subjects (or repeated measures) ANOVA was conducted to compare the effect of (IV) explanation style on (DV) the percentage of task suggestions followed by the children. There was a significant effect of the IV explanation style, Wilks' Lambda = 0.60, $F(2,13) = 4.285, p = .037$. In addition, three paired samples t-tests were used to make post hoc comparisons between conditions. A first paired samples t-test indicated that there was a significant difference in the percentage of task suggestions followed for no explanations ($M = 23\%$, $SD = 28\%$) and cognitive explanations ($M = 7\%$, $SD = 15\%$) conditions; $t(14) = 2.204, p = 0.045$. A second paired samples t-test indicated that there was a significant difference in the percentage of task suggestions followed

for no explanations ($M = 23\%$, $SD = 28\%$) and affective explanations ($M = 11\%$, $SD = 27\%$) conditions; $t(14) = 2.505, p = 0.025$. A third paired samples t-test indicated that there was no significant difference in the percentage of task suggestions followed for cognitive explanations ($M = 7\%$, $SD = 15\%$) and affective explanations ($M = 11\%$, $SD = 27\%$) conditions; $t(14) = -0.501, p = 0.624$. With a LSD test these values are significant; however, if we consider a Bonferroni correction then the significance threshold is 0.0167. So, the ANOVA test shows that explanation style has an effect on the percentage of task suggestions followed by the children; however, the post hoc tests are inconclusive concerning the effect's direction.

We did an additional test where we combined the cognitive and affective conditions and compared the combined (any explanation) group against the no explanation group. A paired samples t-test indicated that there was a significant difference in the percentage of task suggestions followed for no explanations ($M = 23\%$, $SD = 28\%$) and any explanations ($M = 9\%$, $SD = 16\%$) conditions; $t(14) = 2.950, p = 0.011$. This final test indicates that providing *no explanations* for task suggestions correlates with children following the suggested tasks more often.

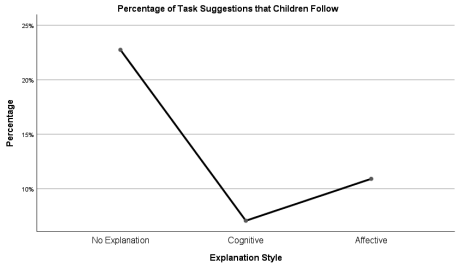


Figure 6.4: The percentage of task suggestions that children follow in the different explanation styles.

6.6. Discussion

The results come from a long-term 'in the wild' study. We recruited children aged 6-14 diagnosed with with T1DM. We are dealing with a real-world system (PAL) which is far more representative than a lab experiment could have been. However, this also means that the experiment was difficult to control. Children could stop the interaction with the system at any point in time. They could potentially request an explanation and close the application before the avatar could present it. Still, the system and the explanations were running robustly during the period of three months.

We found that explanation style influences how often children follow task suggestions. We found no further significant effects. This might be because the exposure of explanations during task suggestions was high. Every time children log-in the system the first thing they saw was a task suggestion which (in the cognitive and affective conditions) is always explained. During the quiz and the tip the explanations were not often explained in a forced manner. Most of the time, the children

would only see a question mark that they could press of their own accord. The results show that children did not press the question marks often. Since children already see an explanation in 20% of the cases, a case might be added in future work where children get *no* forced explanations to prevent potential saturation effects.

We did not expect that the *no explanation* condition would correlate with task suggestions being followed more often. We offer three possible explanations for this. 1) A straightforward explanation is that children simply do not read the longer texts in explained task suggestion (see also figure 6.1 for examples of differently explained task suggestions). This would result in more randomly chosen tasks from the menu. This would mean that the in literature suggested length of explanations [9] is still too long when applying explanations in a long-term experiment with child users. 2) Another possibility is that children *do* read and understand the explanations but they sometimes think they already know what the task is supposed to teach them. For example, if the PAL agent says the child should do a quiz because it teaches the child how to recognise when one might have a hypo, and if the child thinks (s)he already knows this, then the child is more likely to choose another task instead. This would relate to literature about teaching and learning, where it is suggested that explaining the importance of educational material helps students to orient/ plan their behaviour better themselves [30]. This would imply there *is* a positive effect of explanation style on the child's behaviour in the system. 3) The child might sometimes get stubborn from the explanation. Thinking something along the lines of 'I don't feel like practising / doing that!'. Which causes them to choose different tasks.

Future work should determine the underlying mechanism of why certain explanation styles change the users' behaviour in long-term interaction. A possible approach is to (sometimes) 'ask' the users why they chose a particular task after their selection. This was not possible in the here presented work due to limitations imposed by the project; however, it is our recommendation for future long-term experiments in this area. Secondly, the work here indicates that there is insufficient knowledge on when and how affective explanations should be used. When varying the style (but not the content) of your explanations, then (in the long-term) this may only trigger subtle differences in the users. A formal model of *when a particular type of explanation is preferred* is beneficial for further research in this area as this enables testing such a model against randomly chosen styles.

6.7. Conclusion

In this chapter, we presented results from a long-term (2.5 - 3 months) experiment on the effect of explanations on the motivation of children to use an e-health system involving interaction with a virtual robot. We considered cognitive explanations (based on the beliefs and goals of the agent), affective explanations (also using emotions of the agent for generating the explanation), and no explanations (providing no explanations at all for the agent's behaviour). The explanations were implemented in an in-the-wild autonomous health-support application for children (aged 6-14) suffering from T1DM. We found that explanation style influences how

often children follow task suggestions. Specifically, the results indicate that children follow the suggestions more often when *no explanation* is given. We found no other significant effects of explanations in this study. Although no effect was found of cognitive versus affective explanations, this is to our knowledge the first evidence that explanations impact long-term human-agent interaction and system usage. Our results also show that counter-intuitive effects of agent explanations may be expected when used with children, and, that more research is needed to understand why lack of explanations seems to correlate with following task suggestions.

References

[1] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, *Evaluating cognitive and affective intelligent agent explanations in a long-term health-support application for children with type 1 diabetes,* in *2019 Eighth International Conference on Affective Computing and Intelligent Interaction (in press)* (IEEE, 2019).

[2] I. Leite, C. Martinho, and A. Paiva, *Social robots for long-term interaction: a survey,* International Journal of Social Robotics **5**, 291 (2013).

[3] D. N. Lam and K. S. Barber, *Comprehending agent software,* in *Autonomous Agents and Multiagent Systems* (2005) pp. 586–593.

[4] B. Y. Lim, A. K. Dey, and D. Avrahami, *Why and why not explanations improve the intelligibility of context-aware intelligent systems,* in *Human Factors in Computing Systems* (2009) pp. 2119–2128.

[5] S. R. Haynes, M. A. Cohen, and F. E. Ritter, *Designs for explaining intelligent agents,* International Journal of Human-Computer Studies **67**, 90 (2009).

[6] P. Carey, *Data protection: a practical guide to UK and EU law* (Oxford University Press, Inc., 2018).

[7] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, *Explainable agents and robots: Results from a systematic literature review,* in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, 2019) pp. 1078–1088.

[8] M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, *Building explainable artificial intelligence systems,* in *Innovative Applications of Artificial Intelligence* (2006) pp. 1766–1773.

[9] M. Harbers, J. Broekens, K. Van Den Bosch, and J.-J. Meyer, *Guidelines for developing explainable cognitive models,* in *International Conference on Cognitive Modeling* (2010) pp. 85–90.

[10] J. Broekens, M. Harbers, K. Hindriks, K. Van Den Bosch, C. Jonker, and J.-J. Meyer, *Do you get it? user-evaluated explainable bdi agents,* in *Multiagent System Technologies* (Springer, 2010) pp. 28–39.

[11] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, *Enabling robots to communicate their objectives,* Autonomous Robots , 1 (2017).

[12] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, *Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults,* in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on* (IEEE, 2017) pp. 676–682.

[13] P. M. Churchland, *Folk psychology and the explanation of human behavior,* The future of folk psychology: Intentionality and cognitive science , 51 (1991).

[14] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction.* (MIT Press, 2004).

[15] S. A. Döring, *Explaining action by emotion,* The Philosophical Quarterly **53**, 214 (2003).

[16] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, *The role of emotion in self-explanations by cognitive agents,* in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (IEEE, 2017) pp. 88–93.

[17] J. Wang, Y. Wang, C. Wei, N. Yao, A. Yuan, Y. Shan, and C. Yuan, *Smartphone interventions for long-term health management of chronic diseases: an integrative review,* Telemedicine and e-Health **20**, 570 (2014).

[18] J. Li, R. Kizilcec, J. Bailenson, and W. Ju, *Social robots and virtual agents as lecturers for video instruction,* Computers in Human Behavior **55**, 1222 (2016).

[19] M. De Graaf, S. Ben Allouch, and J. Van Dijk, *Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study,* in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2017) pp. 224–233.

[20] N. Wang, D. V. Pynadath, and S. G. Hill, *Trust calibration within a human-robot team: Comparing automatically generated explanations,* in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, 2016) pp. 109–116.

[21] J. D. Lee and K. A. See, *Trust in automation: Designing for appropriate reliance,* Human factors **46**, 50 (2004).

[22] B. M. Muir, *Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems,* Ergonomics **37**, 1905 (1994).

[23] L. R. Ye and P. E. Johnson, *The impact of explanation facilities on user acceptance of expert systems advice,* Mis Quarterly , 157 (1995).

[24] C. Conati and K. VanLehn, *Providing adaptive support to the understanding of instructional material,* in *Proceedings of the 6th international conference on Intelligent user interfaces* (ACM, 2001) pp. 41–47.

[25] D. C. Dennett, *Three kinds of intentional psychology,* in *Reduction, Time and Reality*, edited by R. Healey (Cambridge University Press, Cambridge, 1981) pp. 37–61.

[26] R. Beale and C. Creed, *Affective interaction: How emotional agents affect users,* International journal of human-computer studies **67**, 755 (2009).

[27] M. S. Knowles *et al.*, *The modern practice of adult education*, Vol. 41 (New York Association Press New York, 1970).

[28] S. Lieb and J. Goodlad, *Principles of adult learning,* (2005).

[29] F. C. Keil, *Explanation and understanding,* Annual Review of Psychology **57**, 227 (2006).

[30] J. D. Vermunt and N. Verloop, *Congruence and friction between learning and teaching,* Learning and instruction **9**, 257 (1999).

7

Conclusion

I knew who I was this morning, but I've changed a few times since then.

Lewis Carroll
(Alice Pleasance Liddell in *Alice's Adventures in Wonderland*)

In this Thesis, we focused on designing self-explanations for robots. In most studies, self-explanations are typically based on how humans amongst themselves explain behaviour. However, we have argued that many aspects of how humans explain behaviour have not yet sufficiently been considered for robot self-explanations. Our main research question was:

Main Research Question
Which aspects of human behaviour explanation can be used in the construction of social humanoid robot self-explanations and how should we generate such explanations?

We focused on two aspects of this question: 1) attuning explanations to the receiver; and 2) using emotions in the explanations. We derived five research questions from this main question and addressed these in the respective chapters.

In the introduction of this book, we provided some background information on emotions and explanations by discussing related work in the field. In chapter 2, we discuss the type of social robot system we have addressed in our work. We discuss its functionality and its specific requirements. In chapter 3, we address the issue that a good explanation takes the receiver of the explanation into account. We investigate explanations based on beliefs and desires. We compare the use of different explanation styles on child and adult users and find that personalising explanations is indeed needed. In chapter 4, we address whether and how humans use emotions in their explanations of robot behaviour. We switch from robot self-explanations to explanations that people provide for robot behaviour. We study the usage of emotions specifically. Furthermore, we study if people recognise emotional behaviour of the robot. In chapter 5, we provide a formalisation of the interplay between beliefs, desires, and emotions. In this work, we focus on the simulation of emotions in social robots. We identify a gap in existing models, and address this with a custom formalisation framework for emotions based on cognitive appraisal theory. In chapter 6, we compare emotions-based explanations and (regular) goal-based explanations in a long-term study. We found empirical evidence for an effect of agent explanation on prolonged interaction. However, the effect was not in line with expected effects based on literature. Which introduces a challenge for future work in this area. Finally in the chapter, we discuss the conclusions drawn from the separate chapters. Then, we discuss the limitations of our work and potential directions for future continued work. Finally, we discuss some more general contributions from the thesis as a whole.

7.1. Findings

The first research question was:

Research Question 1; Chapter 2
What are the design principles for a social robot system that must autonomously run for several months?
In the PAL (Personal Assistant for a healthy Lifestyle) project, we strove to do re-

search on human-robot interaction in real world environments. In the wild, it is important to reach autonomous, personalised, long-term interaction [1–5]. Furthermore, we were dealing with different groups of users and focusing on consequential domains (for example, the PAL-project itself focuses on helping children reach self-efficacy regarding diabetes type 1 management). We identified the following four design principles to address these challenges. The system should: (1) be **cloud-based**, (2) be **modular**, (3) have a **common terminology and knowledge-base**, and (4) implement **hybrid artificial intelligence** techniques that all have their own contribution to steering the interaction.

The design principles and resulting system are generic and the implementation was tailored to diabetes self-management for children. The architecture of this system distinguished specific functional modules for the common Knowledge-Base (ontologies), Data-Base, Hybrid Artificial Brain (dialogue manager, action selection and explainable AI), Activities Centre (Timeline, Quiz, Break & Sort, Memory, Tip of the Day, ...), Embodied Conversational Agent (ECA,; i.e. humanoid robot and avatar), and Dashboards (PAL control and PAL inform). The resulting system autonomously interacted with a group of (48) child users, their parents, and their caregivers for two periods of two and a half to three months. The system remained stable and continued to show (more and more) behaviours and support in health education & care. This work can serve as a blueprint for future long-term human-avatar and human-robot interaction studies and thereby facilitate incremental research.

Research Question 2; Chapter 3

What are the differences in preference for goal-based versus belief-based social robot explanations between adults and children?

Social humanoid robots interact over a long period of time with users in complex consequential domains such as healthcare. We argued in the introduction and related work that such systems benefit from the capability to self-explain their behaviour. However, they may have to deal with several types of users and these users might differ in the types of explanations that the robot should give them. In the context of healthcare for diabetic children, we were specifically interacting with both child and adult users.

We implemented a Nao-robot as a belief-desire-intention (BDI)-based agent and explained its actions using two different explanation styles. BDI-based programming is a common way of implementing the high-level reasoning of intelligent agents such as social robots [6, 7]. Two explanation styles have commonly been considered for such implementations: goal-based and belief-based action explanations [8, 9]. We compared the preference for these explanation styles between two user groups. We conducted a user study (19 children, 19 adults) in which the robot performed actions to support type 1 diabetes mellitus management. We investigated the preference of children and adults for goal- versus belief-based action explanations.

We found that adults have a significantly *higher* preference for goal-based

explanations than children. This was first evidence that self-explanations of intelligent agents are perceived differently by children and adults. Research on such preferences is an important step for generating *personalised explanations* in human-robot and human-agent interaction, because it provides input on the form and content such explanations should take.

Research Question 3; Chapter 4

To what extent and in what way do humans use emotions in their explanations of robot behaviour?

Addressing this is important because it: (1) helps us design how robots can explain their own actions; and (2) gives insight into how humans perceive robot behaviour. We discussed both in chapter 4. We particularly focus on emotions because (a) humans use emotions when explaining human behaviour [10], (b) self-explanations of intelligent agents (such as robots) are typically based on how humans explain behaviour [11, 12], but (c) current research on self-explanations by robots has not thoroughly considered emotions yet.

To address this question, we presented filmed behaviours of either a human or a humanoid robot coping with a distressing situation to MTurk participants. The behaviours were modelled to represent several coping styles from the literature [13]. Behaviour in coping styles was chosen because coping strategies are triggered by emotion and are aimed at emotion regulation [14]. We can study the explanations of these coping induced behaviours to find whether people use emotions when explaining robot behaviour (i.e., as existence proof). If people *do* explain robot behaviour using emotions, then this should be observed in behaviours resulting from coping strategies. If people *do not* use emotions for explanations of this behaviour then that is a strong indicator that they also won't use it when the robot shows other types of (semi-)intentional behaviour. We measured (1) whether people could recognise the coping styles; (2) what spontaneous unrestricted explanations people give for the behaviour; and (3) how positive and accepting people were towards the behaviour (i.e., naturalness, appropriateness, warmth, competence, and discomfort). For all these outcomes we considered the influence of the actor type (human versus robot) and scenario (health versus museum).

We show that, and shed light on what way, people use emotions when explaining robot behaviour. They do so with less frequency then they do for human behaviour. Still, only about $1/4th$ of the explanations is devoid of any emotionality (about $1/5th$ for the human behaviour explanations). Furthermore, we found a difference in how people explain *human behaviour* with emotions and how people explain *robot behaviour* with emotions. The perception of coping style correlates with the emotionality of explanations when people explain human behaviour, but not when people explain robot behaviour. This implies that people have slightly simpler models to attribute emotions to robot intentions than for attributing emotions to human intentions. Which implies that we can get away with simpler models for choosing to use emotions in the generation of robot self-explanations as well.

Research Question 4; Chapter 5
How can we incorporate emotion theory into BDI-based agent programming?

Given our findings concerning research question 3, it made sense to further investigate emotions in robot self-explanations. One way of modelling emotions that can be included in explanations is by expanding BDI-based agent programming with a computational model of emotion. Furthermore, BDI-based models are conceptually close to cognitive emotion theory and BDI-based agent programming is an important part of the decision making of hybrid social robot systems as discussed in chapter 2..

In this chapter, we discussed different computational models of emotion and what they lack to enable unambiguous implementation into BDI-based agent programming. These models typically simulate emotions based on cognitive appraisal theory [15]. Cognitive appraisal theory, however, introduces a large set of appraisal processes not specified in enough detail for agent oriented programming.

In this chapter, we discussed a framework based on the belief-desire theory of emotions (BDTE) [16], that enables the computation of emotions for BDI-based agents. We bridged the remaining gap between BDTE and BDI-based agent programming frameworks.

Research Question 5; Chapter 6
What are the effects of cognitive and affective explanations on motivation to use a social robot/ avatar system during long-term interaction?

We address that state of the art explainable artificial intelligence for intelligent agents focuses mainly on explaining an agent's behaviour based on the underlying *beliefs and goals* in *short-term experiments*. However, as argued in this thesis, and as supported by the work concerning research question 3, emotions and emotion words (in addition to beliefs and goals) play a role in intelligent agent (robot/ avatar) self-explanations. Furthermore, research in e-health support systems and human-robot interaction stresses the need for studying long-term interaction with users [1, 17–19].

In this chapter, we report on a long-term experiment in which we tested the effect of cognitive, affective and lack of explanations on children's motivation to use the e-health support system which we described in chapter 2. Children (aged 6-14) suffering from type 1 diabetes mellitus interacted with a virtual robot (avatar) as part of the e-health system over a period of 2.5 - 3 months. Children alternated between the three conditions (cognitive, affective and no explanation). Agent behaviours that were explained to the children included why 1) the agent asks a certain quiz question; 2) the agent provides a specific tip (a short instruction) about diabetes; or, 3) the agent provides a task suggestion, e.g., play a quiz, or, watch a video about diabetes. Their motivation was measured by counting: (1) how often they would continue to play the quiz or (2) ask for an additional tip, (3) how often children would follow the agent's suggestion, additionally and how often they would request

an explanation from the system.

We found that the explanation condition influenced how often children followed task suggestions. Unexpectedly, the results indicate that children follow the suggestions more often when *no explanation* is given. We found no other significant effects of explanations in this study. We discussed three possible explanations for this in chapter 6. We briefly summarise these here. 1) Explanations are additional text to read. Children might simply not read the longer texts in explained task suggestion causing them to choose more randomly from the menu. This would imply explanations need to be even shorter than literature suggests [6], or be given less frequently. 2) Understanding the purpose of the task might help children orient/plan their behaviour better themselves. Which would mean the explanation is indeed helpful. This would relate to literature on education [20]. 3) The child might get stubborn when the robot explains. Thinking something like 'I don't want to do that!'. Which causes them to choose different tasks. Although no difference appeared between cognitive versus affective explanations, this is to our knowledge the first evidence that explanations impact long-term human-agent interaction and system usage. Our results also show that counter-intuitive effects of agent explanations may be expected when used with children and in long-term interaction, and, that more research is needed to understand why lack of explanations seems to correlate with following task suggestions.

7.2. Limitations

We have studied user preferences for- and motivational effects of explanations. However, our techniques for generating the explanations remain somewhat straightforward. This was good for our purposes, because it allowed us to better control the types of explanations given and the effects of these explanations. However, it will be needed in future work to also generate explanations in more complex manners. For example, we now try to cleanly separate beliefs and goals in the generated explanations to find preferences (chapter 3). However, humans would shift between the styles and/or combine the styles in specific situations. So, overall preferences give us insight in the need for personalised explanations and give us some handles on how to personalise them. However, specific situations might still require an explanation in a style that is not generally preferred. We will give some suggestions in the future directions section below (7.3) on how to generate explanations using more elaborate models.

Another limitation of this work is that we did not yet generate emotions in explanations based on the robot's current emotional state as simulated by the computational model of emotion we introduced (CAAF, see chapter 5). There were two reasons for this. First, with CAAF running, we would would have had less control over the exact phrasing of the explanations as provided by the robot. Such a control was important to cleanly compare the different conditions. Second, the emotional state simulated by a computational model of emotion can not directly be used in the explanations. Such models simulate several emotions during interaction. This is good, because it allows the robot to properly interact in complex social interactions. However, we currently do not yet know what emotions should be used in

the explanations, nor when we should use them. This is a very complex problem and requires more research. However, we belief future work should generate emotions in explanations also based on the robot's own emotional state simulated by a well-defined computational model of emotion like CAAF. This will allow for more variety in the explanations and for more intelligent decision mechanisms for what emotions to use (and when to use them) in explanations.

7.3. Future Work

In our thesis, we tested the effects of robot self-explanations and we studied how humans themselves explain robot behaviour. However, we did not tightly combine these studies yet. Robot self-explanations based on insights from how people explain (robot) behaviour can help to improve the self-explanations [11]. Here, however, we want to argue to have future studies that consider an even tighter link between the two research fields. We propose a 2-step research set-up. First, show robot behaviour to lay humans and ask them to explain the behaviour just like we did in chapter 4. Second, show different humans the robot behaviour and let the robot self-explain using the explanations from the first study. In this way, we can study the effects of robot self-explanations on humans without having to construct the explanations ourselves. This relates to the limitation mentioned that more elaborate explanation models are needed.

Furthermore, we argue it is important to have a good conceptual framework to annotate the explanations with. For example, an annotation framework like f-ex [21] to annotate the types of mental constructs (likes, beliefs and desires) used in the explanations. Or, a sentiment miner like LIWC [22] to measure the usage of emotion words in explanations. In this way, we can incrementally learn the effects of explanation styles on things like preference, motivation, comprehensiveness, informativeness, etc within specific contexts and settings.

Note that some methods to annotate the explanations require more time and resources than others. A sentiment miner or other text miner would be an automatic process; whereas a framework like f-ex requires humans to annotate the explanations which can become quite a lot of work when attempting to annotate databases of hundreds, thousands, or even more explanations. It may not be simple to crowd-source such annotations because they do require some knowledge of folk psychology (see also [21] for a description of how to use f-ex), but they do provide far more expressive annotations than text miners can currently give. Still, for automatically annotating and studying explanations it would be valuable to have text mining algorithms that automatically map concepts in explanations like whether a belief or desire was mentioned.

Finally, our measures were user preference for explanation styles in chapter 3 and motivation in chapter 6. Future work should take more complex explanation models that are attuned to the users and that at times use emotions in addition to beliefs and desires and then test these models with regards to, for example, trust in the robot and understanding of what the robot explains to the user. Previous studies showed effects on these measures for systems that explain behaviour [23–26]. However, when the explanation models become more complex than it will

become valuable to see re-evaluate their effect also on these basic measures.

7.4. Overall Contribution

In this thesis, we have made first steps towards developing human-aware explainable artificial intelligence for humanoid robots. We designed and tested the explanations in a real-world ('in the wild') system in a consequential domain (helping children aged 6-14 to become more self-manageable with regards to their illness). The system autonomously interacted with users for two periods of 2.5 - 3 months. Our research shows that it is possible to address interesting and complex research questions in such settings, even considering that there was only a limited group of users (48 children in the final chapter).

Overall, we conclude robot (and avatar) self-explanations must indeed take additional aspects of human behaviour explanation into account. Specifically, we provide evidence supporting:

1. Robot (and avatar) self-explanations should be attuned to the receiver of the explanation.

Which is based on our related work and on chapter 3, where we showed explanations are perceived differently by different types of users). Secondly:

2. Robots (and avatars) should be able to use emotions in their explanations.

Which is based on our discussion of related work and on chapter 4, where we showed humans themselves use emotions when explaining robot behaviour. Furthermore, our final chapter showed that explanation effects occur also in long-term interaction. These effects were not in line with the expectations based on literature, showing the need for more work in this area.

This thesis shows that explainable artificial intelligence (both in the social sciences as well as in human-computer interaction) should consider **individual preferences**, and consider **emotions** in addition to beliefs and desires when **explaining robot or avatar behaviour**.

References

[1] I. Leite, C. Martinho, and A. Paiva, *Social robots for long-term interaction: a survey,* International Journal of Social Robotics **5**, 291 (2013).

[2] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, *et al.*, *Multimodal child-robot interaction: Building social bonds,* Journal of Human-Robot Interaction **1**, 33 (2013).

[3] A. Coninx, P. Baxter, E. Oleari, S. Bellini, B. Bierman, O. B. Henkemans, L. Cañamero, P. Cosi, V. Enescu, R. R. Espinoza, *et al.*, *Towards long-term social child-robot interaction: using multi-activity switching to engage young users,* Journal of Human-Robot Interaction **5**, 32 (2016).

[4] R. Looije, M. A. Neerincx, J. K. Peters, and O. A. Blanson Henkemans, *Integrating robot support functions into varied activities at returning hospital visits,* International Journal of Social Robotics **8**, 483 (2016).

[5] O. A. B. Henkemans, B. P. Bierman, J. Janssen, R. Looije, M. A. Neerincx, M. M. van Dooren, J. L. de Vries, G. J. van der Burg, and S. D. Huisman, *Design and evaluation of a personal robot playing a self-management education game with children with diabetes type 1,* International Journal of Human-Computer Studies **106**, 63 (2017).

[6] M. Harbers, J. Broekens, K. Van Den Bosch, and J.-J. Meyer, *Guidelines for developing explainable cognitive models,* in *International Conference on Cognitive Modeling* (2010) pp. 85–90.

[7] K. V. Hindriks, *Debugging is explaining,* in *International Conference on Principles and Practice of Multi-Agent Systems* (Springer, 2012) pp. 31–45.

[8] D. C. Dennett, *Three kinds of intentional psychology,* in *Reduction, Time and Reality*, edited by R. Healey (Cambridge University Press, Cambridge, 1981) pp. 37–61.

[9] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction.* (MIT Press, 2004).

[10] S. A. Döring, *Explaining action by emotion,* The Philosophical Quarterly **53**, 214 (2003).

[11] T. Miller, *Explanation in artificial intelligence: Insights from the social sciences,* Artificial Intelligence (2018).

[12] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, *Explainable agents and robots: Results from a systematic literature review,* in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, 2019) pp. 1078–1088.

[13] S. Folkman and R. S. Lazarus, *Ways of coping questionnaire* (Consulting Psychologists Press, 1988).

[14] R. S. Lazarus, *Emotion and adaptation.* (Oxford University Press, 1991).

[15] S. Marsella, J. Gratch, and P. Petta, *Computational models of emotion,* A Blueprint for Affective Computing-A sourcebook and manual **11**, 21 (2010).

[16] R. Reisenzein, *Emotions as metarepresentational states of mind: Naturalizing the belief–desire theory of emotion,* Cognitive Systems Research **10**, 6 (2009).

[17] J. Wang, Y. Wang, C. Wei, N. Yao, A. Yuan, Y. Shan, and C. Yuan, *Smartphone interventions for long-term health management of chronic diseases: an integrative review,* Telemedicine and e-Health **20**, 570 (2014).

[18] J. Li, R. Kizilcec, J. Bailenson, and W. Ju, *Social robots and virtual agents as lecturers for video instruction,* Computers in Human Behavior **55**, 1222 (2016).

[19] M. De Graaf, S. Ben Allouch, and J. Van Dijk, *Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study,* in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2017) pp. 224–233.

[20] J. D. Vermunt and N. Verloop, *Congruence and friction between learning and teaching,* Learning and instruction **9**, 257 (1999).

[21] B. F. Malle, *F. EX: A coding scheme for folk explanations of behavior*, Tech. Rep. (Institute of Cognitive and Decision Sciences, University of Oregon (originally published in 1998), 2002).

[22] Y. R. Tausczik and J. W. Pennebaker, *The psychological meaning of words: Liwc and computerized text analysis methods,* Journal of language and social psychology **29**, 24 (2010).

[23] B. M. Muir, *Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems,* Ergonomics **37**, 1905 (1994).

[24] D. N. Lam and K. S. Barber, *Comprehending agent software,* in *Autonomous Agents and Multiagent Systems* (2005) pp. 586–593.

[25] B. Y. Lim, A. K. Dey, and D. Avrahami, *Why and why not explanations improve the intelligibility of context-aware intelligent systems,* in *Human Factors in Computing Systems* (2009) pp. 2119–2128.

[26] S. R. Haynes, M. A. Cohen, and F. E. Ritter, *Designs for explaining intelligent agents,* International Journal of Human-Computer Studies **67**, 90 (2009).

Epilogue

All we can do is decide what to do with the time given us.

J.R.R. Tolkien
(Gandalf the Grey in *The Lord of the Rings*)

That brings us to the end of this book and of my adventure as a PhD Candidate. It marks the end of an era that I greatly enjoyed and where I learned more than I can possibly summarise here. I will write the acknowledgements mainly in Dutch. For my non-Dutch colleagues I will write the appreciation of their help in English.

Ik wil hier aan het einde nog even stilstaan bij iedereen die me heeft geholpen hier te komen. Ten eerste wil ik mijn dagelijkse begeleider Joost bedankten. Ik heb Joost leren kennen bij het vak Affective Computing tijdens mijn master. Hier wist hij met veel enthousiasme de stof uit te leggen, en was hij altijd bereid om een discussie aan te gaan over een willekeurig onderwerp met die vervelende student die altijd wat te vragen had (ik dus). Ik heb met Joost als begeleider mijn master Thesis afgerond. Deze samenwerking is zo goed bevallen dat we er nog maar een PhD project op hebben laten volgen. Iets dat ik zie als 1 van de beste beslissingen van mijn leven.

Verder wil ik mijn Promotor Mark bedanken. De positieve energie en nuchtere kijk zijn echt dingen die je nodig hebt om door te blijven gaan met een project zo gaaf maar soms ook zo vermoeiend als een PhD. Ook wil ik mijn promotor Koen bedanken voor de fijne samenwerking en de scherpe en altijd waardevolle feedback. Joost, Mark, Koen, ik reken mij gelukkig een team van begeleiders te hebben gehad die altijd tijd voor me maakten, en die ook altijd me konden helpen mezelf en het werk tot een hoger niveau te brengen. Bedankt!

(In English.) Special thanks for the independent committee members: Prof.dr. T. Belpaeme, Prof.dr.ir. D.A. Abbink, Prof.dr. C.M. Jonker, and Dr. M.M.A. de Graaf. Thank you for reading and appreciating this work and your vital role in finalising it.

Ook mijn collega's op de afdeling van interactive intelligence wil ik bedanken. Ruud voor de technische ondersteuning met name aangaande de Nao en Pepper robots. Bart voor de technische ondersteuning en de hulp met name tijdens het laatste experiment met het filmen en editen van de filmpjes.

(In English.) My office mates: Rifca, Bernd, Fran, Ding, Thomas, Pietro. Thank you for being great office mates who were always willing to help others and still make time for the 15 o-clock coffee break. Rifca, thank you also for the pleasant times in the PAL meetings. Bernd for always having time for in depth discussions about artificial intelligence or dungeons and dragons. Fran for the pleasant times, coffee breaks, and your never ceasing willingness to help your peers. Ding, I don't know if you remember the graduate school course we followed, but I *will* indeed

remember you as kind-hearted! Thank you for the ping pong matches and pleasant times. Thomas, your sober view on the PhD life often helped me more than you may think. Pietro, thank you for the cheerful tone and for the D&D sessions!

(In English.) I also want to thank all other colleagues. You were great and a big part of the reason I could almost always drive to Delft in a happy mood. Thank you for all the insightful discussions, enjoyable coffee breaks, and memorable period. Vincent, Elie, Malte, Ilir, Myrthe, Rolf, Elena, Miguel, Aleksander, Merijn, Chris, Rijk, Roel, Luciano, Marieke, Ursula, Willem-Paul, Catharine, Catholijn, Wauter, Frans, Joachim, Jasper, Anita and all others!

(In English.) Besides the people from my department, I also want to thank the people from the PAL project. I think we had a great project and collaboration, and I could not have done my research if it was not for all of you. In particular the people with whom I worked together either during the development or during the writing of papers. Bernd, Bert, Antoine, Oya, Rifca (again), Diego, Lorenzo, Paolo, Willeke, Olivier, Michael, Yannis, Sylvia, and even though he is no longer here to read it, I also want to thank Uli for our pleasant work together both in development and in writing papers. Of course, I also thank all the others in the project. I hope to meet all of you again in the future.

Buiten de mensen op de werkvoer hebben ook de mensen die ik van buiten de werkomgeving ken een grote invloed gehad op dit resultaat. Mijn vrienden die ik nu al zo lang ken dat ze zo goed als familie zijn, Jeroll, Joey, Erwin, Nick, Fabian, Soe en alle anderen die het woord ugh als een begrip zien. Peter en Denise de ouders van mijn vriendin. Manon haar zus. Pascale mijn zus, Robin mijn zwager, en hun kleintjes, Liam, en Jonathan. Mijn oom Han. De honden Arthur, Falco, Garvey, Rohan en Beer. De paarden Yolinn, Bell, Dell en Bas. Mijn oma Ooms die eigenlijk een derde ouder is en altijd in me gelooft. Mijn ouders Jan en Joke die op vele manieren een cruciale rol hebben gespeeld in het behalen van mijn PhD. Waarbij het betalen van mijn opleiding voorafgaand aan de PhD nog niet eens het voornaamste is. En Sophie, die altijd aan mijn kant staat. De vrouw met wie ik al 11 jaar mijn leven deel en die dit nog niet weet terwijl ik dit schrijf, maar die ik binnenkort zal vragen mijn vrouw te worden.

Zonder jullie was dit boek er niet geweest.

A

Adjusted Ways of Coping Questionnaire

To do a manipulation check for the coping styles, we ask participants to fill in an adapted version of the ways of coping (WoC) questionnaire. The original questionnaire is somewhat long for an Amazon Mechanical Turk study and was initially designed for recognising coping styles in ones own behaviour. The questions are not all recognisable in someone else (see also table 4.1). To account for this, we developed an adjusted Ways of Coping questionnaire specifically for our purposes.

Firstly, the seek social support questions could be removed since we do not consider that style in our study. Secondly, we picked the three most descriptive sentences per style to include in our questionnaire. We believe that this results in a list that is sufficiently descriptive to capture the different coping styles and short enough for participants in the study. We chose the sentences with the following rationale. (1) the first and second author independently went through the list removing sentences that they deemed unrecognisable in someone else. (2) For every coping style, the three questions with the highest factor loading, amongst the ones that were deemed recognisable by both authors, were included in the list. There was one disagreement in the escape avoidance style concerning sentence 5 (see table 4.1). The eventual decision was to include the sentence in the question list.

Additionally we phrased all the sentences such that they referred directly to Robin. Pronouns were changed based on the condition (him/his versus her/hers versus it/its depending on the actor being male/female/robot respectively). The resulting question list (pronouns in female form) can be seen in table A.1.

Table A.1: The adjusted ways of coping questionnaire to measure perceived coping style in another

Confrontive(C)	1. Robin stood her ground and fought for what she wanted 2. Robin tried to get Bob to change his mind. 3. Robin expressed anger to Bob.
Distancing(D)	1. Robin made light of the situation; refused to get too serious about it. 2. Robin went on as if nothing had happened. 3. Robin looked for the silver lining, so to speak; tried to look on the bright side of things.
Self – Controlling(S – C)	1. Robin tried to keep her feelings to herself. 2. Robin tried not to burn her bridges, but leave things open somewhat. 3. Robin tried not to act too hastily or follow her first hunch.
Accepting – Responsibility(A – R)	1. Robin criticised or lectured herself. 2. Robin realised she brought the problem on herself. 3. Robin apologised or did something to make up.
Escape – Avoidence(E – A)	1. Robin wished that the situation would go away or somehow be over with. 2. Robin had fantasies or wishes about how things might turn out. 3. Robin avoided contact with Bob
Problem – Solving(P – S)	1. Robin knew what had to be done and doubled her efforts to make things work. 2. Robin made a plan of action and followed it. 3. Robin changed something so things would turn out all right.
Possitive – Reappraisal(P – R)	1. Robin changed or grew as a person in a good way. 2. Robin came out of the experience better than when she went in. 3. Robin changed something about herself.

B

Filmed Conversations of Coping Styles

We designed videos of several conversations between two individuals (which for convenience we refer to as 'Bob' and 'Robin'). The content of the conversations is such that in all of them Bob does something that is distressing for Robin. Robin then copes with that in one of the styles discussed in section 4.2.2 and the Ways of Coping [1, 2]. We chose to exclude the 'seek social support' style in our study because we focus on a conversation between two individuals to avoid additional complexity in perspective taking for the participants. Seeking social support would require an additional actor in the scenario. Robin, in some of the video's, was played by a professional actor or actress. In other video's, Robin was acted out by a humanoid robot (the Pepper robot of Softbank which we animated for this purpose). In this appendix, we discuss the design and validation of the conversations in the videos first. Then, we discuss the animation of the robot and the shooting of the videos.

B.0.1. Conversations in Coping Styles

We designed conversations in coping styles based on the Ways of Coping [1]. We validated whether people were able to recognise the dominant coping styles in the conversations via a pilot study. We did two rounds of validation which we discuss here.

Participants were colleagues, friends, and family. They were uninformed of our goal with this study. Both validation rounds had 5 unique participants. In total, there were 10 participants (3 male, 2 female, aged 28-55).

Set-up and Procedure First Validation

This first validation round had 5 (male) participants. We showed people textual prints of the different variations of the conversations and textual prints of the different coping styles as in table 4.1. Participants read all the conversations and all the coping styles of a single scenario (health/ museum). Then, they were shown

Table B.1: Results from the first validation round. Rows show participants with a number, the 'order' in which they were shown the scenarios, and the style they selected per style as intended for the conversation.

Health

Nr.	Order	C	D	S-C	A-R	E-A	P-S	P-R
1	M-H	C	D	S-C	A-R	E-A	P-S	P-R
2	H-M	C	D	S-C	A-R	E-A	P-S	P-R
3	M-H	C	D	S-C	P-R	E-A	P-S	A-R
4	H-M	C	E-A	S-C	A-R	D	P-S	P-R
5	H-M	C	E-A	S-C	A-R	P-R	P-S	D
Agreement:		100%	60%	100%	80%	60%	100%	60%

–

Museum

Nr.	Order	C	D	S-C	A-R	E-A	P-S	P-R
1	M-H	C	D	S-C	A-R	E-A	P-S	P-R
2	H-M	C	E-A	S-C	A-R	D	P-S	P-R
3	M-H	C	D	E-A	A-R	S-C	P-S	P-R
4	H-M	C	D	S-C	A-R	E-A	P-S	P-R
5	H-M	C	D	S-C	P-S	E-A	P-R	A-R
Agreement:		100%	80%	80%	80%	60%	80%	80%

the conversations and styles of the other scenario. The order in which they were shown the scenarios was randomised to correct for learning effects. Participants had to match conversations and styles. They were only allowed to choose one style per conversation. They could take as much time as they wanted. However, all participants finished in about 30 minutes.

During all sessions a researcher was present to answer any questions. It was made clear to the participants that there was no right or wrong answer. Rather we were interested in their subjective perception of the conversations. If participants asked who were having the conversations, then we told them to imagine two persons unknown to each other.

Results 1st Validation

Most agreements were quite good. For example, the confrontive conversations were recognised as such in both the health as the museum scenario by all participants (100% agreement). Other styles proved more difficult and reached 'only' 60% agreement (e.g, distancing).

We argue that 80% agreement is good enough for our purposes. There are always more styles simultaneously used when coping with a situation. Reaching 100% in all cases might prove a lengthy process. Eventually reaching 100% might then even be a random stroke of luck more than anything else. We therefore focused on the styles with 60% agreement and tried to improve those.

Table B.2: Results from the second validation round. Only styles with insufficient agreement were included for this round

		Health				Museum		
Nr.	**Order**	**D**	**A-R**	**E-A**	**P-R**	**D**	**S-C**	**E-A**
6	H-M	D	A-R	E-A	P-R	E-A	S-C	D
7	H-M	D	A-R	E-A	P-R	D	S-C	E-A
8	M-H	D	A-R	E-A	P-R	D	S-C	E-A
9	M-H	D	P-R	E-A	A-R	D	S-C	E-A
10	H-M	D	A-R	E-A	P-R	D	S-C	E-A
Agreement:		100%	80%	100%	80%	80%	100%	80%

Updating the Conversations

We discussed with the participants what brought them to their decisions to find out how we could improve the texts. Based on the resulting recommendations we made a new set of conversations and validated these with 5 different people.

Set-up and Procedure Second Validation

Participants informed us that they matched the clear coping styles first, and then went on to the more difficult to recognise ones. To save our participants and ourselves some time, in this second set we removed the styles and variations with sufficient agreement in the previous validation.

The second validation round had 5 participants (3 male, 2 female, aged 28-55). The styles with 60% agreement and *all* styles that they were confused with were included in the second validation. For example, possitive-reappraisal (P-R) had 60% agreement and was confused with accepting-responsibility (A-R) and distancing (D). Meaning these three styles were included for the second validation. When doing this for all columns in table B.1 this left us with a total of seven conversations.

Results 2nd Validation

For the second validation the agreement went up to at least 80% per style. See table B.2 for an overview of the results.

B.0.2. Making Videos of the Conversations

Using the validated conversations, we continued to translate them into videos where they were played out by actors. We wanted videos of the conversations rather than textual descriptions to make sure that people have the same mental image of the robot. A textual description like 'the Robot did X' might invoke different mental state images for different people. Furthermore, non-verbal characteristics and movements might have an influence on the perception. We chose a Pepper robot from Softbank as embodiment. Pepper is a humanoid robot which makes having dialogue with it seem natural. In addition, its size makes it reasonable to have it in a museum or hospital giving guidance to users.

We hired professional actors (one male one female) to play the role of Robin in the conversations. In addition, we hired a semi-professional actor for the role

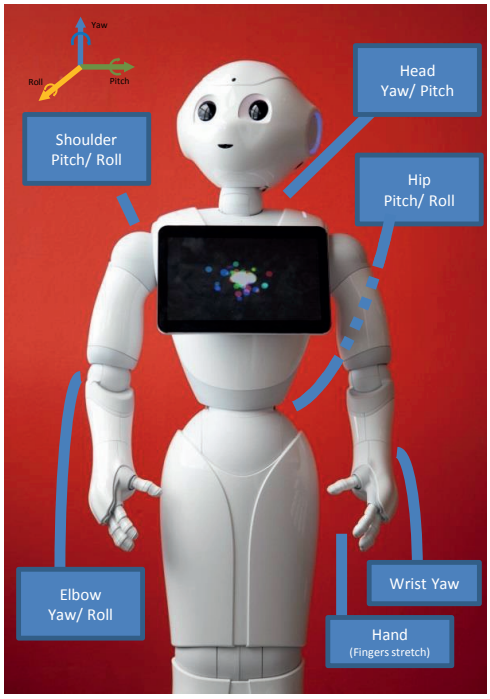


Figure B.1: Pepper degrees of freedom.

of Bob. We gave them descriptions of the coping styles, and they were tasked to make the proper movements and intonations to reflect them. We chose to have Robin be acted out by both a male and a female so that we can check whether differences in perception of the coping styles are due to the human-robot difference or simply due to a more general 'another actor' difference. Figure 4.1 shows eight snapshots of the videos. The videos themselves can be accessed via this link: https://ii.tudelft.nl/ExplainableAI/video/.

The next step was to animate the Pepper robot to move similarly as the actor. However, it is not possible to have the robot move *exactly* the same as the actor. The robot is limited in its movements since: (1) the robot does not have the same degrees of freedom; (2) its proportions are slightly different than that of a human; (3) the tablet blocks arm movements in front of the robot; and finally (4) fast movements cause the engines to make loud noises. See also figure B.1 for the appearance and degrees of freedom of the Pepper robot.

We annotated the videos of one of the actors using the ANVIL software package. For the annotation, we took the limitations of the Pepper robot into account. Annotating subtleties like mouth movement would have been irrelevant, since it is not possible to implement it. We annotated (1) the part of text the movement is a part of; (2) the speed of the movements; and (3) the direction of the arms, hands, head, and body.

Arm movements that were impossible on the robot due to proportional differences or due to the tablet were modelled to resemble the intended gesture as good as possible. For example, pointing could be done by stretching the fingers and pointing the arms in a wider angle around the tablet, rather than pointing a single finger in a straight angle. Finally, very fast movements like nodding 5 times or more within 2 seconds were annotated as such, but were later animated as 2 or 3 slower nods. For the voice, we used the build-in voice of the Pepper robot. This automatically creates intonations with the words. When sentences were not pronounced fluently enough, we tried to improve by adding additional punctuation to the sentences.

References

[1] S. Folkman and R. S. Lazarus, *Ways of coping questionnaire* (Consulting Psychologists Press, 1988).

[2] R. S. Lazarus, *Emotion and adaptation.* (Oxford University Press, 1991).

C

T-values for Coping Style Recognition

C

Table C.1: t-values for coping style recognition.

(a) Human Actor: t-values

Modelled Style	Recognition						
	C	D	S-C	A-R	E-A	P-S	P-R
Confrontive (C)	t(40)=9.3, p=<.0005	t(40)=-6.6, p=<.0005	t(40)=.8, p=.412	t(40)=-6.1, p=<.0005	t(40)=-5.0, p=<.0005	t(40)=4.9, p=<.0005	t(40)=-3.3, p=.002
Distancing (D)	t(34)=-1.1, p=.272	t(34)=2.8, p=.009	t(34)=5.3, p=<.0005	t(34)=-4.9, p=<.0005	t(34)=-1.0, p=.314	t(34)=-.6, p=.521	t(34)=-2.4, p=.021
Self-Controlling (S-C)	t(42)=1.8, p=.086	t(42)=-1.6, p=.114	t(42)=7.2, p=<.0005	t(42)=-8.9, p=<.0005	t(42)=-6.9, p=<.0005	t(42)=6.9, p=<.0005	t(42)=-2.1, p=.045
Accepting-Responsibility (A-R)	t(39)=-9.0, p=<.0005	t(39)=-6.5, p=<.0005	t(39)=3.4, p=.001	t(39)=5.0, p=<.0005	t(39)=-3.4, p=.002	t(39)=5.4, p=<.0005	t(39)=.8, p=.436
Escape-Avoidance (E-A)	t(34)=-5.2, p=<.0005	t(34)=-3.3, p=.002	t(34)=5.9, p=<.0005	t(34)=-1.2, p=.234	t(34)=8.3, p=<.0005	t(34)=-2.8, p=.008	t(34)=-4.7, p=<.0005
Problem-Solving (P-S)	t(40)=3.5, p=.001	t(40)=-1.8, p=.079	t(40)=1.1, p=.264	t(40)=-4.4, p=<.0005	t(40)=-9.1, p=<.0005	t(40)=10.2 p=<.0005	t(40)=-1.2, p=.247
Possitive-Reappraisal (P-R)	t(32)=-4.1, p=<.0005	t(32)=5.0, p=<.0005	t(32)=4.0, p=<.0005	t(32)=-3.2, p=.003	t(32)=-4.2, p=<.0005	t(32)=.7, p=.487	t(32)=.9, p=.396

(b) Robot Actor: t-values

Modelled Style	Recognition						
	C	D	S-C	A-R	E-A	P-S	P-R
Confrontive (C)	t(39)=7.7, p=<.0005	t(39)=-4.7, p=<.0005	t(39)=-2.1, p=.039	t(39)=-5.0, p=<.0005	t(39)=-7.4, p=<.0005	t(39)=8.8, p=<.0005	t(39)=-3.1, p=.003
Distancing (D)	t(37)=-3.6, p=.001	t(37)=5.3, p=<.0005	t(37)=6.1, p=<.0005	t(37)=-2.6, p=.013	t(37)=-7.1, p=<.0005	t(37)=1.0, p=.342	t(37)=-5.9, p=<.0005
Self-Controlling (S-C)	t(45)=.9, p=.370	t(45)=-.7, p=.485	t(45)=6.3, p=<.0005	t(45)=-5.3, p=<.0005	t(45)=-10.9, p=<.0005	t(45)=7.0, p=<.0005	t(45)=-3.0, p=.004
Accepting-Responsibility (A-R)	t(34)=-7.5, p=<.0005	t(34)=-2.7, p=.010	t(34)=2.1, p=.044	t(34)=5.8, p=<.0005	t(34)=-10.3, p=<.0005	t(34)=6.3, p=<.0005	t(34)=1.0, p=.323
Escape-Avoidance (E-A)	t(35)=-3.9, p=<.0005	t(35)=1.4, p=.166	t(35)=3.8, p=.001	t(35)=-2.0, p=.051	t(35)=-1.4, p=.182	t(35)=2.1, p=.044	t(35)=-1.7, p=.097
Problem-Solving (P-S)	t(36)=6.7, p=<.0005	t(36)=-4.6, p=<.0005	t(36)=1.9, p=.064	t(36)=-3.8, p=.001	t(36)=-9.7, p=<.0005	t(36)=9.3, p=<.0005	t(36)=-2.8, p=.008
Possitive-Reappraisal (P-R)	t(46)=-4.6, p=<.0005	t(46)=4.2, p=<.0005	t(46)=4.5, p=<.0005	t(46)=-3.0, p=.005	t(46)=-6.3, p=<.0005	t(46)=1.3, p=.195	t(46)=2.1, p=.042

List of Publications

13. **F. Kaptein, J. Broekens, K.V. Hindriks & M.A. Neerincx**, *Evidence for the Use of Emotion in Human Explanations of Robot and Human Behaviour*, submitted to journal transactions on human robot interaction (THRI)

12. **F. Kaptein, B. Kiefer, A. Cully, O. Celiktutan, B. Bierman, R. Peters, J. Broekens, W. van Vught, M.A. van Bekkum, Y. Demiris & M.A. Neerincx** *A Cloud-Based Robot System for Long-term Interaction: Principles, Implementation, Lessons-Learned* submitted to journal transactions on human robot interaction (THRI)

11. **F. Kaptein, J. Broekens, K.V. Hindriks & M.A. Neerincx**, *Evaluating Cognitive and Affective Intelligent Agent Explanations in a Long-Term Health-Support Application for Children with Type 1 Diabetes*, In 8th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1-7), 2019

10. **B. Dudzik, M. P. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. Heylen, H. Hung, M.A. Neerincx & K.P. Truong**, *Context in Human Emotion Perception for Automatic Affect Detection: A Survey of Audiovisual Databases*, In 8th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 206-212), 2019

9. **M.A. Neerincx, W. van Vught, O. B. Blanson, E. Oleari, J. Broekens, R. Peters, F. Kaptein, Y. Demiris, B. Kiefer & D. Fumagalli**, *Socio-Cognitive Engineering of a Robotic Partner for Child's Diabetes Self-Management*, Frontiers In Robotics and AI, 6, 1-16., 2019

8. **M.A. Neerincx, J. van der Waa, F. Kaptein & J. van Diggelen**, *Using perceptual and cognitive explanations for enhanced human-agent team performance*, In International Conference on Engineering Psychology and Cognitive Ergonomics (pp. 204-214), 2018

7. **F. Kaptein, J. Broekens, K.V. Hindriks & M.A. Neerincx**, *Self-Explanations of a Cognitive Agent by Citing Goals and Emotions*, In Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 81-82), 2017

6. **F. Kaptein, J. Broekens, K.V. Hindriks & M.A. Neerincx**, *The role of emotion in self-explanations by cognitive agents*, In Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 88-93), 2017

5. **F. Kaptein, J. Broekens, K.V. Hindriks & M.A. Neerincx**, *Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults*, In 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (pp. 676-682), 2017

4. **M.A. Van Bekkum, H.U. Krieger, M.A. Neerincx, F. Kaptein, B. Kiefer, R. Peters & S. Racioppa**, *Ontology engineering for the design and implementation of personal pervasive lifestyle support*, in SEMANTiCS (Posters, Demos, SuCCESS), CEUR Workshop Proceedings, 1613-0073, 2016

3. **M.A. Neerincx, F. Kaptein, M.A. Van Bekkum, H.U. Krieger, R. Peters & M. Sapelli**, *Ontologies for social, cognitive and affective agent-based support of child's diabetes self-management*, In Proceedings of ECAI, pp.35-38, workshop on diabetes (DfAI), 2016

2. **H.U. Krieger, R. Peters, B. Kiefer, M.A. Van Bekkum, F. Kaptein & M.A. Neerincx**, *The federated ontology of the pal project interfacing ontologies and integrating time-dependent data*, in Proceedings of the 8th IC3K International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2016

1. **F. Kaptein, J. Broekens, K.V. Hindriks & M.A. Neerincx**, *Caaf: A cognitive affective agent programming framework*, In procedings of the International Conference on Intelligent Virtual Agents (pp. 317-330), 2016